9TH EDITION

Advanced Engineering Mathematics

ERWIN KREYSZIG

Professor of Mathematics
Ohio State University
Columbus, Ohio

WILEY JOHN WILEY & SONS, INC.

Advanced Engineering Mathematics

Systems of Units. Some Important Conversion Factors

The most important systems of units are shown in the table below. The mks system is also known as the *International System of Units* (abbreviated *SI*), and the abbreviations s (instead of sec), g (instead of gm), and N (instead of nt) are also used.

System of units	Length	Mass	Time	Force
cgs system	centimeter (cm)	gram (gm)	second (sec)	dyne
mks system	meter (m)	kilogram (kg)	second (sec)	newton (nt)
Engineering system	foot (ft)	slug	second (sec)	pound (lb)

1 inch (in.) = 2.540000 cm

1 foot (ft) = 12 in. = 30.480000 cm

1 yard (yd) = 3 ft = 91.440000 cm

1 statute mile (mi) = 5280 ft = 1.609344 km

1 nautical mile = 6080 ft = 1.853184 km

1 acre = 4840 yd^2 = 4046.8564 m^2

1 mi^2 = 640 acres = 2.5899881 km^2

1 fluid ounce = 1/128 U.S. gallon = 231/128 $in.^3$ = 29.573730 cm^3

1 U.S. gallon = 4 quarts (liq) = 8 pints (liq) = 128 fl oz = 3785.4118 cm^3

1 British Imperial and Canadian gallon = 1.200949 U.S. gallons = 4546.087 cm^3

1 slug = 14.59390 kg

1 pound (lb) = 4.448444 nt

1 newton (nt) = 10^5 dynes

1 British thermal unit (Btu) = 1054.35 joules

1 joule = 10^7 ergs

1 calorie (cal) = 4.1840 joules

1 kilowatt-hour (kWh) = 3414.4 Btu = $3.6 \cdot 10^6$ joules

1 horsepower (hp) = 2542.48 Btu/h = 178.298 cal/sec = 0.74570 kW

1 kilowatt (kW) = 1000 watts = 3414.43 Btu/h = 238.662 cal/sec

$°F = °C \cdot 1.8 + 32$

$1° = 60' = 3600'' = 0.017453293$ radian

For further details see, for example, D. Halliday, R. Resnick, and J. Walker, *Fundamentals of Physics.* 7th ed., New York: Wiley, 2005. See also AN American National Standard, ASTM/IEEE Standard Metric Practice, Institute of Electrical and Electronics Engineers, Inc., 445 Hoes Lane, Piscataway, N. J. 08854.

Differentiation

$(cu)' = cu'$ (c constant)

$(u + v)' = u' + v'$

$(uv)' = u'v + v'u$

$$\left(\frac{u}{v}\right)' = \frac{u'v - v'u}{v^2}$$

$$\frac{du}{dx} = \frac{du}{dy} \cdot \frac{dy}{dx}$$ (Chain rule)

$(x^n)' = nx^{n-1}$

$(e^x)' = e^x$

$(a^x)' = a^x \ln a$

$(\sin x)' = \cos x$

$(\cos x)' = -\sin x$

$(\tan x)' = \sec^2 x$

$(\cot x)' = -\csc^2 x$

$(\sinh x)' = \cosh x$

$(\cosh x)' = \sinh x$

$$(\ln x)' = \frac{1}{x}$$

$$(\log_a x)' = \frac{\log_a e}{x}$$

$$(\arcsin x)' = \frac{1}{\sqrt{1 - x^2}}$$

$$(\arccos x)' = -\frac{1}{\sqrt{1 - x^2}}$$

$$(\arctan x)' = \frac{1}{1 + x^2}$$

$$(\operatorname{arccot} x)' = -\frac{1}{1 + x^2}$$

Integration

$$\int uv'\,dx = uv - \int u'v\,dx$$

$$\int x^n\,dx = \frac{x^{n+1}}{n+1} + c \qquad (n \neq -1)$$

$$\int \frac{1}{x}\,dx = \ln|x| + c$$

$$\int e^{ax}\,dx = \frac{1}{a}e^{ax} + c$$

$$\int \sin x\,dx = -\cos x + c$$

$$\int \cos x\,dx = \sin x + c$$

$$\int \tan x\,dx = -\ln|\cos x| + c$$

$$\int \cot x\,dx = \ln|\sin x| + c$$

$$\int \sec x\,dx = \ln|\sec x + \tan x| + c$$

$$\int \csc x\,dx = \ln|\csc x - \cot x| + c$$

$$\int \frac{dx}{x^2 + a^2} = \frac{1}{a}\arctan\frac{x}{a} + c$$

$$\int \frac{dx}{\sqrt{a^2 - x^2}} = \arcsin\frac{x}{a} + c$$

$$\int \frac{dx}{\sqrt{x^2 + a^2}} = \operatorname{arcsinh}\frac{x}{a} + c$$

$$\int \frac{dx}{\sqrt{x^2 - a^2}} = \operatorname{arccosh}\frac{x}{a} + c$$

$$\int \sin^2 x\,dx = \tfrac{1}{2}x - \tfrac{1}{4}\sin 2x + c$$

$$\int \cos^2 x\,dx = \tfrac{1}{2}x + \tfrac{1}{4}\sin 2x + c$$

$$\int \tan^2 x\,dx = \tan x - x + c$$

$$\int \cot^2 x\,dx = -\cot x - x + c$$

$$\int \ln x\,dx = x\ln x - x + c$$

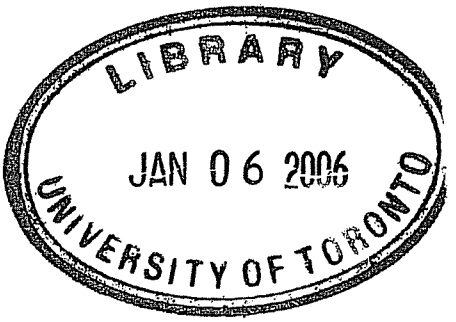
$$\int e^{ax}\sin bx\,dx = \frac{e^{ax}}{a^2 + b^2}(a\sin bx - b\cos bx) + c$$

$$\int e^{ax}\cos bx\,dx = \frac{e^{ax}}{a^2 + b^2}(a\cos bx + b\sin bx) + c$$

Vice President and Publisher: Laurie Rosatone
Editorial Assistant: Daniel Grace
Associate Production Director: Lucille Buonocore
Senior Production Editor: Ken Santor
Media Editor: Stefanie Liebman
Cover Designer: Madelyn Lesure
Cover Photo: © John Sohm/Chromosohm/Photo Researchers

This book was set in Times Roman by GGS Information Services
and printed and bound by Von Hoffmann, Inc.
The cover was printed by Von Hoffmann, Inc.

This book is printed on acid-free paper.



Copyright © 2006 John Wiley & Sons, Inc. All rights reserved.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 646-8600. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008.

To order books or for customer service please, call 1(800)-CALL-WILEY (225-5945).

Kreyszig, Erwin.
Advanced engineering mathematics / Erwin Kreyszig.—9th ed.
p. cm.
Accompanied by instructor's manual.
Includes bibliographical references and index.
ISBN 0-471-48885-2 (cloth : acid-free paper)
1. Mathematical physics. 2. Engineering mathematics. 1. Title.
ISBN-13: 978-0-471-48885-9
ISBN-10: 0-471-48885-2

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

PREFACE

See also http://www.wiley.com/college/kreyszig/

Goal of the Book. Arrangement of Material

This new edition continues the tradition of providing instructors and students with a comprehensive and up-to-date resource for teaching and learning ***engineering mathematics,*** that is, ***applied mathematics*** for engineers and physicists, mathematicians and computer scientists, as well as members of other disciplines. A course in elementary calculus is the sole ***prerequisite.***

The subject matter is arranged into seven parts A–G:

A Ordinary Differential Equations (ODEs) (Chaps. 1–6)
B Linear Algebra. Vector Calculus (Chaps. 7–9)
C Fourier Analysis. Partial Differential Equations (PDEs) (Chaps. 11–12)
D Complex Analysis (Chaps. 13–18)
E Numeric Analysis (Chaps. 19–21)
F Optimization, Graphs (Chaps. 22–23)
G Probability, Statistics (Chaps. 24–25).

This is followed by five appendices:

App. 1 References (ordered by parts)
App. 2 Answers to Odd-Numbered Problems
App. 3 Auxiliary Material (see also inside covers)
App. 4 Additional Proofs
App. 5 Tables of Functions.

This book has helped to pave the way for the present development of engineering mathematics. By a modern approach to those areas A–G, this new edition will prepare the student for the tasks of the present and of the future. The latter can be predicted to some extent by a judicious look at the present trend. Among other features, this trend shows the appearance of more complex production processes, more extreme physical conditions (in space travel, high-speed communication, etc.), and new tasks in robotics and communication systems (e.g., fiber optics and scan statistics on random graphs) and elsewhere. This requires the refinement of existing methods and the creation of new ones.

It follows that students need solid knowledge of basic principles, methods, and results, and a clear view of what engineering mathematics is all about, and that it requires proficiency in all three phases of problem solving:

- **Modeling,** that is, translating a physical or other problem into a mathematical form, into a mathematical *model;* this can be an algebraic equation, a differential equation, a graph, or some other mathematical expression.
- **Solving** the model by selecting and applying a suitable mathematical method, often requiring numeric work on a computer.
- **Interpreting** the mathematical result in physical or other terms to see what it practically means and implies.

It would make no sense to overload students with all kinds of little things that might be of occasional use. Instead they should recognize that mathematics rests on relatively few basic concepts and involves powerful unifying principles. This should give them a firm grasp on the ***interrelations among theory, computing, and*** (physical or other) ***experimentation.***

PARTS AND CHAPTERS OF THE BOOK

PART A
Chaps. 1–6
Ordinary Differential Equations (ODEs)

Chaps. 1–4
Basic Material

Chap. 5
Series Solutions

Chap. 6
Laplace Transforms

PART B
Chaps. 7–10
Linear Algebra. Vector Calculus

Chap. 7
Matrices,
Linear Systems

Chap. 8
Eigenvalue Problems

Chap. 9
Vector Differential
Calculus

Chap. 10
Vector Integral Calculus

PART C
Chaps. 11–12
Fourier Analysis. Partial Differential
Equations (PDEs)

Chap. 11
Fourier Analysis

Chap. 12
Partial Differential Equations

PART D
Chaps. 13–18
Complex Analysis,
Potential Theory

Chaps. 13–17
Basic Material

Chap. 18
Potential Theory

PART E
Chaps. 19–21
Numeric Analysis

Chap. 19
Numerics in
General

Chap. 20
Numeric
Linear Algebra

Chap. 21
Numerics for
ODEs and PDEs

PART F
Chaps. 22–23
Optimization, Graphs

Chap. 22
Linear Programming

Chap. 23
Graphs, Optimization

PART G
Chaps. 24–25
Probability, Statistics

Chap. 24
Data Analysis. Probability Theory

Chap. 25
Mathematical Statistics

GUIDES AND MANUALS

Maple Computer Guide
Mathematica Computer Guide

Student Solutions Manual

Instructor's Manual

General Features of the Book Include:

- **Simplicity of examples,** to make the book teachable—why choose complicated examples when simple ones are as instructive or even better?
- **Independence of chapters,** to provide ***flexibility*** in tailoring courses to special needs.
- **Self-contained presentation,** except for a few clearly marked places where a proof would exceed the level of the book and a reference is given instead.
- **Modern standard notation,** to help students with other courses, *modern* books, and mathematical and engineering journals.

Many sections were rewritten in a more detailed fashion, to make it a ***simpler book.*** This also resulted in a ***better balance between theory and applications.***

Use of Computers

The presentation is ***adaptable to various levels of technology and use of a computer or graphing calculator:*** very little or no use, medium use, or intensive use of a graphing calculator or of an unspecified ***CAS*** (Computer Algebra System, ***Maple, Mathematica, or Matlab*** being popular examples). In either case texts and problem sets form an entity without gaps or jumps. And many problems can be solved by hand or with a computer or both ways. (For ***software,*** see the beginnings of *Part E* on Numeric Analysis and *Part G* on Probability and Statistics.)

More specifically, this new edition on the one hand gives more prominence to tasks the computer ***cannot*** do, notably, modeling and interpreting results. On the other hand, it includes ***CAS projects, CAS problems,*** and ***CAS experiments,*** which ***do require*** a computer and show its power in solving problems that are difficult or impossible to access otherwise. Here our goal is the combination of *intelligent* computer use with high-quality mathematics. This has resulted in a change from a formula-centered teaching and learning of engineering mathematics to a more quantitative, project-oriented, and visual approach. CAS experiments also exhibit the computer as an instrument for observations and experimentations that may become the beginnings of new research, for "proving" or disproving conjectures, or for formalizing empirical relationships that are often quite useful to the engineer as working guidelines. These changes will also help the student in discovering the ***experimental aspect of modern applied mathematics.***

Some ***routine and drill work*** is retained as a necessity for keeping firm contact with the subject matter. In some of it the computer can (but must not) give the student a hand, but there are plenty of problems that are more suitable for pencil-and-paper work.

Major Changes

1. New Problem Sets. Modern engineering mathematics is mostly ***teamwork.*** It usually combines analytic work in the process of modeling and the use of computer algebra and numerics in the process of solution, followed by critical evaluation of results. Our problems—some straightforward, some more challenging, some "thinking problems" not accessible by a CAS, some open-ended—reflect this modern situation with its increased emphasis on qualitative methods and applications, and the problem sets take care of this novel situation by including team projects, CAS projects, and writing projects. The latter will also help the student in writing general reports, as they are required in engineering work quite frequently.

2. Computer Experiments, using the computer as an instrument of **"experimental mathematics"** for exploration and research (see also above). These are mostly open-ended

experiments, demonstrating the use of computers in experimentally finding results, which may be provable afterward or may be valuable heuristic qualitative guidelines to the engineer, in particular in complicated problems.

3. **More on modeling and selecting methods,** tasks that usually cannot be automated.

4. **Student Solutions Manual and Study Guide enlarged,** upon explicit requests of the users. This Manual contains worked-out solutions to carefully selected odd-numbered problems (to which App. 1 gives only the final answers) as well as general comments and hints on studying the text and working further problems, including explanations on the significance and character of concepts and methods in the various sections of the book.

Further Changes, New Features

- Electric circuits moved entirely to Chap. 2, to avoid duplication and repetition
- Second-order ODEs and Higher Order ODEs placed into two separate chapters (2 and 3)
- In Chap. 2, applications presented before variation of parameters
- Series solutions somewhat shortened, without changing the order of sections
- Material on Laplace transforms brought into a better logical order: partial fractions used earlier in a more practical approach, unit step and Dirac's delta put into separate subsequent sections, differentiation and integration of transforms (not of functions!) moved to a later section in favor of practically more important topics
- Second- and third-order determinants made into a separate section for reference throughout the book
- Complex matrices made optional
- Three sections on curves and their application in mechanics combined in a single section
- First two sections on Fourier series combined to provide a better, more direct start
- Discrete and Fast Fourier Transforms included
- Conformal mapping presented in a separate chapter and enlarged
- Numeric analysis updated
- Backward Euler method included
- Stiffness of ODEs and systems discussed
- List of software (in Part E) updated; another list for statistics software added (in Part G)
- References updated, now including about 75 books published or reprinted after 1990

Suggestions for Courses: A Four-Semester Sequence

The material, when taken in sequence, is suitable for four consecutive semester courses, meeting 3–4 hours a week:

1st Semester.	*ODEs* (Chaps. 1–5 or 6)
2nd Semester.	*Linear Algebra. Vector Analysis* (Chaps. 7–10)
3rd Semester.	*Complex Analysis* (Chaps. 13–18)
4th Semester.	*Numeric Methods* (Chaps. 19–21)

Suggestions for Independent One-Semester Courses

The book is also suitable for various independent one-semester courses meeting 3 hours a week. For instance:

Introduction to ODEs (Chaps. 1–2, Sec. 21.1)
Laplace Transforms (Chap. 6)
Matrices and Linear Systems (Chaps. 7–8)
Vector Algebra and Calculus (Chaps. 9–10)
Fourier Series and PDEs (Chaps. 11–12, Secs. 21.4–21.7)
Introduction to Complex Analysis (Chaps. 13–17)
Numeric Analysis (Chaps. 19, 21)
Numeric Linear Algebra (Chap. 20)
Optimization (Chaps. 22–23)
Graphs and Combinatorial Optimization (Chap. 23)
Probability and Statistics (Chaps. 24–25)

Acknowledgments

I am indebted to many of my former teachers, colleagues, and students who helped me directly or indirectly in preparing this book, in particular, the present edition. I profited greatly from discussions with engineers, physicists, mathematicians, and computer scientists, and from their written comments. I want to mention particularly Y. Antipov, D. N. Buechler, S. L. Campbell, R. Carr, P. L. Chambre, V. F. Connolly, Z. Davis, J. Delany, J. W. Dettman, D. Dicker, L. D. Drager, D. Ellis, W. Fox, A. Goriely, R. B. Guenther, J. B. Handley, N. Harbertson, A. Hassen, V. W. Howe, H. Kuhn, G. Lamb, M. T. Lusk, H. B. Mann, I. Marx, K. Millet, J. D. Moore, W. D. Munroe, A. Nadim, B. S. Ng, J. N. Ong, Jr., D. Panagiotis, A. Plotkin, P. J. Pritchard, W. O. Ray, J. T. Scheick, L. F. Shampine, H. A. Smith, J. Todd, H. Unz, A. L. Villone, H. J. Weiss, A. Wilansky, C. H. Wilcox, H. Ya Fan, and A. D. Ziebur, all from the United States, Professors E. J. Norminton and R. Vaillancourt from Canada, and Professors H. Florian and H. Unger from Europe. I can offer here only an inadequate acknowledgment of my gratitude and appreciation.

Special cordial thanks go to Privatdozent Dr. M. Kracht and to Mr. Herbert Kreyszig, MBA, the coauthor of the Student Solutions Manual, who both checked the manuscript in all details and made numerous suggestions for improvements and helped me proofread the galley and page proofs.

Furthermore, I wish to thank John Wiley and Sons (see the list on p. iv) as well as GGS Information Services, in particular Mr. K. Bradley and Mr. J. Nystrom, for their effective cooperation and great care in preparing this new edition.

Suggestions of many readers worldwide were evaluated in preparing this edition. Further comments and suggestions for improving the book will be gratefully received.

ERWIN KREYSZIG

CONTENTS

PART A Ordinary Differential Equations (ODEs) 1

CHAPTER 1 First-Order ODEs 2
1.1 Basic Concepts. Modeling 2
1.2 Geometric Meaning of $y' = f(x, y)$. Direction Fields 9
1.3 Separable ODEs. Modeling 12
1.4 Exact ODEs. Integrating Factors 19
1.5 Linear ODEs. Bernoulli Equation. Population Dynamics 26
1.6 Orthogonal Trajectories. *Optional* 35
1.7 Existence and Uniqueness of Solutions 37
Chapter 1 Review Questions and Problems 42
Summary of Chapter 1 43

CHAPTER 2 Second-Order Linear ODEs 45
2.1 Homogeneous Linear ODEs of Second Order 45
2.2 Homogeneous Linear ODEs with Constant Coefficients 53
2.3 Differential Operators. *Optional* 59
2.4 Modeling: Free Oscillations. (Mass–Spring System) 61
2.5 Euler–Cauchy Equations 69
2.6 Existence and Uniqueness of Solutions. Wronskian 73
2.7 Nonhomogeneous ODEs 78
2.8 Modeling: Forced Oscillations. Resonance 84
2.9 Modeling: Electric Circuits 91
2.10 Solution by Variation of Parameters 98
Chapter 2 Review Questions and Problems 102
Summary of Chapter 2 103

CHAPTER 3 Higher Order Linear ODEs 105
3.1 Homogeneous Linear ODEs 105
3.2 Homogeneous Linear ODEs with Constant Coefficients 111
3.3 Nonhomogeneous Linear ODEs 116
Chapter 3 Review Questions and Problems 122
Summary of Chapter 3 123

CHAPTER 4 Systems of ODEs. Phase Plane. Qualitative Methods 124
4.0 Basics of Matrices and Vectors 124
4.1 Systems of ODEs as Models 130
4.2 Basic Theory of Systems of ODEs 136
4.3 Constant-Coefficient Systems. Phase Plane Method 139
4.4 Criteria for Critical Points. Stability 147
4.5 Qualitative Methods for Nonlinear Systems 151
4.6 Nonhomogeneous Linear Systems of ODEs 159
Chapter 4 Review Questions and Problems 163
Summary of Chapter 4 164

CHAPTER 5 Series Solutions of ODEs. Special Functions 166
5.1 Power Series Method 167
5.2 Theory of the Power Series Method 170

5.3 Legendre's Equation. Legendre Polynomials $P_n(x)$ 177
5.4 Frobenius Method 182
5.5 Bessel's Equation. Bessel Functions $J_\nu(x)$ 189
5.6 Bessel Functions of the Second Kind $Y_\nu(x)$ 198
5.7 Sturm–Liouville Problems. Orthogonal Functions 203
5.8 Orthogonal Eigenfunction Expansions 210
Chapter 5 Review Questions and Problems 217
Summary of Chapter 5 218

CHAPTER 6 Laplace Transforms 220
6.1 Laplace Transform. Inverse Transform. Linearity. s-Shifting 221
6.2 Transforms of Derivatives and Integrals. ODEs 227
6.3 Unit Step Function. t-Shifting 233
6.4 Short Impulses. Dirac's Delta Function. Partial Fractions 241
6.5 Convolution. Integral Equations 248
6.6 Differentiation and Integration of Transforms. 254
6.7 Systems of ODEs 258
6.8 Laplace Transform: General Formulas 264
6.9 Table of Laplace Transforms 265
Chapter 6 Review Questions and Problems 267
Summary of Chapter 6 269

PART B Linear Algebra. Vector Calculus 271

CHAPTER 7 Linear Algebra: Matrices, Vectors, Determinants. Linear Systems 272
7.1 Matrices, Vectors: Addition and Scalar Multiplication 272
7.2 Matrix Multiplication 278
7.3 Linear Systems of Equations. Gauss Elimination 287
7.4 Linear Independence. Rank of a Matrix. Vector Space 296
7.5 Solutions of Linear Systems: Existence, Uniqueness 302
7.6 For Reference: Second- and Third-Order Determinants 306
7.7 Determinants. Cramer's Rule 308
7.8 Inverse of a Matrix. Gauss–Jordan Elimination 315
7.9 Vector Spaces, Inner Product Spaces. Linear Transformations. *Optional* 323
Chapter 7 Review Questions and Problems 330
Summary of Chapter 7 331

CHAPTER 8 Linear Algebra: Matrix Eigenvalue Problems 333
8.1 Eigenvalues, Eigenvectors 334
8.2 Some Applications of Eigenvalue Problems 340
8.3 Symmetric, Skew-Symmetric, and Orthogonal Matrices 345
8.4 Eigenbases. Diagonalization. Quadratic Forms 349
8.5 Complex Matrices and Forms. *Optional* 356
Chapter 8 Review Questions and Problems 362
Summary of Chapter 8 363

CHAPTER 9 Vector Differential Calculus. Grad, Div, Curl 364
9.1 Vectors in 2-Space and 3-Space 364
9.2 Inner Product (Dot Product) 371
9.3 Vector Product (Cross Product) 377
9.4 Vector and Scalar Functions and Fields. Derivatives 384
9.5 Curves. Arc Length. Curvature. Torsion 389
9.6 Calculus Review: Functions of Several Variables. *Optional* 400
9.7 Gradient of a Scalar Field. Directional Derivative 403
9.8 Divergence of a Vector Field 410
9.9 Curl of a Vector Field 414
Chapter 9 Review Questions and Problems 416
Summary of Chapter 9 417

CHAPTER 10 Vector Integral Calculus. Integral Theorems 420
10.1 Line Integrals 420
10.2 Path Independence of Line Integrals 426
10.3 Calculus Review: Double Integrals. *Optional* 433
10.4 Green's Theorem in the Plane 439
10.5 Surfaces for Surface Integrals 445
10.6 Surface Integrals 449
10.7 Triple Integrals. Divergence Theorem of Gauss 458
10.8 Further Applications of the Divergence Theorem 463
10.9 Stokes's Theorem 468
Chapter 10 Review Questions and Problems 473
Summary of Chapter 10 474

PART C Fourier Analysis. Partial Differential Equations (PDEs) 477

CHAPTER 11 Fourier Series, Integrals, and Transforms 478
11.1 Fourier Series 478
11.2 Functions of Any Period $p = 2L$ 487
11.3 Even and Odd Functions. Half-Range Expansions 490
11.4 Complex Fourier Series. *Optional* 496
11.5 Forced Oscillations 499
11.6 Approximation by Trigonometric Polynomials 502
11.7 Fourier Integral 506
11.8 Fourier Cosine and Sine Transforms 513
11.9 Fourier Transform. Discrete and Fast Fourier Transforms 518
11.10 Tables of Transforms 529
Chapter 11 Review Questions and Problems 532
Summary of Chapter 11 533

CHAPTER 12 Partial Differential Equations (PDEs) 535
12.1 Basic Concepts 535
12.2 Modeling: Vibrating String, Wave Equation 538
12.3 Solution by Separating Variables. Use of Fourier Series 540
12.4 D'Alembert's Solution of the Wave Equation. Characteristics 548
12.5 Heat Equation: Solution by Fourier Series 552

12.6 Heat Equation: Solution by Fourier Integrals and Transforms 562
12.7 Modeling: Membrane, Two-Dimensional Wave Equation 569
12.8 Rectangular Membrane. Double Fourier Series 571
12.9 Laplacian in Polar Coordinates. Circular Membrane. Fourier–Bessel Series 579
12.10 Laplace's Equation in Cylindrical and Spherical Coordinates. Potential 587
12.11 Solution of PDEs by Laplace Transforms 594
Chapter 12 Review Questions and Problems 597
Summary of Chapter 12 598

PART D Complex Analysis 601

CHAPTER 13 Complex Numbers and Functions 602
13.1 Complex Numbers. Complex Plane 602
13.2 Polar Form of Complex Numbers. Powers and Roots 607
13.3 Derivative. Analytic Function 612
13.4 Cauchy–Riemann Equations. Laplace's Equation 618
13.5 Exponential Function 623
13.6 Trigonometric and Hyperbolic Functions 626
13.7 Logarithm. General Power 630
Chapter 13 Review Questions and Problems 634
Summary of Chapter 13 635

CHAPTER 14 Complex Integration 637
14.1 Line Integral in the Complex Plane 637
14.2 Cauchy's Integral Theorem 646
14.3 Cauchy's Integral Formula 654
14.4 Derivatives of Analytic Functions 658
Chapter 14 Review Questions and Problems 662
Summary of Chapter 14 663

CHAPTER 15 Power Series, Taylor Series 664
15.1 Sequences, Series, Convergence Tests 664
15.2 Power Series 673
15.3 Functions Given by Power Series 678
15.4 Taylor and Maclaurin Series 683
15.5 Uniform Convergence. *Optional* 691
Chapter 15 Review Questions and Problems 698
Summary of Chapter 15 699

CHAPTER 16 Laurent Series. Residue Integration 701
16.1 Laurent Series 701
16.2 Singularities and Zeros. Infinity 707
16.3 Residue Integration Method 712
16.4 Residue Integration of Real Integrals 718
Chapter 16 Review Questions and Problems 726
Summary of Chapter 16 727

CHAPTER 17 Conformal Mapping 728
17.1 Geometry of Analytic Functions: Conformal Mapping 729
17.2 Linear Fractional Transformations 734
17.3 Special Linear Fractional Transformations 737

17.4 Conformal Mapping by Other Functions 742
17.5 Riemann Surfaces. *Optional* 746
Chapter 17 Review Questions and Problems 747
Summary of Chapter 17 748

CHAPTER 18 Complex Analysis and Potential Theory 749
18.1 Electrostatic Fields 750
18.2 Use of Conformal Mapping. Modeling 754
18.3 Heat Problems 757
18.4 Fluid Flow 761
18.5 Poisson's Integral Formula for Potentials 768
18.6 General Properties of Harmonic Functions 771
Chapter 18 Review Questions and Problems 775
Summary of Chapter 18 776

PART E Numeric Analysis 777
Software 778

CHAPTER 19 Numerics in General 780
19.1 Introduction 780
19.2 Solution of Equations by Iteration 787
19.3 Interpolation 797
19.4 Spline Interpolation 810
19.5 Numeric Integration and Differentiation 817
Chapter 19 Review Questions and Problems 830
Summary of Chapter 19 831

CHAPTER 20 Numeric Linear Algebra 833
20.1 Linear Systems: Gauss Elimination 833
20.2 Linear Systems: LU-Factorization, Matrix Inversion 840
20.3 Linear Systems: Solution by Iteration 845
20.4 Linear Systems: Ill-Conditioning, Norms 851
20.5 Least Squares Method 859
20.6 Matrix Eigenvalue Problems: Introduction 863
20.7 Inclusion of Matrix Eigenvalues 866
20.8 Power Method for Eigenvalues 872
20.9 Tridiagonalization and QR-Factorization 875
Chapter 20 Review Questions and Problems 883
Summary of Chapter 20 884

CHAPTER 21 Numerics for ODEs and PDEs 886
21.1 Methods for First-Order ODEs 886
21.2 Multistep Methods 898
21.3 Methods for Systems and Higher Order ODEs 902
21.4 Methods for Elliptic PDEs 909
21.5 Neumann and Mixed Problems. Irregular Boundary 917
21.6 Methods for Parabolic PDEs 922
21.7 Method for Hyperbolic PDEs 928
Chapter 21 Review Questions and Problems 930
Summary of Chapter 21 932

PART F Optimization, Graphs 935

CHAPTER 22 Unconstrained Optimization. Linear Programming 936
22.1 Basic Concepts. Unconstrained Optimization 936
22.2 Linear Programming 939
22.3 Simplex Method 944
22.4 Simplex Method: Difficulties 947
Chapter 22 Review Questions and Problems 952
Summary of Chapter 22 953

CHAPTER 23 Graphs. Combinatorial Optimization 954
23.1 Graphs and Digraphs 954
23.2 Shortest Path Problems. Complexity 959
23.3 Bellman's Principle. Dijkstra's Algorithm 963
23.4 Shortest Spanning Trees: Greedy Algorithm 966
23.5 Shortest Spanning Trees: Prim's Algorithm 970
23.6 Flows in Networks 973
23.7 Maximum Flow: Ford–Fulkerson Algorithm 979
23.8 Bipartite Graphs. Assignment Problems 982
Chapter 23 Review Questions and Problems 987
Summary of Chapter 23 989

PART G Probability, Statistics 991

CHAPTER 24 Data Analysis. Probability Theory 993
24.1 Data Representation. Average. Spread 993
24.2 Experiments, Outcomes, Events 997
24.3 Probability 1000
24.4 Permutations and Combinations 1006
24.5 Random Variables. Probability Distributions 1010
24.6 Mean and Variance of a Distribution 1016
24.7 Binomial, Poisson, and Hypergeometric Distributions 1020
24.8 Normal Distribution 1026
24.9 Distributions of Several Random Variables 1032
Chapter 24 Review Questions and Problems 1041
Summary of Chapter 24 1042

CHAPTER 25 Mathematical Statistics 1044
25.1 Introduction. Random Sampling 1044
25.2 Point Estimation of Parameters 1046
25.3 Confidence Intervals 1049
25.4 Testing Hypotheses. Decisions 1058
25.5 Quality Control 1068
25.6 Acceptance Sampling 1073
25.7 Goodness of Fit. χ^2-Test 1076
25.8 Nonparametric Tests 1080
25.9 Regression. Fitting Straight Lines. Correlation 1083
Chapter 25 Review Questions and Problems 1092
Summary of Chapter 25 1093

APPENDIX 1 References A1

APPENDIX 2 Answers to Odd-Numbered Problems A4

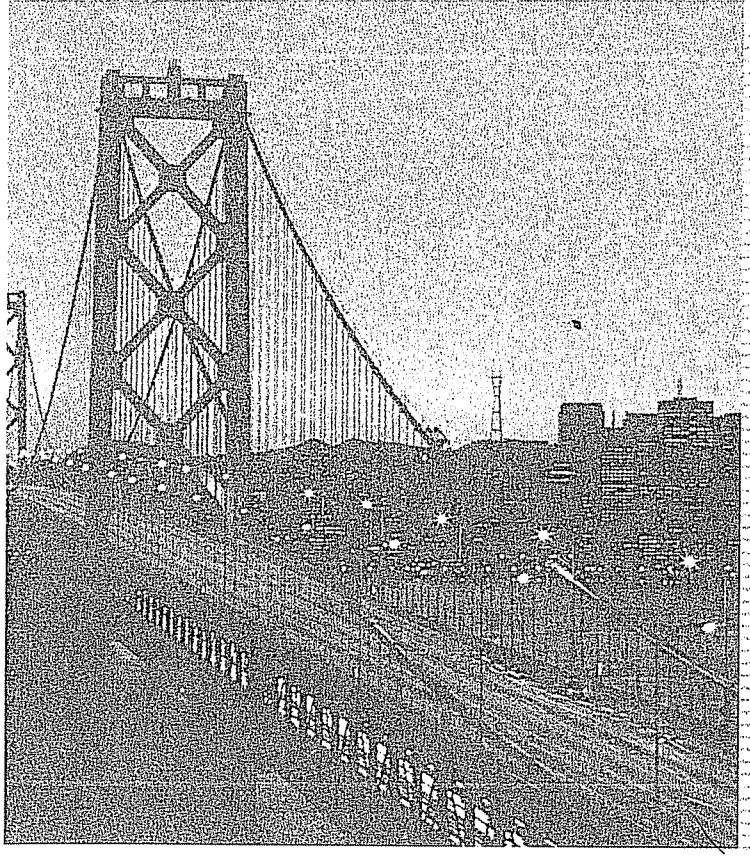
APPENDIX 3 Auxiliary Material A60
A3.1 Formulas for Special Functions A60
A3.2 Partial Derivatives A66
A3.3 Sequences and Series A69
A3.4 Grad, Div, Curl, ∇^2 in Curvilinear Coordinates A71

APPENDIX 4 Additional Proofs A74

APPENDIX 5 Tables A94

PHOTO CREDITS P1

INDEX I1



PART A

Ordinary Differential Equations (ODEs)

CHAPTER 1 First-Order ODEs

CHAPTER 2 Second-Order Linear ODEs

CHAPTER 3 Higher Order Linear ODEs

CHAPTER 4 Systems of ODEs. Phase Plane. Qualitative Methods

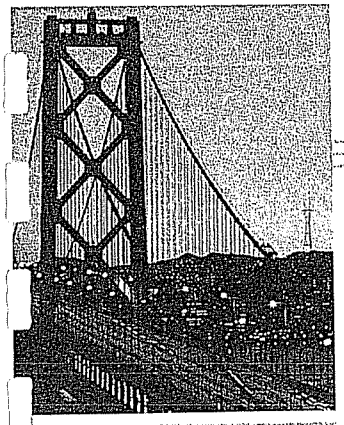
CHAPTER 5 Series Solutions of ODEs. Special Functions

CHAPTER 6 Laplace Transforms

Differential equations are of basic importance in engineering mathematics because many physical laws and relations appear mathematically in the form of a differential equation. In Part A we shall consider various physical and geometric problems that lead to differential equations, with emphasis on ***modeling,*** that is, the transition from the physical situation to a "mathematical model." In this chapter the model will be a differential equation, and as we proceed we shall explain the most important standard methods for solving such equations.

Part A concerns ***ordinary*** **differential equations (ODEs),** whose unknown functions depend on a *single* variable. *Partial* differential equations (PDEs), involving unknown functions of *several* variables, follow in Part C.

ODEs are very well suited for computers. ***Numeric methods for ODEs can be studied directly after Chaps.*** **1** ***or*** **2.** See Secs. 21.1–21.3, which are independent of the other sections on numerics.



CHAPTER 1

First-Order ODEs

In this chapter we begin our program of studying ordinary differential equations (ODEs) by deriving them from physical or other problems **(modeling),** solving them by standard methods, and interpreting solutions and their graphs in terms of a given problem. Questions of existence and uniqueness of solutions will also be discussed (in Sec. 1.7).

We begin with the simplest ODEs, called ODEs ***of the first order*** because they involve only the first derivative of the unknown function, no higher derivatives. Our usual notation for the unknown function will be $y(x)$, or $y(t)$ if the independent variable is time t.

If you wish, use your computer algebra system (CAS) for checking solutions, but make sure that you gain a conceptual understanding of the basic terms, such as ODE, direction field, and initial value problem.

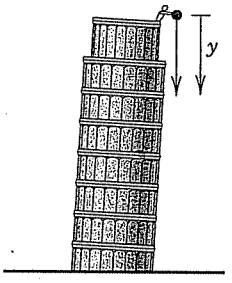
COMMENT. ***Numerics for first-order ODEs can be studied immediately after this chapter.*** See Secs. 21.1–21.2, which are independent of other sections on numerics.

Prerequisite: Integral calculus.
Sections that may be omitted in a shorter course: 1.6, 1.7.
References and Answers to Problems: App. 1 Part A, and App. 2

1.1 Basic Concepts. Modeling

If we want to solve an engineering problem (usually of a physical nature), we first have to formulate the problem as a mathematical expression in terms of variables, functions, equations, and so forth. Such an expression is known as a mathematical **model** of the given problem. The process of setting up a model, solving it mathematically, and interpreting the result in physical or other terms is called *mathematical modeling* or, briefly, **modeling.** We shall illustrate this process by various examples and problems because modeling requires experience. (Your computer may help you in solving but hardly in setting up models.)

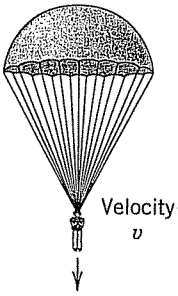
Since many physical concepts, such as velocity and acceleration, are derivatives, a model is very often an equation containing derivatives of an unknown function. Such a model is called a **differential equation.** Of course, we then want to find a solution (a function that satisfies the equation), explore its properties, graph it, find values of it, and interpret it in physical terms so that we can understand the behavior of the physical system in our given problem. However, before we can turn to methods of solution we must first define basic concepts needed throughout this chapter.



Falling stone

$y'' = g = const.$

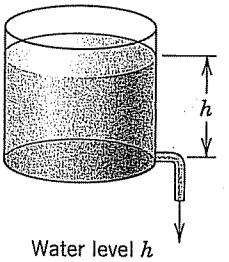
(Sec. 1.1)



Parachutist

$mv' = mg - bv^2$

(Sec. 1.2)

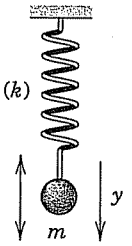


Water level h

Outflowing water

$h' = -k\sqrt{h}$

(Sec. 1.3)

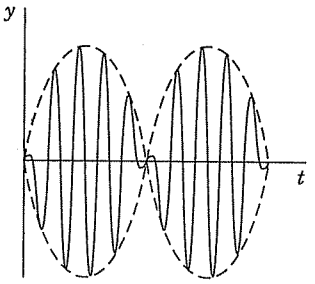


Displacement y

Vibrating mass on a spring

$my'' + ky = 0$

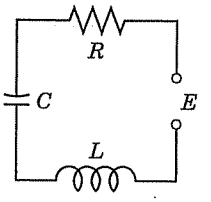
(Secs. 2.4, 2.8)



Beats of a vibrating system

$y'' + \omega_0^2 y = \cos\ \omega t, \quad \omega_0 \approx \omega$

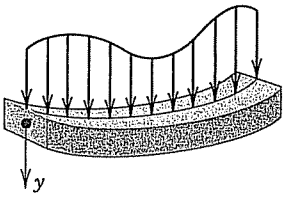
(Sec. 2.8)



Current I in an RLC circuit

$LI'' + RI' + \frac{1}{C}I = E'$

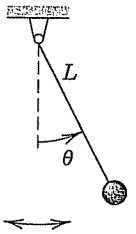
(Sec. 2.9)



Deformation of a beam

$EIy^{iv} = f(x)$

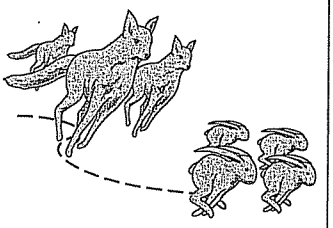
(Sec. 3.3)



Pendulum

$L\theta'' + g \sin \theta = 0$

(Sec. 4.5)



Lotka–Volterra predator–prey model

$y_1' = ay_1 - by_1y_2$

$y_2' = ky_1y_2 - ly_2$

(Sec. 4.5)

Fig. 1. Some applications of differential equations

An **ordinary differential equation (ODE)** is an equation that contains one or several derivatives of an unknown function, which we usually call $y(x)$ (or sometimes $y(t)$ if the independent variable is time t). The equation may also contain y itself, known functions of x (or t), and constants. For example,

$$y' = \cos x, \tag{1}$$

$$y'' + 9y = 0, \tag{2}$$

$$x^2y'''y' + 2e^xy'' = (x^2 + 2)y^2 \tag{3}$$

are ordinary differential equations (ODEs). The term *ordinary* distinguishes them from *partial differential equations* (PDEs), which involve partial derivatives of an unknown function of *two or more* variables. For instance, a PDE with unknown function u of two variables x and y is

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0.$$

PDEs are more complicated than ODEs; they will be considered in Chap. 12.

An ODE is said to be of **order** n if the nth derivative of the unknown function y is the highest derivative of y in the equation. The concept of order gives a useful classification into ODEs of first order, second order, and so on. Thus, (1) is of first order, (2) of second order, and (3) of third order.

In this chapter we shall consider **first-order ODEs.** Such equations contain only the first derivative y' and may contain y and any given functions of x. Hence we can write them as

$$F(x, y, y') = 0 \tag{4}$$

or often in the form

$$y' = f(x, y).$$

This is called the *explicit form,* in contrast with the *implicit form* (4). For instance, the implicit ODE $x^{-3}y' - 4y^2 = 0$ (where $x \neq 0$) can be written explicitly as $y' = 4x^3y^2$.

Concept of Solution

A function

$$y = h(x)$$

is called a **solution** of a given ODE (4) on some open interval $a < x < b$ if $h(x)$ is defined and differentiable throughout the interval and is such that the equation becomes an identity if y and y' are replaced with h and h', respectively. The curve (the graph) of h is called a **solution curve.**

Here, **open interval** $a < x < b$ means that the endpoints a and b are not regarded as points belonging to the interval. Also, $a < x < b$ includes *infinite intervals* $-\infty < x < b$, $a < x < \infty$, $-\infty < x < \infty$ (the real line) as special cases.

EXAMPLE 1 **Verification of Solution**

$y = h(x) = c/x$ (c an arbitrary constant, $x \neq 0$) is a solution of $xy' = -y$. To verify this, differentiate, $y' = h'(x) = -c/x^2$, and multiply by x to get $xy' = -c/x = -y$. Thus, $xy' = -y$, the given ODE.

EXAMPLE 2 **Solution Curves**

The ODE $y' = dy/dx = \cos x$ can be solved directly by integration on both sides. Indeed, using calculus, we obtain $y = \int \cos x \, dx = \sin x + c$, where c is an arbitrary constant. This is a ***family of solutions.*** Each value of c, for instance, 2.75 or 0 or -8, gives one of these curves. Figure 2 shows some of them, for $c = -3, -2, -1, 0, 1, 2, 3, 4$.

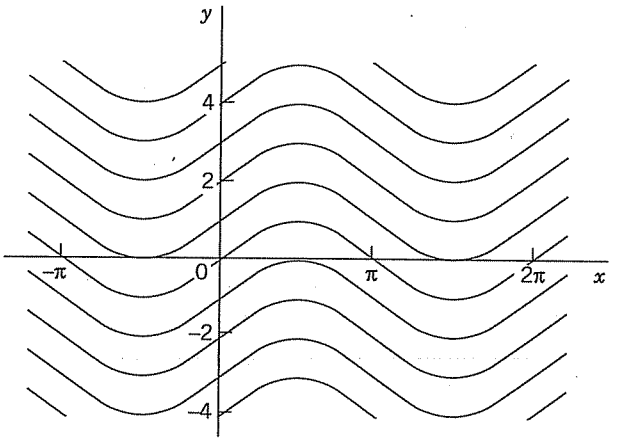


Fig. 2. Solutions $y = \sin x + c$ of the ODE $y' = \cos x$

EXAMPLE 3 **Exponential Growth, Exponential Decay**

From calculus we know that $y = ce^{3t}$ (c any constant) has the derivative (chain rule!)

$$y' = \frac{dy}{dt} = 3ce^{3t} = 3y.$$

This shows that y is a solution of $y' = 3y$. Hence this ODE can model **exponential growth,** for instance, of animal populations or colonies of bacteria. It also applies to humans for small populations in a large country (e.g., the United States in early times) and is then known as ***Malthus's law.***[1] We shall say more about this topic in Sec. 1.5.

Similarly, $y' = -0.2y$ (with a minus on the right!) has the solution $y = ce^{-0.2t}$. Hence this ODE models **exponential decay,** for instance, of a radioactive substance (see Example 5). Figure 3 shows solutions for some positive c. Can you find what the solutions look like for negative c?

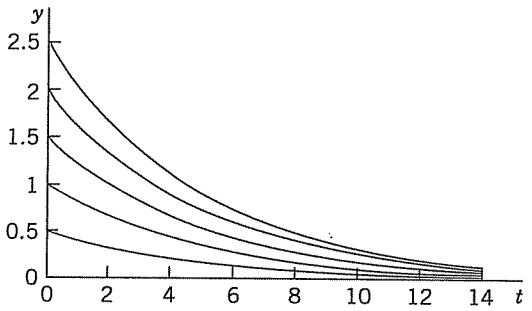


Fig. 3. Solutions of $y' = -0.2y$ in Example 3

[1]Named after the English pioneer in classic economics, THOMAS ROBERT MALTHUS (1766–1834).

We see that each ODE in these examples has a solution that contains an arbitrary constant c. Such a solution containing an arbitrary constant c is called a **general solution** of the ODE.

(We shall see that c is sometimes not completely arbitrary but must be restricted to some interval to avoid complex expressions in the solution.)

We shall develop methods that will give general solutions *uniquely* (perhaps except for notation). Hence we shall say ***the*** general solution of a given ODE (instead of ***a*** general solution).

Geometrically, the general solution of an ODE is a family of infinitely many solution curves, one for each value of the constant c. If we choose a specific c (e.g., $c = 6.45$ or 0 or -2.01) we obtain what is called a **particular solution** of the ODE. A particular solution does not contain any arbitrary constants.

In most cases, general solutions exist, and every solution not containing an arbitrary constant is obtained as a particular solution by assigning a suitable value to c. Exceptions to these rules occur but are of minor interest in applications; see Prob. 16 in Problem Set 1.1.

Initial Value Problem

In most cases the unique solution of a given problem, hence a particular solution, is obtained from a general solution by an **initial condition** $y(x_0) = y_0$, with given values x_0 and y_0, that is used to determine a value of the arbitrary constant c. Geometrically this condition means that the solution curve should pass through the point (x_0, y_0) in the xy-plane. An ODE together with an initial condition is called an **initial value problem.** Thus, if the ODE is explicit, $y' = f(x, y)$, the initial value problem is of the form

$$y' = f(x, y), \qquad y(x_0) = y_0. \tag{5}$$

EXAMPLE 4 Initial Value Problem

Solve the initial value problem

$$y' = \frac{dy}{dx} = 3y, \qquad y(0) = 5.7.$$

Solution. The general solution is $y(x) = ce^{3x}$; see Example 3. From this solution and the initial condition we obtain $y(0) = ce^0 = c = 5.7$. Hence the initial value problem has the solution $y(x) = 5.7e^{3x}$. This is a particular solution. ■

Modeling

The general importance of modeling to the engineer and physicist was emphasized at the beginning of this section. We shall now consider a basic physical problem that will show the typical steps of modeling in detail: Step 1 the transition from the physical situation (the physical system) to its mathematical formulation (its mathematical model); Step 2 the solution by a mathematical method; and Step 3 the physical interpretation of the result. This may be the easiest way to obtain a first idea of the nature and purpose of differential equations and their applications. Realize at the outset that your ***computer*** (your ***CAS***) may perhaps give you a hand in Step 2, but Steps 1 and 3 are basically your work. And Step 2

requires a solid knowledge and good understanding of solution methods available to you—*you* have to choose the method for your work by hand or by the computer. Keep this in mind, and always check computer results for errors (which may result, for instance, from false inputs).

EXAMPLE 5 Radioactivity. Exponential Decay

Given an amount of a radioactive substance, say, 0.5 g (gram), find the amount present at any later time.

Physical Information. Experiments show that at each instant a radioactive substance decomposes at a rate proportional to the the amount present.

Step 1. Setting up a mathematical model (a differential equation) of the physical process. Denote by $y(t)$ the amount of substance still present at any time t. By the physical law, the time rate of change $y'(t) = dy/dt$ is proportional to $y(t)$. Denote the constant of proportionality by k. Then

$$\frac{dy}{dt} = ky. \tag{6}$$

The value of k is known from experiments for various radioactive substances (e.g., $k = -1.4 \cdot 10^{-11}\text{sec}^{-1}$, approximately, for radium ${}_{88}\text{Ra}^{226}$). *k is negative* because $y(t)$ decreases with time. The given initial amount is 0.5 g. Denote the corresponding time by $t = 0$. Then the initial condition is $y(0) = 0.5$. This is the instant at which the process begins; this motivates the term *initial condition* (which, however, is also used more generally when the independent variable is not time or when you choose a t other than $t = 0$). Hence the model of the process is the **initial value problem**

$$\frac{dy}{dt} = ky, \qquad y(0) = 0.5. \tag{7}$$

Step 2. Mathematical solution. As in Example 3 we conclude that the ODE (6) models exponential decay and has the general solution (with arbitrary constant c but definite given k)

$$y(t) = ce^{kt}. \tag{8}$$

We now use the initial condition to determine c. Since $y(0) = c$ from (8), this gives $y(0) = c = 0.5$. Hence the particular solution governing this process is

$$y(t) = 0.5e^{kt} \tag{9}$$

(Fig. 4).

Always check your result—it may involve human or computer errors! Verify by differentiation (chain rule!) that your solution (9) satisfies (7) as well as $y(0) = 0.5$:

$$\frac{dy}{dt} = 0.5ke^{kt} = k \cdot 0.5e^{kt} = ky, \qquad y(0) = 0.5e^0 = 0.5.$$

Step 3. Interpretation of result. Formula (9) gives the amount of radioactive substance at time t. It starts from the correct given initial amount and decreases with time because k (the constant of proportionality, depending on the kind of substance) is negative. The limit of y as $t \to \infty$ is zero. ■

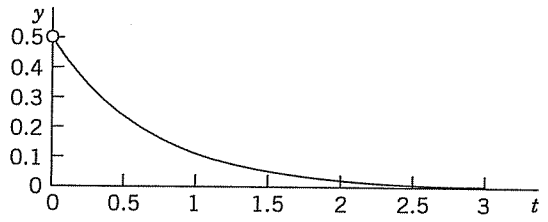


Fig. 4. Radioactivity (Exponential decay, $y = 0.5\,e^{kt}$, with $k = -1.5$ as an example)

EXAMPLE 6 A Geometric Application

Geometric problems may also lead to initial value problems. For instance, find the curve through the point (1, 1) in the xy-plane having at each of its points the slope $-y/x$.

Solution. The slope y' should equal $-y/x$. This gives the ODE $y' = -y/x$. Its general solution is $y = c/x$ (see Example 1). This is a family of hyperbolas with the coordinate axes as asymptotes.

Now, for the curve to pass through (1, 1), we must have $y = 1$ when $x = 1$. Hence the initial condition is $y(1) = 1$. From this condition and $y = c/x$ we get $y(1) = c/1 = 1$; that is, $c = 1$. This gives the particular solution $y = 1/x$ (drawn somewhat thicker in Fig. 5). ▩

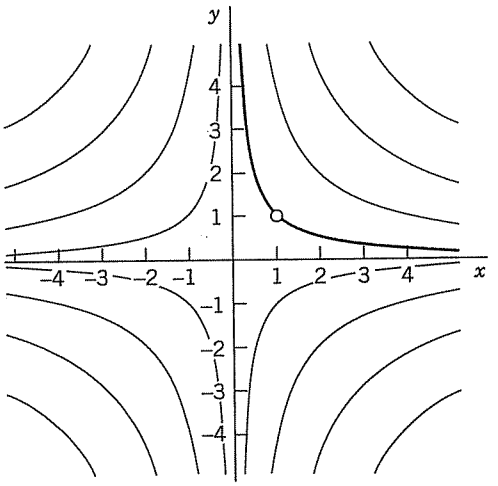


Fig. 5. Solutions of $y' = -y/x$ (hyperbolas)

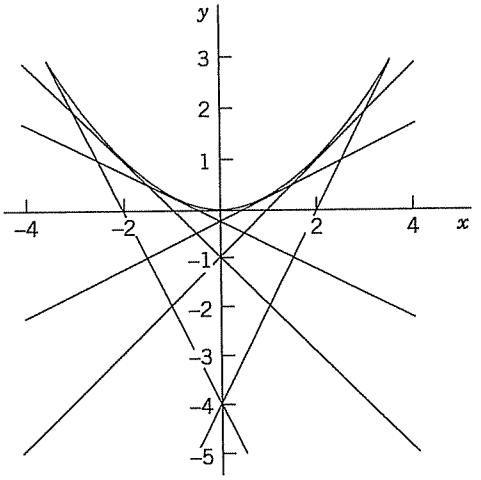


Fig. 6. Particular solutions and singular solution in Problem 16

PROBLEM SET 1.1

1–4 CALCULUS

Solve the ODE by integration.

1. $y' = -\sin \pi x$
2. $y' = e^{-3x}$
3. $y' = xe^{x^2/2}$
4. $y' = \cosh 4x$

5–9 VERIFICATION OF SOLUTION

State the order of the ODE. Verify that the given function is a solution. (a, b, c are arbitrary constants.)

5. $y' = 1 + y^2$, $y = \tan (x + c)$
6. $y'' + \pi^2 y = 0$, $y = a \cos \pi x + b \sin \pi x$
7. $y'' + 2y' + 10y = 0$, $y = 4e^{-x} \sin 3x$
8. $y' + 2y = 4(x + 1)^2$, $y = 5e^{-2x} + 2x^2 + 2x + 1$
9. $y''' = \cos x$, $y = -\sin x + ax^2 + bx + c$

10–14 INITIAL VALUE PROBLEMS

Verify that y is a solution of the ODE. Determine from y the particular solution satisfying the given initial condition. Sketch or graph this solution.

10. $y' = 0.5y$, $y = ce^{0.5x}$, $y(2) = 2$
11. $y' = 1 + 4y^2$, $y = \frac{1}{2} \tan (2x + c)$, $y(0) = 0$
12. $y' = y - x$, $y = ce^x + x + 1$, $y(0) = 3$
13. $y' + 2xy = 0$, $y = ce^{-x^2}$, $y(1) = 1/e$
14. $y' = y \tan x$, $y = c \sec x$, $y(0) = \frac{1}{2}\pi$

15. **(Existence)** **(A)** Does the ODE $y'^2 = -1$ have a (real) solution?

    (B) Does the ODE $|y'| + |y| = 0$ have a general solution?

16. **(Singular solution)** An ODE may sometimes have an additional solution that cannot be obtained from the general solution and is then called a *singular solution.* The ODE $y'^2 - xy' + y = 0$ is of the kind. Show by differentiation and substitution that it has the general solution $y = cx - c^2$ and the singular solution $y = x^2/4$. Explain Fig. 6.

17–22 MODELING, APPLICATIONS

The following problems will give you a first impression of modeling. Many more problems on modeling follow throughout this chapter.

17. **(Falling body)** If we drop a stone, we can assume air resistance ("drag") to be negligible. Experiments show that under that assumption the acceleration $y'' = d^2y/dt^2$ of this motion is constant (equal to the so-called acceleration of gravity $g = 9.80$ m/sec^2 $= 32$ ft/sec^2). State this as an ODE for $y(t)$, the distance fallen as a function of time t. Solve the ODE to get the familiar law of free fall, $y = gt^2/2$.

18. (Falling body) If in Prob. 17 the stone starts at $t = 0$ from initial position y_0 with initial velocity $v = v_0$, show that the solution is $y = gt^2/2 + v_0 t + y_0$. How long does a fall of 100 m take if the body falls from rest? A fall of 200 m? (Guess first.)

19. (Airplane takeoff) If an airplane has a run of 3 km, starts with a speed 6 m/sec, moves with constant acceleration, and makes the run in 1 min, with what speed does it take off?

20. (Subsonic flight) The efficiency of the engines of subsonic airplanes depends on air pressure and usually is maximum near about 36 000 ft. Find the air pressure $y(x)$ at this height without calculation. *Physical information.* The rate of change $y'(x)$ is proportional to the pressure, and at 18 000 ft the pressure has decreased to half its value y_0 at sea level.

21. (Half-life) The half-life of a radioactive substance is the time in which half of the given amount disappears. Hence it measures the rapidity of the decay. What is the half-life of radium ${}_{88}\mathrm{Ra}^{226}$ (in years) in Example 5?

22. (Interest rates) Show by algebra that the investment $y(t)$ from a deposit y_0 after t years at an interest rate r is

$$y_a(t) = y_0[1 + r]^t \quad \text{(Interest compounded annually)}$$

$$y_d(t) = y_0[1 + (r/365)]^{365t}$$

(Interest compounded daily).

Recall from calculus that

$$[1 + (1/n)]^n \to e \text{ as } n \to \infty;$$

hence $[1 + (r/n)]^{nt} \to e^{rt}$; thus

$$y_c(t) = y_0 e^{rt} \quad \text{(Interest compounded continuously).}$$

What ODE does the last function satisfy? Let the initial investment be \$1000 and $r = 6\%$. Compute the value of the investment after 1 year and after 5 years using each of the three formulas. Is there much difference?

1.2 Geometric Meaning of $y' = f(x, y)$. Direction Fields

A first-order ODE

$$y' = f(x, y) \tag{1}$$

has a simple geometric interpretation. From calculus you know that the derivative $y'(x)$ of $y(x)$ is the slope of $y(x)$. Hence a solution curve of (1) that passes through a point (x_0, y_0) must have at that point the slope $y'(x_0)$ equal to the value of f at that point; that is,

$$y'(x_0) = f(x_0, y_0).$$

Read this paragraph again before you go on, and think about it.

It follows that you can indicate directions of solution curves of (1) by drawing short straight-line segments (*lineal elements*) in the xy-plane (as in Fig. 7a) and then fitting (approximate) solution curves through the direction field (or *slope field*) thus obtained. This method is important for two reasons.

1. You need not solve (1). This is essential because many ODEs have complicated solution formulas or none at all.
2. The method shows, in graphical form, the whole family of solutions and their typical properties. The accuracy is somewhat limited, but in most cases this does not matter.

Let us illustrate this method for the ODE

$$y' = xy. \tag{2}$$

Direction Fields by a CAS (Computer Algebra System). A CAS plots lineal elements at the points of a square grid, as in Fig. 7a for (2), into which you can fit solution curves. Decrease the mesh size of the grid in regions where $f(x, y)$ varies rapidly.

Direction Fields by Using Isoclines (the Older Method). Graph the curves $f(x, y) = k = const$, called *isoclines* (meaning *curves of equal inclination*). For (2) these are the hyperbolas $f(x, y) = xy = k = const$ (and the coordinate axes) in Fig. 7b. By (1), these are the curves along which the derivative y' is constant. These are not yet solution curves—don't get confused. Along each isocline draw many parallel line elements of the corresponding slope k. This gives the direction field, into which you can now graph approximate solution curves.

We mention that for the ODE (2) in Fig. 7 we would not need the method, because we shall see in the next section that ODEs such as (2) can easily be solved exactly. For the time being, let us verify by substitution that (2) has the general solution

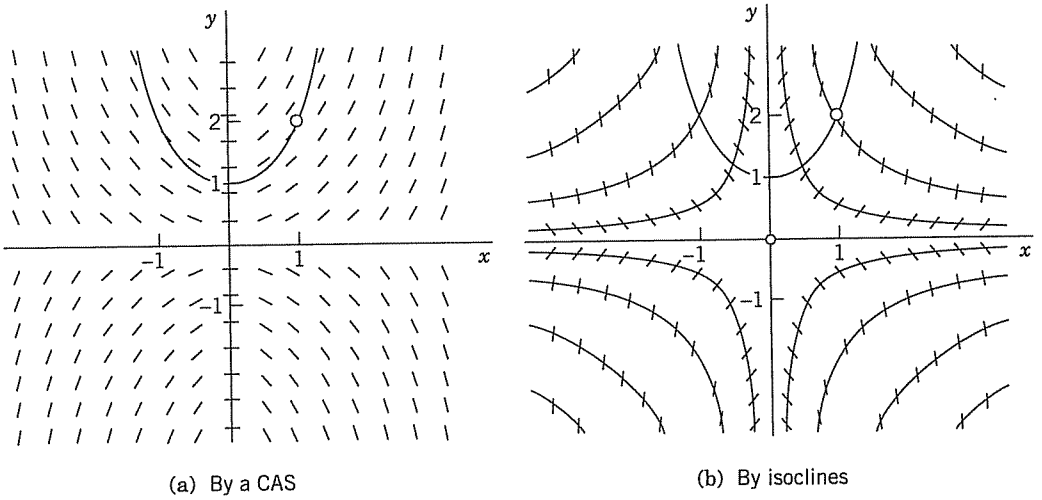
$$y(x) = ce^{x^2/2} \qquad (c \text{ arbitrary}).$$

Indeed, by differentiation (chain rule!) we get $y' = x(ce^{x^2/2}) = xy$. Of course, knowing the solution, we now have the advantage of obtaining a feel for the accuracy of the method by comparing with the exact solution. The particular solution in Fig. 7 through $(x, y) = (1, 2)$ must satisfy $y(1) = 2$. Thus, $2 = ce^{1/2}$, $c = 2/\sqrt{e} = 1.213$, and the particular solution is $y(x) = 1.213e^{x^2/2}$.

A famous ODE for which we *do* need direction fields is

$$y' = 0.1(1 - x^2) - \frac{x}{y}. \tag{3}$$

(It is related to the van der Pol equation of electronics, which we shall discuss in Sec. 4.5.) The direction field in Fig. 8 shows lineal elements generated by the computer. We have also added the isoclines for $k = -5, -3, \frac{1}{4}, 1$ as well as three typical solution curves, one that is (almost) a circle and two spirals approaching it from inside and outside.



(a) By a CAS (b) By isoclines

Fig. 7. Direction field of $y' = xy$

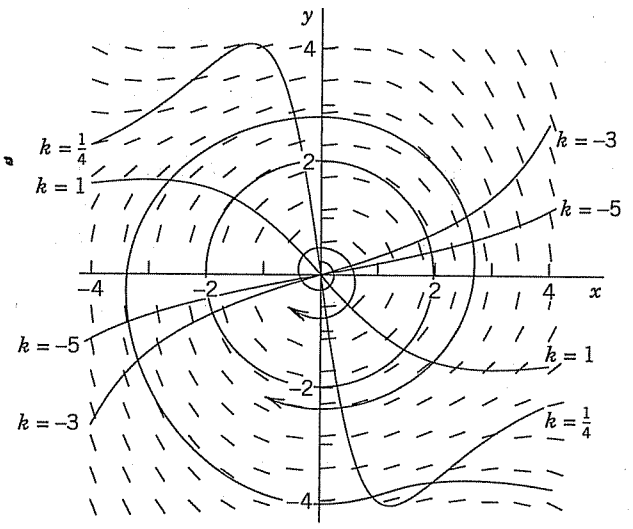


Fig. 8. Direction field of $y' = 0.1(1 - x^2) - \dfrac{x}{y}$

On Numerics

Direction fields give "all" solutions, but with limited accuracy. If we need accurate numeric values of a solution (or of several solutions) for which we have no formula, we can use a **numeric method.** If you want to get an idea of how these methods work, go to Sec. 21.1 and study the first two pages on the **Euler–Cauchy method,** which is typical of more accurate methods later in that section, notably of the classical **Runge–Kutta method.** It would make little sense to interrupt the present flow of ideas by including such methods here; indeed, it would be a duplication of the material in Sec. 21.1. For an excursion to that section you need no extra prerequisites; Sec. 1.1 just discussed is sufficient.

PROBLEM SET 1.2

1–10 DIRECTION FIELDS, SOLUTION CURVES

Graph a direction field (by a CAS or by hand). In the field graph approximate solution curves through the given point or points (x, y) by hand.

1. $y' = e^x - y$, (0, 0), (0, 1)
2. $4yy' = -9x$, (2, 2)
3. $y' = 1 + y^2$, $(\frac{1}{4}\pi, 1)$
4. $y' = y - 2y^2$, (0, 0), (0, 0.25), (0, 0.5), (0, 1)
5. $y' = x^2 - 1/y$, (1, −2)
6. $y' = 1 + \sin y$, (−1, 0), (1, −4)
7. $y' = y^3 + x^3$, (0, 1)
8. $y' = 2xy + 1$, (−1, 2), (0, 0), (1, −2)
9. $y' = y \tanh x - 2$, (−1, −2), (1, 0), (1, 2)
10. $y' = e^{y/x}$, (1, 1), (2, 2), (3, 3)

11–15 ACCURACY

Direction fields are very useful because you can see solutions (as many as you want) without solving the ODE, which may be difficult or impossible in terms of a formula. To get a feel for the accuracy of the method, graph a field, sketch solution curves in it, and compare them with the exact solutions.

11. $y' = \sin \frac{1}{2}\pi x$
12. $y' = 1/x^2$
13. $y' = -2y$ (Sol. $y = ce^{-2x}$)
14. $y' = 3y/x$ (Sol. $y = cx^3$)
15. $y' = -\ln x$

16–18 MOTIONS

A body moves on a straight line, with velocity as given, and $y(t)$ is its distance from a fixed point 0 and t time. Find a model of the motion (an ODE). Graph a direction field.

In it sketch a solution curve corresponding to the given initial condition.

16. Velocity equal to the reciprocal of the distance, $y(1) = 1$

17. Product of velocity and distance equal to $-t$, $y(3) = -3$

18. Velocity plus distance equal to the square of time, $y(0) = 6$

19. (Skydiver) Two forces act on a parachutist, the attraction by the earth mg (m = mass of person plus equipment, $g = 9.8$ m/sec^2 the acceleration of gravity) and the air resistance, assumed to be proportional to the square of the velocity $v(t)$. Using **Newton's second law** of motion (mass × acceleration = resultant of the forces), set up a model (an ODE for $v(t)$). Graph a direction field (choosing m and the constant of proportionality equal to 1). Assume that the parachute opens when $v = 10$ m/sec. Graph the corresponding solution in the field. What is the limiting velocity?

20. CAS PROJECT. Direction Fields. Discuss direction fields as follows.

(a) Graph a direction field for the ODE $y' = 1 - y$ and in it the solution satisfying $y(0) = 5$ showing **exponential approach.** Can you see the limit of any solution directly from the ODE? For what initial condition will the solution be increasing? Constant? Decreasing?

(b) What do the solution curves of $y' = -x^3/y^3$ look like, as concluded from a direction field. How do they seem to differ from circles? What are the isoclines? What happens to those curves when you drop the minus on the right? Do they look similar to familiar curves? First, guess.

(c) Compare, as best as you can, the old and the computer methods, their advantages and disadvantages. Write a short report.

1.3 Separable ODEs. Modeling

Many practically useful ODEs can be reduced to the form

$$g(y)y' = f(x) \tag{1}$$

by purely algebraic manipulations. Then we can integrate on both sides with respect to x, obtaining

$$\int g(y)\, y'\, dx = \int f(x)\, dx + c. \tag{2}$$

On the left we can switch to y as the variable of integration. By calculus, $y'\, dx = dy$, so that

$$\int g(y)\, dy = \int f(x)\, dx + c. \tag{3}$$

If f and g are continuous functions, the integrals in (3) exist, and by evaluating them we obtain a general solution of (1). This method of solving ODEs is called the **method of separating variables,** and (1) is called a **separable equation,** because in (3) the variables are now separated: x appears only on the right and y only on the left.

EXAMPLE 1 A Separable ODE

The ODE $y' = 1 + y^2$ is separable because it can be written

$$\frac{dy}{1 + y^2} = dx. \qquad \text{By integration,} \qquad \arctan y = x + c \qquad \text{or} \qquad y = \tan(x + c).$$

It is very important to introduce the constant of integration immediately when the integration is performed. If we wrote arctan $y = x$, then $y = \tan x$, and *then* introduced c, we would have obtained $y = \tan x + c$, which is not a solution (when $c \neq 0$). Verify this. ■

Modeling

The importance of modeling was emphasized in Sec. 1.1, and separable equations yield various useful models. Let us discuss this in terms of some typical examples.

EXAMPLE 2 **Radiocarbon Dating**[2]

In September 1991 the famous Iceman (Oetzi), a mummy from the Neolithic period of the Stone Age found in the ice of the Oetztal Alps (hence the name "Oetzi") in Southern Tyrolia near the Austrian–Italian border, caused a scientific sensation. When did Oetzi approximately live and die if the ratio of carbon ${}_6C^{14}$ to carbon ${}_6C^{12}$ in this mummy is 52.5% of that of a living organism?

Physical Information. In the atmosphere and in living organisms, the ratio of radioactive carbon ${}_6C^{14}$ (made radioactive by cosmic rays) to ordinary carbon ${}_6C^{12}$ is constant. When an organism dies, its absorption of ${}_6C^{14}$ by breathing and eating terminates. Hence one can estimate the age of a fossil by comparing the radioactive carbon ratio in the fossil with that in the atmosphere. To do this, one needs to know the half-life of ${}_6C^{14}$, which is 5715 years (*CRC Handbook of Chemistry and Physics,* 83rd ed., Boca Raton: CRC Press, 2002, page 11–52, line 9).

Solution. ***Modeling.*** Radioactive decay is governed by the ODE $y' = ky$ (see Sec. 1.1, Example 5). By separation and integration (where t is time and y_0 is the initial ratio of ${}_6C^{14}$ to ${}_6C^{12}$)

$$\frac{dy}{y} = k\,dt, \qquad \ln|y| = kt + c, \qquad y = y_0 e^{kt}.$$

Next we use the half-life $H = 5715$ to determine k. When $t = H$, half of the original substance is still present. Thus,

$$y_0 e^{kH} = 0.5y_0, \qquad e^{kH} = 0.5, \qquad k = \frac{\ln 0.5}{H} = -\frac{0.693}{5715} = -0.0001213.$$

Finally, we use the ratio 52.5% for determining the time t when Oetzi died (actually, was killed),

$$e^{kt} = e^{-0.0001213t} = 0.525, \qquad t = \frac{\ln 0.525}{-0.0001213} = 5312. \qquad \textit{Answer:} \quad \text{About 5300 years ago.}$$

Other methods show that radiocarbon dating values are usually too small. According to recent research, this is due to a variation in that carbon ratio because of industrial pollution and other factors, such as nuclear testing. ■

EXAMPLE 3 **Mixing Problem**

Mixing problems occur quite frequently in chemical industry. We explain here how to solve the basic model involving a single tank. The tank in Fig. 9 contains 1000 gal of water in which initially 100 lb of salt is dissolved. Brine runs in at a rate of 10 gal/min, and each gallon contains 5 lb of dissoved salt. The mixture in the tank is kept uniform by stirring. Brine runs out at 10 gal/min. Find the amount of salt in the tank at any time t.

Solution. ***Step 1. Setting up a model.*** Let $y(t)$ denote the amount of salt in the tank at time t. Its time rate of change is

$$y' = \text{Salt inflow rate} - \text{Salt outflow rate} \qquad \textbf{"Balance law".}$$

5 lb times 10 gal gives an inflow of 50 lb of salt. Now, the outflow is 10 gal of brine. This is 10/1000 = 0.01 (= 1%) of the total brine content in the tank, hence 0.01 of the salt content $y(t)$, that is, $0.01y(t)$. Thus the model is the ODE

$$y' = 50 - 0.01y = -0.01(y - 5000). \tag{4}$$

[2]Method by WILLARD FRANK LIBBY (1908–1980), American chemist, who was awarded for this work the 1960 Nobel Prize in chemistry.

Step 2. Solution of the model. The ODE (4) is separable. Separation, integration, and taking exponents on both sides gives

$$\frac{dy}{y - 5000} = -0.01\,dt, \qquad \ln|y - 5000| = -0.01t + c^*, \qquad y - 5000 = ce^{-0.01t}.$$

Initially the tank contains 100 lb of salt. Hence $y(0) = 100$ is the initial condition that will give the unique solution. Substituting $y = 100$ and $t = 0$ in the last equation gives $100 - 5000 = ce^0 = c$. Hence $c = -4900$. Hence the amount of salt in the tank at time t is

$$y(t) = 5000 - 4900e^{-0.01t}. \tag{5}$$

This function shows an exponential approach to the limit 5000 lb; see Fig. 9. Can you explain physically that $y(t)$ should increase with time? That its limit is 5000 lb? Can you see the limit directly from the ODE?

The model discussed becomes more realistic in problems on pollutants in lakes (see Problem Set 1.5, Prob. 27) or drugs in organs. These types of problems are more difficult because the mixing may be imperfect and the flow rates (in and out) may be different and known only very roughly. ■

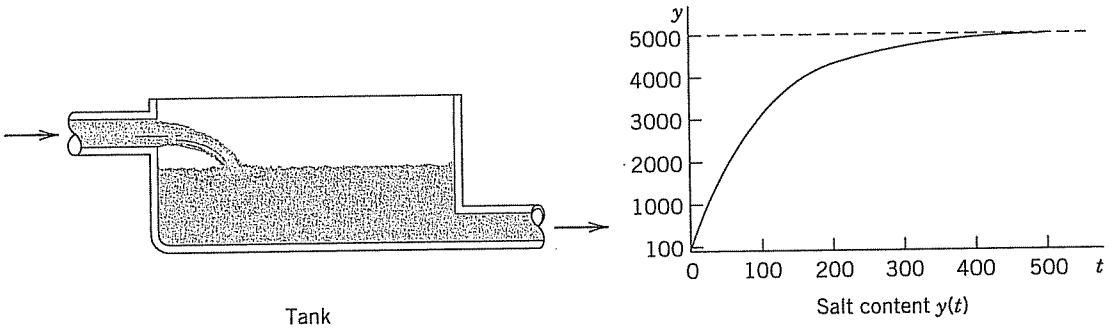


Fig. 9. Mixing problem in Example 3

EXAMPLE 4 **Heating an Office Building (Newton's Law of Cooling[3])**

Suppose that in Winter the daytime temperature in a certain office building is maintained at 70°F. The heating is shut off at 10 P.M. and turned on again at 6 A.M. On a certain day the temperature inside the building at 2 A.M. was found to be 65°F. The outside temperature was 50°F at 10 P.M. and had dropped to 40°F by 6 A.M. What was the temperature inside the building when the heat was turned on at 6 A.M.?

Physical information. Experiments show that the time rate of change of the temperature T of a body B (which conducts heat well, as, for example, a copper ball does) is proportional to the difference between T and the temperature of the surrounding medium **(Newton's law of cooling).**

Solution. ***Step 1. Setting up a model.*** Let $T(t)$ be the temperature inside the building and T_A the outside temperature (assumed to be constant in Newton's law). Then by Newton's law,

$$\frac{dT}{dt} = k(T - T_A). \tag{6}$$

Such experimental laws are derived under idealized assumptions that rarely hold exactly. However, even if a model seems to fit the reality only poorly (as in the present case), it may still give valuable qualitative information. To see how good a model is, the engineer will collect experimental data and compare them with calculations from the model.

---

[3]Sir ISAAC NEWTON (1642–1727), great English physicist and mathematician, became a professor at Cambridge in 1669 and Master of the Mint in 1699. He and the German mathematician and philosopher GOTTFRIED WILHELM LEIBNIZ (1646–1716) invented (independently) the differential and integral calculus. Newton discovered many basic physical laws and created the method of investigating physical problems by means of calculus. His *Philosophiae naturalis principia mathematica* (*Mathematical Principles of Natural Philosophy,* 1687) contains the development of classical mechanics. His work is of greatest importance to both mathematics and physics.

Step 2. General solution. We cannot solve (6) because we do not know T_A, just that it varied between 50°F and 40°F, so we follow the ***Golden Rule:*** *If you cannot solve your problem, try to solve a simpler one.* We solve (6) with the unknown function T_A replaced with the average of the two known values, or 45°F. For physical reasons we may expect that this will give us a reasonable approximate value of T in the building at 6 A.M.

For constant $T_A = 45$ (or any other *constant* value) the ODE (6) is separable. Separation, integration, and taking exponents gives the general solution

$$\frac{dT}{T-45} = k\,dt, \qquad \ln|T-45| = kt + c^*, \qquad T(t) = 45 + ce^{kt} \qquad (c = e^{c^*}).$$

Step 3. Particular solution. We choose 10 P.M. to be $t = 0$. Then the given initial condition is $T(0) = 70$ and yields a particular solution, call it T_p. By substitution,

$$T(0) = 45 + ce^0 = 70, \qquad c = 70 - 45 = 25, \qquad T_p(t) = 45 + 25e^{kt}.$$

Step 4. Determination of k. We use $T(4) = 65$, where $t = 4$ is 2 A.M. Solving algebraically for k and inserting k into $T_p(t)$ gives (Fig. 10)

$$T_p(4) = 45 + 25e^{4k} = 65, \qquad e^{4k} = 0.8, \qquad k = \tfrac{1}{4}\ln 0.8 = -0.056, \qquad T_p(t) = 45 + 25e^{-0.056t}.$$

Step 5. Answer and interpretation. 6 A.M. is $t = 8$ (namely, 8 hours after 10 P.M.), and

$$T_p(8) = 45 + 25e^{-0.056 \cdot 8} = 61[°\text{F}].$$

Hence the temperature in the building dropped 9°F, a result that looks reasonable. ▩

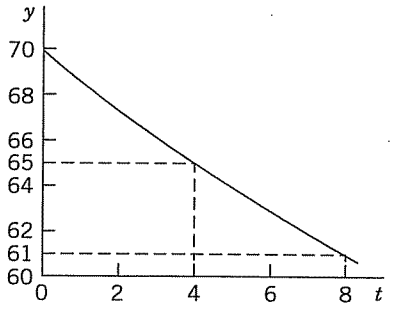


Fig. 10. Particular solution (temperature) in Example 4

EXAMPLE 5 **Leaking Tank. Outflow of Water Through a Hole (Torricelli's Law)**

This is another prototype engineering problem that leads to an ODE. It concerns the outflow of water from a cylindrical tank with a hole at the bottom (Fig. 11). You are asked to find the height of the water in the tank at any time if the tank has diameter 2 m, the hole has diameter 1 cm, and the initial height of the water when the hole is opened is 2.25 m. When will the tank be empty?

Physical information. Under the influence of gravity the outflowing water has velocity

$$v(t) = 0.600\sqrt{2gh(t)} \qquad \textbf{(Torricelli's law}^4\textbf{)}, \tag{7}$$

where $h(t)$ is the height of the water above the hole at time t, and $g = 980$ cm/sec^2 $= 32.17$ ft/sec^2 is the acceleration of gravity at the surface of the earth.

Solution. ***Step 1. Setting up the model.*** To get an equation, we relate the decrease in water level $h(t)$ to the outflow. The volume ΔV of the outflow during a short time Δt is

$$\Delta V = Av\,\Delta t \qquad (A = \text{Area of hole}).$$

[4]EVANGELISTA TORRICELLI (1608–1647), Italian physicist, pupil and successor of GALILEO GALILEI (1564–1642) at Florence. The "contraction factor" 0.600 was introduced by J. C. BORDA in 1766 because the stream has a smaller cross section than the area of the hole.

ΔV must equal the change ΔV^* of the volume of the water in the tank. Now

$$\Delta V^* = -B\,\Delta h \qquad (B = \text{Cross-sectional area of tank})$$

where Δh (> 0) is the decrease of the height $h(t)$ of the water. The minus sign appears because the volume of the water in the tank decreases. Equating ΔV and ΔV^* gives

$$-B\,\Delta h = Av\,\Delta t.$$

We now express v according to Torricelli's law and then let Δt (the length of the time interval considered) approach 0—this is a ***standard way*** of obtaining an ODE as a model. That is, we have

$$\frac{\Delta h}{\Delta t} = -\frac{A}{B}\,v = -\frac{A}{B}\,0.600\sqrt{2gh(t)},$$

and by letting $\Delta t \to 0$ we obtain the ODE

$$\frac{dh}{dt} = -26.56\,\frac{A}{B}\,\sqrt{h},$$

where $26.56 = 0.600\,\sqrt{2 \cdot 980}$. This is our model, a first-order ODE.

Step 2. General solution. Our ODE is separable. A/B is constant. Separation and integration gives

$$\frac{dh}{\sqrt{h}} = -26.56\,\frac{A}{B}\,dt \qquad \text{and} \qquad 2\sqrt{h} = c^* - 26.56\,\frac{A}{B}\,t.$$

Dividing by 2 and squaring gives $h = (c - 13.28At/B)^2$. Inserting $13.28A/B = 13.28 \cdot 0.5^2\pi/100^2\pi = 0.000332$ yields the general solution

$$h(t) = (c - 0.000332t)^2.$$

Step 3. Particular solution. The initial height (the initial condition) is $h(0) = 225$ cm. Substitution of $t = 0$ and $h = 225$ gives from the general solution $c^2 = 225$, $c = 15.00$ and thus the particular solution (Fig. 11)

$$h_p(t) = (15.00 - 0.000332t)^2.$$

Step 4. Tank empty. $h_p(t) = 0$ if $t = 15.00/0.000332 = 45\,181$ [sec] $= 12.6$ [hours].

Here you see distinctly the ***importance of the choice of units***—we have been working with the Cgs system, in which time is measured in seconds! We used $g = 980$ cm/sec^2.

Step 5. Checking. Check the result.

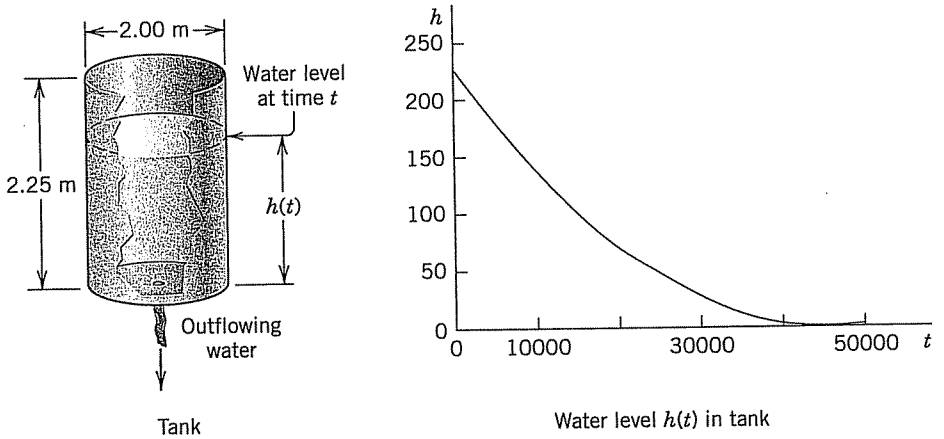


Fig. 11. Example 5. Outflow from a cylindrical tank ("leaking tank"). Torricelli's law

Extended Method: Reduction to Separable Form

Certain nonseparable ODEs can be made separable by transformations that introduce for y a new unknown function. We discuss this technique for a class of ODEs of practical

importance, namely, for equations

(8) $$y' = f\left(\frac{y}{x}\right).$$

Here, f is any (differentiable) function of y/x, such as $\sin(y/x)$, $(y/x)^4$, and so on. (Such an ODE is sometimes called a *homogeneous ODE*, a term we shall not use but reserve for a more important purpose in Sec. 1.5.)

The form of such an ODE suggests that we set $y/x = u$; thus,

(9) $$y = ux \qquad \text{and by product differentiation} \qquad y' = u'x + u.$$

Substitution into $y' = f(y/x)$ then gives $u'x + u = f(u)$ or $u'x = f(u) - u$. We see that this can be separated:

(10) $$\frac{du}{f(u) - u} = \frac{dx}{x}.$$

EXAMPLE 6 Reduction to Separable Form

Solve

$$2xyy' = y^2 - x^2.$$

Solution. To get the usual explicit form, divide the given equation by $2xy$,

$$y' = \frac{y^2 - x^2}{2xy} = \frac{y}{2x} - \frac{x}{2y}.$$

Now substitute y and y' from (9) and then simplify by subtracting u on both sides,

$$u'x + u = \frac{u}{2} - \frac{1}{2u}, \qquad u'x = -\frac{u}{2} - \frac{1}{2u} = \frac{-u^2 - 1}{2u}.$$

You see that in the last equation you can now separate the variables,

$$\frac{2u\,du}{1 + u^2} = -\frac{dx}{x}. \qquad \text{By integration,} \qquad \ln(1 + u^2) = -\ln|x| + c^* = \ln\left|\frac{1}{x}\right| + c^*.$$

Take exponents on both sides to get $1 + u^2 = c/x$ or $1 + (y/x)^2 = c/x$. Multiply the last equation by x^2 to obtain (Fig. 12)

$$x^2 + y^2 = cx. \qquad \text{Thus} \qquad \left(x - \frac{c}{2}\right)^2 + y^2 = \frac{c^2}{4}.$$

This general solution represents a family of circles passing through the origin with centers on the x-axis. ■

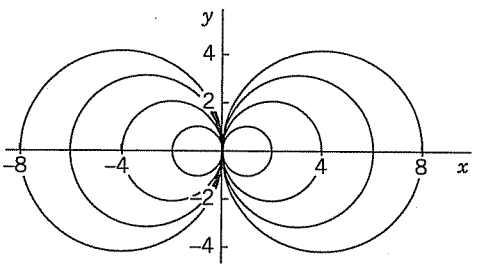


Fig. 12. General solution (family of circles) in Example 6

PROBLEM SET 1.2

1. **(Constant of integration)** An arbitrary constant of integration must be introduced immediately when the integration is performed. Why is this important? Give an example of your own.

2–9 GENERAL SOLUTION

Find a general solution. Show the steps of derivation. Check your answer by substitution.

2. $y' + (x + 2)y^2 = 0$
3. $y' = 2 \sec 2y$
4. $y' = (y + 9x)^2 \quad (y + 9x = v)$
5. $yy' + 36x = 0$
6. $y' = (4x^2 + y^2)/(xy)$
7. $y' \sin \pi x = y \cos \pi x$
8. $xy' = \frac{1}{2}y^2 + y$
9. $y'e^{\pi x} = y^2 + 1$

10–19 INITIAL VALUE PROBLEMS

Find the particular solution. Show the steps of derivation, beginning with the general solution. (L, R, b are constants.)

10. $yy' + 4x = 0, \; y(0) = 3$
11. $dr/dt = -2tr, \; r(0) = r_0$
12. $2xyy' = 3y^2 + x^2, \; y(1) = 2$
13. $L \, dI/dt + RI = 0, \; I(0) = I_0$
14. $y' = y/x + (2x^3/y)\cos(x^2), \; y(\sqrt{\pi/2}) = \sqrt{\pi}$
15. $e^{2x}y' = 2(x + 2)y^3, \; y(0) = 1/\sqrt{5} \approx 0.45$
16. $xy' = y + 4x^5 \cos^2(y/x), \; y(2) = 0$
17. $y'x \ln x = y, \; y(3) = \ln 81$
18. $dr/d\theta = b[(dr/d\theta)\cos\theta + r \sin\theta], \; r(\frac{1}{2}\pi) = \pi, \; 0 < b < 1$
19. $yy' = (x - 1)e^{-y^2}, \; y(0) = 1$
20. **(Particular solution)** Introduce limits of integration in (3) such that y obtained from (3) satisfies the initial condition $y(x_0) = y_0$. Try the formula out on Prob. 19.

21–36 APPLICATIONS, MODELING

21. **(Curves)** Find all curves in the xy-plane whose tangents all pass through a given point (a, b).
22. **(Curves)** Show that any (nonvertical) straight line through the origin of the xy-plane intersects all solution curves of $y' = g(y/x)$ at the same angle.
23. **(Exponential growth)** If the growth rate of the amount of yeast at any time t is proportional to the amount present at that time and doubles in 1 week, how much yeast can be expected after 2 weeks? After 4 weeks?
24. **(Population model)** If in a population of bacteria the birth rate and death rate are proportional to the number of individuals present, what is the population as a function of time? Figure out the limiting situation for increasing time and interpret it.
25. **(Radiocarbon dating)** If a fossilized tree is claimed to be 4000 years old, what should be its ${}_6C^{14}$ content expressed as a percent of the ratio of ${}_6C^{14}$ to ${}_6C^{12}$ in a living organism?
26. **(Gompertz growth in tumors)** The Gompertz model is $y' = -Ay \ln y \; (A > 0)$, where $y(t)$ is the mass of tumor cells at time t. The model agrees well with clinical observations. The declining growth rate with increasing $y > 1$ corresponds to the fact that cells in the interior of a tumor may die because of insufficient oxygen and nutrients. Use the ODE to discuss the growth and decline of solutions (tumors) and to find constant solutions. Then solve the ODE.
27. **(Dryer)** If wet laundry loses half of its moisture during the first 5 minutes of drying in a dryer and if the rate of loss of moisture is proportional to the moisture content, when will the laundry be practically dry, say, when will it have lost 95% of its moisture? First guess.
28. **(Alibi?)** Jack, arrested when leaving a bar, claims that he has been inside for at least half an hour (which would provide him with an alibi). The police check the water temperature of his car (parked near the entrance of the bar) at the instant of arrest and again 30 minutes later, obtaining the values 190°F and 110°F, respectively. Do these results give Jack an alibi? (Solve by inspection.)
29. **(Law of cooling)** A thermometer, reading 10°C, is brought into a room whose temperature is 23°C. Two minutes later the thermometer reading is 18°C. How long will it take until the reading is practically 23°C, say, 22.8°C? First guess.
30. **(Torricelli's law)** How does the answer in Example 5 (the time when the tank is empty) change if the diameter of the hole is doubled? First guess.
31. **(Torricelli's law)** Show that (7) looks reasonable inasmuch as $\sqrt{2gh(t)}$ is the speed a body gains if it falls a distance h (and air resistance is neglected).
32. **(Rope)** To tie a boat in a harbor, how many times must a rope be wound around a bollard (a vertical rough cylindrical post fixed on the ground) so that a man holding one end of the rope can resist a force exerted by the boat one thousand times greater than the man can exert? First guess. Experiments show that the change ΔS of the force S in a small portion of the rope is proportional to S and to the small angle $\Delta\phi$ in Fig. 13. Take the proportionality constant 0.15.

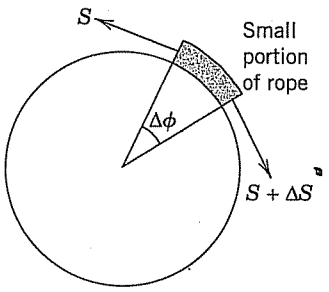


Fig. 13. Problem 32

33. (Mixing) A tank contains 800 gal of water in which 200 lb of salt is dissolved. Two gallons of fresh water runs in per minute, and 2 gal of the mixture in the tank, kept uniform by stirring, runs out per minute. How much salt is left in the tank after 5 hours?

34. WRITING PROJECT. Exponential Increase, Decay, Approach. Collect, order, and present all the information on the ODE $y' = ky$ and its applications from the text and the problems. Add examples of your own.

35. CAS EXPERIMENT. Graphing Solutions. A CAS can usually graph solutions even if they are given by integrals that cannot be evaluated by the usual methods of calculus. Show this as follows.

(A) Graph the curves for the seven initial value problems $y' = e^{-x^2/2}$, $y(0) = 0, \pm 1, \pm 2, \pm 3$, common axes. Are these curves congruent? Why?

(B) Experiment with approximate curves of nth partial sums of the Maclaurin series obtained by termwise integration of that of y in (A); graph them and describe qualitatively the accuracy for a fixed interval $0 \leqq x \leqq b$ and increasing n, and then for fixed n and increasing b.

(C) Experiment with $y' = \cos(x^2)$ as in (B).

(D) Find an initial value problem with solution $y = e^{x^2}\int_0^x e^{-t^2}\,dt$ and experiment with it as in (B).

36. TEAM PROJECT. Torricelli's Law. Suppose that the tank in Example 5 is hemispherical, of radius R, initially full of water, and has an outlet of 5 cm^2 cross-sectional area at the bottom. (Make a sketch.) Set up the model for outflow. Indicate what portion of your work in Example 5 you can use (so that it can become part of the general method independent of the shape of the tank). Find the time t to empty the tank (a) for any R, (b) for $R = 1$ m. Plot t as function of R. Find the time when $h = R/2$ (a) for any R, (b) for $R = 1$ m.

1.4 Exact ODEs. Integrating Factors

We remember from calculus that if a function $u(x, y)$ has continuous partial derivatives, its **differential** (also called its *total differential*) is

$$du = \frac{\partial u}{\partial x}\,dx + \frac{\partial u}{\partial y}\,dy.$$

From this it follows that if $u(x, y) = c = const$, then $du = 0$.

For example, if $u = x + x^2y^3 = c$, then

$$du = (1 + 2xy^3)\,dx + 3x^2y^2\,dy = 0$$

or

$$y' = \frac{dy}{dx} = -\frac{1 + 2xy^3}{3x^2y^2},$$

an ODE that we can solve by going backward. This idea leads to a powerful solution method as follows.

A first-order ODE $M(x, y) + N(x, y)y' = 0$, written as (use $dy = y'\,dx$ as in Sec. 1.3)

(1) $$M(x, y)\,dx + N(x, y)\,dy = 0$$

is called an **exact differential equation** if the **differential form** $M(x, y)\,dx + N(x, y)\,dy$ is **exact,** that is, this form is the differential

$$du = \frac{\partial u}{\partial x}\,dx + \frac{\partial u}{\partial y}\,dy \tag{2}$$

of some function $u(x, y)$. Then (1) can be written

$$du = 0.$$

By integration we immediately obtain the general solution of (1) in the form

$$u(x, y) = c. \tag{3}$$

This is called an **implicit solution,** in contrast with a solution $y = h(x)$ as defined in Sec. 1.1, which is also called an *explicit solution,* for distinction. Sometimes an implicit solution can be converted to explicit form. (Do this for $x^2 + y^2 = 1$.) If this is not possible, your CAS may graph a figure of the **contour lines** (3) of the function $u(x, y)$ and help you in understanding the solution.

Comparing (1) and (2), we see that (1) is an exact differential equation if there is some function $u(x, y)$ such that

$$\text{(a)}\quad \frac{\partial u}{\partial x} = M, \qquad \text{(b)}\quad \frac{\partial u}{\partial y} = N. \tag{4}$$

From this we can derive a formula for checking whether (1) is exact or not, as follows.

Let M and N be continuous and have continuous first partial derivatives in a region in the xy-plane whose boundary is a closed curve without self-intersections. Then by partial differentiation of (4) (see App. 3.2 for notation),

$$\frac{\partial M}{\partial y} = \frac{\partial^2 u}{\partial y\,\partial x}, \qquad \frac{\partial N}{\partial x} = \frac{\partial^2 u}{\partial x\,\partial y}.$$

By the assumption of continuity the two second partial derivatives are equal. Thus

$$\frac{\partial M}{\partial y} = \frac{\partial N}{\partial x}. \tag{5}$$

This condition is not only necessary but also sufficient for (1) to be an exact differential equation. (We shall prove this in Sec. 10.2 in another context. Some calculus books (e.g., Ref. [GR11] also contain a proof.)

If (1) is exact, the function $u(x, y)$ can be found by inspection or in the following systematic way. From (4a) we have by integration with respect to x

(6) $$u = \int M\,dx + k(y);$$

in this integration, y is to be regarded as a constant, and $k(y)$ plays the role of a "constant" of integration. To determine $k(y)$, we derive $\partial u/\partial y$ from (6), use (4b) to get dk/dy, and integrate dk/dy to get k.

Formula (6) was obtained from (4a). Instead of (4a) we may equally well use (4b). Then instead of (6) we first have by integration with respect to y

(6*) $$u = \int N\,dy + l(x).$$

To determine $l(x)$, we derive $\partial u/\partial x$ from (6*), use (4a) to get dl/dx, and integrate. We illustrate all this by the following typical examples.

EXAMPLE 1 An Exact ODE

Solve

(7) $$\cos(x + y)\,dx + (3y^2 + 2y + \cos(x + y))\,dy = 0.$$

Solution. ***Step 1. Test for exactness.*** Our equation is of the form (1) with

$$M = \cos(x + y),$$

$$N = 3y^2 + 2y + \cos(x + y).$$

Thus

$$\frac{\partial M}{\partial y} = -\sin(x + y),$$

$$\frac{\partial N}{\partial x} = -\sin(x + y).$$

From this and (5) we see that (7) is exact.

Step 2. Implicit general solution. From (6) we obtain by integration

(8) $$u = \int M\,dx + k(y) = \int \cos(x + y)\,dx + k(y) = \sin(x + y) + k(y).$$

To find $k(y)$, we differentiate this formula with respect to y and use formula (4b), obtaining

$$\frac{\partial u}{\partial y} = \cos(x + y) + \frac{dk}{dy} = N = 3y^2 + 2y + \cos(x + y).$$

Hence $dk/dy = 3y^2 + 2y$. By integration, $k = y^3 + y^2 + c^*$. Inserting this result into (8) and observing (3), we obtain the *answer*

$$u(x, y) = \sin(x + y) + y^3 + y^2 = c.$$

Step 3. Checking an implicit solution. We can check by differentiating the implicit solution $u(x, y) = c$ implicitly and see whether this leads to the given ODE (7):

(9) $$du = \frac{\partial u}{\partial x}\,dx + \frac{\partial u}{\partial y}\,dy = \cos(x + y)\,dx + (\cos(x + y) + 3y^2 + 2y)\,dy = 0.$$

This completes the check. ■

EXAMPLE 2 An Initial Value Problem

Solve the initial value problem

$$(\cos y \sinh x + 1)\,dx - \sin y \cosh x\,dy = 0, \qquad y(1) = 2. \tag{10}$$

Solution. You may verify that the given ODE is exact. We find u. For a change, let us use (6*),

$$u = -\int \sin y \cosh x\,dy + l(x) = \cos y \cosh x + l(x).$$

From this, $\partial u/\partial x = \cos y \sinh x + dl/dx = M = \cos y \sinh x + 1$. Hence $dl/dx = 1$. By integration, $l(x) = x + c^*$. This gives the general solution $u(x, y) = \cos y \cosh x + x = c$. From the initial condition, $\cos 2 \cosh 1 + 1 = 0.358 = c$. Hence the answer is $\cos y \cosh x + x = 0.358$. Figure 14 shows the particular solutions for $c = 0$, 0.358 (thicker curve), 1, 2, 3. Check that the answer satisfies the ODE. (Proceed as in Example 1.) Also check that the initial condition is satisfied. ▩

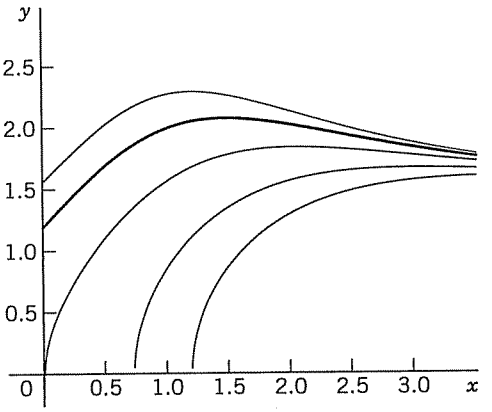


Fig. 14. Particular solu.ions in Example 2

EXAMPLE 3 WARNING! Breakdown in the Case of Nonexactness

The equation $-y\,dx + x\,dy = 0$ is not exact because $M = -y$ and $N = x$, so that in (5), $\partial M/\partial y = -1$ but $\partial N/\partial x = 1$. Let us show that in such a case the present method does not work. From (6),

$$u = \int M\,dx + k(y) = -xy + k(y), \qquad \text{hence} \qquad \frac{\partial u}{\partial y} = -x + \frac{dk}{dy}.$$

Now, $\partial u/\partial y$ should equal $N = x$, by (4b). However, this is impossible because $k(y)$ can depend only on y. Try (6*); it will also fail. Solve the equation by another method that we have discussed. ▩

Reduction to Exact Form. Integrating Factors

The ODE in Example 3 is $-y\,dx + x\,dy = 0$. It is not exact. However, if we multiply it by $1/x^2$, we get an exact equation [check exactness by (5)!],

$$\frac{-y\,dx + x\,dy}{x^2} = -\frac{y}{x^2}\,dx + \frac{1}{x}\,dy = d\left(\frac{y}{x}\right) = 0. \tag{11}$$

Integration of (11) then gives the general solution $y/x = c = const.$

This example gives the idea. All we did was multiply a given nonexact equation, say,

$$P(x, y)\,dx + Q(x, y)\,dy = 0, \tag{12}$$

by a function F that, in general, will be a function of both x and y. The result was an equation

$$FP\,dx + FQ\,dy = 0 \tag{13}$$

that is exact, so we can solve it as just discussed. Such a function $F(x, y)$ is then called an **integrating factor** of (12).

EXAMPLE 4 **Integrating Factor**

The integrating factor in (11) is $F = 1/x^2$. Hence in this case the exact equation (13) is

$$FP\,dx + FQ\,dy = \frac{-y\,dx + x\,dy}{x^2} = d\left(\frac{y}{x}\right) = 0. \qquad \text{Solution} \qquad \frac{y}{x} = c.$$

These are straight lines $y = cx$ through the origin.

It is remarkable that we can readily find other integrating factors for the equation $-y\,dx + x\,dy = 0$, namely, $1/y^2$, $1/(xy)$, and $1/(x^2 + y^2)$, because

$$\frac{-y\,dx + x\,dy}{y^2} = d\left(\frac{x}{y}\right), \quad \frac{-y\,dx + x\,dy}{xy} = -d\left(\ln\frac{x}{y}\right), \quad \frac{-y\,dx + x\,dy}{x^2 + y^2} = d\left(\arctan\frac{y}{x}\right). \tag{14}$$

■

How to Find Integrating Factors

In simpler cases we may find integrating factors by inspection or perhaps after some trials, keeping (14) in mind. In the general case, the idea is the following.

For $M\,dx + N\,dy = 0$ the exactness condition (4) is $\partial M/\partial y = \partial N/\partial x$. Hence for (13), $FP\,dx + FQ\,dy = 0$, the exactness condition is

$$\frac{\partial}{\partial y}(FP) = \frac{\partial}{\partial x}(FQ). \tag{15}$$

By the product rule, with subscripts denoting partial derivatives, this gives

$$F_y P + FP_y = F_x Q + FQ_x.$$

In the general case, this would be complicated and useless. So we follow the ***Golden Rule:*** *If you cannot solve your problem, try to solve a simpler one*—the result may be useful (and may also help you later on). Hence we look for an integrating factor depending only on ***one*** variable; fortunately, in many practical cases, there are such factors, as we shall see. Thus, let $F = F(x)$. Then $F_y = 0$, and $F_x = F' = dF/dx$, so that (15) becomes

$$FP_y = F'Q + FQ_x.$$

Dividing by FQ and reshuffling terms, we have

$$\frac{1}{F}\frac{dF}{dx} = R, \qquad \text{where} \qquad R = \frac{1}{Q}\left(\frac{\partial P}{\partial y} - \frac{\partial Q}{\partial x}\right). \tag{16}$$

This proves the following theorem.

THEOREM 1

Integrating Factor $F(x)$

If (12) *is such that the right side* R *of* (16), *depends only on* x, *then* (12) *has an integrating factor* $F = F(x)$, *which is obtained by integrating* (16) *and taking exponents on both sides,*

(17) $$F(x) = \exp \int R(x)\, dx.$$

Similarly, if $F^* = F^*(y)$, then instead of (16) we get

(18) $$\frac{1}{F^*}\frac{dF^*}{dy} = R^*, \qquad \text{where} \qquad R^* = \frac{1}{P}\left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right)$$

and we have the companion

THEOREM 2

Integrating Factor $F^*(y)$

If (12) *is such that the right side* R^* *of* (18) *depends only on* y, *then* (12) *has an integrating factor* $F^* = F^*(y)$, *which is obtained from* (18) *in the form*

(19) $$F^*(y) = \exp \int R^*(y)\, dy.$$

EXAMPLE 5 **Application of Theorems 1 and 2. Initial Value Problem**

Using Theorem 1 or 2, find an integrating factor and solve the initial value problem

(20) $$(e^{x+y} + ye^y)\, dx + (xe^y - 1)\, dy = 0, \quad y(0) = -1$$

Solution. ***Step 1. Nonexactness.*** The exactness check fails:

$$\frac{\partial P}{\partial y} = \frac{\partial}{\partial y}(e^{x+y} + ye^y) = e^{x+y} + e^y + ye^y \qquad \text{but} \qquad \frac{\partial Q}{\partial x} = \frac{\partial}{\partial x}(xe^y - 1) = e^y.$$

Step 2. Integrating factor. General solution. Theorem 1 fails because R [the right side of (16)] depends on both x and y,

$$R = \frac{1}{Q}\left(\frac{\partial P}{\partial y} - \frac{\partial Q}{\partial x}\right) = \frac{1}{xe^y - 1}(e^{x+y} + e^y + ye^y - e^y).$$

Try Theorem 2. The right side of (18) is

$$R^* = \frac{1}{P}\left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) = \frac{1}{e^{x+y} + ye^y}(e^y - e^{x+y} - e^y - ye^y) = -1.$$

Hence (19) gives the integrating factor $F^*(y) = e^{-y}$. From this result and (20) you get the exact equation

$$(e^x + y)\, dx + (x - e^{-y})\, dy = 0.$$

Test for exactness; you will get 1 on both sides of the exactness condition. By integration, using (4a),

$$u = \int (e^x + y)\, dx = e^x + xy + k(y).$$

Differentiate this with respect to y and use (4b) to get

$$\frac{\partial u}{\partial y} = x + \frac{dk}{dy} = N = x - e^{-y}, \qquad \frac{dk}{dy} = -e^{-y}, \qquad k = e^{-y} + c^*.$$

Hence the general solution is

$$u(x, y) = e^x + xy + e^{-y} = c.$$

Step 3. Particular solution. The initial condition $y(0) = 1$ gives $u(0, -1) = 1 + 0 + e = 3.72$. Hence the answer is $e^x + xy + e^{-y} = 1 + e = 3.72$. Figure 15 shows several particular solutions obtained as level curves of $u(x, y) = c$, obtained by a CAS, a convenient way in cases in which it is impossible or difficult to cast a solution into explicit form. Note the curve that (nearly) satisfies the initial condition.

Step 4. Checking. Check by substitution that the answer satisfies the given equation as well as the initial condition. ■

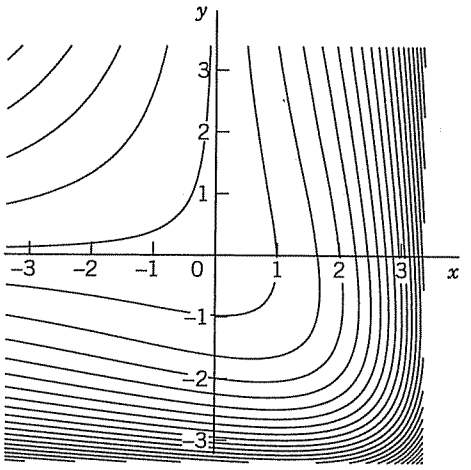


Fig. 15. Particular solutions in Example 5

PROBLEM SET 1.4

1–20 EXACT ODEs. INTEGRATING FACTORS

Test for exactness. If exact, solve. If not, use an integrating factor as given or find it by inspection or from the theorems in the text. Also, if an initial condition is given, determine the corresponding particular solution.

1. $x^3\,dx + y^3\,dy = 0$

2. $(x - y)(dx - dy) = 0$

3. $-\pi \sin \pi x \sinh y\,dx + \cos \pi x \cosh y\,dy = 0$

4. $(e^y - ye^x)\,dx + (xe^y - e^x)\,dy = 0$

5. $9x\,dx + 4y\,dy = 0$

6. $e^x(\cos y\,dx - \sin y\,dy) = 0$

7. $e^{-2\theta}\,dr - 2re^{-2\theta}\,d\theta = 0$

8. $(2x + 1/y - y/x^2)\,dx + (2y + 1/x - x/y^2)\,dy = 0$

9. $(-y/x^2 + 2\cos 2x)\,dx + (1/x - 2\sin 2y)\,dy = 0$

10. $-2xy \sin(x^2)\,dx + \cos(x^2)\,dy = 0$

11. $-y\,dx + x\,dy = 0$

12. $(e^{x+y} - y)\,dx + (xe^{x+y} + 1)\,dy = 0$

13. $-3y\,dx + 2x\,dy = 0, \quad F(x, y) = y/x^4$

14. $(x^4 + y^2)\,dx - xy\,dy = 0, \quad y(2) = 1$

15. $e^{2x}(2\cos y\,dx - \sin y\,dy) = 0, \quad y(0) = 0$

16. $-\sin xy\,(y\,dx + x\,dy) = 0, \quad y(1) = \pi$

17. $(\cos \omega x + \omega \sin \omega x)\,dx + e^x\,dy = 0, \; y(0) = 1$

18. $(\cos xy + x/y)\,dx + (1 + (x/y)\cos xy)\,dy = 0$

19. $e^{-y}\,dx + e^{-x}(-e^{-y} + 1)\,dy = 0, \quad F = e^{x+y}$

20. $(\sin y \cos y + x\cos^2 y)\,dx + x\,dy = 0$

21. Under what conditions for the constants A, B, C, D is $(Ax + By)\,dx + (Cx + Dy)\,dy = 0$ exact? Solve the exact equation.

22. CAS PROJECT. Graphing Particular Solutions Graph particular solutions of the following ODE, proceeding as explained.

$$(21) \qquad y \cos x \, dx + \frac{1}{y} \, dy = 0$$

(a) Test for exactness. If necessary, find an integrating factor. Find the general solution $u(x, y) = c$.

(b) Solve (21) by separating variables. Is this simpler than **(a)**?

(c) Graph contours $u(x, y) = c$ by your CAS. (Cf. Fig. 16.)

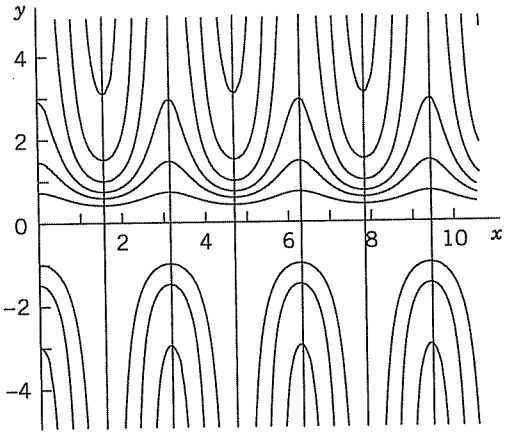


Fig. 16. Particular solutions in CAS Project 22

(d) In another graph show the solution curves satisfying $y(0) = \pm 1, \pm 2, \pm 3, \pm 4$. Compare the quality of **(c)** and **(d)** and comment.

(e) Do the same steps for another nonexact ODE of your choice.

23. WRITING PROJECT. Working Backward. Start from solutions $u(x, y) = c$ of your choice, find a corresponding exact ODE, destroy exactness by a multiplication or division. This should give you a feel for the form of ODEs you can reach by the method of integrating factors. (Working backward is useful in other areas, too; Euler and other great masters frequently did it.)

24. TEAM PROJECT. Solution by Several Methods. Show this as indicated. Compare the amount of work.

(A) $e^y(\sinh x \, dx + \cosh x \, dy) = 0$ as an exact ODE and by separation.

(B) $(1 + 2x) \cos y \, dx + dy/\cos y = 0$ by Theorem 2 and by separation.

(C) $(x^2 + y^2) \, dx - 2xy \, dy = 0$ by Theorem 1 or 2 and by separation with $v = y/x$.

(D) $3x^2 y \, dx + 4x^3 \, dy = 0$ by Theorems 1 and 2 and by separation.

(E) Search the text and the problems for further ODEs that can be solved by more than one of the methods discussed so far. Make a list of these ODEs. Find further cases of your own.

1.5 Linear ODEs. Bernoulli Equation. Population Dynamics

Linear ODEs or ODEs that can be transformed to linear form are models of various phenomena, for instance, in physics, biology, population dynamics, and ecology, as we shall see. A first-order ODE is said to be ***linear*** if it can be written

$$(1) \qquad y' + p(x)y = r(x).$$

The defining feature of this equation is that it is linear in both the unknown function y and its derivative $y' = dy/dx$, whereas p and r may be ***any*** given functions of x. If in an application the independent variable is time, we write t instead of x.

If the first term is $f(x)y'$ (instead of y'), divide the equation by $f(x)$ to get the **"standard form"** (1), with y' as the first term, which is practical.

For instance, $y' \cos x + y \sin x = x$ is a linear ODE, and its standard form is $y' + y \tan x = x \sec x$.

The function $r(x)$ on the right may be a force, and the solution $y(x)$ a displacement in a motion or an electrical current or some other physical quantity. In engineering, $r(x)$ is frequently called the **input,** and $y(x)$ is called the **output** or the *response* to the input (and, if given, to the initial condition).

Homogeneous Linear ODE. We want to solve (1) in some interval $a < x < b$, call it J, and we begin with the simpler special case that $r(x)$ is zero for all x in J. (This is sometimes written $r(x) \equiv 0$.) Then the ODE (1) becomes

(2) $$y' + p(x)y = 0$$

and is called **homogeneous.** By separating variables and integrating we then obtain

$$\frac{dy}{y} = -p(x)\,dx, \qquad \text{thus} \qquad \ln|y| = -\int p(x)\,dx + c^*.$$

Taking exponents on both sides, we obtain the general solution of the homogeneous ODE (2),

(3) $$y(x) = ce^{-\int p(x)\,dx} \qquad (c = \pm e^{c^*} \quad \text{when} \quad y \gtrless 0);$$

here we may also choose $c = 0$ and obtain the **trivial solution** $y(x) = 0$ for all x in that interval.

Nonhomogeneous Linear ODE. We now solve (1) in the case that $r(x)$ in (1) is not everywhere zero in the interval J considered. Then the ODE (1) is called **nonhomogeneous.** It turns out that in this case, (1) has a pleasant property; namely, it has an integrating factor depending only on x. We can find this factor $F(x)$ by Theorem 1 in the last section. For this purpose we write (1) as

$$(py - r)\,dx + dy = 0.$$

This is $P\,dx + Q\,dy = 0$, where $P = py - r$ and $Q = 1$. Hence the right side of (16) in Sec. 1.4 is simply $1(p - 0) = p$, so that (16) becomes

$$\frac{1}{F}\frac{dF}{dx} = p(x).$$

Separation and integration gives

$$\frac{dF}{F} = p\,dx \qquad \text{and} \qquad \ln|F| = \int p\,dx.$$

Taking exponents on both sides, we obtain the desired integrating factor $F(x)$,

$$F(x) = e^{\int p\,dx}.$$

We now multiply (1) on both sides by this F. Then by the product rule,

$$e^{\int p\,dx}(y' + py) = (e^{\int p\,dx}y)' = e^{\int p\,dx}r.$$

By integrating the second and third of these three expressions with respect to x we get

$$e^{\int p\,dx}y = \int e^{\int p\,dx}r\,dx + c.$$

Dividing this equation by $e^{\int p\,dx}$ and denoting the exponent $\int p\,dx$ by h, we obtain

(4) $$y(x) = e^{-h}\left(\int e^{h}r\,dx + c\right), \qquad h = \int p(x)\,dx.$$

(The constant of integration in h does not matter; see Prob. 2.) Formula (4) is the general solution of (1) in the form of an integral. Solving (1) is now reduced to the evaluation of an integral. In cases in which this cannot be done by the usual methods of calculus, one may have to use a numeric method for integrals (Sec. 19.5) or for the ODE itself (Sec. 21.1).

The structure of (4) is interesting. The only quantity depending on a given initial condition is c. Accordingly, writing (4) as a sum of two terms,

$$(4^*) \qquad y(x) = e^{-h}\int e^h r\,dx + ce^{-h},$$

we see the following:

(5) Total Output = Response to the Input r + Response to the Initial Data.

EXAMPLE 1 First-Order ODE, General Solution

Solve the linear ODE

$$y' - y = e^{2x}.$$

Solution. Here,

$$p = -1, \qquad r = e^{2x}, \qquad h = \int p\,dx = -x$$

and from (4) we obtain the general solution

$$y(x) = e^x\left(\int e^{-x}e^{2x}\,dx + c\right) = e^x(e^x + c) = ce^x + e^{2x}.$$

From (4*) and (5) we see that the response to the input is e^{2x}.

In simpler cases, such as the present, we may not need the general formula (4), but may wish to proceed directly, multiplying the given equation by $e^h = e^{-x}$. This gives

$$(y' - y)e^{-x} = (ye^{-x})' = e^{2x}e^{-x} = e^x.$$

Integrating on both sides, we obtain the same result as before:

$$ye^{-x} = e^x + c, \qquad \text{hence} \qquad y = e^{2x} + ce^x.$$

EXAMPLE 2 First-Order ODE, Initial Value Problem

Solve the initial value problem

$$y' + y\tan x = \sin 2x, \qquad y(0) = 1.$$

Solution. Here $p = \tan x$, $r = \sin 2x = 2\sin x\cos x$, and

$$\int p\,dx = \int \tan x\,dx = \ln|\sec x|.$$

From this we see that in (4),

$$e^h = \sec x, \qquad e^{-h} = \cos x, \qquad e^h r = (\sec x)(2\sin x\cos x) = 2\sin x,$$

and the general solution of our equation is

$$y(x) = \cos x\left(2\int \sin x\,dx + c\right) = c\cos x - 2\cos^2 x.$$

From this and the initial condition, $1 = c\cdot 1 - 2\cdot 1^2$; thus $c = 3$ and the solution of our initial value problem is $y = 3\cos x - 2\cos^2 x$. Here $3\cos x$ is the response to the initial data, and $-2\cos^2 x$ is the response to the input $\sin 2x$.

EXAMPLE 3 Hormone Level

Assume that the level of a certain hormone in the blood of a patient varies with time. Suppose that the time rate of change is the difference between a sinusoidal input of a 24-hour period from the thyroid gland and a continuous removal rate proportional to the level present. Set up a model for the hormone level in the blood and find its general solution. Find the particular solution satisfying a suitable initial condition.

Solution. ***Step 1. Setting up a model.*** Let $y(t)$ be the hormone level at time t. Then the removal rate is $Ky(t)$. The input rate is $A + B\cos(2\pi t/24)$, where A is the average input rate, and $A \geqq B$ to make the input nonnegative. (The constants A, B, and K can be determined by measurements.) Hence the model is

$$y'(t) = \text{In} - \text{Out} = A + B\cos\left(\tfrac{1}{12}\pi t\right) - Ky(t) \qquad \text{or} \qquad y' + Ky = A + B\cos\left(\tfrac{1}{12}\pi t\right).$$

The initial condition for a particular solution y_{part} is $y_{part}(0) = y_0$ with $t = 0$ suitably chosen, e.g., 6:00 A.M.

Step 2. General solution. In (4) we have $p = K = const$, $h = Kt$, and $r = A + B\cos\left(\frac{1}{12}\pi t\right)$. Hence (4) gives the general solution

$$y(t) = e^{-Kt}\int e^{Kt}\left(A + B\cos\frac{\pi t}{12}\right)dt + ce^{-Kt}$$

$$= e^{-Kt}e^{Kt}\left[\frac{A}{K} + \frac{B}{144K^2 + \pi^2}\left(144K\cos\frac{\pi t}{12} + 12\pi\sin\frac{\pi t}{12}\right)\right] + ce^{-Kt}$$

$$= \frac{A}{K} + \frac{B}{144K^2 + \pi^2}\left(144K\cos\frac{\pi t}{12} + 12\pi\sin\frac{\pi t}{12}\right) + ce^{-Kt}.$$

The last term decreases to 0 as t increases, practically after a short time and regardless of c (that is, of the initial condition). The other part of $y(t)$ is called the **steady-state solution** because it consists of constant and periodic terms. The entire solution is called the **transient-state solution** because it models the transition from rest to the steady state. These terms are used quite generally for physical and other systems whose behavior depends on time.

Step 3. Particular solution. Setting $t = 0$ in $y(t)$ and choosing $y_0 = 0$, we have

$$y(0) = \frac{A}{K} + \frac{B}{144K^2 + \pi^2}\cdot 144K + c = 0, \qquad \text{thus} \qquad c = -\frac{A}{K} - \frac{B}{144K^2 + \pi^2}\cdot 144K.$$

Inserting this result into $y(t)$, we obtain the particular solution

$$y_{part}(t) = \frac{A}{K} + \frac{B}{144K^2 + \pi^2}\left(144K\cos\frac{\pi t}{12} + 12\pi\sin\frac{\pi t}{12}\right) - \left(\frac{A}{K} + \frac{144KB}{144K^2 + \pi^2}\right)e^{-Kt}$$

with the steady-state part as before. To plot y_{part} we must specify values for the constants, say, $A = B = 1$ and $K = 0.05$. Figure 17 shows this solution. Notice that the transition period is relatively short (although K is small), and the curve soon looks sinusoidal; this is the response to the input $A + B\cos\left(\frac{1}{12}\pi t\right) = 1 + \cos\left(\frac{1}{12}\pi t\right)$. ▩

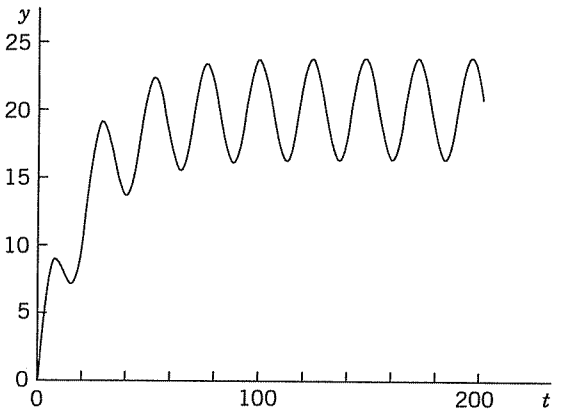


Fig. 17. Particular solution in Example 3

Reduction to Linear Form. Bernoulli Equation

Numerous applications can be modeled by ODEs that are nonlinear but can be transformed to linear ODEs. One of the most useful ones of these is the **Bernoulli equation**[5]

$$y' + p(x)y = g(x)y^a \qquad (a \text{ any real number}). \tag{6}$$

If $a = 0$ or $a = 1$, Equation (6) is linear. Otherwise it is nonlinear. Then we set

$$u(x) = [y(x)]^{1-a}.$$

We differentiate this and substitute y' from (6), obtaining

$$u' = (1 - a)y^{-a}y' = (1 - a)y^{-a}(gy^a - py).$$

Simplification gives

$$u' = (1 - a)(g - py^{1-a}),$$

where $y^{1-a} = u$ on the right, so that we get the linear ODE

$$u' + (1 - a)pu = (1 - a)g. \tag{7}$$

For further ODEs reducible to linear from, see Ince's classic [A11] listed in App. 1. See also Team Project 44 in Problem Set 1.5.

EXAMPLE 4 **Logistic Equation**

Solve the following Bernoulli equation, known as the **logistic equation** (or **Verhulst equation**[6]):

$$y' = Ay - By^2 \tag{8}$$

Solution. Write (8) in the form (6), that is,

$$y' - Ay = -By^2$$

to see that $a = 2$, so that $u = y^{1-a} = y^{-1}$. Differentiate this u and substitute y' from (8),

$$u' = -y^{-2}y' = -y^{-2}(Ay - By^2) = B - Ay^{-1}.$$

The last term is $-Ay^{-1} = -Au$. Hence we have obtained the linear ODE

[5]JAKOB BERNOULLI (1654–1705), Swiss mathematician, professor at Basel, also known for his contribution to elasticity theory and mathematical probability. The method for solving Bernoulli's equation was discovered by the Leibniz in 1696. Jakob Bernoulli's students included his nephew NIKLAUS BERNOULLI (1687–1759), who contributed to probability theory and infinite series, and his youngest brother JOHANN BERNOULLI (1667–1748), who had profound influence on the development of calculus, became Jakob's successor at Basel, and had among his students GABRIEL CRAMER (see Sec. 7.7) and LEONHARD EULER (see Sec. 2.5). His son DANIEL BERNOULLI (1700–1782) is known for his basic work in fluid flow and the kinetic theory of gases.

[6]PIERRE-FRANÇOIS VERHULST, Belgian statistician, who introduced Eq. (8) as a model for human population growth in 1838.

$$u' + Au = B.$$

The general solution is [by (4)]

$$u = ce^{-At} + B/A.$$

Since $u = 1/y$, this gives the general solution of (8),

(9) $$y = \frac{1}{u} = \frac{1}{ce^{-At} + B/A}$$ (Fig. 18).

Directly from (8) we see that $y \equiv 0$ ($y(t) = 0$ for all t) is also a solution. ▩

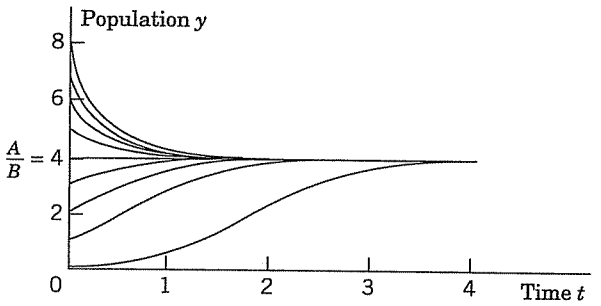


Fig. 18. Logistic population model. Curves (9) in Example 4 with $A/B = 4$

Population Dynamics

The logistic equation (8) plays an important role in **population dynamics,** a field that models the evolution of populations of plants, animals, or humans over time t. If $B = 0$, then (8) is $y' = dy/dt = Ay$. In this case its solution (9) is $y = (1/c)e^{At}$ and gives exponential growth, as for a small population in a large country (the United States in early times!). This is called ***Malthus's law.*** (See also Example 3 in Sec. 1.1.)

The term $-By^2$ in (8) is a "braking term" that prevents the population from growing without bound. Indeed, if we write $y' = Ay[1 - (B/A)y]$, we see that if $y < A/B$, then $y' > 0$, so that an initially small population keeps growing as long as $y < A/B$. But if $y > A/B$, then $y' < 0$ and the population is decreasing as long as $y > A/B$. The limit is the same in both cases, namely, A/B. See Fig. 18.

We see that in the logistic equation (8) the independent variable t does not occur explicitly. An ODE $y' = f(t, y)$ in which t does not occur explicitly is of the form

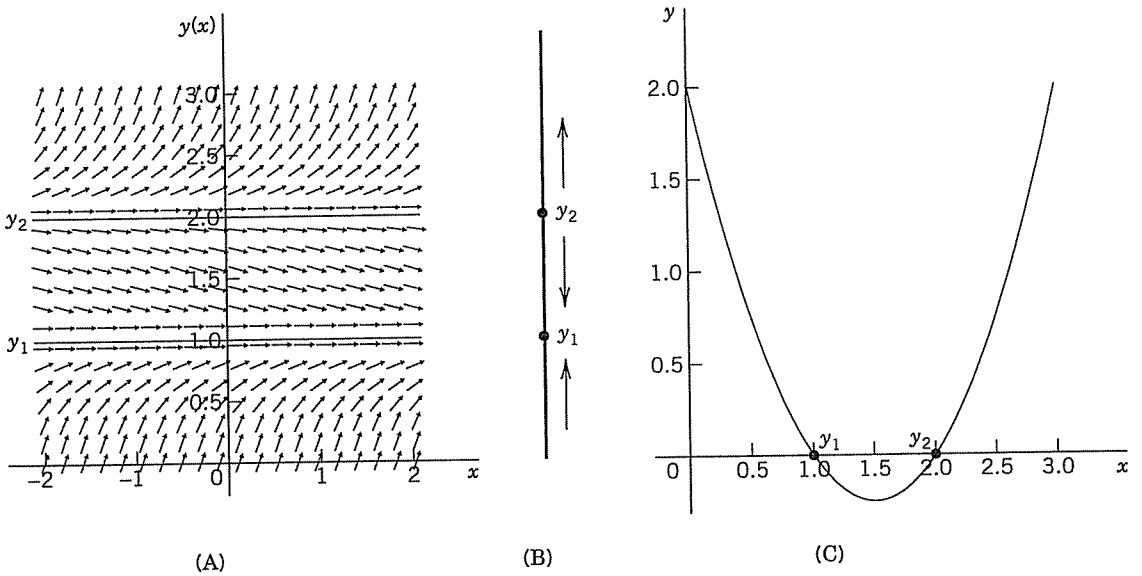
(10) $$y' = f(y)$$

and is called an **autonomous ODE.** Thus the logistic equation (8) is autonomous.

Equation (10) has constant solutions, called **equilibrium solutions** or **equilibrium points.** These are determined by the zeros of $f(y)$, because $f(y) = 0$ gives $y' = 0$ by (10); hence $y = const$. These zeros are known as **critical points** of (10). An equilibrium solution is called **stable** if solutions close to it for some t remain close to it for all further t. It is called **unstable** if solutions initially close to it do not remain close to it as t increases. For instance, $y = 0$ in Fig. 18 is an unstable equilibrium solution, and $y = 4$ is a stable one.

EXAMPLE 5 Stable and Unstable Equilibrium Solutions. "Phase Line Plot"

The ODE $y' = (y - 1)(y - 2)$ has the stable equilibrium solution $y_1 = 1$ and the unstable $y_2 = 2$, as the direction field in Fig. 19 suggests. The values y_1 and y_2 are the zeros of the parabola $f(y) = (y - 1)(y - 2)$ in the figure. Now, since the ODE is autonomous, we can "condense" the direction field to a "phase line plot" giving y_1 and y_2, and the direction (upward or downward) of the arrows in the field, and thus giving information about the stability or instability of the equilibrium solutions. ▩



(A) (B) (C)

Fig. 19. Example 5. (A) Direction field. (B) "Phase line". (C) Parabola $f(y)$

A few further population models will be discussed in the problem set. For some more details of population dynamics, see C. W. Clark, *Mathematical Bioeconomics*, New York, Wiley, 1976.

Further important applications of linear ODEs follow in the next section.

PROBLEM SET 1.5

1. **(CAUTION!)** Show that $e^{-\ln x} = 1/x$ (not $-x$) and $e^{-\ln(\sec x)} = \cos x$.
2. **(Integration constant)** Give a reason why in (4) you may choose the constant of integration in $\int p\,dx$ to be zero.

3–17 GENERAL SOLUTION. INITIAL VALUE PROBLEMS

Find the general solution. If an initial condition is given, find also the corresponding particular solution and graph or sketch it. (Show the details of your work.)

3. $y' + 3.5y = 2.8$
4. $y' = 4y + x$
5. $y' + 1.25y = 5, \quad y(0) = 6.6$
6. $x^2y' + 3xy = 1/x, \quad y(1) = -1$
7. $y' + ky = e^{2kx}$
8. $y' + 2y = 4\cos 2x, \quad y(\frac{1}{4}\pi) = 2$
9. $y' = 6(y - 2.5)\tanh 1.5x$
10. $y' + 4x^2y = (4x^2 - x)e^{-x^2/2}$
11. $y' + 2y\sin 2x = 2e^{\cos 2x}, \quad y(0) = 0$
12. $y'\tan x = 2y - 8, \quad y(\frac{1}{2}\pi) = 0$
13. $y' + 4y\cot 2x = 6\cos 2x, \quad y(\frac{1}{4}\pi) = 2$
14. $y' + y\tan x = e^{-0.01x}\cos x, \quad y(0) = 0$
15. $y' + y/x^2 = 2xe^{1/x}, \quad y(1) = 13.86$
16. $y'\cos^2 x + 3y = 1, \quad y(\frac{1}{4}\pi) = \frac{4}{3}$
17. $x^3y' + 3x^2y = 5\sinh 10x$

18–24 NONLINEAR ODEs

Using a method of this section or separating variables, find the general solution. If an initial condition is given, find also the particular solution and sketch or graph it.

18. $y' + y = y^2, \quad y(0) = -1$

19. $y' = 5.7y - 6.5y^2$

20. $(x^2 + 1)y' = -\tan y, \quad y(0) = \frac{1}{2}\pi$

21. $y' + (x + 1)y = e^{x^2}y^3, \quad y(0) = 0.5$

22. $y' \sin 2y + x \cos 2y = 2x$

23. $2yy' + y^2 \sin x = \sin x, \quad y(0) = \sqrt{2}$

24. $y' + x^2y = (e^{-x^3} \sinh x)/(3y^2)$

25–36 FURTHER APPLICATIONS

25. (Investment programs) Bill opens a retirement savings account with an initial amount y_0 and then adds $\$k$ to the account at the beginning of every year until retirement at age 65. Assume that the interest is compounded continuously at the same rate R over the years. Set up a model for the balance in the account and find the general solution as well as the particular solution, letting $t = 0$ be the instant when the account is opened. How much money will Bill have in the account at age 65 if he starts at 25 and invests \$1000 initially as well as annually, and the interest rate R is 6%? How much should he invest initially and annually (same amounts) to obtain the same final balance as before if he starts at age 45? First, guess.

26. (Mixing problem) A tank (as in Fig. 9 in Sec. 1.3) contains 1000 gal of water in which 200 lb of salt is dissolved. 50 gal of brine, each gallon containing $(1 + \cos t)$ lb of dissolved salt, runs into the tank per minute. The mixture, kept uniform by stirring, runs out at the same rate. Find the amount of salt in the tank at any time t (Fig. 20).

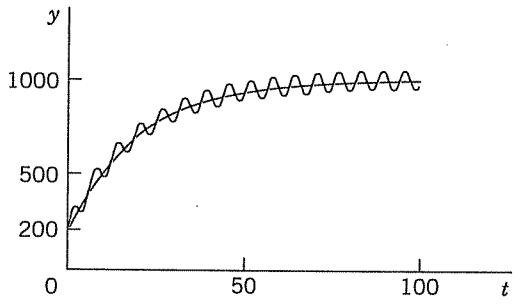


Fig. 20. Amount of salt $y(t)$ in the tank in Problem 26

27. (Lake Erie) Lake Erie has a water volume of about 450 km^3 and a flow rate (in and out) of about 175 km^3 per year. If at some instant the lake has pollution concentration $p = 0.04\%$, how long, approximately, will it take to decrease it to $p/2$, assuming that the inflow is much cleaner, say, it has pollution concentration $p/4$, and the mixture is uniform (an assumption that is only very imperfectly true)? First, guess.

28. (Heating and cooling of a building) Heating and cooling of a building can be modeled by the ODE

$$T' = k_1(T - T_a) + k_2(T - T_w) + P,$$

where $T = T(t)$ is the temperature in the building at time t, T_a the outside temperature, T_w the temperature wanted in the building, and P the rate of increase of T due to machines and people in the building, and k_1 and k_2 are (negative) constants. Solve this ODE, assuming $P = const$, $T_w = const$, and T_a varying sinusoidally over 24 hours, say, $T_a = A - C \cos(2\pi/24)t$. Discuss the effect of each term of the equation on the solution.

29. (Drug injection) Find and solve the model for drug injection into the bloodstream if, beginning at $t = 0$, a constant amount A g/min is injected and the drug is simultaneously removed at a rate proportional to the amount of the drug present at time t.

30. (Epidemics) A model for the spread of contagious diseases is obtained by assuming that the rate of spread is proportional to the number of contacts between infected and noninfected persons, who are assumed to move freely among each other. Set up the model. Find the equilibrium solutions and indicate their stability or instability. Solve the ODE. Find the limit of the proportion of infected persons as $t \to \infty$ and explain what it means.

31. (Extinction vs. unlimited growth) If in a population $y(t)$ the death rate is proportional to the population, and the birth rate is proportional to the chance encounters of meeting mates for reproduction, what will the model be? Without solving, find out what will eventually happen to a small initial population. To a large one. Then solve the model.

32. (Harvesting renewable resources. Fishing) Suppose that the population $y(t)$ of a certain kind of fish is given by the logistic equation (8), and fish are caught at a rate Hy proportional to y. Solve this so-called *Schaefer model.* Find the equilibrium solutions y_1 and y_2 (> 0) when $H < A$. The expression $Y = Hy_2$ is called the **equilibrium harvest** or **sustainable yield** *corresponding to H.* Why?

33. (Harvesting) In Prob. 32 find and graph the solution satisfying $y(0) = 2$ when (for simplicity) $A = B = 1$ and $H = 0.2$. What is the limit? What does it mean? What if there were no fishing?

34. (Intermittent harvesting) In Prob. 32 assume that you fish for 3 years, then fishing is banned for the next 3 years. Thereafter you start again. And so on. This is called *intermittent harvesting.* Describe qualitatively how the population will develop if intermitting is

continued periodically. Find and graph the solution for the first 9 years, assuming that $A = B = 1$, $H = 0.2$, and $y(0) = 2$.

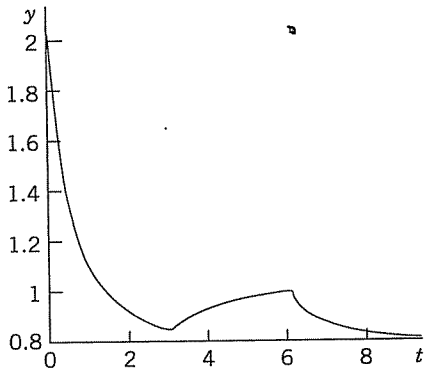


Fig. 21. Fish population in Problem 34

35. **(Harvesting)** If a population of mice (in multiples of 1000) follows the logistic law with $A = 1$ and $B = 0.25$, and if owls catch at a time rate of 10% of the population present, what is the model, its equilibrium harvest for that catch, and its solution?

36. **(Harvesting)** Do you save work in Prob. 34 if you first transform the ODE to a linear ODE? Do this transformation. Solve the resulting ODE. Does the resulting $y(t)$ agree with that in Prob. 34?

37–40 GENERAL PROPERTIES OF LINEAR ODEs

These properties are of practical and theoretical importance because they enable us to obtain new solutions from given ones. Thus in modeling, whenever possible, we prefer linear ODEs over nonlinear ones, which have no similar properties.

Show that nonhomogeneous linear ODEs (1) and homogeneous linear ODEs (2) have the following properties. Illustrate each property by a calculation for two or three equations of your choice. Give proofs.

37. The sum $y_1 + y_2$ of two solutions y_1 and y_2 of the homogeneous equation (2) is a solution of (2), and so is a scalar multiple ay_1 for any constant a. These properties are not true for (1)!

38. $y = 0$ (that is, $y(x) = 0$ for all x, also written $y(x) \equiv 0$) is a solution of (2) [not of (1) if $r(x) \neq 0$!], called the **trivial solution.**

39. The sum of a solution of (1) and a solution of (2) is a solution of (1).

40. The difference of two solutions of (1) is a solution of (2).

41. If y_1 is a solution of (1), what can you say about cy_1?

42. If y_1 and y_2 are solutions of $y_1' + py_1 = r_1$ and $y_2' + py_2 = r_2$, respectively (with the same p!), what can you say about the sum $y_1 + y_2$?

43. **CAS EXPERIMENT. (a)** Solve the ODE $y' - y/x = -x^{-1} \cos(1/x)$. Find an initial condition for which the arbitrary constant is zero. Graph the resulting particular solution, experimenting to obtain a good figure near $x = 0$.

(b) Generalizing (a) from $n = 1$ to arbitrary n, solve the ODE $y' - ny/x = -x^{n-2} \cos(1/x)$. Find an initial condition as in (a), and experiment with the graph.

44. **TEAM PROJECT. Riccati Equation, Clairaut Equation.** A **Riccati equation** is of the form

$$y' + p(x)y = g(x)y^2 + h(x). \tag{11}$$

A **Clairaut equation** is of the form

$$y = xy' + g(y'). \tag{12}$$

(a) Apply the transformation $y = Y + 1/u$ to the Riccati equation (11), where Y is a solution of (11), and obtain for u the linear ODE $u' + (2Yg - p)u = -g$. Explain the effect of the transformation by writing it as $y = Y + v$, $v = 1/u$.

(b) Show that $y = Y = x$ is a solution of $y' - (2x^3 + 1)y = -x^2y^2 - x^4 - x + 1$ and solve this Riccati equation, showing the details.

(c) Solve $y' + (3 - 2x^2 \sin x)y = -y^2 \sin x + 2x + 3x^2 - x^4 \sin x$, using (and verifying) that $y = x^2$ is a solution.

(d) By working "backward" from the u-equation find further Riccati equations that have relatively simple solutions.

(e) Solve the Clairaut equation $y = xy' + 1/y'$. *Hint.* Differentiate this ODE with respect to x.

(f) Solve the Clairaut equation $y'^2 - xy' + y = 0$ in Prob. 16 of Problem Set 1.1.

(g) Show that the Clairaut equation (12) has as solutions a family of straight lines $y = cx + g(c)$ and a singular solution determined by $g'(s) = -x$, where $s = y'$, that forms the envelope of that family.

45. **(Variation of parameter)** Another method of obtaining (4) results from the following idea. Write (3) as cy^*, where y^* is the exponential function, which is a solution of the homogeneous linear ODE $y^{*\prime} + py^* = 0$. Replace the arbitrary constant c in (3) with a function u to be determined so that the resulting function $y = uy^*$ is a solution of the nonhomogeneous linear ODE $y' + py = r$.

46. **TEAM PROJECT. Transformations of ODEs.** We have transformed ODEs to separable form, to exact form, and to linear form. The purpose of such transformations is an extension of solution methods to larger classes of ODEs. Describe the key idea of each of these transformations and give three typical examples of your choice for each transformation, showing each step (not just the transformed ODE).

1.6 Orthogonal Trajectories. *Optional*

An important type of problem in physics or geometry is to find a family of curves that intersect a given family of curves at right angles. The new curves are called **orthogonal trajectories** of the given curves (and conversely). Examples are curves of equal temperature (*isotherms*) and curves of heat flow, curves of equal altitude (*contour lines*) on a map and curves of steepest descent on that map, curves of equal potential (*equipotential curves,* curves of equal voltage—the concentric circles in Fig. 22), and curves of electric force (the straight radial segments in Fig. 22).

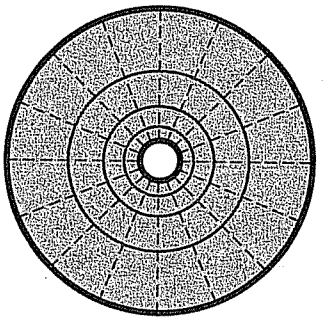


Fig. 22. Equipotential lines and curves of electric force (dashed) between two concentric (black) circles (cylinders in space)

Here the **angle of intersection** between two curves is defined to be the angle between the tangents of the curves at the intersection point. *Orthogonal* is another word for *perpendicular.*

In many cases orthogonal trajectories can be found by using ODEs, as follows. Let

$$G(x, y, c) = 0 \tag{1}$$

be a given family of curves in the xy-plane, where each curve is specified by some value of c. This is called a **one-parameter family of curves,** and c is called the *parameter* of the family. For instance, a one-parameter family of quadratic parabolas is given by (Fig. 23)

$$y = cx^2 \qquad \text{or, written as in (1),} \qquad G(x, y, c) = y - cx^2 = 0.$$

Step 1. Find an ODE for which the given family is a general solution. Of course, this ODE must no longer contain the parameter c. In our example we solve algebraically for c and then differentiate and simplify; thus,

$$\frac{y}{x^2} = c, \qquad \frac{y'x^2 - 2xy}{x^4} = 0,$$

hence

$$y' = \frac{2y}{x}.$$

The last of these equations is the ODE of the given family of curves. It is of the form

$$y' = f(x, y). \tag{2}$$

Step 2. Write down the ODE of the orthogonal trajectories, that is, the ODE whose general solution gives the orthogonal trajectories of the given curves. This ODE is

$$\tilde{y}' = -\frac{1}{f(x, \tilde{y})} \tag{3}$$

with the same f as in (2). Why? Well, a given curve passing through a point (x_0, y_0) has slope $f(x_0, y_0)$ at that point, by (2). The trajectory through (x_0, y_0) has slope $-1/f(x_0, y_0)$ by (3). The product of these slopes is -1, as we see. From calculus it is known that this is the condition for orthogonality (perpendicularity) of two straight lines (the tangents at (x_0, y_0)), hence of the curve and its orthogonal trajectory at (x_0, y_0).

Step 3. Solve (3).

For our parabolas $y = cx^2$ we have $y' = 2y/x$. Hence their orthogonal trajectories are obtained from $\tilde{y}' = -x/2\tilde{y}$ or $2\tilde{y}\tilde{y}' + x = 0$. By integration, $\tilde{y}^2 + \frac{1}{2}x^2 = c^*$. These are the ellipses in Fig. 23 with semi-axes $\sqrt{2c^*}$ and $\sqrt{c^*}$. Here, $c^* > 0$ because $c^* = 0$ gives just the origin, and $c^* < 0$ gives no real solution at all.

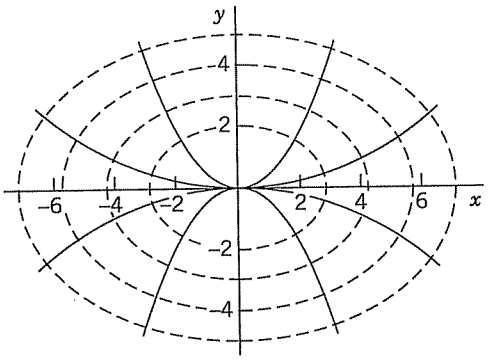


Fig. 23. Parabolas and orthogonal trajectories (ellipses) in the text

PROBLEM SET 6

1–12 ORTHOGONAL TRAJECTORIES

Sketch or graph some of the given curves. Guess what their orthogonal trajectories may look like. Find these trajectories.

(Show the details of your work.)

1. $y = 4x + c$
2. $y = c/x$
3. $y = cx$
4. $y^2 = 2x^2 + c$
5. $x^2y = c$
6. $y = ce^{-3x}$
7. $y = ce^{x^2/2}$
8. $x^2 - y^2 = c$
9. $4x^2 + y^2 = c$
10. $x = c\sqrt{y}$
11. $x = ce^{y/4}$
12. $x^2 + (y - c)^2 = c^2$

13–15 OTHER FORMS OF THE ODEs (2) AND (3)

13. **(y as independent variable)** Show that (3) may be written $dx/d\tilde{y} = -f(x, \tilde{y})$. Use this form to find the orthogonal trajectories of $y = 2x + ce^{-x}$.

14. (Family $g(x, y) = c$) Show that if a family is given as $g(x, y) = c$, then the orthogonal trajectories can be obtained from the following ODE, and use the latter to solve Prob. 6 written in the form $g(x, y) = c$.

$$\frac{d\tilde{y}}{dx} = \frac{\partial g/\partial \tilde{y}}{\partial g/\partial x}$$

15. (Cauchy–Riemann equations) Show that for a family $u(x, y) = c = const$ the orthogonal trajectories $v(x, y) = c^* = const$ can be obtained from the following *Cauchy–Riemann equations* (which are basic in complex analysis in Chap. 13) and use them to find the orthogonal trajectories of $e^x \sin y = const$. (Here, subscripts denote partial derivatives.)

$$u_x = v_y, \qquad u_y = -v_x$$

16–20 APPLICATIONS

16. (Fluid flow) Suppose that the *streamlines* of the flow (paths of the particles of the fluid) in Fig. 24 are $\Psi(x, y) = xy = const$. Find their orthogonal trajectories (called **equipotential lines,** for reasons given in Sec. 18.4).

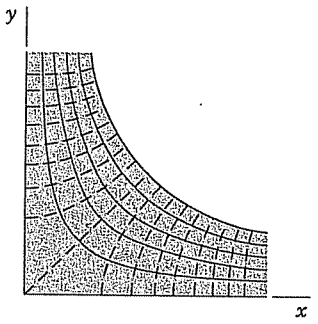


Fig. 24. Flow in a channel in Problem 16

17. (Electric field) Let the electric equipotential lines (curves of constant potential) between two concentric cylinders (Fig. 22) be given by $u(x, y) = x^2 + y^2 = c$. Use the method in the text to find their orthogonal trajectories (the curves of electric force).

18. (Electric field) The lines of electric force of two opposite charges of the same strength at $(-1, 0)$ and $(1, 0)$ are the circles through $(-1, 0)$ and $(1, 0)$. Show that these circles are given by $x^2 + (y - c)^2 = 1 + c^2$. Show that the **equipotential lines** (orthogonal trajectories of those circles) are the circles given by $(x + c^*)^2 + \tilde{y}^2 = c^{*2} - 1$ (dashed in Fig. 25).

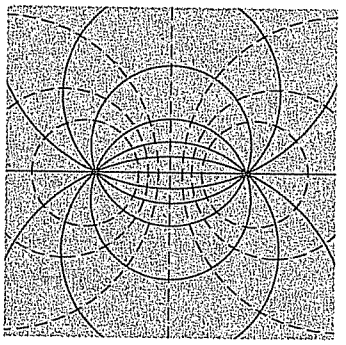


Fig. 25. Electric field in Problem 18

19. (Temperature field) Let the **isotherms** (curves of constant temperature) in a body in the upper half-plane $y > 0$ be given by $4x^2 + 9y^2 = c$. Find the orthogonal trajectories (the curves along which heat will flow in regions filled with heat-conducting material and free of heat sources or heat sinks).

20. TEAM PROJECT. Conic Sections. (A) State the main steps of the present method of obtaining orthogonal trajectories.

(B) Find conditions under which the orthogonal trajectories of families of ellipses $x^2/a^2 + y^2/b^2 = c$ are again conic sections. Illustrate your result graphically by sketches or by using your CAS. What happens if $a \to 0$? If $b \to 0$?

(C) Investigate families of hyperbolas $x^2/a^2 - y^2/b^2 = c$ in a similar fashion.

(D) Can you find more complicated curves for which you get ODEs that you can solve? Give it a try.

1.7 Existence and Uniqueness of Solutions

The initial value problem

$$|y'| + |y| = 0, \qquad y(0) = 1$$

has no solution because $y = 0$ (that is, $y(x) = 0$ for all x) is the only solution of the ODE. The initial value problem

$$y' = 2x, \qquad y(0) = 1$$

has precisely one solution, namely, $y = x^2 + 1$. The initial value problem

$$xy' = y - 1, \qquad y(0) = 1$$

has infinitely many solutions, namely, $y = 1 + cx$, where c is an arbitrary constant because $y(0) = 1$ for all c.

From these examples we see that an **initial value problem**

$$y' = f(x, y), \qquad y(x_0) = y_0 \tag{1}$$

may have no solution, precisely one solution, or more than one solution. This fact leads to the following two fundamental questions.

Problem of Existence

Under what conditions does an initial value problem of the form (1) *have at least one solution (hence one or several solutions)?*

Problem of Uniqueness

Under what conditions does that problem have at most one solution (hence excluding the case that is has more than one solution)?

Theorems that state such conditions are called **existence theorems** and **uniqueness theorems,** respectively.

Of course, for our simple examples we need no theorems because we can solve these examples by inspection; however, for complicated ODEs such theorems may be of considerable practical importance. Even when you are sure that your physical or other system behaves uniquely, occasionally your model may be oversimplified and may not give a faithful picture of the reality.

THEOREM 1

Existence Theorem

Let the right side $f(x, y)$ of the ODE in the initial value problem

$$y' = f(x, y), \qquad y(x_0) = y_0 \tag{1}$$

be continuous at all points (x, y) in some rectangle

$$R\text{: } |x - x_0| < a, \qquad |y - y_0| < b \tag{Fig. 26}$$

and **bounded** *in R; that is, there is a number K such that*

$$|f(x, y)| \leqq K \qquad \text{for all } (x, y) \text{ in } R. \tag{2}$$

Then the initial value problem (1) *has at least one solution $y(x)$. This solution exists at least for all x in the subinterval $|x - x_0| < \alpha$ of the interval $|x - x_0| < a$; here, α is the smaller of the two numbers a and b/K.*

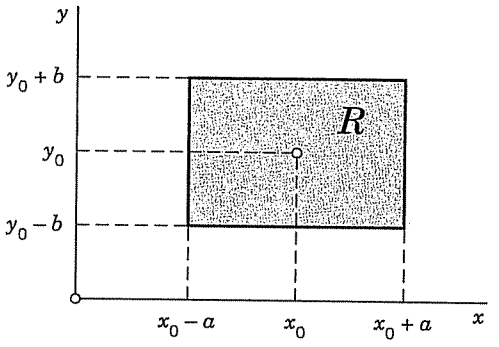


Fig. 26. Rectangle R in the existence and uniqueness theorems

(*Example of Boundedness.* The function $f(x, y) = x^2 + y^2$ is bounded (with $K = 2$) in the square $|x| < 1$, $|y| < 1$. The function $f(x, y) = \tan(x + y)$ is not bounded for $|x + y| < \pi/2$. Explain!)

THEOREM 2

Uniqueness Theorem

Let f and its partial derivative $f_y = \partial f/\partial y$ be continuous for all (x, y) in the rectangle R (Fig. 26) *and bounded, say,*

(3) (a) $|f(x, y)| \leqq K$, (b) $|f_y(x, y)| \leqq M$ *for all (x, y) in R.*

Then the initial value problem (1) *has at most one solution $y(x)$. Thus, by Theorem* 1, *the problem has precisely one solution. This solution exists at least for all x in that subinterval $|x - x_0| < \alpha$.*

Understanding These Theorems

These two theorems take care of almost all practical cases. Theorem 1 says that if $f(x, y)$ is continuous in some region in the xy-plane containing the point (x_0, y_0), then the initial value problem (1) has at least one solution.

Theorem 2 says that if, moreover, the partial derivative $\partial f/\partial y$ of f with respect to y exists and is continuous in that region, then (1) can have at most one solution; hence, by Theorem 1, it has precisely one solution.

Read again what you have just read—these are entirely new ideas in our discussion.

Proofs of these theorems are beyond the level of this book (see Ref. [A11] in App. 1); however, the following remarks and examples may help you to a good understanding of the theorems.

Since $y' = f(x, y)$, the condition (2) implies that $|y'| \leqq K$; that is, the slope of any solution curve $y(x)$ in R is at least $-K$ and at most K. Hence a solution curve that passes through the point (x_0, y_0) must lie in the colored region in Fig. 27 on the next page bounded by the lines l_1 and l_2 whose slopes are $-K$ and K, respectively. Depending on the form of R, two different cases may arise. In the first case, shown in Fig. 27a, we have $b/K \geqq a$ and therefore $\alpha = a$ in the existence theorem, which then asserts that the solution exists for all x between $x_0 - a$ and $x_0 + a$. In the second case, shown in Fig. 27b, we have $b/K < a$. Therefore, $\alpha = b/K < a$, and all we can conclude from the theorems is that the solution exists for all x between $x_0 - b/K$ and $x_0 + b/K$. For larger or smaller x's the solution curve may leave the rectangle R, and since nothing is assumed about f outside R, nothing can be concluded about the solution for those larger or smaller x's; that is, for such x's the solution may or may not exist—we don't know.

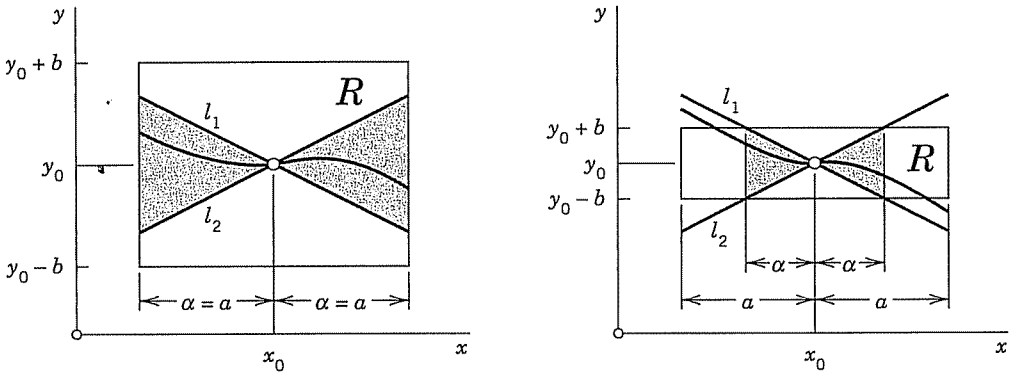


Fig. 27. The condition (2) of the existence theorem. (a) First case. (b) Second case

Let us illustrate our discussion with a simple example. We shall see that our choice of a rectangle R with a large base (a long x-interval) will lead to the case in Fig. 27b.

EXAMPLE 1 **Choice of a Rectangle**

Consider the initial value problem

$$y' = 1 + y^2, \qquad y(0) = 0$$

and take the rectangle R; $|x| < 5$, $|y| < 3$. Then $a = 5$, $b = 3$, and

$$|f(x, y)| = |1 + y^2| \leqq K = 10,$$

$$\left|\frac{\partial f}{\partial y}\right| = 2|y| \leqq M = 6,$$

$$\alpha = \frac{b}{K} = 0.3 < a.$$

Indeed, the solution of the problem is $y = \tan x$ (see Sec. 1.3, Example 1). This solution is discontinuous at $\pm\pi/2$, and there is no *continuous* solution valid in the entire interval $|x| < 5$ from which we started. ■

The conditions in the two theorems are sufficient conditions rather than necessary ones, and can be lessened. In particular, by the mean value theorem of differential calculus we have

$$f(x, y_2) - f(x, y_1) = (y_2 - y_1)\left.\frac{\partial f}{\partial y}\right|_{y=\tilde{y}}$$

where (x, y_1) and (x, y_2) are assumed to be in R, and $\tilde{y}$ is a suitable value between y_1 and y_2. From this and (3b) it follows that

$$|f(x, y_2) - f(x, y_1)| \leqq M|y_2 - y_1|. \tag{4}$$

It can be shown that (3b) may be replaced by the weaker condition (4), which is known as a **Lipschitz condition.**[7] However, continuity of $f(x, y)$ is not enough to guarantee the *uniqueness* of the solution. This may be illustrated by the following example.

EXAMPLE 2 Nonuniqueness

The initial value problem

$$y' = \sqrt{|y|}, \qquad y(0) = 0$$

has the two solutions

$$y \equiv 0 \qquad \text{and} \qquad y^* = \begin{cases} x^2/4 & \text{if } x \geqq 0 \\ -x^2/4 & \text{if } x < 0 \end{cases}$$

although $f(x, y) = \sqrt{|y|}$ is continuous for all y. The Lipschitz condition (4) is violated in any region that includes the line $y = 0$, because for $y_1 = 0$ and positive y_2 we have

$$\text{(5)} \qquad \frac{|f(x, y_2) - f(x, y_1)|}{|y_2 - y_1|} = \frac{\sqrt{y_2}}{y_2} = \frac{1}{\sqrt{y_2}}, \qquad (\sqrt{y_2} > 0)$$

and this can be made as large as we please by choosing y_2 sufficiently small, whereas (4) requires that the quotient on the left side of (5) should not exceed a fixed constant M. ■

PROBLEM SET 1.7

1. **(Vertical strip)** If the assumptions of Theorems 1 and 2 are satisfied not merely in a rectangle but in a vertical infinite strip $|x - x_0| < a$, in what interval will the solution of (1) exist?
2. **(Existence?)** Does the initial value problem $(x - 1)y' = 2y$, $y(1) = 1$ have a solution? Does your result contradict our present theorems?
3. **(Common points)** Can two solution curves of the same ODE have a common point in a rectangle in which the assumptions of the present theorems are satisfied?
4. **(Change of initial condition)** What happens in Prob. 2 if you replace $y(1) = 1$ with $y(1) = k$?
5. **(Linear ODE)** If p and r in $y' + p(x)y = r(x)$ are continuous for all x in an interval $|x - x_0| \leqq a$, show that $f(x, y)$ in this ODE satisfies the conditions of our present theorems, so that a corresponding initial value problem has a unique solution. Do you actually need these theorems for this ODE?
6. **(Three possible cases)** Find all initial conditions such that $(x^2 - 4x)y' = (2x - 4)y$ has no solution, precisely one solution, and more than one solution.
7. **(Length of x-interval)** In most cases the solution of an initial value problem (1) exists in an x-interval larger than that guaranteed by the present theorems. Show this fact for $y' = 2y^2$, $y(1) = 1$ by finding the best possible α (choosing b optimally) and comparing the result with the actual solution.
8. **PROJECT. Lipschitz Condition. (A)** State the definition of a Lipschitz condition. Explain its relation to the existence of a partial derivative. Explain its significance in our present context. Illustrate your statements by examples of your own.

   (B) Show that for a ***linear*** ODE $y' + p(x)y = r(x)$ with continuous p and r in $|x - x_0| \leqq a$ a Lipschitz condition holds. This is remarkable because it means that for a *linear* ODE the continuity of $f(x, y)$ guarantees not only the existence but also the uniqueness of the solution of an initial value problem. (Of course, this also follows directly from (4) in Sec. 1.5.)

   (C) Discuss the uniqueness of solution for a few simple ODEs that you can solve by one of the methods considered, and find whether a Lipschitz condition is satisfied.
9. **(Maximum α)** What is the largest possible α in Example 1 in the text?
10. **CAS PROJECT. Picard Iteration. (A)** Show that by integrating the ODE in (1) and observing the initial condition you obtain

    $$\text{(6)} \qquad y(x) = y_0 + \int_{x_0}^{x} f(t, y(t))\, dt.$$

[7]RUDOLF LIPSCHITZ (1832–1903), German mathematician. Lipschitz and similar conditions are important in modern theories, for instance, in partial differential equations.

This form (6) of (1) suggests **Picard's iteration method**[8], which is defined by

$$(7)\quad y_n(x) = y_0 + \int_{x_0}^{x} f(t, y_{n-1}(t))\, dt, \quad n = 1, 2, \cdots.$$

It gives approximations $y_1, y_2, y_3, \cdots$ of the unknown solution y of (1). Indeed, you obtain y_1 by substituting $y = y_0$ on the right and integrating—this is the first step—, then y_2 by substituting $y = y_1$ on the right and integrating—this is the second step—, and so on. Write a program of the iteration that gives a printout of the first approximations $y_0, y_1, \cdots, y_N$ as well as their graphs on common axes. Try your program on two initial value problems of your own choice.

(B) Apply the iteration to $y' = x + y$, $y(0) = 0$. Also solve the problem exactly.

(C) Apply the iteration to $y' = 2y^2$, $y(0) = 1$. Also solve the problem exactly.

(D) Find all solutions of $y' = 2\sqrt{y}$, $y(1) = 0$. Which of them does Picard's iteration approximate?

(E) Experiment with the conjecture that Picard's iteration converges to the solution of the problem for any initial choice of y in the integrand in (7) (leaving y_0 outside the integral as it is). Begin with a simple ODE and see what happens. When you are reasonably sure, take a slightly more complicated ODE and give it a try.

CHAPTER 1 REVIEW QUESTIONS AND PROBLEMS

1. Explain the terms *ordinary differential equation (ODE), partial differential equation (PDE), order, general solution,* and *particular solution.* Give examples. Why are these concepts of importance?
2. What is an initial condition? How is this condition used in an initial value problem?
3. What is a homogeneous linear ODE? A nonhomogeneous linear ODE? Why are these equations simpler than nonlinear ODEs?
4. What do you know about direction fields and their practical importance?
5. Give examples of mechanical problems that lead to ODEs.
6. Why do electric circuits lead to ODEs?
7. Make a list of the solution methods considered. Explain each method with a few short sentences and illustrate it by a typical example.
8. Can certain ODEs be solved by more than one method? Give three examples.
9. What are integrating factors? Explain the idea. Give examples.
10. Does every first-order ODE have a solution? A general solution? What do you know about uniqueness of solutions?

11–14 DIRECTION FIELDS

Graph a direction field (by a CAS or by hand) and sketch some of the solution curves. Solve the ODE exactly and compare.

11. $y' = 1 + 4y^2$ **12.** $y' = 3y - 2x$

13. $y' = 4y - y^2$ **14.** $y' = 16x/y$

15–26 GENERAL SOLUTION

Find the general solution. Indicate which method in this chapter you are using. Show the details of your work.

15. $y' = x^2(1 + y^2)$

16. $y' = x(y - x^2 + 1)$

17. $yy' + xy^2 = x$

18. $-\pi \sin \pi x \cosh 3y\, dx + 3 \cos \pi x \sinh 3y\, dy = 0$

19. $y' + y \sin x = \sin x$ **20.** $y' - y = 1/y$

21. $3 \sin 2y\, dx + 2x \cos 2y\, dy = 0$

22. $xy' = x \tan (y/x) + y$

23. $(y \cos xy - 2x)\, dx + (x \cos xy + 2y)\, dy = 0$

24. $xy' = (y - 2x)^2 + y$ (Set $y - 2x = z$.)

25. $\sin (y - x)\, dx + [\cos (y - x) - \sin (y - x)]\, dy = 0$

26. $xy' = (y/x)^3 + y$

27–32 INITIAL VALUE PROBLEMS

Solve the following initial value problems. Indicate the method used. Show the details of your work.

27. $yy' + x = 0, \quad y(3) = 4$

28. $y' - 3y = -12y^2, \quad y(0) = 2$

29. $y' = 1 + y^2, \quad y(\frac{1}{4}\pi) = 0$

30. $y' + \pi y = 2b \cos \pi x, \quad y(0) = 0$

31. $(2xy^2 - \sin x)\, dx + (2 + 2x^2 y)\, dy = 0, \; y(0) = 1$

32. $[2y + y^2/x + e^x(1 + 1/x)]\, dx + (x + 2y)\, dy = 0, \; y(1) = 1$

[8] EMILE PICARD (1856–1941), French mathematician, also known for his important contributions to complex analysis (see Sec. 16.2 for his famous theorem). Picard used his method to prove Theorems 1 and 2 as well as the convergence of the sequence (7) to the solution of (1). In precomputer times the iteration was of little *practical* value because of the integrations.

33–43 APPLICATIONS, MODELING

33. (Heat flow) If the isotherms in a region are $x^2 - y^2 = c$, what are the curves of heat flow (assuming orthogonality)?

34. (Law of cooling) A thermometer showing 10°C is brought into a room whose temperature is 25°C. After 5 minutes it shows 20°C. When will the thermometer practically reach the room temperature, say, 24.9°C?

35. (Half-life) If 10% of a radioactive substance disintegrates in 4 days, what is its half-life?

36. (Half-life) What is the half-life of a substance if after 5 days, 0.020 g is present and after 10 days, 0.015 g?

37. (Half-life) When will 99% of the substance in Prob. 35 have disintegrated?

38. (Air circulation) In a room containing 20 000 ft^3 of air, 600 ft^3 of fresh air flows in per minute, and the mixture (made practically uniform by circulating fans) is exhausted at a rate of 600 cubic feet per minute (cfm). What is the amount of fresh air $y(t)$ at any time if $y(0) = 0$? After what time will 90% of the air be fresh?

39. (Electric field) If the equipotential lines in a region of the xy-plane are $4x^2 + y^2 = c$, what are the curves of the electrical force? Sketch both families of curves.

40. (Chemistry) In a bimolecular reaction $A + B \to M$, a moles per liter of a substance A and b moles per liter of a substance B are combined. Under constant temperature the rate of reaction is

$$y' = k(a - y)(b - y) \qquad \textbf{(Law of mass action);}$$

that is, y' is proportional to the product of the concentrations of the substances that are reacting, where $y(t)$ is the number of moles per liter which have reacted after time t. Solve this ODE, assuming that $a \neq b$.

41. (Population) Find the population $y(t)$ if the birth rate is proportional to $y(t)$ and the death rate is proportional to the square of $y(t)$.

42. (Curves) Find all curves in the first quadrant of the xy-plane such that for every tangent, the segment between the coordinate axes is bisected by the point of tangency. (Make a sketch.)

43. (Optics) Lambert's law of absorption[9] states that the absorption of light in a thin transparent layer is proportional to the thickness of the layer and to the amount of light incident on that layer. Formulate this law as an ODE and solve it.

SUMMARY OF CHAPTER 1
First-Order ODEs

This chapter concerns **ordinary differential equations (ODEs) of first order** and their applications. These are equations of the form

(1) $\qquad F(x, y, y') = 0 \qquad$ or in explicit form $\qquad y' = f(x, y)$

involving the derivative $y' = dy/dx$ of an unknown function y, given functions of x, and, perhaps, y itself. If the independent variable x is time, we denote it by t.

In Sec. 1.1 we explained the basic concepts and the process of **modeling,** that is, of expressing a physical or other problem in some mathematical form and solving it. Then we discussed the method of direction fields (Sec. 1.2), solution methods and models (Secs. 1.3–1.6), and, finally, ideas on existence and uniqueness of solutions (Sec. 1.7).

[9]JOHANN HEINRICH LAMBERT (1728–1777), German physicist and mathematician.

A first-order ODE usually has a **general solution,** that is, a solution involving an arbitrary constant, which we denote by c. In applications we usually have to find a unique solution by determining a value of c from an **initial condition** $y(x_0) = y_0$. Together with the ODE this is called an **initial value problem**

(2) $$y' = f(x, y), \qquad y(x_0) = y_0 \quad (x_0, y_0 \text{ given numbers})$$

and its solution is a **particular solution** of the ODE. Geometrically, a general solution represents a family of curves, which can be graphed by using **direction fields** (Sec. 1.2). And each particular solution corresponds to one of these curves.

A **separable ODE** is one that we can put into the form

(3) $$g(y)\,dy = f(x)\,dx \qquad \text{(Sec. 1.3)}$$

by algebraic manipulations (possibly combined with transformations, such as $y/x = u$) and solve by integrating on both sides.

An **exact ODE** is of the form

(4) $$M(x, y)\,dx + N(x, y)\,dy = 0 \qquad \text{(Sec. 1.4)}$$

where $M\,dx + N\,dy$ is the **differential**

$$du = u_x\,dx + u_y\,dy$$

of a function $u(x, y)$, so that from $du = 0$ we immediately get the implicit general solution $u(x, y) = c$. This method extends to nonexact ODEs that can be made exact by multiplying them by some function $F(x, y)$, called an **integrating factor** (Sec. 1.4).

Linear ODEs

(5) $$y' + p(x)y = r(x)$$

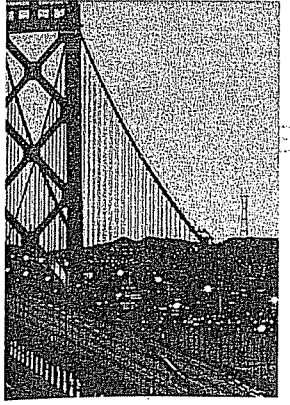
are very important. Their solutions are given by the integral formula (4), Sec. 1.5. Certain nonlinear ODEs can be transformed to linear form in terms of new variables. This holds for the **Bernoulli equation**

$$y' + p(x)y = g(x)y^a \qquad \text{(Sec. 1.5).}$$

Applications and **modeling** are discussed throughout the chapter, in particular in Secs. 1.1, 1.3, 1.5 (**population dynamics,** etc.), and 1.6 (**trajectories**).

Picard's **existence** and **uniqueness theorems** are explained in Sec. 1.7 (and *Picard's iteration* in Problem Set 1.7).

Numeric methods for first-order ODEs can be studied in Secs. 21.1 and 21.2 immediately after this chapter, as indicated in the chapter opening.



CHAPTER 2

Second-Order Linear ODEs

Ordinary differential equations (ODEs) may be divided into two large classes, **linear ODEs** and **nonlinear ODEs.** Whereas nonlinear ODEs of second (and higher) order generally are difficult to solve, linear ODEs are much simpler because various properties of their solutions can be characterized in a general way, and there are standard methods for solving many of these equations.

Linear ODEs ***of the second order*** are the most important ones because of their applications in mechanical and electrical engineering (Secs. 2.4, 2.8, 2.9). And their theory is typical of that of all linear ODEs, but the formulas are simpler than for higher order equations. Also the transition to higher order (in Chap. 3) will be almost immediate.

This chapter includes the derivation of general and particular solutions, the latter in connection with initial value problems.

(Boundary value problems follow in Chap. 5, which also contains solution methods for Legendre's, Bessel's, and the hypergeometric equations.)

COMMENT. ***Numerics for second-order ODEs can be studied immediately after this chapter.*** See Sec. 21.3, which is independent of other sections in Chaps. 19–21.

Prerequisite: Chap. 1, in particular, Sec. 1.5.
Sections that may be omitted in a shorter course: 2.3, 2.9, 2.10.
References and Answers to Problems: App. 1 Part A, and App. 2.

2.1 Homogeneous Linear ODEs of Second Order

We have already considered first-order linear ODEs (Sec. 1.5) and shall now define and discuss linear ODEs of second order. These equations have important engineering applications, especially in connection with mechanical and electrical vibrations (Secs. 2.4, 2.8, 2.9) as well as in wave motion, heat conduction, and other parts of physics, as we shall see in Chap. 12.

A second-order ODE is called **linear** if it can be written

$$y'' + p(x)y' + q(x)y = r(x) \tag{1}$$

and **nonlinear** if it cannot be written in this form.

The distinctive feature of this equation is that it is *linear in y and its derivatives,* whereas the functions p, q, and r on the right may be any given functions of x. If the equation begins with, say, $f(x)y''$, then divide by $f(x)$ to have the **standard form** (1) with y'' as the first term, which is practical.

If $r(x) \equiv 0$ (that is, $r(x) = 0$ for all x considered; read "$r(x)$ is identically zero"), then (1) reduces to

$$y'' + p(x)y' + q(x)y = 0 \tag{2}$$

and is called **homogeneous.** If $r(x) \not\equiv 0$, then (1) is called **nonhomogeneous.** This is similar to Sec. 1.5.

For instance, a nonhomogeneous linear ODE is

$$y'' + 25y = e^{-x} \cos x,$$

and a homogeneous linear ODE is

$$xy'' + y' + xy = 0, \qquad \text{in standard form} \qquad y'' + \frac{1}{x}y' + y = 0.$$

An example of a nonlinear ODE is

$$y''y + y'^2 = 0.$$

The functions p and q in (1) and (2) are called the **coefficients** of the ODEs.

Solutions are defined similarly as for first-order ODEs in Chap. 1. A function

$$y = h(x)$$

is called a *solution* of a (linear or nonlinear) second-order ODE on some open interval I if h is defined and twice differentiable throughout that interval and is such that the ODE becomes an identity if we replace the unknown y by h, the derivative y' by h', and the second derivative y'' by h''. Examples are given below.

Homogeneous Linear ODEs: Superposition Principle

Sections 2.1–2.6 will be devoted to **homogeneous** linear ODEs (2) and the remaining sections of the chapter to nonhomogeneous linear ODEs.

Linear ODEs have a rich solution structure. For the homogeneous equation the backbone of this structure is the *superposition principle* or *linearity principle,* which says that we can obtain further solutions from given ones by adding them or by multiplying them with any constants. Of course, this is a great advantage of homogeneous linear ODEs. Let us first discuss an example.

EXAMPLE 1 **Homogeneous Linear ODEs: Superposition of Solutions**

The functions $y = \cos x$ and $y = \sin x$ are solutions of the homogeneous linear ODE

$$y'' + y = 0$$

for all x. We verify this by differentiation and substitution. We obtain $(\cos x)'' = -\cos x$; hence

$$y'' + y = (\cos x)'' + \cos x = -\cos x + \cos x = 0.$$

Similarly for $y = \sin x$ (verify!). We can go an important step further. We multiply $\cos x$ by any constant, for instance, 4.7, and $\sin x$ by, say, -2, and take the sum of the results, claiming that it is a solution. Indeed, differentiation and substitution gives

$$(4.7 \cos x - 2 \sin x)'' + (4.7 \cos x - 2 \sin x) = -4.7 \cos x + 2 \sin x + 4.7 \cos x - 2 \sin x = 0.$$

In this example we have obtained from y_1 (= cos x) and y_2 (= sin x) a function of the form

$$y = c_1y_1 + c_2y_2 \qquad (c_1, c_2 \text{ arbitrary constants}). \tag{3}$$

This is called a **linear combination** of y_1 and y_2. In terms of this concept we can now formulate the result suggested by our example, often called the **superposition principle** or **linearity principle.**

THEOREM 1

Fundamental Theorem for the Homogeneous Linear ODE (2)

For a homogeneous linear ODE (2), *any linear combination of two solutions on an open interval I is again a solution of* (2) *on I. In particular, for such an equation, sums and constant multiples of solutions are again solutions.*

PROOF Let y_1 and y_2 be solutions of (2) on I. Then by substituting $y = c_1y_1 + c_2y_2$ and its derivatives into (2), and using the familiar rule $(c_1y_1 + c_2y_2)' = c_1y_1' + c_2y_2'$, etc., we get

$$\begin{aligned} y'' + py' + qy &= (c_1y_1 + c_2y_2)'' + p(c_1y_1 + c_2y_2)' + q(c_1y_1 + c_2y_2) \\ &= c_1y_1'' + c_2y_2'' + p(c_1y_1' + c_2y_2') + q(c_1y_1 + c_2y_2) \\ &= c_1(y_1'' + py_1' + qy_1) + c_2(y_2'' + py_2' + qy_2) = 0, \end{aligned}$$

since in the last line, $(\cdots) = 0$ because y_1 and y_2 are solutions, by assumption. This shows that y is a solution of (2) on I. ▩

CAUTION! Don't forget that this highly important theorem holds for *homogeneous linear* ODEs only but ***does not hold*** for nonhomogeneous linear or nonlinear ODEs, as the following two examples illustrate.

EXAMPLE 2 **A Nonhomogeneous Linear ODE**

Verify by substitution that the functions $y = 1 + \cos x$ and $y = 1 + \sin x$ are solutions of the nonhomogeneous linear ODE

$$y'' + y = 1,$$

but their sum is not a solution. Neither is, for instance, $2(1 + \cos x)$ or $5(1 + \sin x)$. ▩

EXAMPLE 3 **A Nonlinear ODE**

Verify by substitution that the functions $y = x^2$ and $y = 1$ are solutions of the nonlinear ODE

$$y''y - xy' = 0,$$

but their sum is not a solution. Neither is $-x^2$, so you cannot even multiply by -1! ▩

Initial Value Problem. Basis. General Solution

Recall from Chap. 1 that for a first-order ODE, an *initial value problem* consists of the ODE and one *initial condition* $y(x_0) = y_0$. The initial condition is used to determine the *arbitrary constant* c in the *general solution* of the ODE. This results in a unique solution, as we need it in most applications. That solution is called a *particular solution* of the ODE. These ideas extend to second-order equations as follows.

For a second-order homogeneous linear ODE (2) an **initial value problem** consists of (2) and two **initial conditions**

$$y(x_0) = K_0, \qquad y'(x_0) = K_1. \tag{4}$$

These conditions prescribe given values K_0 and K_1 of the solution and its first derivative (the slope of its curve) at the same given $x = x_0$ in the open interval considered.

The conditions (4) are used to determine the two arbitrary constants c_1 and c_2 in a **general solution**

$$y = c_1 y_1 + c_2 y_2 \tag{5}$$

of the ODE; here, y_1 and y_2 are suitable solutions of the ODE, with "suitable" to be explained after the next example. This results in a unique solution, passing through the point (x_0, K_0) with K_1 as the tangent direction (the slope) at that point. That solution is called a **particular solution** of the ODE (2).

EXAMPLE 4 **Initial Value Problem**

Solve the initial value problem

$$y'' + y = 0, \qquad y(0) = 3.0, \qquad y'(0) = -0.5.$$

Solution. ***Step 1. General solution.*** The functions $\cos x$ and $\sin x$ are solutions of the ODE (by Example 1), and we take

$$y = c_1 \cos x + c_2 \sin x.$$

This will turn out to be a general solution as defined below.

Step 2. Particular solution. We need the derivative $y' = -c_1 \sin x + c_2 \cos x$. From this and the initial values we obtain, since $\cos 0 = 1$ and $\sin 0 = 0$,

$$y(0) = c_1 = 3.0 \qquad \text{and} \qquad y'(0) = c_2 = -0.5.$$

This gives as the solution of our initial value problem the particular solution

$$y = 3.0 \cos x - 0.5 \sin x.$$

Figure 28 shows that at $x = 0$ it has the value 3.0 and the slope -0.5, so that its tangent intersects the x-axis at $x = 3.0/0.5 = 6.0$. (The scales on the axes differ!) ■

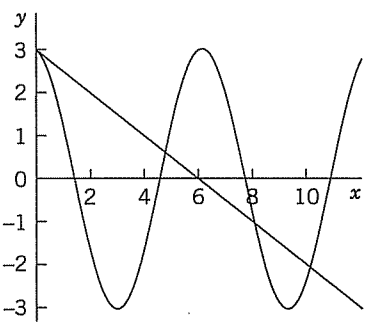


Fig. 28. Particular solution and initial tangent in Example 4

Observation. Our choice of y_1 and y_2 was general enough to satisfy both initial conditions. Now let us take instead two proportional solutions $y_1 = \cos x$ and

$y_2 = k \cos x$, so that $y_1/y_2 = 1/k = const.$ Then we can write $y = c_1y_1 + c_2y_2$ in the form

$$y = c_1 \cos x + c_2(k \cos x) = C \cos x \qquad \text{where} \qquad C = c_1 + c_2k.$$

Hence we are no longer able to satisfy two initial conditions with only one arbitrary constant C. Consequently, in defining the concept of a general solution, we must exclude proportionality. And we see at the same time why the concept of a general solution is of importance in connection with initial value problems.

DEFINITION

General Solution, Basis, Particular Solution

A **general solution** of an ODE (2) on an open interval I is a solution (5) in which y_1 and y_2 are solutions of (2) on I that are not proportional, and c_1 and c_2 are arbitrary constants. These y_1, y_2 are called a **basis** (or a **fundamental system**) of solutions of (2) on I.

A **particular solution** of (2) on I is obtained if we assign specific values to c_1 and c_2 in (5).

For the definition of an *interval* see Sec. 1.1. Also, c_1 and c_2 must sometimes be restricted to some interval in order to avoid complex expressions in the solution. Furthermore, as usual, y_1 and y_2 are called *proportional* on I if for all x on I,

$$(6) \qquad (a) \quad y_1 = ky_2 \qquad \text{or} \qquad (b) \quad y_2 = ly_1$$

where k and l are numbers, zero or not. (Note that (a) implies (b) if and only if $k \neq 0$).

Actually, we can reformulate our definition of a basis by using a concept of general importance. Namely, two functions y_1 and y_2 are called **linearly independent** on an interval I where they are defined if

$$(7) \qquad k_1y_1(x) + k_2y_2(x) = 0 \qquad \text{everywhere on } I \text{ implies} \qquad k_1 = 0 \text{ and } k_2 = 0.$$

And y_1 and y_2 are called **linearly dependent** on I if (7) also holds for some constants k_1, k_2 not both zero. Then if $k_1 \neq 0$ or $k_2 \neq 0$, we can divide and see that y_1 and y_2 are proportional,

$$y_1 = -\frac{k_2}{k_1} y_2 \qquad \text{or} \qquad y_2 = -\frac{k_1}{k_2} y_1.$$

In contrast, in the case of linear *independence* these functions are not proportional because then we cannot divide in (7). This gives the following

DEFINITION

Basis (Reformulated)

A **basis** of solutions of (2) on an open interval I is a pair of linearly independent solutions of (2) on I.

If the coefficients p and q of (2) are continuous on some open interval I, then (2) has a general solution. It yields the unique solution of any initial value problem (2), (4). It

includes all solutions of (2) on I; hence (2) has no *singular solutions* (solutions not obtainable from of a general solution; see also Problem Set 1.1). All this will be shown in Sec. 2.6.

EXAMPLE 5 Basis, General Solution, Particular Solution

$\cos x$ and $\sin x$ in Example 4 form a basis of solutions of the ODE $y'' + y = 0$ for all x because their quotient is $\cot x \neq const$ (or $\tan x \neq const$). Hence $y = c_1 \cos x + c_2 \sin x$ is a general solution. The solution $y = 3.0 \cos x - 0.5 \sin x$ of the initial value problem is a particular solution. ■

EXAMPLE 6 Basis, General Solution, Particular Solution

Verify by substitution that $y_1 = e^x$ and $y_2 = e^{-x}$ are solutions of the ODE $y'' - y = 0$. Then solve the initial value problem

$$y'' - y = 0, \qquad y(0) = 6, \qquad y'(0) = -2.$$

Solution. $(e^x)'' - e^x = 0$ and $(e^{-x})'' - e^{-x} = 0$ shows that e^x and e^{-x} are solutions. They are not proportional, $e^x/e^{-x} = e^{2x} \neq const$. Hence e^x, e^{-x} form a basis for all x. We now write down the corresponding general solution and its derivative and equate their values at 0 to the given initial conditions,

$$y = c_1e^x + c_2e^{-x}, \qquad y' = c_1e^x - c_2e^{-x}, \qquad y(0) = c_1 + c_2 = 6, \qquad y'(0) = c_1 - c_2 = -2.$$

By addition and subtraction, $c_1 = 2$, $c_2 = 4$, so that the *answer* is $y = 2e^x + 4e^{-x}$. This is the particular solution satisfying the two initial conditions. ■

Find a Basis if One Solution Is Known. Reduction of Order

It happens quite often that one solution can be found by inspection or in some other way. Then a second linearly independent solution can be obtained by solving a first-order ODE. This is called the method of **reduction of order.**[1] We first show this method for an example and then in general.

EXAMPLE 7 Reduction of Order if a Solution Is Known. Basis

Find a basis of solutions of the ODE

$$(x^2 - x)y'' - xy' + y = 0.$$

Solution. Inspection shows that $y_1 = x$ is a solution because $y_1' = 1$ and $y_1'' = 0$, so that the first term vanishes identically and the second and third terms cancel. The idea of the method is to substitute

$$y = uy_1 = ux, \qquad y' = u'x + u, \qquad y'' = u''x + 2u'$$

into the ODE. This gives

$$(x^2 - x)(u''x + 2u') - x(u'x + u) + ux = 0.$$

ux and $-xu$ cancel and we are left with the following ODE, which we divide by x, order, and simplify,

$$(x^2 - x)(u''x + 2u') - x^2u' = 0, \qquad (x^2 - x)u'' + (x - 2)u' = 0.$$

---

[1]Credited to the great mathematician JOSEPH LOUIS LAGRANGE (1736–1813), who was born in Turin, of French extraction, got his first professorship when he was 19 (at the Military Academy of Turin), became director of the mathematical section of the Berlin Academy in 1766, and moved to Paris in 1787. His important major work was in the calculus of variations, celestial mechanics, general mechanics (*Mécanique analytique,* Paris, 1788), differential equations, approximation theory, algebra, and number theory.

This ODE is of first order in $v = u'$, namely, $(x^2 - x)v' + (x - 2)v = 0$. Separation of variables and integration gives

$$\frac{dv}{v} = -\frac{x-2}{x^2-x}\,dx = \left(\frac{1}{x-1} - \frac{2}{x}\right)dx, \qquad \ln|v| = \ln|x-1| - 2\ln|x| = \ln\frac{|x-1|}{x^2}.$$

We need no constant of integration because we want to obtain a particular solution; similarly in the next integration. Taking exponents and integrating again, we obtain

$$v = \frac{x-1}{x^2} = \frac{1}{x} - \frac{1}{x^2}, \qquad u = \int v\,dx = \ln|x| + \frac{1}{x}, \qquad \text{hence} \qquad y_2 = ux = x\ln|x| + 1.$$

Since $y_1 = x$ and $y_2 = x \ln |x| + 1$ are linearly independent (their quotient is not constant), we have obtained a basis of solutions, valid for all positive x. ■

In this example we applied **reduction of order** to a homogeneous linear ODE [see (2)]

$$y'' + p(x)y' + q(x)y = 0.$$

Note that we now take the ODE in standard form, with y'', not $f(x)y''$—this is essential in applying our subsequent formulas. We assume a solution y_1 of (2) on an open interval I to be known and want to find a basis. For this we need a second linearly independent solution y_2 of (2) on I. To get y_2, we substitute

$$y = y_2 = uy_1, \qquad y' = y_2' = u'y_1 + uy_1', \qquad y'' = y_2'' = u''y_1 + 2u'y_1' + uy_1''$$

into (2). This gives

$$u''y_1 + 2u'y_1' + uy_1'' + p(u'y_1 + uy_1') + quy_1 = 0. \tag{8}$$

Collecting terms in u'', u', and u, we have

$$u''y_1 + u'(2y_1' + py_1) + u(y_1'' + py_1' + qy_1) = 0.$$

Now comes the main point. Since y_1 is a solution of (2), the expression in the last parentheses is zero. Hence u is gone, and we are left with an ODE in u' and u''. We divide this remaining ODE by y_1 and set $u' = U$, $u'' = U'$,

$$u'' + u'\,\frac{2y_1' + py_1}{y_1} = 0, \qquad \text{thus} \qquad U' + \left(\frac{2y_1'}{y_1} + p\right)U = 0.$$

This is the desired first-order ODE, the reduced ODE. Separation of variables and integration gives

$$\frac{dU}{U} = -\left(\frac{2y_1'}{y_1} + p\right)dx \qquad \text{and} \qquad \ln|U| = -2\ln|y_1| - \int p\,dx.$$

By taking exponents we finally obtain

$$U = \frac{1}{y_1^2}\,e^{-\int p\,dx}. \tag{9}$$

Here $U = u'$, so that $u = \int U\,dx$. Hence the desired second solution is

$$y_2 = y_1 u = y_1 \int U\,dx.$$

The quotient $y_2/y_1 = u = \int U\,dx$ cannot be constant (since $U > 0$), so that y_1 and y_2 form a basis of solutions.

PROBLEM SET 2.1

1–6 GENERAL SOLUTION. INITIAL VALUE PROBLEM

(More in the next problem set.) Verify by substitution that the given functions form a basis. Solve the given initial value problem. (Show the details of your work.)

1. $y'' - 16y = 0$, e^{4x}, e^{-4x}, $y(0) = 3, y'(0) = 8$
2. $y'' + 25y = 0$, $\cos 5x, \sin 5x$, $y(0) = 0.8$, $y'(0) = -6.5$
3. $y'' + 2y' + 2y = 0$, $e^{-x}\cos x, e^{-x}\sin x$, $y(0) = 1, y'(0) = -1$
4. $y'' - 6y' + 9y = 0$, e^{3x}, xe^{3x}, $y(0) = -1.4$, $y'(0) = 4.6$
5. $x^2y'' + xy' - 4y = 0$, x^2, x^{-2}, $y(1) = 11$, $y'(1) = -6$
6. $x^2y'' - 7xy' + 15y = 0$, x^3, x^5, $y(1) = 0.4$, $y'(1) = 1.0$

7–14 LINEAR INDEPENDENCE AND DEPENDENCE

Are the following functions linearly independent on the given interval?

7. x, $x \ln x$ $(0 < x < 10)$
8. $3x^2$, $2x^n$ $(0 < x < 1)$
9. e^{ax}, e^{-ax} (any interval)
10. $\cos^2 x$, $\sin^2 x$ (any interval)
11. $\ln x$, $\ln x^2$ $(x > 0)$
12. $x - 2$, $x + 2$ $(-2 < x < 2)$
13. $5 \sin x \cos x$, $3 \sin 2x$ $(x > 0)$
14. 0, $\sinh \pi x$ $(x > 0)$

REDUCTION OF ORDER is important because it gives a simpler ODE. A second-order ODE $F(x, y, y', y'') = 0$, linear or not, can be reduced to first order if y does not occur explicitly (Prob. 15) or if x does not occur explicitly (Prob. 16) or if the ODE is homogeneous linear and we know a solution (see the text).

15. **(Reduction)** Show that $F(x, y', y'') = 0$ can be reduced to first order in $z = y'$ (from which y follows by integration). Give two examples of your own.
16. **(Reduction)** Show that $F(y, y', y'') = 0$ can be reduced to a first-order ODE with y as the independent variable and $y'' = (dz/dy)z$, where $z = y'$; derive this by the chain rule. Give two examples.

17–22

Reduce to first order and solve (showing each step in detail).

17. $y'' = ky'$
18. $y'' = 1 + y'^2$
19. $yy'' = 4y'^2$
20. $xy'' + 2y' + xy = 0$, $y_1 = x^{-1}\cos x$
21. $y'' + y'^3 \sin y = 0$
22. $(1 - x^2)y'' - 2xy' + 2y = 0$, $y_1 = x$

23. **(Motion)** A small body moves on a straight line. Its velocity equals twice the reciprocal of its acceleration. If at $t = 0$ the body has distance 1 m from the origin and velocity 2 m/sec, what are its distance and velocity after 3 sec?
24. **(Hanging cable)** It can be shown that the curve $y(x)$ of an inextensible flexible homogeneous cable hanging between two fixed points is obtained by solving $y'' = k\sqrt{1 + y'^2}$, where the constant k depends on the weight. This curve is called a *catenary* (from Latin *catena* = the chain). Find and graph $y(x)$, assuming $k = 1$ and those fixed points are $(-1, 0)$ and $(1, 0)$ in a vertical xy-plane.
25. **(Curves)** Find and sketch or graph the curves passing through the origin with slope 1 for which the second derivative is proportional to the first.
26. **WRITING PROJECT. General Properties of Solutions of Linear ODEs.** Write a short essay (with proofs and simple examples of your own) that includes the following.

    (a) The superposition principle.

    (b) $y \equiv 0$ is a solution of the homogeneous equation (2) (called the **trivial solution**).

    (c) The sum $y = y_1 + y_2$ of a solution y_1 of (1) and y_2 of (2) is a solution of (1).

    (d) Explore possibilities of making further general statements on solutions of (1) and (2) (sums, differences, multiples).
27. **CAS PROJECT. Linear Independence.** Write a program for testing linear independence and dependence. Try it out on some of the problems in this problem set and on examples of your own.

2.2 Homogeneous Linear ODEs with Constant Coefficients

We shall now consider second-order homogeneous linear ODEs whose coefficients a and b are constant,

$$y'' + ay' + by = 0. \tag{1}$$

These equations have important applications, especially in connection with mechanical and electrical vibrations, as we shall see in Secs. 2.4, 2.8, and 2.9.

How to solve (1)? We remember from Sec. 1.5 that the solution of the *first-order* linear ODE with a constant coefficient k

$$y' + ky = 0$$

is an exponential function $y = ce^{-kx}$. This gives us the idea to try as a solution of (1) the function

$$y = e^{\lambda x}. \tag{2}$$

Substituting (2) and its derivatives

$$y' = \lambda e^{\lambda x} \qquad \text{and} \qquad y'' = \lambda^2 e^{\lambda x}$$

into our equation (1), we obtain

$$(\lambda^2 + a\lambda + b)e^{\lambda x} = 0.$$

Hence if λ is a solution of the important **characteristic equation** (or *auxiliary equation*)

$$\lambda^2 + a\lambda + b = 0 \tag{3}$$

then the exponential function (2) is a solution of the ODE (1). Now from elementary algebra we recall that the roots of this quadratic equation (3) are

$$\lambda_1 = \tfrac{1}{2}(-a + \sqrt{a^2 - 4b}), \qquad \lambda_2 = \tfrac{1}{2}(-a - \sqrt{a^2 - 4b}). \tag{4}$$

(3) and (4) will be basic because our derivation shows that the functions

$$y_1 = e^{\lambda_1 x} \qquad \text{and} \qquad y_2 = e^{\lambda_2 x} \tag{5}$$

are solutions of (1). Verify this by substituting (5) into (1).

From algebra we further know that the quadratic equation (3) may have three kinds of roots, depending on the sign of the discriminant $a^2 - 4b$, namely,

(Case I) *Two real roots if* $a^2 - 4b > 0$,
(Case II) *A real double root if* $a^2 - 4b = 0$,
(Case III) *Complex conjugate roots if* $a^2 - 4b < 0$.

Case I. Two Distinct Real Roots λ_1 and λ_2

In this case, a basis of solutions of (1) on any interval is

$$y_1 = e^{\lambda_1 x} \qquad \text{and} \qquad y_2 = e^{\lambda_2 x}$$

because y_1 and y_2 are defined (and real) for all x and their quotient is not constant. The corresponding general solution is

$$y = c_1 e^{\lambda_1 x} + c_2 e^{\lambda_2 x}. \tag{6}$$

EXAMPLE 1 **General Solution in the Case of Distinct Real Roots**

We can now solve $y'' - y = 0$ in Example 6 of Sec. 2.1 systematically. The characteristic equation is $\lambda^2 - 1 = 0$. Its roots are $\lambda_1 = 1$ and $\lambda_2 = -1$. Hence a basis of solutions is e^x and e^{-x} and gives the same general solution as before,

$$y = c_1 e^x + c_2 e^{-x}.$$

EXAMPLE 2 **Initial Value Problem in the Case of Distinct Real Roots**

Solve the initial value problem

$$y'' + y' - 2y = 0, \qquad y(0) = 4, \qquad y'(0) = -5.$$

Solution. ***Step 1. General solution.*** The characteristic equation is

$$\lambda^2 + \lambda - 2 = 0.$$

Its roots are

$$\lambda_1 = \tfrac{1}{2}(-1 + \sqrt{9}) = 1 \qquad \text{and} \qquad \lambda_2 = \tfrac{1}{2}(-1 - \sqrt{9}) = -2$$

so that we obtain the general solution

$$y = c_1 e^x + c_2 e^{-2x}.$$

Step 2. Particular solution. Since $y'(x) = c_1 e^x - 2c_2 e^{-2x}$, we obtain from the general solution and the initial conditions

$$y(0) = c_1 + c_2 = 4,$$
$$y'(0) = c_1 - 2c_2 = -5.$$

Hence $c_1 = 1$ and $c_2 = 3$. This gives the *answer* $y = e^x + 3e^{-2x}$. Figure 29 shows that the curve begins at $y = 4$ with a negative slope (-5, but note that the axes have different scales!), in agreement with the initial conditions.

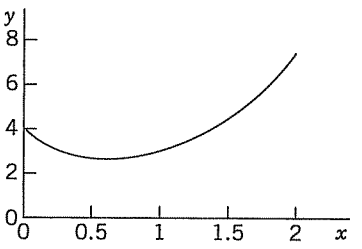


Fig. 29. Solution in Example 2

Case II. Real Double Root $\lambda = -a/2$

If the discriminant $a^2 - 4b$ is zero, we see directly from (4) that we get only one root, $\lambda = \lambda_1 = \lambda_2 = -a/2$, hence only one solution,

$$y_1 = e^{-(a/2)x}.$$

To obtain a second independent solution y_2 (needed for a basis), we use the method of reduction of order discussed in the last section, setting $y_2 = uy_1$. Substituting this and its derivatives $y_2' = u'y_1 + uy_1'$ and y_2'' into (1), we first have

$$(u''y_1 + 2u'y_1' + uy_1'') + a(u'y_1 + uy_1') + buy_1 = 0.$$

Collecting terms in u'', u', and u, as in the last section, we obtain

$$u''y_1 + u'(2y_1' + ay_1) + u(y_1'' + ay_1' + by_1) = 0.$$

The expression in the last parentheses is zero, since y_1 is a solution of (1). The expression in the first parentheses is zero, too, since

$$2y_1' = -ae^{-ax/2} = -ay_1.$$

We are thus left with $u''y_1 = 0$. Hence $u'' = 0$. By two integrations, $u = c_1x + c_2$. To get a second independent solution $y_2 = uy_1$, we can simply choose $c_1 = 1$, $c_2 = 0$ and take $u = x$. Then $y_2 = xy_1$. Since these solutions are not proportional, they form a basis. Hence in the case of a double root of (3) a basis of solutions of (1) on any interval is

$$e^{-ax/2}, \qquad xe^{-ax/2}.$$

The corresponding general solution is

$$y = (c_1 + c_2x)e^{-ax/2}. \tag{7}$$

Warning. If λ is a *simple* root of (4), then $(c_1 + c_2x)e^{\lambda x}$ with $c_2 \neq 0$ is ***not*** a solution of (1).

EXAMPLE 3 **General Solution in the Case of a Double Root**

The characteristic equation of the ODE $y'' + 6y' + 9y = 0$ is $\lambda^2 + 6\lambda + 9 = (\lambda + 3)^2 = 0$. It has the double root $\lambda = -3$. Hence a basis is e^{-3x} and xe^{-3x}. The corresponding general solution is $y = (c_1 + c_2x)e^{-3x}$. ▩

EXAMPLE 4 **Initial Value Problem in the Case of a Double Root**

Solve the initial value problem

$$y'' + y' + 0.25y = 0, \qquad y(0) = 3.0, \quad y'(0) = -3.5.$$

Solution. The characteristic equation is $\lambda^2 + \lambda + 0.25 = (\lambda + 0.5)^2 = 0$. It has the double root $\lambda = -0.5$. This gives the general solution

$$y = (c_1 + c_2x)e^{-0.5x}.$$

We need its derivative

$$y' = c_2e^{-0.5x} - 0.5(c_1 + c_2x)e^{-0.5x}.$$

From this and the initial conditions we obtain

$$y(0) = c_1 = 3.0, \qquad y'(0) = c_2 - 0.5c_1 = -3.5; \qquad \text{hence} \qquad c_2 = -2.$$

The particular solution of the initial value problem is $y = (3 - 2x)e^{-0.5x}$. See Fig. 30. ■

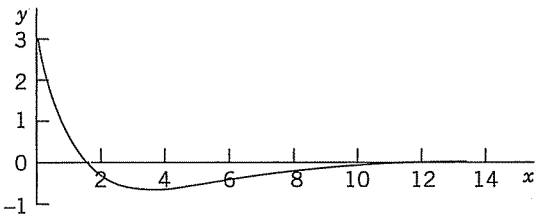


Fig. 30. Solution in Example 4

Case III. Complex Roots $-\frac{1}{2}a + i\omega$ and $-\frac{1}{2}a - i\omega$

This case occurs if the discriminant $a^2 - 4b$ of the characteristic equation (3) is negative. In this case, the roots of (3) and thus the solutions of the ODE (1) come at first out complex. However, we show that from them we can obtain a basis of *real* solutions

$$(8) \qquad y_1 = e^{-ax/2}\cos \omega x, \qquad y_2 = e^{-ax/2}\sin \omega x \qquad (\omega > 0)$$

where $\omega^2 = b - \frac{1}{4}a^2$. It can be verified by substitution that these are solutions in the present case. We shall derive them systematically after the two examples by using the complex exponential function. They form a basis on any interval since their quotient $\cot \omega x$ is not constant. Hence a real general solution in Case III is

$$(9) \qquad y = e^{-ax/2}(A\cos \omega x + B \sin \omega x) \qquad (A, B \text{ arbitrary}).$$

EXAMPLE 5 Complex Roots. Initial Value Problem

Solve the initial value problem

$$y'' + 0.4y' + 9.04y = 0, \qquad y(0) = 0, \qquad y'(0) = 3.$$

Solution. ***Step 1. General solution.*** The characteristic equation is $\lambda^2 + 0.4\lambda + 9.04 = 0$. It has the roots $-0.2 \pm 3i$. Hence $\omega = 3$, and a general solution (9) is

$$y = e^{-0.2x}(A\cos 3x + B \sin 3x).$$

Step 2. Particular solution. The first initial condition gives $y(0) = A = 0$. The remaining expression is $y = Be^{-0.2x}\sin 3x$. We need the derivative (chain rule!)

$$y' = B(-0.2e^{-0.2x}\sin 3x + 3e^{-0.2x}\cos 3x).$$

From this and the second initial condition we obtain $y'(0) = 3B = 3$. Hence $B = 1$. Our solution is

$$y = e^{-0.2x}\sin 3x.$$

Figure 31 shows y and the curves of $e^{-0.2x}$ and $-e^{-0.2x}$ (dashed), between which the curve of y oscillates. Such "damped vibrations" (with $x = t$ being time) have important mechanical and electrical applications, as we shall soon see (in Sec. 2.4). ■

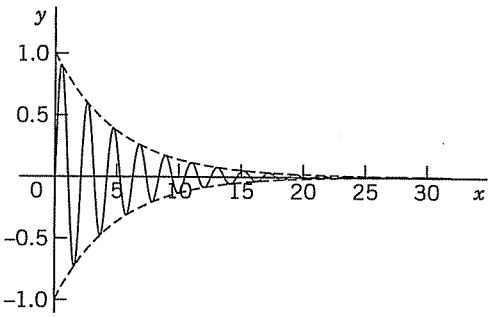


Fig. 31. Solution in Example 5

EXAMPLE 6 Complex Roots

A general solution of the ODE

$$y'' + \omega^2 y = 0 \qquad (\omega \text{ constant, not zero})$$

is

$$y = A \cos \omega x + B \sin \omega x.$$

With $\omega = 1$ this confirms Example 4 in Sec. 2.1. ■

Summary of Cases I–III

Case	Roots of (2)	Basis of (1)	General Solution of (1)
I	Distinct real λ_1, λ_2	$e^{\lambda_1 x}, e^{\lambda_2 x}$	$y = c_1 e^{\lambda_1 x} + c_2 e^{\lambda_2 x}$
II	Real double root $\lambda = -\frac{1}{2}a$	$e^{-ax/2}, xe^{-ax/2}$	$y = (c_1 + c_2 x)e^{-ax/2}$
III	Complex conjugate $\lambda_1 = -\frac{1}{2}a + i\omega$, $\lambda_2 = -\frac{1}{2}a - i\omega$	$e^{-ax/2} \cos \omega x$, $e^{-ax/2} \sin \omega x$	$y = e^{-ax/2}(A \cos \omega x + B \sin \omega x)$

It is very interesting that in applications to mechanical systems or electrical circuits, these three cases correspond to three different forms of motion or flows of current, respectively. We shall discuss this basic relation between theory and practice in detail in Sec. 2.4 (and again in Sec. 2.8).

Derivation in Case III. Complex Exponential Function

If verification of the solutions in (8) satisfies you, skip the systematic derivation of these real solutions from the complex solutions by means of the complex exponential function e^z of a complex variable $z = r + it$. We write $r + it$, not $x + iy$ because x and y occur in the ODE. The definition of e^z in terms of the real functions e^r, $\cos t$, and $\sin t$ is

(10) $$e^z = e^{r+it} = e^r e^{it} = e^r(\cos t + i \sin t).$$

This is motivated as follows. For real $z = r$, hence $t = 0$, $\cos 0 = 1$, $\sin 0 = 0$, we get the real exponential function e^r. It can be shown that $e^{z_1+z_2} = e^{z_1}e^{z_2}$, just as in real. (Proof in Sec. 13.5.) Finally, if we use the Maclaurin series of e^z with $z = it$ as well as $i^2 = -1$, $i^3 = -i$, $i^4 = 1$, etc., and reorder the terms as shown (this is permissible, as can be proved), we obtain the series

$$\begin{aligned} e^{it} &= 1 + it + \frac{(it)^2}{2!} + \frac{(it)^3}{3!} + \frac{(it)^4}{4!} + \frac{(it)^5}{5!} + \cdots \\ &= 1 - \frac{t^2}{2!} + \frac{t^4}{4!} - + \cdots + i\left(t - \frac{t^3}{3!} + \frac{t^5}{5!} - + \cdots\right) \\ &= \cos t + i \sin t. \end{aligned}$$

(Look up these real series in your calculus book if necessary.) We see that we have obtained the formula

$$e^{it} = \cos t + i \sin t, \tag{11}$$

called the **Euler formula.** Multiplication by e^r gives (10).

For later use we note that $e^{-it} = \cos(-t) + i \sin(-t) = \cos t - i \sin t$, so that by addition and subtraction of this and (11),

$$\cos t = \frac{1}{2}(e^{it} + e^{-it}), \qquad \sin t = \frac{1}{2i}(e^{it} - e^{-it}). \tag{12}$$

After these comments on the definition (10), let us now turn to Case III.

In Case III the radicand $a^2 - 4b$ in (4) is negative. Hence $4b - a^2$ is positive and, using $\sqrt{-1} = i$, we obtain in (4)

$$\tfrac{1}{2}\sqrt{a^2 - 4b} = \tfrac{1}{2}\sqrt{-(4b - a^2)} = \sqrt{-(b - \tfrac{1}{4}a^2)} = i\sqrt{b - \tfrac{1}{4}a^2} = i\omega$$

with ω defined as in (8). Hence in (4),

$$\lambda_1 = \tfrac{1}{2}a + i\omega \qquad \text{and, similarly,} \qquad \lambda_2 = \tfrac{1}{2}a - i\omega.$$

Using (10) with $r = -\frac{1}{2}ax$ and $t = \omega x$, we thus obtain

$$\begin{aligned} e^{\lambda_1 x} &= e^{-(a/2)x + i\omega x} = e^{-(a/2)x}(\cos \omega x + i \sin \omega x) \\ e^{\lambda_2 x} &= e^{-(a/2)x - i\omega x} = e^{-(a/2)x}(\cos \omega x - i \sin \omega x). \end{aligned}$$

We now add these two lines and multiply the result by $\frac{1}{2}$. This gives y_1 as in (8). Then we subtract the second line from the first and multiply the result by $1/(2i)$. This gives y_2 as in (8). These results obtained by addition and multiplication by constants are again solutions, as follows from the superposition principle in Sec. 2.1. This concludes the derivation of these real solutions in Case III.

PROBLEM SET 2.2

1–14 GENERAL SOLUTION

Find a general solution. Check your answer by substitution.

1. $y'' - 6y' - 7y = 0$
2. $10y'' - 7y' + 1.2y = 0$
3. $4y'' - 20y' + 25y = 0$
4. $y'' + 4\pi y' + 4\pi^2 y = 0$
5. $100y'' + 20y' - 99y = 0$
6. $y'' + 2y' + 5y = 0$
7. $y'' - y' + 2.5y = 0$
8. $y'' + 2.6y' + 1.69y = 0$
9. $y'' - 2y' - 5.25y = 0$
10. $y'' - 2y = 0$
11. $y'' + 9\pi^2 y = 0$
12. $y'' + 2.4y' + 4.0y = 0$
13. $y'' - 144y = 0$
14. $y'' + y' - 0.96y = 0$

15–20 FIND ODE

Find an ODE $y'' + ay' + by = 0$ for the given basis.

15. e^{2x}, e^{x}
16. $e^{0.5x}$, $e^{-3.5x}$
17. $e^{x\sqrt{3}}$, $xe^{x\sqrt{3}}$
18. 1, e^{-3x}
19. e^{4x}, e^{-4x}
20. $e^{(-1+i)x}$, $e^{-(1+i)x}$

21–32 INITIAL VALUE PROBLEMS

Solve the initial value problem. Check that your answer satisfies the ODE as well as the initial conditions. (Show the details of your work.)

21. $y'' - 2y' - 3y = 0,\ y(0) = 2,\ y'(0) = 14$
22. $y'' + 2y' + y = 0,\ y(0) = 4,\ y'(0) = -6$
23. $y'' + 4y' + 5y = 0,\ y(0) = 2,\ y'(0) = -5$
24. $10y'' - 50y' + 65y = 0,\ y(0) = 1.5,\ y'(0) = 1.5$
25. $y'' + \pi y' = 0,\ y(0) = 3,\ y'(0) = -\pi$
26. $10y'' + 18y' + 5.6y = 0,\ y(0) = 4,\ y'(0) = -3.8$
27. $10y'' + 5y' + 0.625y = 0,\ y(0) = 2,\ y'(0) = -4.5$
28. $y'' - 9y = 0,\ y(0) = -2,\ y'(0) = -12$
29. $20y'' + 4y' + y = 0,\ y(0) = 3.2,\ y'(0) = 0$
30. $y'' + 2ky' + (k^2 + \omega^2)y = 0,\ y(0) = 1,\ y'(0) = -k$
31. $y'' - 25y = 0,\ y(0) = 0,\ y'(0) = 40$
32. $y'' - 2y' - 24y = 0,\ y(0) = 0,\ y'(0) = 20$

33. **(Instability)** Solve $y'' - y = 0$ for the initial conditions $y(0) = 1, y'(0) = -1$. Then change the initial conditions to $y(0) = 1.001, y'(0) = -0.999$ and explain why this small change of 0.001 at $x = 0$ causes a large change later, e.g., 22 at $x = 10$.

34. **TEAM PROJECT. General Properties of Solutions**

(A) Coefficient formulas. Show how a and b in (1) can be expressed in terms of λ_1 and λ_2. Explain how these formulas can be used in constructing equations for given bases.

(B) Root zero. Solve $y'' + 4y' = 0$ (i) by the present method, and (ii) by reduction to first order. Can you explain why the result must be the same in both cases? Can you do the same for a general ODE $y'' + ay' = 0$?

(C) Double root. Verify directly that $xe^{\lambda x}$ with $\lambda = -a/2$ is a solution of (1) in the case of a double root. Verify and explain why $y = e^{-2x}$ is a solution of $y'' - y' - 6y = 0$ but xe^{-2x} is not.

(D) Limits. Double roots should be limiting cases of distinct roots λ_1, λ_2 as, say, $\lambda_2 \to \lambda_1$. Experiment with this idea. (Remember l'Hôpital's rule from calculus.) Can you arrive at $xe^{\lambda_1 x}$? Give it a try.

35. **(Verification)** Show by substitution that y_1 in (8) is a solution of (1).

2.3 Differential Operators. *Optional*

This short section can be omitted without interrupting the flow of ideas; it will not be used in the sequel (except for the notations Dy, D^2y, etc., for y', y'', etc.).

Operational calculus means the technique and application of operators. Here, an **operator** is a transformation that transforms a function into another function. Hence differential calculus involves an operator, the **differential operator** D, which transforms a (differentiable) function into its derivative. In operator notation we write

$$(1) \qquad Dy = y' = \frac{dy}{dx}.$$

Similarly, for the higher derivatives we write $D^2y = D(Dy) = y''$, and so on. For example, $D \sin = \cos$, $D^2 \sin = -\sin$, etc.

For a homogeneous linear ODE $y'' + ay' + by = 0$ with constant coefficients we can now introduce the **second-order differential operator**

$$L = P(D) = D^2 + aD + bI,$$

where I is the **identity operator** defined by $Iy = y$. Then we can write that ODE as

$$Ly = P(D)y = (D^2 + aD + bI)y = 0. \tag{2}$$

P suggests "polynomial." L is a **linear operator.** By definition this means that if Ly and Lw exist (this is the case if y and w are twice differentiable), then $L(cy + kw)$ exists for any constants c and k, and

$$L(cy + kw) = cLy + kLw.$$

Let us show that from (2) we reach agreement with the results in Sec. 2.2. Since $(De^{\lambda})(x) = \lambda e^{\lambda x}$ and $(D^2e^{\lambda})(x) = \lambda^2 e^{\lambda x}$, we obtain

$$\begin{aligned} Le^{\lambda}(x) &= P(D)e^{\lambda}(x) = (D^2 + aD + bI)e^{\lambda}(x) \\ &= (\lambda^2 + a\lambda + b)e^{\lambda x} = P(\lambda)e^{\lambda x} = 0. \end{aligned} \tag{3}$$

This confirms our result of Sec. 2.2 that $e^{\lambda x}$ *is a solution of the ODE* (2) *if and only if* λ *is a solution of the characteristic equation* $P(\lambda) = 0$.

$P(\lambda)$ is a polynomial in the usual sense of algebra. If we replace λ by the operator D, we obtain the "operator polynomial" $P(D)$. *The point of this operational calculus is that* $P(D)$ *can be treated just like an algebraic quantity.* In particular, we can factor it.

EXAMPLE 1 **Factorization, Solution of an ODE**

Factor $P(D) = D^2 - 3D - 40I$ and solve $P(D)y = 0$.

Solution. $D^2 - 3D - 40I = (D - 8I)(D + 5I)$ because $I^2 = I$. Now $(D - 8I)y = y' - 8y = 0$ has the solution $y_1 = e^{8x}$. Similarly, the solution of $(D + 5I)y = 0$ is $y_2 = e^{-5x}$. This is a basis of $P(D)y = 0$ on any interval. From the factorization we obtain the ODE, as expected,

$$\begin{aligned} (D - 8I)(D + 5I)y = (D - 8I)(y' + 5y) &= D(y' + 5y) - 8(y' + 5y) \\ &= y'' + 5y' - 8y' - 40y = y'' - 3y' - 40y = 0. \end{aligned}$$

Verify that this agrees with the result of our method in Sec. 2.2. This is not unexpected because we factored $P(D)$ in the same way as the characteristic polynomial $P(\lambda) = \lambda^2 - 3\lambda - 40$. ■

It was essential that L in (2) has *constant* coefficients. Extension of operator methods to variable-coefficient ODEs is more difficult and will not be considered here.

If operational methods were limited to the simple situations illustrated in this section, it would perhaps not be worth mentioning. Actually, the power of the operator approach appears in more complicated engineering problems, as we shall see in Chap. 6.

PROBLEM SET 2.3

1–5 APPLICATION OF DIFFERENTIAL OPERATORS

Apply the given operator to the given functions. (Show all steps in detail.)

1. $(D - I)^2$; e^x, xe^x, $\sin x$
2. $8D^2 + 2D - I$; $\cosh \frac{1}{2}x$, $\sinh \frac{1}{2}x$, $e^{x/2}$
3. $D - 0.4I$; $2x^3 - 1$, $e^{0.4x}$, $xe^{0.4x}$
4. $(D + 5I)(D - I)$; $e^{-5x} \sin x$, e^{5x}, x^2
5. $(D - 4I)(D + 3I)$; $x^3 - x^2$, $\sin 4x$, e^{-3x}

6–13 GENERAL SOLUTION

Factor as in the text and solve. (Show the details.)

6. $(D^2 - 5.5D + 6.66I)y = 0$
7. $(D + 2I)^2 y = 0$
8. $(D^2 - 0.49I)y = 0$
9. $(D^2 + 6D + 13I)y = 0$
10. $(10D^2 + 2D + 1.7I)y = 0$
11. $(D^2 + 4.1D + 3.1I)y = 0$
12. $(4D^2 + 4\pi D + \pi^2 I)y = 0$
13. $(D^2 + 17.64\omega^2 I)y = 0$
14. **(Double root)** If $D^2 + aD + bI$ has distinct roots μ and λ, show that a particular solution is $y = (e^{\mu x} - e^{\lambda x})/(\mu - \lambda)$. Obtain from this a solution $xe^{\lambda x}$ by letting $\mu \to \lambda$ and applying l'Hôpital's rule.
15. **(Linear operator)** Illustrate the linearity of L in (2) by taking $c = 4$, $k = -6$, $y = e^{2x}$, and $w = \cos 2x$. Prove that L is linear.
16. **(Definition of linearity)** Show that the definition of linearity in the text is equivalent to the following. If $L[y]$ and $L[w]$ exist, then $L[y + w]$ exists and $L[cy]$ and $L[kw]$ exist for all constants c and k, and $L[y + w] = L[y] + L[w]$ as well as $L[cy] = cL[y]$ and $L[kw] = kL[w]$.

2.4 Modeling: Free Oscillations (Mass–Spring System)

Linear ODEs with constant coefficients have important applications in mechanics, as we show now (and in Sec. 2.8), and in electric circuits (to be shown in Sec. 2.9). In this section we consider a basic mechanical system, a mass on an elastic spring ("mass-spring system," Fig. 32), which moves up and down. Its model will be a homogeneous linear ODE.

Setting Up the Model

We take an ordinary spring that resists compression as well extension and suspend it vertically from a fixed support, as shown in Fig. 32. At the lower end of the spring we

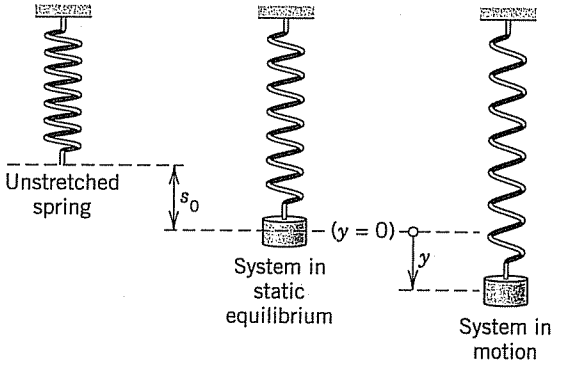


Fig. 32. Mechanical mass–spring system

attach a body of mass m. We assume m to be so large that we can neglect the mass of the spring. If we pull the body down a certain distance and then release it, it starts moving. We assume that it moves strictly vertically.

How can we obtain the motion of the body, say, the displacement $y(t)$ as function of time t? Now, this motion is determined by **Newton's second law**

$$\text{(1)} \qquad \text{Mass} \times \text{Acceleration} = my'' = \text{Force}$$

where $y'' = d^2y/dt^2$ and "Force" is the resultant of all the forces acting on the body.

(For systems of units and conversion factors, see the inside of the front cover.)

We choose the ***downward direction as the positive direction,*** thus regarding downward forces as positive and upward forces as negative.

Consider Fig. 32. The spring is first unstretched. We now attach the body. This stretches the spring by an amount s_0 shown in the figure. It causes an upward force F_0 in the spring. Experiments show that F_0 is proportional to the stretch s_0, say,

$$\text{(2)} \qquad F_0 = -ks_0 \qquad \textbf{(Hooke's law[2]).}$$

k (> 0) is called the **spring constant** (or *spring modulus*). The minus sign indicates that F_0 points upward, in our negative direction. Stiff springs have large k. (Explain!)

The extension s_0 is such that F_0 in the spring balances the weight $W = mg$ of the body (where $g = 980$ cm/sec^2 $= 32.17$ ft/sec^2 is the gravitational constant). Hence $F_0 + W = -ks_0 + mg = 0$. These forces will not affect the motion. Spring and body are again at rest. This is called the **static equilibrium** of the system (Fig. 32b). We measure the displacement $y(t)$ of the body from this 'equilibrium point' as the origin $y = 0$, downward positive and upward negative.

From the position $y = 0$ we pull the body downward. This further stretches the spring by some amount $y > 0$ (the distance we pull it down). By Hooke's law this causes an (additional) upward force F_1 in the spring,

$$F_1 = -ky.$$

F_1 is a **restoring force.** It has the tendency to *restore* the system, that is, to pull the body back to $y = 0$.

Undamped System: ODE and Solution

Every system has damping—otherwise it would keep moving forever. But practically, the effect of damping may often be negligible, for example, for the motion of an iron ball on a spring during a few minutes. Then F_1 is the only force in (1) causing the motion. Hence (1) gives the model $my'' = -ky$ or

$$\text{(3)} \qquad my'' + ky = 0.$$

[2]ROBERT HOOKE (1635–1703), English physicist, a forerunner of Newton with respect to the law of gravitation.

By the method in Sec. 2.2 (see Example 6) we obtain as a general solution

$$(4) \qquad y(t) = A \cos \omega_0 t + B \sin \omega_0 t, \qquad \omega_0 = \sqrt{\frac{k}{m}}.$$

The corresponding motion is called a **harmonic oscillation.**

Since the trigonometric functions in (4) have the period $2\pi/\omega_0$, the body executes $\omega_0/2\pi$ cycles per second. This is the **frequency** of the oscillation, which is also called the **natural frequency** of the system. It is measured in cycles per second. Another name for cycles/sec is hertz (Hz).[3]

The sum in (4) can be combined into a phase-shifted cosine with amplitude $C = \sqrt{A^2 + B^2}$ and phase angle $\delta = \arctan(B/A)$,

$$(4^*) \qquad y(t) = C \cos(\omega_0 t - \delta).$$

To verify this, apply the addition formula for the cosine [(6) in App. 3.1] to (4*) and then compare with (4). Equation (4) is simpler in connection with initial value problems, whereas (4*) is physically more informative because it exhibits the amplitude and phase of the oscillation.

Figure 33 shows typical forms of (4) and (4*), all corresponding to some positive initial displacement $y(0)$ [which determines $A = y(0)$ in (4)] and different initial velocities $y'(0)$ [which determine $B = y'(0)/\omega_0$].

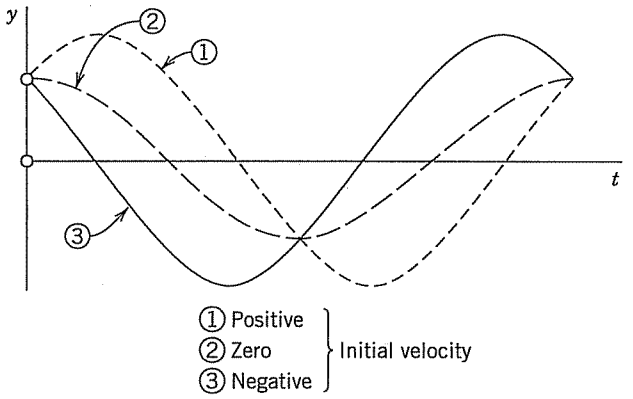


Fig. 33. Harmonic oscillations

EXAMPLE 1 Undamped Motion. Harmonic Oscillation

If an iron ball of weight $W = 98$ nt (about 22 lb) stretches a spring 1.09 m (about 43 in.), how many cycles per minute will this mass–spring system execute? What will its motion be if we pull down the weight an additional 16 cm (about 6 in.) and let it start with zero initial velocity?

Solution. Hooke's law (2) with W as the force and 1.09 meter as the stretch gives $W = 1.09k$; thus $k = W/1.09 = 98/1.09 = 90$ [kg/sec^2] $= 90$ [nt/meter]. The mass is $m = W/g = 98/9.8 = 10$ [kg]. This gives the frequency $\omega_0/(2\pi) = \sqrt{k/m}/(2\pi) = 3/(2\pi) = 0.48$ [Hz] $= 29$ [cycles/min].

[3]HEINRICH HERTZ (1857–1894), German physicist, who discovered electromagnetic waves, as the basis of wireless communication developed by GUGLIELMO MARCONI (1874–1937), Italian physicist (Nobel prize in 1909).

From (4) and the initial conditions, $y(0) = A = 0.16$ [meter] and $y'(0) = \omega_0 B = 0$. Hence the motion is

$$y(t) = 0.16 \cos 3t \text{ [meter]} \quad \text{or} \quad 0.52 \cos 3t \text{ [ft]} \quad \text{(Fig. 34).}$$

If you have a chance of experimenting with a mass–spring system, don't miss it. You will be surprised about the good agreement between theory and experiment, usually within a fraction of one percent if you measure carefully.

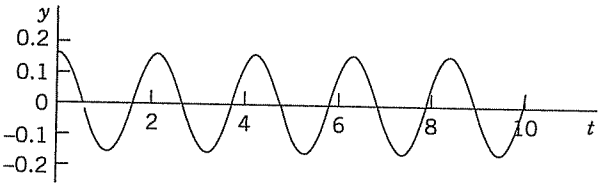


Fig. 34. Harmonic oscillation in Example 1

Damped System: ODE and Solutions

We now add a **damping force**

$$F_2 = -cy'$$

to our model $my'' = -ky$, so that we have $my'' = -ky - cy'$ or

(5) $$my'' + cy' + ky = 0.$$

Physically this can be done by connecting the body to a dashpot; see Fig. 35. We assume this new force to be proportional to the velocity $y' = dy/dt$, as shown. This is generally a good approximation, at least for small velocities.

c is called the **damping constant.** We show that c is positive. If at some instant, y' is positive, the body is moving downward (which is the positive direction). Hence the damping force $F_2 = -cy'$, always acting *against* the direction of motion, must be an upward force, which means that it must be negative, $F_2 = -cy' < 0$, so that $-c < 0$ and $c > 0$. For an upward motion, $y' < 0$ and we have a downward $F_2 = -cy > 0$; hence $-c < 0$ and $c > 0$, as before.

The ODE (5) is homogeneous linear and has constant coefficients. Hence we can solve it by the method in Sec. 2.2. The characteristic equation is (divide (5) by m)

$$\lambda^2 + \frac{c}{m}\lambda + \frac{k}{m} = 0.$$

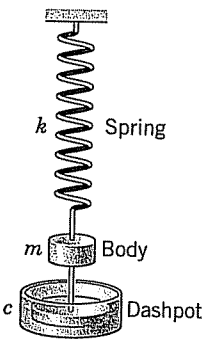


Fig. 35. Damped system

By the usual formula for the roots of a quadratic equation we obtain, as in Sec. 2.2,

$$\text{(6)} \quad \lambda_1 = -\alpha + \beta, \quad \lambda_2 = -\alpha - \beta, \quad \text{where} \quad \alpha = \frac{c}{2m} \quad \text{and} \quad \beta = \frac{1}{2m}\sqrt{c^2 - 4mk}.$$

It is now most interesting that depending on the amount of damping (much, medium, or little) there will be three types of motion corresponding to the three Cases I, II, II in Sec. 2.2:

Case I.	$c^2 > 4mk$.	*Distinct real roots* λ_1, λ_2.	**(Overdamping)**
Case II.	$c^2 = 4mk$.	*A real double root.*	**(Critical damping)**
Case III.	$c^2 < 4mk$.	*Complex conjugate roots.*	**(Underdamping)**

Discussion of the Three Cases

Case I. Overdamping

If the damping constant c is so large that $c^2 > 4mk$, then λ_1 and λ_2 are distinct real roots. In this case the corresponding general solution of (5) is

$$\text{(7)} \quad y(t) = c_1 e^{-(\alpha-\beta)t} + c_2 e^{-(\alpha+\beta)t}.$$

We see that in this case, damping takes out energy so quickly that the body does not oscillate. For $t > 0$ both exponents in (7) are negative because $\alpha > 0$, $\beta > 0$, and $\beta^2 = \alpha^2 - k/m < \alpha^2$. Hence both terms in (7) approach zero as $t \to \infty$. Practically speaking, after a sufficiently long time the mass will be at rest at the static equilibrium position ($y = 0$). Figure 36 shows (7) for some typical initial conditions.

Case II. Critical Damping

Critical damping is the border case between nonoscillatory motions (Case I) and oscillations (Case III). It occurs if the characteristic equation has a double root, that is, if $c^2 = 4mk$,

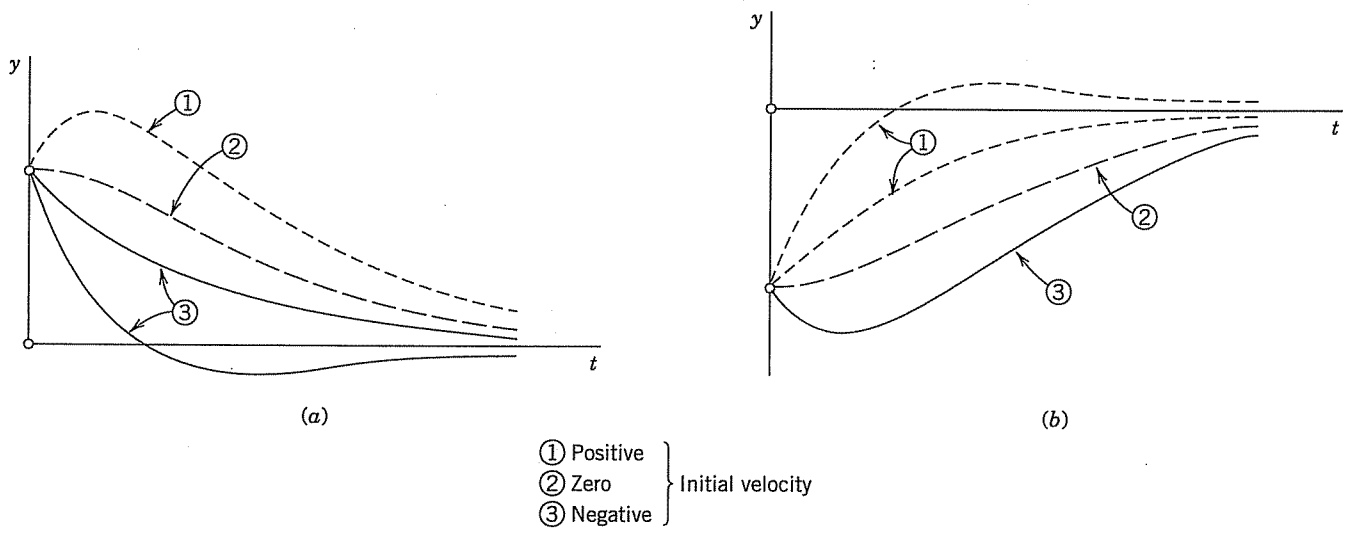


Fig. 36. Typical motions (7) in the overdamped case
(*a*) Positive initial displacement
(*b*) Negative initial displacement

so that $\beta = 0$, $\lambda_1 = \lambda_2 = -\alpha$. Then the corresponding general solution of (5) is

$$y(t) = (c_1 + c_2 t)e^{-\alpha t}. \tag{8}$$

This solution can pass through the equilibrium position $y = 0$ at most once because $e^{-\alpha t}$ is never zero and $c_1 + c_2 t$ can have at most one positive zero. If both c_1 and c_2 are positive (or both negative), it has no positive zero, so that y does not pass through 0 at all. Figure 37 shows typical forms of (8). Note that they look almost like those in the previous figure.

Case III. Underdamping

This is the most interesting case. It occurs if the damping constant c is so small that $c^2 < 4mk$. Then β in (6) is no longer real but pure imaginary, say,

$$\beta = i\omega^* \qquad \text{where} \qquad \omega^* = \frac{1}{2m}\sqrt{4mk - c^2} = \sqrt{\frac{k}{m} - \frac{c^2}{4m^2}} \quad (> 0). \tag{9}$$

(We write ω^* to reserve ω for driving and electromotive forces in Secs. 2.8 and 2.9.) The roots of the characteristic equation are now complex conjugate,

$$\lambda_1 = -\alpha + i\omega^*, \qquad \lambda_2 = -\alpha - i\omega^*$$

with $\alpha = c/(2m)$, as given in (6). Hence the corresponding general solution is

$$y(t) = e^{-\alpha t}(A \cos \omega^* t + B \sin \omega^* t) = Ce^{-\alpha t} \cos (\omega^* t - \delta) \tag{10}$$

where $C^2 = A^2 + B^2$ and $\tan \delta = B/A$, as in (4*).

This represents **damped oscillations.** Their curve lies between the dashed curves $y = Ce^{-\alpha t}$ and $y = -Ce^{-\alpha t}$ in Fig. 38, touching them when $\omega^* t - \delta$ is an integer multiple of π because these are the points at which $\cos(\omega^* t - \delta)$ equals 1 or -1.

The frequency is $\omega^*/(2\pi)$ Hz (hertz, cycles/sec). From (9) we see that the smaller c (> 0) is, the larger is ω^* and the more rapid the oscillations become. If c approaches 0, then ω^* approaches $\omega_0 = \sqrt{k/m}$, giving the harmonic oscillation (4), whose frequency $\omega_0/(2\pi)$ is the natural frequency of the system.

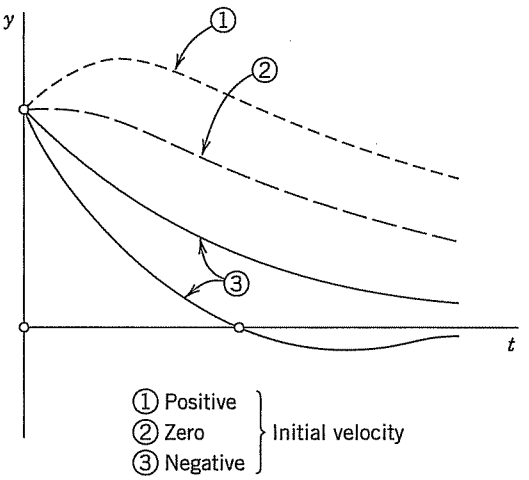


Fig. 37. Critical damping [see (8)]

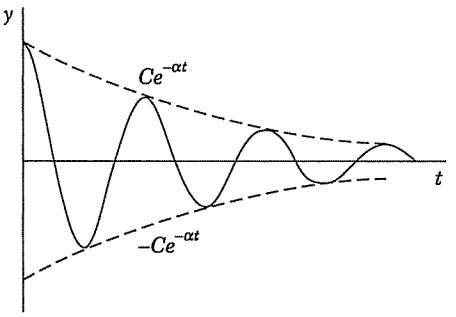


Fig. 38. Damped oscillation in Case III [see (10)]

EXAMPLE 2 The Three Cases of Damped Motion

How does the motion in Example 1 change if we change the damping constant c to one of the following three values, with $y(0) = 0.16$ and $y'(0) = 0$ as before?

$$\text{(I) } c = 100 \text{ kg/sec}, \qquad \text{(II) } c = 60 \text{ kg/sec}, \qquad \text{(III) } c = 10 \text{ kg/sec}.$$

Solution. It is interesting to see how the behavior of the system changes due to the effect of the damping, which takes energy from the system, so that the oscillations decrease in amplitude (Case III) or even disappear (Cases II and I).

(I) With $m = 10$ and $k = 90$, as in Example 1, the model is the initial value problem

$$10y'' + 100y' + 90y = 0, \qquad y(0) = 0.16 \text{ [meter]}, \qquad y'(0) = 0.$$

The characteristic equation is $10\lambda^2 + 100\lambda + 90 = 10(\lambda + 9)(\lambda + 1) = 0$. It has the roots -9 and -1. This gives the general solution

$$y = c_1e^{-9t} + c_2e^{-t}. \qquad \text{We also need} \qquad y' = -9c_1e^{-9t} - c_2e^{-t}.$$

The initial conditions give $c_1 + c_2 = 0.16$, $-9c_1 - c_2 = 0$. The solution is $c_1 = -0.02$, $c_2 = 0.18$. Hence in the overdamped case the solution is

$$y = -0.02e^{-9t} + 0.18e^{-t}.$$

It approaches 0 as $t \to \infty$. The approach is rapid; after a few seconds the solution is practically 0, that is, the iron ball is at rest.

(II) The model is as before, with $c = 60$ instead of 100. The characteristic equation now has the form $10\lambda^2 + 60\lambda + 90 = 10(\lambda + 3)^2 = 0$. It has the double root -3. Hence the corresponding general solution is

$$y = (c_1 + c_2t)e^{-3t}. \qquad \text{We also need} \qquad y' = (c_2 - 3c_1 - 3c_2t)e^{-3t}.$$

The initial conditions give $y(0) = c_1 = 0.16$, $y'(0) = c_2 - 3c_1 = 0$, $c_2 = 0.48$. Hence in the critical case the solution is

$$y = (0.16 + 0.48t)e^{-3t}.$$

It is always positive and decreases to 0 in a monotone fashion.

(III) The model now is $10y'' + 10y' + 90y = 0$. Since $c = 10$ is smaller than the critical c, we shall get oscillations. The characteristic equation is $10\lambda^2 + 10\lambda + 90 = 10\left[(\lambda + \frac{1}{2})^2 + 9 - \frac{1}{4}\right] = 0$. It has the complex roots [see (4) in Sec. 2.2 with $a = 1$ and $b = 9$]

$$\lambda = -0.5 \pm \sqrt{0.5^2 - 9} = -0.5 \pm 2.96i.$$

This gives the general solution

$$y = e^{-0.5t}(A \cos 2.96t + B \sin 2.96t).$$

Thus $y(0) = A = 0.16$. We also need the derivative

$$y' = e^{-0.5t}(-0.5A \cos 2.96t - 0.5B \sin 2.96t - 2.96A \sin 2.96t + 2.96B \cos 2.96t).$$

Hence $y'(0) = -0.5A + 2.96B = 0$, $B = 0.5A/2.96 = 0.027$. This gives the solution

$$y = e^{-0.5t}(0.16 \cos 2.96t + 0.027 \sin 2.96t) = 0.162e^{-0.5t} \cos (2.96t - 0.17).$$

We see that these damped oscillations have a smaller frequency than the harmonic oscillations in Example 1 by about 1% (since 2.96 is smaller than 3.00 by about 1%). Their amplitude goes to zero. See Fig. 39. ■

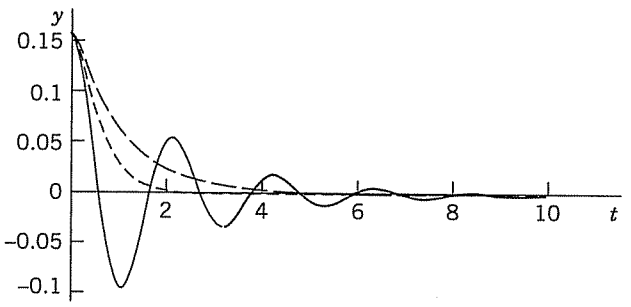


Fig. 39. The three solutions in Example 2

This section concerned ***free motions*** of mass-spring systems. Their models are *homogeneous* linear ODEs. *Nonhomogeneous* linear ODEs will arise as models of **forced motions,** that is, motions under the influence of a "driving force". We shall study them in Sec. 2.8, after we have learned how to solve those ODEs.

PROBLEM SET 2.4

1–8 MOTION WITHOUT DAMPING (HARMONIC OSCILLATIONS)

1. **(Initial value problem)** Find the harmonic motion (4) that starts from y_0 with initial velocity v_0. Graph or sketch the solutions for $\omega_0 = \pi$, $y_0 = 1$, and various v_0 of your choice on common axes. At what t-values do all these curves intersect? Why?
2. **(Spring combinations)** Find the frequency of vibration of a ball of mass $m = 3$ kg on a spring of modulus (i) $k_1 = 27$ nt/m, (ii) $k_2 = 75$ nt/m, (iii) on these springs in parallel (see Fig. 40), (iv) in series, that is, the ball hangs on one spring, which in turn hangs on the other spring.
3. **(Pendulum)** Find the frequency of oscillation of a pendulum of length L (Fig. 41), neglecting air resistance and the weight of the rod, and assuming θ to be so small that $\sin\theta$ practically equals θ.
4. **(Frequency)** What is the frequency of a harmonic oscillation if the static equilibrium position of the ball is 10 cm lower than the lower end of the spring before the ball is attached?
5. **(Initial velocity)** Could you make a harmonic oscillation move faster by giving the body a greater initial push?
6. **(Archimedian principle)** This principle states that the buoyancy force equals the weight of the water displaced by the body (partly or totally submerged). The cylindrical buoy of diameter 60 cm in Fig. 42 is floating in water with its axis vertical. When depressed downward in the water and released, it vibrates with period 2 sec. What is its weight?

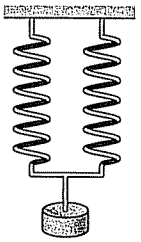


Fig. 40. Parallel springs (Problem 2)

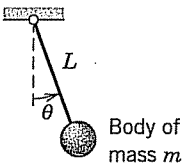


Fig. 41. Pendulum (Problem 3)

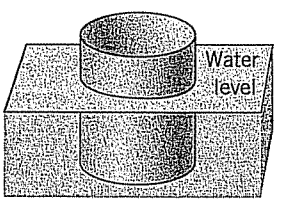


Fig. 42. Buoy (Problem 6)

7. **(Frequency)** How does the frequency of a harmonic motion change if we take (i) a spring of three times the modulus, (ii) a heavier ball?
8. **TEAM PROJECT. Harmonic Motions in Different Physical Systems.** Different physical or other systems may have the same or similar models, thus showing the ***unifying power of mathematical methods.*** Illustrate this for the systems in Figs. 43–45.

(a) Flat spring (Fig. 43). The spring is horizontally clamped at one end, and a body of weight 25 nt (about 5.6 lb) is attached at the other end. Find the motion of the system, assuming that its static equilibrium is 2 cm below the horizontal line, we let the system start from this position with initial velocity 15 cm/sec, and damping is negligible.

(b) Torsional vibrations (Fig. 44). Undamped torsional vibrations (rotations back and forth) of a wheel attached to an elastic thin rod are modeled by the ODE $I_0\theta'' + K\theta = 0$, where θ is the angle measured from the state of equilibrium, I_0 is the polar moment of inertia of the wheel about its center, and K is the torsional stiffness of the rod. Solve this ODE for $K/I_0 = 17.64\ \text{sec}^{-2}$, initial angle 45°, and initial angular velocity $15°\ \text{sec}^{-1}$.

(c) Water in a tube (Fig. 45). What is the frequency of vibration of 5 liters of water (about 1.3 gal) in a U-shaped tube of diameter 4 cm, neglecting friction?

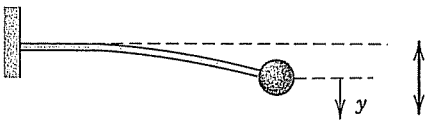


Fig. 43. Flat spring (Project 8a)

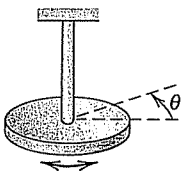


Fig. 44. Torsional vibrations (Project 8b)

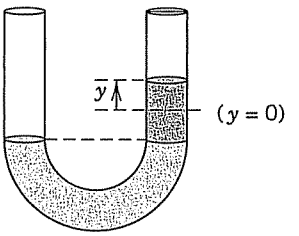


Fig. 45. Tube (Project 8c)

9–17 DAMPED MOTION

9. **(Frequency)** Find an approximation formula for ω^* in terms of ω_0 by applying the binomial theorem in (9) and retaining only the first two terms. How good is the approximation in Example 2, III?

10. **(Extrema)** Find the location of the maxima and minima of $y = e^{-2t}\cos t$ obtained approximately from a graph of y and compare it with the exact values obtained by calculation.

11. **(Maxima)** Show that the maxima of an underdamped motion occur at equidistant t-values and find the distance.

12. **(Logarithmic decrement)** Show that the ratio of two consecutive maximum amplitudes of a damped oscillation (10) is constant, and the natural logarithm of this ratio, called the *logarithmic decrement*, equals $\Delta = 2\pi\alpha/\omega^*$. Find Δ for the solutions of $y'' + 2y' + 5y = 0$.

13. **(Shock absorber)** What is the smallest value of the damping constant of a shock absorber in the suspension of a wheel of a car (consisting of a spring and an absorber) that will provide (theoretically) an oscillation-free ride if the mass of the car is 2000 kg and the spring constant equals 4500 kg/sec^2?

14. **(Damping constant)** Consider an underdamped motion of a body of mass $m = 2$ kg. If the time between two consecutive maxima is 2 sec and the maximum amplitude decreases to $\frac{1}{4}$ of its initial value after 15 cycles, what is the damping constant of the system?

15. **(Initial value problem)** Find the critical motion (8) that starts from y_0 with initial velocity v_0. Graph solution curves for $\alpha = 1$, $y_0 = 1$ and several v_0 such that (i) the curve does not intersect the t-axis, (ii) it intersects it at $t = 1, 2, \cdots, 5$, respectively.

16. **(Initial value problem)** Find the overdamped motion (7) that starts from y_0 with initial velocity v_0.

17. **(Overdamping)** Show that in the overdamped case, the body can pass through $y = 0$ at most once.

18. **CAS PROJECT. Transition Between Cases I, II, III.** Study this transition in terms of graphs of typical solutions. (Cf. Fig. 46.)

(a) ***Avoiding unnecessary generality is part of good modeling.*** Decide that the initial value problems (A) and (B),

(A) $y'' + cy' + y = 0, \quad y(0) = 1, \quad y'(0) = 0$

(B) the same with different c and $y'(0) = -2$ (instead of 0), will give practically as much information as a problem with other m, k, $y(0)$, $y'(0)$.

(b) ***Consider*** **(A).** Choose suitable values of c, perhaps better ones than in Fig. 46 for the transition from Case III to II and I. Guess c for the curves in the figure.

(c) ***Time to go to rest.*** Theoretically, this time is infinite (why?). Practically, the system is at rest when its motion has become very small, say, less than 0.1% of the initial displacement (this choice being up to us), that is in our case,

(11) $|y(t)| < 0.001$ for all t greater than some t_1.

In engineering constructions, damping can often be varied without too much trouble. Experimenting with your graphs, find empirically a relation between t_1 and c.

(d) ***Solve*** **(A)** ***analytically.*** Give a reason why the solution c of $y(t_2) = -0.001$, with t_2 the solution of $y'(t) = 0$, will give you the best possible c satisfying (11).

(e) Consider (B) empirically as in (a) and (b). What is the main difference between (B) and (A)?

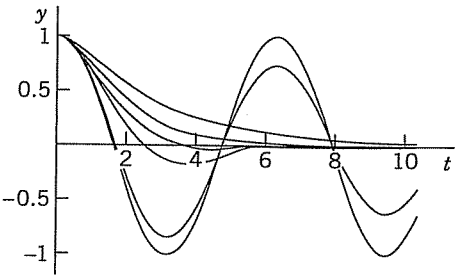


Fig. 46. CAS Project 18

2.5 Euler–Cauchy Equations

Euler–Cauchy equations[4] are ODEs of the form

(1) $$x^2y'' + axy' + by = 0$$

[4]LEONHARD EULER (1707–1783) was an enormously creative Swiss mathematician. He made fundamental contributions to almost all branches of mathematics and its application to physics. His important books on algebra and calculus contain numerous basic results of his own research. The great French mathematician AUGUSTIN LOUIS CAUCHY (1789–1857) is the father of modern analysis. He is the creator of complex analysis and had great influence on ODEs, PDEs, infinite series, elasticity theory, and optics.

with given constants a and b and unknown $y(x)$. We substitute

(2) $$y = x^m$$

and its derivatives $y' = mx^{m-1}$ and $y'' = m(m-1)x^{m-2}$ into (1). This gives

$$x^2m(m-1)x^{m-2} + axmx^{m-1} + bx^m = 0.$$

We now see that (2) was a rather natural choice because we have obtained a common factor x^m. Dropping it, we have the auxiliary equation $m(m-1) + am + b = 0$ or

(3) $$m^2 + (a-1)m + b = 0. \qquad \text{(Note: } a-1\text{, not } a.\text{)}$$

Hence $y = x^m$ is a solution of (1) if and only if m is a root of (3). The roots of (3) are

(4) $$m_1 = \tfrac{1}{2}(1-a) + \sqrt{\tfrac{1}{4}(1-a)^2 - b}, \qquad m_2 = \tfrac{1}{2}(1-a) - \sqrt{\tfrac{1}{4}(1-a)^2 - b}.$$

Case I. If the roots m_1 and m_2 are real and different, then solutions are

$$y_1(x) = x^{m_1} \qquad \text{and} \qquad y_2(x) = x^{m_2}$$

They are linearly independent since their quotient is not constant. Hence they constitute a basis of solutions of (1) for all x for which they are real. The corresponding general solution for all these x is

(5) $$y = c_1x^{m_1} + c_2x^{m_2} \qquad (c_1,\ c_2 \text{ arbitrary}).$$

EXAMPLE 1 General Solution in the Case of Different Real Roots

The Euler–Cauchy equation

$$x^2y'' + 1.5xy' - 0.5y = 0$$

has the auxiliary equation

$$m^2 + 0.5m - 0.5 = 0. \qquad \text{(Note: 0.5, not 1.5!)}$$

The roots are 0.5 and -1. Hence a basis of solutions for all positive x is $y_1 = x^{0.5}$ and $y_2 = 1/x$ and gives the general solution

$$y = c_1\sqrt{x} + \frac{c_2}{x} \qquad (x > 0). \qquad \blacksquare$$

Case II. Equation (4) shows that the auxiliary equation (3) has a double root $m_1 = \frac{1}{2}(1-a)$ if and only if $(1-a)^2 - 4b = 0$. The Euler–Cauchy equation (1) then has the form

(6) $$x^2y'' + axy' + \tfrac{1}{4}(1-a)^2y = 0.$$

A solution is $y_1 = x^{(1-a)/2}$. To obtain a second linearly independent solution, we apply the method of reduction of order from Sec. 2.1 as follows. Starting from $y_2 = uy_1$, we obtain for u the expression (9), Sec. 2.1, namely,

$$u = \int U\,dx \qquad \text{where} \qquad U = \frac{1}{y_1^2}\exp\left(-\int p\,dx\right).$$

Here it is crucial that p is taken from the ODE written in standard form, in our case,

$$(6^*) \qquad y'' + \frac{a}{x}y' + \frac{(1-a)^2}{4x^2}y = 0.$$

This shows that $p = a/x$ (not ax). Hence its integral is $a \ln x = \ln (x^a)$, the exponential function in U is $1/x^a$, and division by ${y_1}^2 = x^{1-a}$ gives $U = 1/x$, and $u = \ln x$ by integration.

Thus, in this "critical case," a basis of solutions for positive x is $y_1 = x^m$ and $y_2 = x^m \ln x$, where $m = \frac{1}{2}(1 - a)$. Linear independence follows from the fact that the quotient of these solutions is not constant. Hence, for all x for which y_1 and y_2 are defined and real, a general solution is

$$(7) \qquad y = (c_1 + c_2 \ln x)x^m, \qquad m = \tfrac{1}{2}(1 - a).$$

EXAMPLE 2 **General Solution in the Case of a Double Root**

The Euler–Cauchy equation $x^2y'' - 5xy' + 9y = 0$ has the auxiliary equation $m^2 - 6m + 9 = 0$. It has the double root $m = 3$, so that a general solution for all positive x is

$$y = (c_1 + c_2 \ln x)x^3.$$

Case III. The case of complex roots is of minor practical importance, and it suffices to present an example that explains the derivation of real solutions from complex ones.

EXAMPLE 3 **Real General Solution in the Case of Complex Roots**

The Euler–Cauchy equation

$$x^2y'' + 0.6xy' + 16.04y = 0$$

has the auxiliary equation $m^2 - 0.4m + 16.04 = 0$. The roots are complex conjugate, $m_1 = 0.2 + 4i$ and $m_2 = 0.2 - 4i$, where $i = \sqrt{-1}$. (We know from algebra that if a polynomial with real coefficients has complex roots, these are always conjugate.) Now use the trick of writing $x = e^{\ln x}$ and obtain

$$x^{m_1} = x^{0.2+4i} = x^{0.2}(e^{\ln x})^{4i} = x^{0.2}e^{(4 \ln x)i},$$
$$x^{m_2} = x^{0.2-4i} = x^{0.2}(e^{\ln x})^{-4i} = x^{0.2}e^{-(4 \ln x)i}.$$

Next apply Euler's formula (11) in Sec. 2.2 with $t = 4 \ln x$ to these two formulas. This gives

$$x^{m_1} = x^{0.2}[\cos (4 \ln x) + i \sin (4 \ln x)],$$
$$x^{m_2} = x^{0.2}[\cos (4 \ln x) - i \sin (4 \ln x)].$$

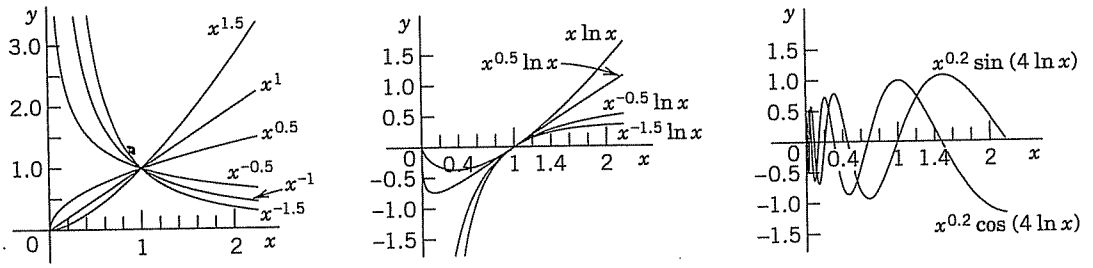
Add these two formulas, so that the sine drops out, and divide the result by 2. Then subtract the second formula from the first, so that the cosine drops out, and divide the result by $2i$. This yields

$$x^{0.2} \cos (4 \ln x) \qquad \text{and} \qquad x^{0.2} \sin (4 \ln x)$$

respectively. By the superposition principle in Sec. 2.2 these are solutions of the Euler–Cauchy equation (1). Since their quotient $\cot (4 \ln x)$ is not constant, they are linearly independent. Hence they form a basis of solutions, and the corresponding real general solution for all positive x is

$$(8) \qquad y = x^{0.2}[A \cos (4 \ln x) + B \sin (4 \ln x)].$$

Figure 47 shows typical solution curves in the three cases discussed, in particular the basis functions in Examples 1 and 3.



Case I: Real roots Case II: Double root Case III: Complex roots

Fig. 47. Euler–Cauchy equations

EXAMPLE 4 Boundary Value Problem. Electric Potential Field Between Two Concentric Spheres

Find the electrostatic potential $v = v(r)$ between two concentric spheres of radii $r_1 = 5$ cm and $r_2 = 10$ cm kept at potentials $v_1 = 110$ V and $v_2 = 0$, respectively.

Physical Information. $v(r)$ is a solution of the Euler–Cauchy equation $rv'' + 2v' = 0$, where $v' = dv/dr$.

Solution. The auxiliary equation is $m^2 + m = 0$. It has the roots 0 and -1. This gives the general solution $v(r) = c_1 + c_2/r$. From the "boundary conditions" (the potentials on the spheres) we obtain

$$v(5) = c_1 + \frac{c_2}{5} = 110, \qquad v(10) = c_1 + \frac{c_2}{10} = 0.$$

By subtraction, $c_2/10 = 110$, $c_2 = 1100$. From the second equation, $c_1 = -c_2/10 = -110$. *Answer:* $v(r) = -110 + 1100/r$ V. Figure 48 shows that the potential is not a straight line, as it would be for a potential between two parallel plates. For example, on the sphere of radius 7.5 cm it is not 110/2 = 55 V, but considerably less. (What is it?) ▩

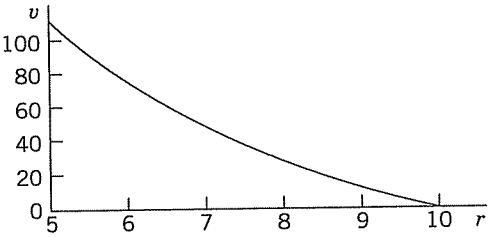


Fig. 48. Potential $v(r)$ in Example 4

PROBLEM SET 2.5

1–10 GENERAL SOLUTION

Find a real general solution, showing the details of your work.

1. $x^2y'' - 6y = 0$
2. $4x^2y'' + 4xy' - y = 0$
3. $x^2y'' - 7xy' + 16y = 0$
4. $x^2y'' + 3xy' + y = 0$
5. $x^2y'' - xy' + 2y = 0$
6. $2x^2y'' + 4xy' + 5y = 0$
7. $(10x^2D^2 - 20xD + 22.4I)y = 0$
8. $(4x^2D^2 + I)y = 0$
9. $(100x^2D^2 + 9I)y = 0$
10. $(10x^2D^2 + 6xD + 0.5I)y = 0$

11–15 INITIAL VALUE PROBLEM

Solve and graph the solution, showing the details of your work.

11. $x^2y'' - 4xy' + 6y = 0$, $y(1) = 1$, $y'(1) = 0$
12. $x^2y'' + 3xy' + y = 0$, $y(1) = 4$, $y'(1) = -2$
13. $(x^2D^2 + 2xD + 100.25I)y = 0$, $y(1) = 2$, $y'(1) = -11$
14. $(x^2D^2 - 2xD + 2.25I)y = 0$, $y(1) = 2.2$, $y'(1) = 2.5$
15. $(xD^2 + 4D)y = 0$, $y(1) = 12$, $y'(1) = -6$

16. TEAM PROJECT. Double Root

(A) Derive a second linearly independent solution of (1) by reduction of order; but instead of using (9), Sec. 2.1, perform all steps directly for the present ODE (1).

(B) Obtain $x^m \ln x$ by considering the solutions x^m and x^{m+s} of a suitable Euler–Cauchy equation and letting $s \to 0$.

(C) Verify by substitution that $x^m \ln x$, $m = (1 - a)/2$, is a solution in the critical case.

(D) Transform the Euler–Cauchy equation (1) into an ODE with constant coefficients by setting $x = e^t$ $(x > 0)$.

(E) Obtain a second linearly independent solution of the Euler–Cauchy equation in the "critical case" from that of a constant-coefficient ODE.

2.6 Existence and Uniqueness of Solutions. Wronskian

In this section we shall discuss the general theory of homogeneous linear ODEs

$$y'' + p(x)y' + q(x)y = 0 \tag{1}$$

with continuous, but otherwise arbitrary ***variable coefficients*** p and q. This will concern the existence and form of a general solution of (1) as well as the uniqueness of the solution of initial value problems consisting of such an ODE and two initial conditions

$$y(x_0) = K_0, \qquad y'(x_0) = K_1 \tag{2}$$

with given x_0, K_0, and K_1.

The two main results will be Theorem 1, stating that such an initial value problem always has a solution which is unique, and Theorem 4, stating that a general solution

$$y = c_1 y_1 + c_2 y_2 \qquad (c_1, c_2 \text{ arbitrary}) \tag{3}$$

includes all solutions. Hence *linear* ODEs with continuous coefficients have no *"singular solutions"* (solutions not obtainable from a general solution).

Clearly, no such theory was needed for constant-coefficient or Euler–Cauchy equations because everything resulted explicitly from our calculations.

Central to our present discussion is the following theorem.

THEOREM 1

Existence and Uniqueness Theorem for Initial Value Problems

If $p(x)$ and $q(x)$ are continuous functions on some open interval I (see Sec. 1.1) *and x_0 is in I, then the initial value problem consisting of* (1) *and* (2) *has a unique solution $y(x)$ on the interval I.*

The proof of existence uses the same prerequisites as the existence proof in Sec. 1.7 and will not be presented here; it can be found in Ref. [A11] listed in App. 1. Uniqueness proofs are usually simpler than existence proofs. But for Theorem 1, even the uniqueness proof is long, and we give it as an additional proof in App. 4.

Linear Independence of Solutions

Remember from Sec. 2.1 that a general solution on an open interval I is made up from a **basis** y_1, y_2 on I, that is, from a pair of linearly independent solutions on I. Here we call y_1, y_2 **linearly independent** on I if the equation

$$(4) \qquad k_1y_1(x) + k_2y_2(x) = 0 \quad \text{on } I \qquad \text{implies} \qquad k_1 = 0, \quad k_2 = 0.$$

We call y_1, y_2 **linearly dependent** on I if this equation also holds for constants k_1, k_2 not both 0. In this case, and only in this case, y_1 and y_2 are proportional on I, that is (see Sec. 2.1),

$$(5) \qquad \text{(a)} \quad y_1 = ky_2 \qquad \text{or} \qquad \text{(b)} \quad y_2 = ly_1 \qquad \text{for all } x \text{ on } I.$$

For our discussion the following criterion of linear independence and dependence of solutions will be helpful.

THEOREM 2

Linear Dependence and Independence of Solutions

Let the ODE (1) *have continuous coefficients* $p(x)$ *and* $q(x)$ *on an open interval* I. *Then two solutions* y_1 *and* y_2 *of* (1) *on* I *are linearly dependent on* I *if and only if their* **"Wronskian"**

$$(6) \qquad W(y_1, y_2) = y_1y_2' - y_2y_1'$$

is 0 *at some* x_0 *in* I. *Furthermore, if* $W = 0$ *at an* $x = x_0$ *in* I, *then* $W \equiv 0$ *on* I; *hence if there is an* x_1 *in* I *at which* W *is not* 0, *then* y_1, y_2 *are linearly independent on* I.

PROOF **(a)** Let y_1 and y_2 be linearly dependent on I. Then (5a) or (5b) holds on I. If (5a) holds, then

$$W(y_1, y_2) = y_1y_2' - y_2y_1' = ky_2y_2' - y_2ky_2' = 0.$$

Similarly if (5b) holds.

(b) Conversely, we let $W(y_1, y_2) = 0$ for some $x = x_0$ and show that this implies linear dependence of y_1 and y_2 on I. We consider the linear system of equations in the unknowns k_1, k_2

$$(7) \qquad \begin{aligned} k_1y_1(x_0) + k_2y_2(x_0) &= 0 \\ k_1y_1'(x_0) + k_2y_2'(x_0) &= 0. \end{aligned}$$

To eliminate k_2, multiply the first equation by y_2' and the second by $-y_2$ and add the resulting equations. This gives

$$k_1y_1(x_0)y_2'(x_0) - k_1y_1'(x_0)y_2(x_0) = k_1W(y_1(x_0), y_2(x_0)) = 0.$$

Similarly, to eliminate k_1, multiply the first equation by $-y_1'$ and the second by y_1 and add the resulting equations. This gives

$$k_2W(y_1(x_0), y_2(x_0)) = 0.$$

If W were not 0 at x_0, we could divide by W and conclude that $k_1 = k_2 = 0$. Since W is 0, division is not possible, and the system has a solution for which k_1 and k_2 are not both 0. Using *these numbers* k_1, k_2, we introduce the function

$$y(x) = k_1 y_1(x) + k_2 y_2(x).$$

Since (1) is homogeneous linear, Fundamental Theorem 1 in Sec. 2.1 (the superposition principle) implies that this function is a solution of (1) on I. From (7) we see that it satisfies the initial conditions $y(x_0) = 0$, $y'(x_0) = 0$. Now another solution of (1) satisfying the same initial conditions is $y^* \equiv 0$. Since the coefficients p and q of (1) are continuous, Theorem 1 applies and gives uniqueness, that is, $y \equiv y^*$, written out

$$k_1 y_1 + k_2 y_2 \equiv 0 \qquad \text{on } I.$$

Now since k_1 and k_2 are not both zero, this means linear dependence of y_1, y_2 on I.

(c) We prove the last statement of the theorem. If $W(x_0) = 0$ at an x_0 in I, we have linear dependence of y_1, y_2 on I by part (b), hence $W \equiv 0$ by part (a) of this proof. Hence in the case of linear dependence it cannot happen that $W(x_1) \neq 0$ at an x_1 in I. If it does happen, it thus implies linear independence as claimed. ∎

Remark. Determinants. Students familiar with second-order determinants may have noticed that

$$W(y_1, y_2) = \begin{vmatrix} y_1 & y_2 \\ y_1' & y_2' \end{vmatrix} = y_1 y_2' - y_2 y_1'.$$

This determinant is called the *Wronski determinant*[5] or, briefly, the **Wronskian,** of two solutions y_1 and y_2 of (1), as has already been mentioned in (6). Note that its four entries occupy the same positions as in the linear system (7).

EXAMPLE 1 **Illustration of Theorem 2**

The functions $y_1 = \cos \omega x$ and $y_2 = \sin \omega x$ are solutions of $y'' + \omega^2 y = 0$. Their Wronskian is

$$W(\cos \omega x, \sin \omega x) = \begin{vmatrix} \cos \omega x & \sin \omega x \\ -\omega \sin \omega x & \omega \cos \omega x \end{vmatrix} = y_1 y_2' - y_2 y_1' = \omega \cos^2 \omega x + \omega \sin^2 \omega x = \omega.$$

Theorem 2 shows that these solutions are linearly independent if and only if $\omega \neq 0$. Of course, we can see this directly from the quotient $y_2/y_1 = \tan \omega x$. For $\omega = 0$ we have $y_2 \equiv 0$, which implies linear dependence (why?). ∎

EXAMPLE 2 **Illustration of Theorem 2 for a Double Root**

A general solution of $y'' - 2y' + y = 0$ on any interval is $y = (c_1 + c_2 x)e^x$. (Verify!). The corresponding Wronskian is not 0, which shows linear independence of e^x and xe^x on any interval. Namely,

$$W(x, xe^x) = \begin{vmatrix} e^x & xe^x \\ e^x & (x+1)e^x \end{vmatrix} = (x+1)e^{2x} - xe^{2x} = e^{2x} \neq 0.$$ ∎

[5]Introduced by WRONSKI (JOSEF MARIA HÖNE, 1776–1853), Polish mathematician.

A General Solution of (1) Includes All Solutions

This will be our second main result, as announced at the beginning. Let us start with existence.

THEOREM 3

Existence of a General Solution

If $p(x)$ and $q(x)$ are continuous on an open interval I, then (1) *has a general solution on I.*

PROOF By Theorem 1, the ODE (1) has a solution $y_1(x)$ on I satisfying the initial conditions

$$y_1(x_0) = 1, \qquad y_1'(x_0) = 0$$

and a solution $y_2(x)$ on I satisfying the initial conditions

$$y_2(x_0) = 0, \qquad y_2'(x_0) = 1.$$

The Wronskian of these two solutions has at $x = x_0$ the value

$$W(y_1(0), y_2(0)) = y_1(x_0)y_2'(x_0) - y_2(x_0)y_1'(x_0) = 1.$$

Hence, by Theorem 2, these solutions are linearly independent on I. They form a basis of solutions of (1) on I, and $y = c_1y_1 + c_2y_2$ with arbitrary c_1, c_2 is a general solution of (1) on I, whose existence we wanted to prove. ∎

We finally show that a general solution is as general as it can possibly be.

THEOREM 4

A General Solution Includes All Solutions

If the ODE (1) *has continuous coefficients $p(x)$ and $q(x)$ on some open interval I, then every solution $y = Y(x)$ of* (1) *on I is of the form*

$$Y(x) = C_1y_1(x) + C_2y_2(x) \tag{8}$$

where y_1, y_2 is any basis of solutions of (1) *on I and C_1, C_2 are suitable constants.*

Hence (1) *does not have* **singular solutions** *(that is, solutions not obtainable from a general solution).*

PROOF Let $y = Y(x)$ be any solution of (1) on I. Now, by Theorem 3 the ODE (1) has a general solution

$$y(x) = c_1y_1(x) + c_2y_2(x) \tag{9}$$

on I. We have to find suitable values of c_1, c_2 such that $y(x) = Y(x)$ on I. We choose any x_0 in I and show first that we can find values of c_1, c_2 such that we reach agreement at x_0, that is, $y(x_0) = Y(x_0)$ and $y'(x_0) = Y'(x_0)$. Written out in terms of (9), this becomes

$$\begin{aligned} &\text{(a)} \quad c_1y_1(x_0) + c_2y_2(x_0) = Y(x_0) \\ &\text{(b)} \quad c_1y_1'(x_0) + c_2y_2'(x_0) = Y'(x_0). \end{aligned} \tag{10}$$

We determine the unknowns c_1 and c_2. To eliminate c_2, we multiply (10a) by $y_2'(x_0)$ and (10b) by $-y_2(x_0)$ and add the resulting equations. This gives an equation for c_1. Then we multiply (10a) by $-y_1'(x_0)$ and (10b) by $y_1(x_0)$ and add the resulting equations. This gives an equation for c_2. These new equations are as follows, where we take the values of y_1, y_1', y_2, y_2', Y, Y' at x_0:

$$c_1(y_1y_2' - y_2y_1') = c_1W(y_1, y_2) = Yy_2' - y_2Y'$$
$$c_2(y_1y_2' - y_2y_1') = c_2W(y_1, y_2) = y_1Y' - Yy_1'.$$

Since y_1, y_2 is a basis, the Wronskian W in these equations is not 0, and we can solve for c_1 and c_2. We call the (unique) solution $c_1 = C_1$, $c_2 = C_2$. By substituting it into (9) we obtain from (9) the particular solution

$$y^*(x) = C_1y_1(x) + C_2y_2(x).$$

Now since C_1, C_2 is a solution of (10), we see from (10) that

$$y^*(x_0) = Y(x_0), \qquad y^{*\prime}(x_0) = Y'(x_0).$$

From the uniqueness stated in Theorem 1 this implies that y^* and Y must be equal everywhere on I, and the proof is complete. ■

Looking back at he content of this section, we see that homogeneous linear ODEs with continuous variable coefficients have a conceptually and structurally rather transparent existence and uniqueness theory of solutions. Important in itself, this theory will also provide the foundation of an investigation of nonhomogeneous linear ODEs, whose theory and engineering applications we shall study in the remaining four sections of this chapter.

PROBLEM SET 2.6

1–17 BASES OF SOLUTIONS. CORRESPONDING ODEs. WRONSKIANS

Find an ODE (1) for which the given functions are solutions. Show linear independence (a) by considering quotients, (b) by Theorem 2.

1. $e^{0.5x}$, $e^{-0.5x}$
2. $\cos \pi x$, $\sin \pi x$
3. e^{kx}, xe^{kx}
4. x^3, x^{-2}
5. $x^{0.25}$, $x^{0.25} \ln x$
6. $e^{3.4x}$, $e^{-2.5x}$
7. $\cos(2 \ln x)$, $\sin(2 \ln x)$
8. e^{-2x}, xe^{-2x}
9. $x^{1.5}$, $x^{-0.5}$
10. x^{-3}, $x^{-3} \ln x$
11. $\cosh 2.5x$, $\sinh 2.5x$
12. $e^{-2x} \cos \omega x$, $e^{-2x} \sin \omega x$
13. $e^{-x} \cos 0.8x$, $e^{-x} \sin 0.8x$
14. $x^{-1} \cos(\ln x)$, $x^{-1} \sin(\ln x)$
15. $e^{-2.5x} \cos 0.3x$, $e^{-2.5x} \sin 0.3x$
16. $e^{-kx} \cos \pi x$, $e^{-kx} \sin \pi x$
17. $e^{-3.8\pi x}$, $xe^{-3.8\pi x}$

18. **TEAM PROJECT. Consequences of the Present Theory.** This concerns some noteworthy general properties of solutions. Assume that the coefficients p and q of the ODE (1) are continuous on some open interval I, to which the subsequent statements refer.

(A) Solve $y'' - y = 0$ (a) by exponential functions, (b) by hyperbolic functions. How are the constants in the corresponding general solutions related?

(B) Prove that the solutions of a basis cannot be 0 at the same point.

(C) Prove that the solutions of a basis cannot have a maximum or minimum at the same point.

(D) Express $(y_2/y_1)'$ by a formula involving the Wronskian W. Why is it likely that such a formula should exist? Use it to find W in Prob. 10.

(E) Sketch $y_1(x) = x^3$ if $x \geqq 0$ and 0 if $x < 0$, $y_2(x) = 0$ if $x \geqq 0$ and x^3 if $x < 0$. Show linear independence on $-1 < x < 1$. What is their Wronskian? What Euler–Cauchy equation do y_1, y_2 satisfy? Is there a contradiction to Theorem 2?

(F) Prove **Abel's formula**[6]

$$W(y_1(x), y_2(x)) = c \exp\left[-\int_{x_0}^{x} p(t)\,dt\right]$$

where $c = W(y_1(x_0), y_2(x_0))$. Apply it to Prob. 12. *Hint:* Write (1) for y_1 and for y_2. Eliminate q algebraically from these two ODEs, obtaining a first-order linear ODE. Solve it.

2.7 Nonhomogeneous ODEs

Method of Undetermined Coefficients

In this section we proceed from homogeneous to nonhomogeneous linear ODEs

$$y'' + p(x)y' + q(x)y = r(x) \tag{1}$$

where $r(x) \not\equiv 0$. We shall see that a "general solution" of (1) is the sum of a general solution of the corresponding homogeneous ODE

$$y'' + p(x)y' + q(x)y = 0 \tag{2}$$

and a "particular solution" of (1). These two new terms "general solution of (1)" and "particular solution of (1)" are defined as follows.

DEFINITION

General Solution, Particular Solution

A **general solution** of the nonhomogeneous ODE (1) on an open interval I is a solution of the form

$$y(x) = y_h(x) + y_p(x); \tag{3}$$

here, $y_h = c_1y_1 + c_2y_2$ is a general solution of the homogeneous ODE (2) on I and y_p is any solution of (1) on I containing no arbitrary constants.

A **particular solution** of (1) on I is a solution obtained from (3) by assigning specific values to the arbitrary constants c_1 and c_2 in y_h.

Our task is now twofold, first to justify these definitions and then to develop a method for finding a solution y_p of (1).

Accordingly, we first show that a general solution as just defined satisfies (1) and that the solutions of (1) and (2) are related in a very simple way.

THEOREM 1

Relations of Solutions of (1) to Those of (2)

(a) *The sum of a solution y of* (1) *on some open interval I and a solution $\tilde{y}$ of* (2) *on I is a solution of* (1) *on I. In particular,* (3) *is a solution of* (1) *on I.*

(b) *The difference of two solutions of* (1) *on I is a solution of* (2) *on I.*

[6]NIELS HENRIK ABEL (1802–1829), Norwegian mathematician.

PROOF **(a)** Let $L[y]$ denote the left side of (1). Then for any solutions y of (1) and $\tilde{y}$ of (2) on I,

$$L[y + \tilde{y}] = L[y] + L[\tilde{y}] = r + 0 = r.$$

(b) For any solutions y and y^* of (1) on I we have $L[y - y^*] = L[y] - L[y^*] = r - r = 0$. ∎

Now for *homogeneous ODEs* (2) we know that general solutions include all solutions. We show that the same is true for nonhomogeneous ODEs (1).

THEOREM 2

A General Solution of a Nonhomogeneous ODE Includes All Solutions

If the coefficients $p(x)$, $q(x)$, and the function $r(x)$ in (1) *are continuous on some open interval I, then every solution of* (1) *on I is obtained by assigning suitable values to the arbitrary constants c_1 and c_2 in a general solution* (3) *of* (1) *on I.*

PROOF Let y^* be any solution of (1) on I and x_0 any x in I. Let (3) be any general solution of (1) on I. This solution exists. Indeed, $y_h = c_1y_1 + c_2y_2$ exists by Theorem 3 in Sec. 2.6 because of the continuity assumption, and y_p exists according to a construction to be shown in Sec. 2.10. Now, by Theorem 1(b) just proved, the difference $Y = y^* - y_p$ is a solution of (2) on I. At x_0 we have

$$Y(x_0) = y^*(x_0) - y_p(x_0), \qquad Y'(x_0) = y^{*\prime}(x_0) - y_p'(x_0).$$

Theorem 1 in Sec. 2.6 implies that for these conditions, as for any other initial conditions in I, there exists a unique particular solution of (2) obtained by assigning suitable values to c_1, c_2 in y_h. From this and $y^* = Y + y_p$ the statement follows. ∎

Method of Undetermined Coefficients

Our discussion suggests the following. *To solve the nonhomogeneous ODE* (1) *or an initial value problem for* (1), *we have to solve the homogeneous ODE* (2) *and find any solution y_p of* (1), so that we obtain a general solution (3) of (1).

How can we find a solution y_p of (1)? One method is the so-called **method of undetermined coefficients.** It is much simpler than another, more general method (to be discussed in Sec. 2.10). Since it applies to models of vibrational systems and electric circuits to be shown in the next two sections, it is frequently used in engineering.

More precisely, the method of undetermined coefficients is suitable for linear ODEs with ***constant coefficients a and b***

$$y'' + ay' + by = r(x) \tag{4}$$

when $r(x)$ is an exponential function, a power of x, a cosine or sine, or sums or products of such functions. These functions have derivatives similar to $r(x)$ itself. This gives the idea. We choose a form for y_p similar to $r(x)$, but with unknown coefficients to be determined by substituting that y_p and its derivatives into the ODE. Table 2.1 on p. 80 shows the choice of y_p for practically important forms of $r(x)$. Corresponding rules are as follows.

Choice Rules for the Method of Undetermined Coefficients

(a) **Basic Rule.** *If $r(x)$ in (4) is one of the functions in the first column in Table 2.1, choose y_p in the same line and determine its undetermined coefficients by substituting y_p and its derivatives into (4).*

(b) **Modification Rule.** *If a term in your choice for y_p happens to be a solution of the homogeneous ODE corresponding to (4), multiply your choice of y_p by x (or by x^2 if this solution corresponds to a double root of the characteristic equation of the homogeneous ODE).*

(c) **Sum Rule.** *If $r(x)$ is a sum of functions in the first column of Table 2.1, choose for y_p the sum of the functions in the corresponding lines of the second column.*

The Basic Rule applies when $r(x)$ is a single term. The Modification Rule helps in the indicated case, and to recognize such a case, we have to solve the homogeneous ODE first. The Sum Rule follows by noting that the sum of two solutions of (1) with $r = r_1$ and $r = r_2$ (and the same left side!) is a solution of (1) with $r = r_1 + r_2$. (Verify!)

The method is self-correcting. A false choice for y_p or one with too few terms will lead to a contradiction. A choice with too many terms will give a correct result, with superfluous coefficients coming out zero.

Let us illustrate Rules (a)–(c) by the typical Examples 1–3.

Table 2.1 Method of Undetermined Coefficients

Term in $r(x)$	Choice for $y_p(x)$
$ke^{\gamma x}$	$Ce^{\gamma x}$
kx^n $(n = 0, 1, \cdots)$	$K_n x^n + K_{n-1}x^{n-1} + \cdots + K_1 x + K_0$
$k \cos \omega x$ $k \sin \omega x$	$K \cos \omega x + M \sin \omega x$
$ke^{\alpha x} \cos \omega x$ $ke^{\alpha x} \sin \omega x$	$e^{\alpha x}(K \cos \omega x + M \sin \omega x)$

EXAMPLE 1 Application of the Basic Rule (a)

Solve the initial value problem

$$y'' + y = 0.001x^2, \qquad y(0) = 0, \qquad y'(0) = 1.5. \tag{5}$$

Solution. ***Step 1. General solution of the homogeneous ODE.*** The ODE $y'' + y = 0$ has the general solution

$$y_h = A \cos x + B \sin x.$$

Step 2. Solution y_p of the nonhomogeneous ODE. We first try $y_p = Kx^2$. Then $y_p'' = 2K$. By substitution, $2K + Kx^2 = 0.001x^2$. For this to hold for all x, the coefficient of each power of x (x^2 and x^0) must be the same on both sides; thus $K = 0.001$ and $2K = 0$, a contradiction.

The second line in Table 2.1 suggests the choice

$$y_p = K_2x^2 + K_1x + K_0. \qquad \text{Then} \qquad y_p'' + y_p = 2K_2 + K_2x^2 + K_1x + K_0 = 0.001x^2.$$

Equating the coefficients of x^2, x, x^0 on both sides, we have $K_2 = 0.001$, $K_1 = 0$, $2K_2 + K_0 = 0$. Hence $K_0 = -2K_2 = -0.002$. This gives $y_p = 0.001x^2 - 0.002$, and

$$y = y_h + y_p = A \cos x + B \sin x + 0.001x^2 - 0.002.$$

Step 3. Solution of the initial value problem. Setting $x = 0$ and using the first initial condition gives $y(0) = A - 0.002 = 0$, hence $A = 0.002$. By differentiation and from the second initial condition,

$$y' = y_h' + y_p' = -A \sin x + B \cos x + 0.002x \qquad \text{and} \qquad y'(0) = B = 1.5.$$

This gives the answer (Fig. 49)

$$y = 0.002 \cos x + 1.5 \sin x + 0.001x^2 - 0.002.$$

Figure 49 shows y as well as the quadratic parabola y_p about which y is oscillating, practically like a sine curve since the cosine term is smaller by a factor of about 1/1000. ■

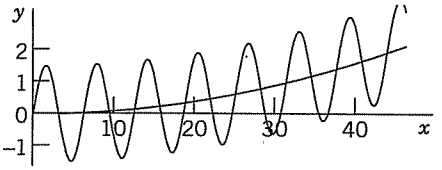


Fig. 49. Solution in Example 1

EXAMPLE 2 Application of the Modification Rule (b)

Solve the initial value problem

$$y'' + 3y' + 2.25y = -10\, e^{-1.5x}, \qquad y(0) = 1, \qquad y'(0) = 0. \tag{6}$$

Solution. ***Step 1. General solution of the homogeneous ODE.*** The characteristic equation of the homogeneous ODE is $\lambda^2 + 3\lambda + 2.25 = (\lambda + 1.5)^2 = 0$. Hence the homogeneous ODE has the general solution

$$y_h = (c_1 + c_2 x)e^{-1.5x}.$$

Step 2. Solution y_p of the nonhomogeneous ODE. The function $e^{-1.5x}$ on the right would normally require the choice $Ce^{-1.5x}$. But we see from y_h that this function is a solution of the homogeneous ODE, which corresponds to a *double root* of the characteristic equation. Hence, according to the Modification Rule we have to multiply our choice function by x^2. That is, we choose

$$y_p = Cx^2e^{-1.5x}. \qquad \text{Then} \qquad y_p' = C(2x - 1.5x^2)e^{-1.5x}, \qquad y_p'' = C(2 - 3x - 3x + 2.25x^2)e^{-1.5x}.$$

We substitute these expressions into the given ODE and omit the factor $e^{-1.5x}$. This yields

$$C(2 - 6x + 2.25x^2) + 3C(2x - 1.5x^2) + 2.25Cx^2 = -10.$$

Comparing the coefficients of x^2, x, x^0 gives $0 = 0$, $0 = 0$, $2C = -10$, hence $C = -5$. This gives the solution $y_p = -5x^2e^{-1.5x}$. Hence the given ODE has the general solution

$$y = y_h + y_p = (c_1 + c_2 x)e^{-1.5x} - 5x^2e^{-1.5x}.$$

Step 3. Solution of the initial value problem. Setting $x = 0$ in y and using the first initial condition, we obtain $y(0) = c_1 = 1$. Differentiation of y gives

$$y' = (c_2 - 1.5c_1 - 1.5c_2x)e^{-1.5x} - 10xe^{-1.5x} + 7.5x^2e^{-1.5x}.$$

From this and the second initial condition we have $y'(0) = c_2 - 1.5c_1 = 0$. Hence $c_2 = 1.5c_1 = 1.5$. This gives the answer (Fig. 50)

$$y = (1 + 1.5x)\, e^{-1.5x} - 5x^2e^{-1.5x} = (1 + 1.5x - 5x^2)e^{-1.5x}.$$

The curve begins with a horizontal tangent, crosses the x-axis at $x = 0.6217$ (where $1 + 1.5x - 5x^2 = 0$) and approaches the axis from below as x increases. ■

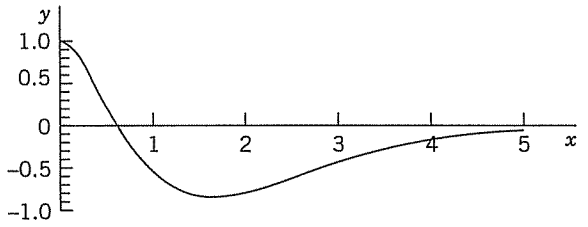


Fig. 50. Solution in Example 2

EXAMPLE 3 Application of the Sum Rule (c)

Solve the initial value problem

$$(7) \qquad y'' + 2y' + 5y = e^{0.5x} + 40\cos 10x - 190\sin 10x, \qquad y(0) = 0.16, \qquad y'(0) = 40.08.$$

Solution. ***Step 1. General solution of the homogeneous ODE.*** The characteristic equation

$$\lambda^2 + 2\lambda + 5 = (\lambda + 1 + 2i)(\lambda + 1 - 2i) = 0$$

shows that a real general solution of the homogeneous ODE is

$$y_h = e^{-x}(A\cos 2x + B\sin 2x).$$

Step 2. Solution of the nonhomogeneous ODE. We write $y_p = y_{p1} + y_{p2}$, where y_{p1} corresponds to the exponential term and y_{p2} to the sum of the other two terms. We set

$$y_{p1} = Ce^{0.5x}. \qquad \text{Then} \qquad y'_{p1} = 0.5Ce^{0.5x} \qquad \text{and} \qquad y''_{p1} = 0.25Ce^{0.5x}.$$

Substitution into the given ODE and omission of the exponential factor gives $(0.25 + 2 \cdot 0.5 + 5)C = 1$, hence $C = 1/6.25 = 0.16$, and $y_{p1} = 0.16e^{0.5x}$.

We now set $y_{p2} = K\cos 10x + M\sin 10x$, as in Table 2.1, and obtain

$$y'_{p2} = -10K\sin 10x + 10M\cos 10x, \qquad y''_{p2} = -100K\cos 10x - 100M\sin 10x.$$

Substitution into the given ODE gives for the cosine terms and for the sine terms

$$-100K + 2 \cdot 10M + 5K = 40, \qquad -100M - 2 \cdot 10K + 5M = -190$$

or, by simplification,

$$-95K + 20M = 40, \qquad -20K - 95M = -190.$$

The solution is $K = 0$, $M = 2$. Hence $y_{p2} = 2\sin 10x$. Together,

$$y = y_h + y_{p1} + y_{p2} = e^{-x}(A\cos 2x + B\sin 2x) + 0.16e^{0.5x} + 2\sin 10x.$$

Step 3. Solution of the initial value problem. From y and the first initial condition, $y(0) = A + 0.16 = 0.16$, hence $A = 0$. Differentiation gives

$$y' = e^{-x}(-A\cos 2x - B\sin 2x - 2A\sin 2x + 2B\cos 2x) + 0.08e^{0.5x} + 20\cos 10x.$$

From this and the second initial condition we have $y'(0) = -A + 2B + 0.08 + 20 = 40.08$, hence $B = 10$. This gives the solution (Fig. 51)

$$y = 10e^{-x}\sin 2x + 0.16e^{0.5x} + 2\sin 10x.$$

The first term goes to 0 relatively fast. When $x = 4$, it is practically 0, as the dashed curves $\pm 10e^{-x} + 0.16e^{0.5x}$ show. From then on, the last term, $2\sin 10x$, gives an oscillation about $0.16e^{0.5x}$, the monotone increasing dashed curve.

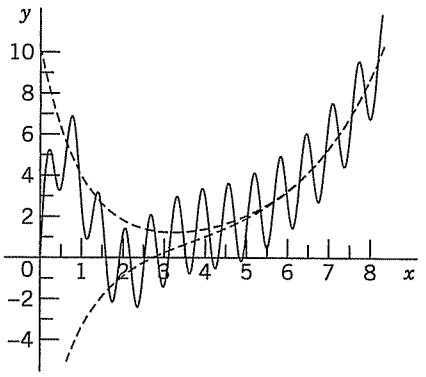


Fig. 51. Solution in Example 3

Stability. The following is important. If (and only if) all the roots of the characteristic equation of the homogeneous ODE $y'' + ay' + by = 0$ in (4) are negative, or have a negative real part, then a general solution y_h of this ODE goes to 0 as $x \to \infty$, so that the **"transient solution"** $y = y_h + y_p$ of (4) approaches the **"steady-state solution"** y_p. In this case the nonhomogeneous ODE and the physical or other system modeled by the ODE are called **stable;** otherwise they are called **unstable.** For instance, the ODE in Example 1 is unstable.

Basic applications follow in the next two sections.

PROBLEM SET 2.7

1–14 GENERAL SOLUTIONS OF NONHOMOGENEOUS ODEs

Find a (real) general solution. Which rule are you using? (Show each step of your calculation.)

1. $y'' + 3y' + 2y = 30e^{2x}$
2. $y'' + 4y' + 3.75y = 109 \cos 5x$
3. $y'' - 16y = 19.2e^{4x} + 60e^{x}$
4. $y'' + 9y = \cos x + \frac{1}{3}\cos 3x$
5. $y'' + y' - 6y = 6x^3 - 3x^2 + 12x$
6. $y'' + 4y' + 4y = e^{-2x}\sin 2x$
7. $y'' + 6y' + 73y = 80e^{x}\cos 4x$
8. $y'' + 10y' + 25y = 100 \sinh 5x$
9. $y'' - 0.16y = 32 \cosh 0.4x$
10. $y'' + 4y' + 6.25y = 3.125(x + 1)^2$
11. $y'' + 1.44y = 24 \cos 1.2x$
12. $y'' + 9y = 18x + 36 \sin 3x$
13. $y'' + 4y' + 5y = 25x^2 + 13 \sin 2x$
14. $y'' + 2y' + y = 2x \sin x$

15–20 INITIAL VALUE PROBLEMS FOR NONHOMOGENEOUS ODEs

Solve the initial value problem. State which rules you are using. Show each step of your calculation in detail.

15. $y'' + 4y = 16 \cos 2x, \quad y(0) = 0, \quad y'(0) = 0$
16. $y'' - 3y' + 2.25y = 27(x^2 - x), \quad y(0) = 20, \quad y'(0) = 30$
17. $y'' + 0.2y' + 0.26y = 1.22e^{0.5x}, \quad y(0) = 3.5, \quad y'(0) = 0.35$
18. $y'' - 2y' = 12e^{2x} - 8e^{-2x}, \quad y(0) = -2, \quad y'(0) = 12$
19. $y'' - y' - 12y = 144x^3 + 12.5, \quad y(0) = 5, \quad y'(0) = -0.5$
20. $y'' + 2y' + 10y = 17 \sin x - 37 \sin 3x, \quad y(0) = 6.6, \quad y'(0) = -2.2$

21. **WRITING PROJECT. Initial Value Problem.** Write out all the details of Example 3 in your own words. Discuss Fig. 51 in more detail. Why is it that some of the "half-waves" do not reach the dashed curves, whereas others preceding them (and, of course, all later ones) excede the dashed curves?

22. **TEAM PROJECT. Extensions of the Method of Undetermined Coefficients. (a)** Extend the method to products of the function in Table 2.1. **(b)** Extend the method to Euler–Cauchy equations. Comment on the practical significance of such extensions.

23. **CAS PROJECT. Structure of Solutions of Initial Value Problems.** Using the present method, find, graph, and discuss the solutions y of initial value problems of your own choice. Explore effects on solutions caused by

changes of initial conditions. Graph y_p, y, $y - y_p$ separately, to see the separate effects. Find a problem in which (a) the part of y resulting from y_h decreases to zero, (b) increases, (c) is not present in the answer y. Study a problem with $y(0) = 0$, $y'(0) = 0$. Consider a problem in which you need the Modification Rule (a) for a simple root, (b) for a double root. Make sure that your problems cover all three Cases I, II, III (see Sec. 2.2).

2.8 Modeling: Forced Oscillations. Resonance

In Sec. 2.4 we considered vertical motions of a mass-spring system (vibration of a mass m on an elastic spring, as in Figs. 32 and 52) and modeled it by the ***homogeneous*** linear ODE

$$my'' + cy' + ky = 0. \tag{1}$$

Here $y(t)$ as a function of time t is the displacement of the body of mass m from rest. These were **free motions,** that is, motions in the absence of *external forces* (outside forces) caused solely by *internal forces,* forces within the system. These are the force of inertia my'', the damping force cy' (if $c > 0$), and the spring force ky acting as a restoring force.

We now extend our model by including an external force, call it $r(t)$, on the right. Then we have

$$my'' + cy' + ky = r(t). \tag{2*}$$

Mechanically this means that at each instant t the resultant of the internal forces is in equilibrium with $r(t)$. The resulting motion is called a **forced motion** with **forcing function** $r(t)$, which is also known as **input** or **driving force,** and the solution $y(t)$ to be obtained is called the **output** or the **response** *of the system to the driving force.*

Of special interest are periodic external forces, and we shall consider a driving force of the form

$$r(t) = F_0 \cos \omega t \qquad (F_0 > 0,\ \omega > 0).$$

Then we have the nonhomogeneous ODE

$$my'' + cy' + ky = F_0 \cos \omega t. \tag{2}$$

Its solution will familiarize us with further interesting facts fundamental in engineering mathematics, in particular with resonance.

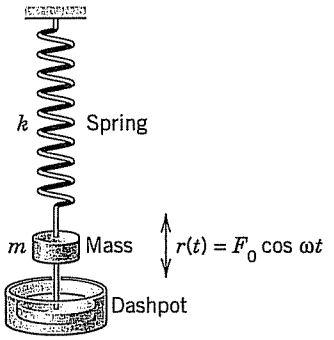


Fig. 52. Mass on a spring

Solving the Nonhomogeneous ODE (2)

From Sec. 2.7 we know that a general solution of (2) is the sum of a general solution y_h of the homogeneous ODE (1) plus any solution y_p of (2). To find y_p, we use the method of undetermined coefficients (Sec. 2.7), starting from

$$y_p(t) = a \cos \omega t + b \sin \omega t. \tag{3}$$

By differentiating this function (chain rule!) we obtain

$$y_p' = -\omega a \sin \omega t + \omega b \cos \omega t,$$
$$y_p'' = -\omega^2 a \cos \omega t - \omega^2 b \sin \omega t.$$

Substituting y_p, y_p', and y_p'' into (2) and collecting the cosine and the sine terms, we get

$$[(k - m\omega^2)a + \omega cb] \cos \omega t + [-\omega ca + (k - m\omega^2)b] \sin \omega t = F_0 \cos \omega t.$$

The cosine terms on both sides must be equal, and the coefficient of the sine term on the left must be zero since there is no sine term on the right. This gives the two equations

$$\begin{aligned} (k - m\omega^2)a + \quad \omega cb \quad &= F_0 \\ -\omega ca \quad + (k - m\omega^2)b \quad &= 0 \end{aligned} \tag{4}$$

for determining the unknown coefficients a and b. This is a linear system. We can solve it by elimination. To eliminate b, multiply the first equation by $k - m\omega^2$ and the second by $-\omega c$ and add the results, obtaining

$$(k - m\omega^2)^2 a + \omega^2 c^2 a = F_0(k - m\omega^2).$$

Similarly, to eliminate a, multiply the first equation by ωc and the second by $k - m\omega^2$ and add to get

$$\omega^2 c^2 b + (k - m\omega^2)^2 b = F_0 \omega c.$$

If the factor $(k - m\omega^2)^2 + \omega^2 c^2$ is not zero, we can divide by this factor and solve for a and b,

$$a = F_0 \frac{k - m\omega^2}{(k - m\omega^2)^2 + \omega^2 c^2}, \qquad b = F_0 \frac{\omega c}{(k - m\omega^2)^2 + \omega^2 c^2}.$$

If we set $\sqrt{k/m} = \omega_0$ (> 0) as in Sec. 2.4, then $k = m\omega_0^2$ and we obtain

$$a = F_0 \frac{m(\omega_0^2 - \omega^2)}{m^2(\omega_0^2 - \omega^2)^2 + \omega^2 c^2}, \qquad b = F_0 \frac{\omega c}{m^2(\omega_0^2 - \omega^2)^2 + \omega^2 c^2}. \tag{5}$$

We thus obtain the general solution of the nonhomogeneous ODE (2) in the form

$$y(t) = y_h(t) + y_p(t). \tag{6}$$

Here y_h is a general solution of the homogeneous ODE (1) and y_p is given by (3) with coefficients (5).

We shall now discuss the behavior of the mechanical system, distinguishing between the two cases $c = 0$ (no damping) and $c > 0$ (damping). These cases will correspond to two basically different types of output.

Case 1. Undamped Forced Oscillations. Resonance

If the damping of the physical system is so small that its effect can be neglected over the time interval considered, we can set $c = 0$. Then (5) reduces to $a = F_0/[m(\omega_0^2 - \omega^2)]$ and $b = 0$. Hence (3) becomes (use $\omega_0^2 = k/m$)

$$(7) \qquad y_p(t) = \frac{F_0}{m(\omega_0^2 - \omega^2)} \cos \omega t = \frac{F_0}{k[1 - (\omega/\omega_0)^2]} \cos \omega t.$$

Here we must assume that $\omega^2 \neq \omega_0^2$; physically, the frequency $\omega/(2\pi)$ [cycles/sec] of the driving force is different from the *natural frequency* $\omega_0/(2\pi)$ of the system, which is the frequency of the free undamped motion [see (4) in Sec. 2.4]. From (7) and from (4*) in Sec. 2.4 we have the general solution of the "undamped system"

$$(8) \qquad y(t) = C \cos(\omega_0 t - \delta) + \frac{F_0}{m(\omega_0^2 - \omega^2)} \cos \omega t.$$

We see that this output is a ***superposition of two harmonic oscillations*** *of the frequencies just mentioned.*

Resonance. We discuss (7). We see that the maximum amplitude of y_p is (put $\cos \omega t = 1$)

$$(9) \qquad a_0 = \frac{F_0}{k} \rho \qquad \text{where} \qquad \rho = \frac{1}{1 - (\omega/\omega_0)^2}.$$

a_0 depends on ω and ω_0. If $\omega \to \omega_0$, then ρ and a_0 tend to infinity. This excitation of large oscillations by matching input and natural frequencies ($\omega = \omega_0$) is called **resonance.** ρ is called the **resonance factor** (Fig. 53), and from (9) we see that $\rho/k = a_0/F_0$ is the ratio of the amplitudes of the particular solution y_p and of the input $F_0 \cos \omega t$. We shall see later in this section that resonance is of basic importance in the study of vibrating systems.

In the case of resonance the nonhomogeneous ODE (2) becomes

$$(10) \qquad y'' + \omega_0^2 y = \frac{F_0}{m} \cos \omega_0 t.$$

Then (7) is no longer valid, and from the Modification Rule in Sec. 2.7 we conclude that a particular solution of (10) is of the form

$$y_p(t) = t(a \cos \omega_0 t + b \sin \omega_0 t).$$

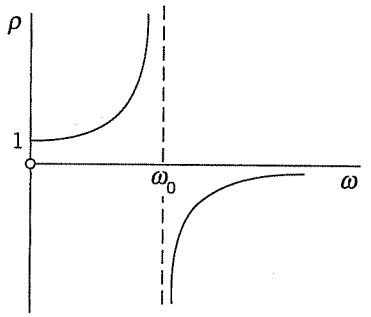


Fig. 53. Resonance factor $\rho(\omega)$

By substituting this into (10) we find $a = 0$ and $b = F_0/(2m\omega_0)$. Hence (Fig. 54)

(11)
$$y_p(t) = \frac{F_0}{2m\omega_0}\, t \sin \omega_0 t.$$

We see that because of the factor t the amplitude of the vibration becomes larger and larger. Practically speaking, systems with very little damping may undergo large vibrations that can destroy the system. We shall return to this practical aspect of resonance later in this section.

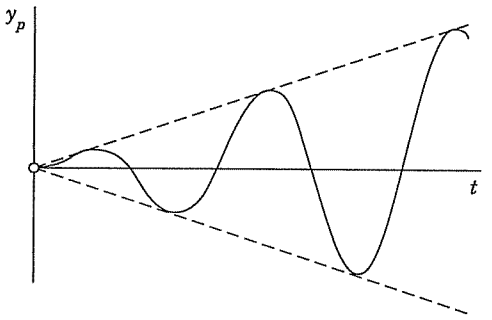


Fig. 54. Particular solution in the case of resonance

Beats. Another interesting and highly important type of oscillation is obtained if ω is close to ω_0. Take, for example, the particular solution [see (8)]

(12)
$$y(t) = \frac{F_0}{m(\omega_0^2 - \omega^2)} (\cos \omega t - \cos \omega_0 t) \qquad (\omega \neq \omega_0).$$

Using (12) in App. 3.1, we may write this as

$$y(t) = \frac{2F_0}{m(\omega_0^2 - \omega^2)} \sin\left(\frac{\omega_0 + \omega}{2}\, t\right) \sin\left(\frac{\omega_0 - \omega}{2}\, t\right).$$

Since ω is close to ω_0, the difference $\omega_0 - \omega$ is small. Hence the period of the last sine function is large, and we obtain an oscillation of the type shown in Fig. 55, the dashed curve resulting from the first sine factor. This is what musicians are listening to when they ***tune*** their instruments.

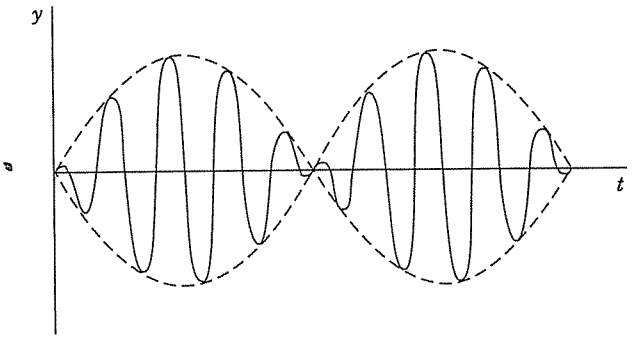


Fig. 55. Forced undamped oscillation when the difference of the input and natural frequencies is small **("beats")**

Case 2. Damped Forced Oscillations

If the damping of the mass-spring system is not negligibly small, we have $c > 0$ and a damping term cy' in (1) and (2). Then the general solution y_h of the homogeneous ODE (1) approaches zero as t goes to infinity, as we know from Sec. 2.4. Practically, it is zero after a sufficiently long time. Hence the **"transient solution"** (6) of (2), given by $y = y_h + y_p$, approaches the **"steady-state solution"** y_p. This proves the following.

THEOREM 1

Steady-State Solution

After a sufficiently long time the output of a damped vibrating system under a purely sinusoidal driving force [see (2)] will practically be a harmonic oscillation whose frequency is that of the input.

Amplitude of the Steady-State Solution. Practical Resonance

Whereas in the undamped case the amplitude of y_p approaches infinity as ω approaches ω_0, this will not happen in the damped case. In this case the amplitude will always be finite. But it may have a maximum for some ω depending on the damping constant c. This may be called **practical resonance.** It is of great importance because if c is not too large, then some input may excite oscillations large enough to damage or even destroy the system. Such cases happened, in particular in earlier times when less was known about resonance. Machines, cars, ships, airplanes, bridges, and high-rising buildings are vibrating mechanical systems, and it is sometimes rather difficult to find constructions that are completely free of undesired resonance effects, caused, for instance, by an engine or by strong winds.

To study the amplitude of y_p as a function of ω, we write (3) in the form

$$y_p(t) = C^* \cos(\omega t - \eta). \tag{13}$$

C^* is called the **amplitude** of y_p and η the **phase angle** or **phase lag** because it measures the lag of the output behind the input. According to (5), these quantities are

$$C^*(\omega) = \sqrt{a^2 + b^2} = \frac{F_0}{\sqrt{m^2(\omega_0^2 - \omega^2)^2 + \omega^2 c^2}},$$

$$\tan \eta(\omega) = \frac{b}{a} = \frac{\omega c}{m(\omega_0^2 - \omega^2)}. \tag{14}$$

Let us see whether $C^*(\omega)$ has a maximum and, if so, find its location and then its size. We denote the radicand in the second root in C^* by R. Equating the derivative of C^* to zero, we obtain

$$\frac{dC^*}{d\omega} = F_0\left(-\frac{1}{2}R^{-3/2}\right)[2m^2(\omega_0{}^2 - \omega^2)(-2\omega) + 2\omega c^2].$$

The expression in the brackets [. . .] is zero if

$$c^2 = 2m^2(\omega_0{}^2 - \omega^2) \qquad (\omega_0{}^2 = k/m). \tag{15}$$

By reshuffling terms we have

$$2m^2\omega^2 = 2m^2\omega_0{}^2 - c^2 = 2mk - c^2.$$

The right side of this equation becomes negative if $c^2 > 2mk$, so that then (15) has no real solution and C^* decreases monotone as ω increases, as the lowest curve in Fig. 56 on p. 90 shows. If c is smaller, $c^2 < 2mk$, then (15) has a real solution $\omega = \omega_{\max}$, where

$$\omega_{\max}^2 = \omega_0{}^2 - \frac{c^2}{2m^2}. \tag{15*}$$

From (15*) we see that this solution increases as c decreases and approaches ω_0 as c approaches zero. See also Fig. 56.

The size of $C^*(\omega_{\max})$ is obtained from (14), with $\omega^2 = \omega_{\max}^2$ given by (15*). For this ω^2 we obtain in the second radicand in (14) from (15*)

$$m^2(\omega_0{}^2 - \omega_{\max}^2)^2 = \frac{c^4}{4m^2} \qquad \text{and} \qquad \omega_{\max}^2 c^2 = \left(\omega_0{}^2 - \frac{c^2}{2m^2}\right)c^2.$$

The sum of the right sides of these two formulas is

$$(c^4 + 4m^2\omega_0{}^2c^2 - 2c^4)/(4m^2) = c^2(4m^2\omega_0{}^2 - c^2)/(4m^2).$$

Substitution into (14) gives

$$C^*(\omega_{\max}) = \frac{2mF_0}{c\sqrt{4m^2\omega_0{}^2 - c^2}}. \tag{16}$$

We see that $C^*(\omega_{\max})$ is always finite when $c > 0$. Furthermore, since the expression

$$c^2 4m^2\omega_0{}^2 - c^4 = c^2(4mk - c^2)$$

in the denominator of (16) decreases monotone to zero as c^2 ($< 2mk$) goes to zero, the maximum amplitude (16) increases monotone to infinity, in agreement with our result in Case 1. Figure 56 shows the **amplification** C^*/F_0 (ratio of the amplitudes of output and input) as a function of ω for $m = 1$, $k = 1$, hence $\omega_0 = 1$, and various values of the damping constant c.

Figure 57 shows the phase angle (the lag of the output behind the input), which is less than $\pi/2$ when $\omega < \omega_0$, and greater than $\pi/2$ for $\omega > \omega_0$.

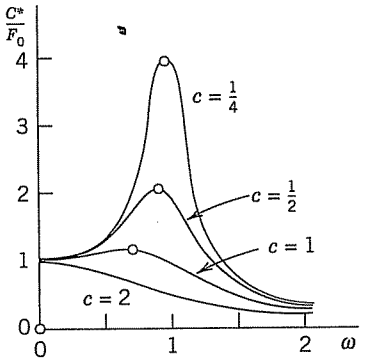


Fig. 56. Amplification C^*/F_0 as a function of ω for $m = 1$, $k = 1$, and various values of the damping constant c

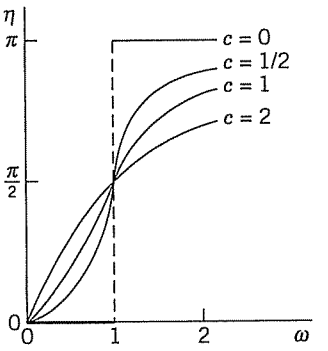


Fig. 57. Phase lag η as a function of ω for $m = 1$, $k = 1$, thus $\omega_0 = 1$, and various values of the damping constant c

PROBLEM SET 2.8

1–8 STEADY-STATE SOLUTIONS

Find the steady-state oscillation of the mass–spring system modeled by the given ODE. Show the details of your calculations.

1. $y'' + 6y' + 8y = 130 \cos 3t$
2. $4y'' + 8y' + 13y = 8 \sin 1.5t$
3. $y'' + y' + 4.25y = 221 \cos 4.5t$
4. $y'' + 4y' + 5y = \cos t - \sin t$
5. $(D^2 + 2D + I)y = -\sin 2t$
6. $(D^2 + 4D + 3I)y = \cos t + \frac{1}{3} \cos 3t$
7. $(D^2 + 6D + 18I)y = \cos 3t - 3 \sin 3t$
8. $(D^2 + 2D + 10I)y = -25 \sin 4t$

9–14 TRANSIENT SOLUTIONS

Find the transient motion of the mass–spring system modeled by the given ODE. (Show the details of your work.)

9. $y'' + 2y' + 0.75y = 13 \sin t$
10. $y'' + 4y' + 4y = \cos 4t$
11. $4y'' + 12y' + 9y = 75 \sin 3t$
12. $(D^2 + 5D + 4I)y = \sin 2t$
13. $(D^2 + 3D + 3.25I)y = 13 - 39 \cos 2t$
14. $(D^2 + 2D + 5I)y = 1 + \sin t$

15–20 INITIAL VALUE PROBLEMS

Find the motion of the mass–spring system modeled by the ODE and initial conditions. Sketch or graph the solution curve. In addition, sketch or graph the curve of $y - y_p$ to see when the system practically reaches the steady state.

15. $y'' + 2y' + 26y = 13 \cos 3t$, $y(0) = 1$, $y'(0) = 0.4$
16. $y'' + 64y = \cos t$, $y(0) = 0$, $y'(0) = 1$
17. $y'' + 6y' + 8y = 4 \sin 2t$, $y(0) = 0.7$, $y'(0) = -11.8$
18. $(D^2 + 2D + I)y = 75(\sin t - \frac{1}{2} \sin 2t + \frac{1}{3} \sin 3t)$, $y(0) = 0$, $y'(0) = 1$
19. $(4D^2 + 12D + 13I)y = 12 \cos t - 6 \sin t$, $y(0) = 1$, $y'(0) = -1$
20. $y'' + 25y = 99 \cos 4.9t$, $y(0) = 2$, $y'(0) = 0$
21. **(Beats)** Derive the formula after (12) from (12). Can there be beats if the system has damping?
22. **(Beats)** How does the graph of the solution in Prob. 20 change if you change (a) $y(0)$, (b) the frequency of the driving force?
23. **WRITING PROJECT. Free and Forced Vibrations.** Write a condensed report of 2–3 pages on the most important facts about free and forced vibrations.
24. **CAS EXPERIMENT. Undamped Vibrations.** **(a)** Solve the initial value problem $y'' + y = \cos \omega t$, $\omega^2 \neq 1$, $y(0) = 0$, $y'(0) = 0$. Show that the solution can be written

$$y(t) = \frac{2}{1 - \omega^2} \sin\left[\frac{1}{2}(1 + \omega)t\right] \times \sin\left[\frac{1}{2}(1 - \omega)t\right]. \tag{17}$$

(b) Experiment with (17) by changing ω to see the change of the curves from those for small ω (> 0) to beats, to resonance and to large values of ω (see Fig. 58).

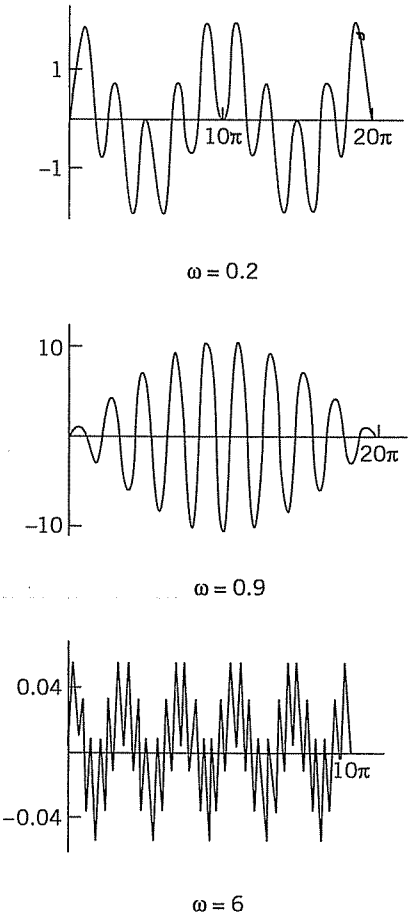


Fig. 58. Typical solution curves in CAS Experiment 24

25. TEAM PROJECT. Practical Resonance. (a) Give a detailed derivation of the crucial formula (16).

(b) By considering dC^*/dc show that $C^*(\omega_{\max})$ increases as c ($\leqq \sqrt{2mk}$) decreases.

(c) Illustrate practical resonance with an ODE of your own in which you vary c, and sketch or graph corresponding curves as in Fig. 56.

(d) Take your ODE with c fixed and an input of two terms, one with frequency close to the practical resonance frequency and the other not. Discuss and sketch or graph the output.

(e) Give other applications (not in the book) in which resonance is important.

26. (Gun barrel) Solve

$$y'' + y = \begin{cases} 1 - t^2/\pi^2 & \text{if } 0 \leqq t \leqq \pi \\ 0 & \text{if } t > \pi, \end{cases}$$

$$y(0) = y'(0) = 0.$$

This models an undamped system on which a force F acts during some interval of time (see Fig. 59), for instance, the force on a gun barrel when a shell is fired, the barrel being braked by heavy springs (and then damped by a dashpot, which we disregard for simplicity). *Hint.* At π both y and y' must be continuous.

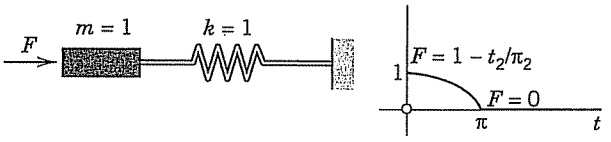


Fig. 59. Problem 26

2.9 Modeling: Electric Circuits

Designing good models is a task the computer cannot do. Hence setting up models has become an important task in modern applied mathematics. The best way to gain experience is to consider models from various fields. Accordingly, modeling electric circuits to be discussed will be *profitable for all students,* not just for electrical engineers and computer scientists.

We have just seen that linear ODEs have important applications in mechanics (see also Sec. 2.4). Similarly, they are models of electric circuits, as they occur as portions of large networks in computers and elsewhere. The circuits we shall consider here are basic building blocks of such networks. They contain three kinds of components, namely, resistors, inductors, and capacitors. Figure 60 on p. 92 shows such an ***RLC*-circuit,** as they are called. In it a resistor of resistance R Ω (ohms), an inductor of inductance L H (henrys), and a capacitor of capacitance C F (farads) are wired in series as shown, and connected to an electromotive force $E(t)$ V (volts) (a generator, for instance), sinusoidal as in Fig. 60, or of some other kind. R, L, C, and E are given and we want to find the current $I(t)$ A (amperes) in the circuit.

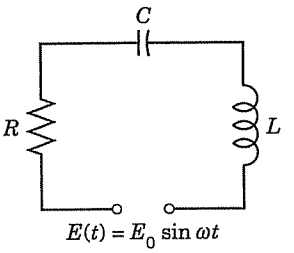


Fig. 60. *RLC*-circuit

An ODE for the current $I(t)$ in the *RLC*-circuit in Fig. 60 is obtained from the following law (which is the analog of Newton's second law, as we shall see later).

Kirchhoff's Voltage Law (KVL).[7] *The voltage (the electromotive force) impressed on a closed loop is equal to the sum of the voltage drops across the other elements of the loop.*

In Fig. 60 the circuit is a closed loop, and the impressed voltage $E(t)$ equals the sum of the voltage drops across the three elements R, L, C of the loop.

Voltage Drops. Experiments show that a current I flowing through a resistor, inductor or capacitor causes a voltage drop (voltage difference, measured in volts) at the two ends; these drops are

RI **(Ohm's law)** Voltage drop for a resistor of resistance R ohms (Ω),

$LI' = L\dfrac{dI}{dt}$ Voltage drop for an inductor of inductance L henrys (H),

$\dfrac{Q}{C}$. Voltage drop for a capacitor of capacitance C farads (F).

Here Q coulombs is the charge on the capacitor, related to the current by

$$I(t) = \frac{dQ}{dt}, \qquad \text{equivalently,} \qquad Q(t) = \int I(t)\,dt.$$

This is summarized in Fig. 61.

According to KVL we thus have in Fig. 60 for an *RLC*-circuit with electromotive force $E(t) = E_0 \sin \omega t$ (E_0 constant) as a model the **"integro-differential equation"**

$$(1') \qquad LI' + RI + \frac{1}{C}\int I\,dt = E(t) = E_0 \sin \omega t.$$

---

[7]GUSTAV ROBERT KIRCHHOFF (1824–1887), German physicist. Later we shall also need **Kirchhoff's current law (KCL):**

At any point of a circuit, the sum of the inflowing currents is equal to the sum of the outflowing currents.

The units of measurement of electrical quantities are named after ANDRÉ MARIE AMPÈRE (1775–1836), French physicist, CHARLES AUGUSTIN DE COULOMB (1736–1806), French physicist and engineer, MICHAEL FARADAY (1791–1867), English physicist, JOSEPH HENRY (1797–1878), American physicist, GEORG SIMON OHM (1789–1854), German physicist, and ALESSANDRO VOLTA (1745–1827), Italian physicist.

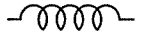
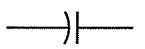
Name	Symbol	Notation		Unit	Voltage Drop
Ohm's resistor		R	Ohm's resistance	ohms (Ω)	RI
Inductor		L	Inductance	henrys (H)	$L\frac{dI}{dt}$
Capacitor		C	Capacitance	farads (F)	Q/C

Fig. 61. Elements in an *RLC*-circuit

To get rid of the integral, we differentiate $(1')$ with respect to t, obtaining

$$\textbf{(1)} \qquad LI'' + RI' + \frac{1}{C}I = E'(t) = E_0\omega \cos \omega t.$$

This shows that the current in an *RLC*-circuit is obtained as the solution of this nonhomogeneous second-order ODE (1) with constant coefficients.

From $(1')$, using $I = Q'$, hence $I' = Q''$, we also have directly

$$(1'') \qquad LQ'' + RQ' + \frac{1}{C}Q = E_0 \sin \omega t.$$

But in most practical problems the current $I(t)$ is more important than the charge $Q(t)$, and for this reason we shall concentrate on (1) rather than on $(1'')$.

Solving the ODE (1) for the Current. Discussion of Solution

A general solution of (1) is the sum $I = I_h + I_p$, where I_h is a general solution of the homogeneous ODE corresponding to (1) and I_p is a particular solution of (1). We first determine I_p by the method of undetermined coefficients, proceeding as in the previous section. We substitute

$$\textbf{(2)} \qquad I_p = a \cos \omega t + b \sin \omega t$$

$$I_p' = \omega(-a \sin \omega t + b \cos \omega t)$$

$$I_p'' = \omega^2(-a \cos \omega t - b \sin \omega t)$$

into (1). Then we collect the cosine terms and equate them to $E_0\omega \cos \omega t$ on the right, and we equate the sine terms to zero because there is no sine term on the right,

$$L\omega^2(-a) + R\omega b + a/C = E_0\omega \qquad \text{(Cosine terms)}$$

$$L\omega^2(-b) + R\omega(-a) + b/C = 0 \qquad \text{(Sine terms).}$$

To solve this system for a and b, we first introduce a combination of L and C, called the **reactance**

$$\textbf{(3)} \qquad S = \omega L - \frac{1}{\omega C}.$$

Dividing the previous two equations by ω, ordering them, and substituting S gives

$$-Sa + Rb = E_0$$

$$-Ra - Sb = 0.$$

We now eliminate b by multiplying the first equation by S and the second by R, and adding. Then we eliminate a by multiplying the first equation by R and the second by $-S$, and adding. This gives

$$-(S^2 + R^2)a = E_0 S, \qquad (R^2 + S^2)b = E_0 R.$$

In any practical case the resistance R is different from zero, so that we can solve for a and b,

$$a = \frac{-E_0 S}{R^2 + S^2}, \qquad b = \frac{E_0 R}{R^2 + S^2}. \tag{4}$$

Equation (2) with coefficients a and b given by (4) is the desired particular solution I_p of the nonhomogeneous ODE (1) governing the current I in an RLC-circuit with sinusoidal electromotive force.

Using (4), we can write I_p in terms of "physically visible" quantities, namely, amplitude I_0 and phase lag θ of the current behind the electromotive force, that is,

$$I_p(t) = I_0 \sin(\omega t - \theta) \tag{5}$$

where [see (14) in App. A3.1]

$$I_0 = \sqrt{a^2 + b^2} = \frac{E_0}{\sqrt{R^2 + S^2}}, \qquad \tan\theta = -\frac{a}{b} = \frac{S}{R}.$$

The quantity $\sqrt{R^2 + S^2}$ is called the **impedance.** Our formula shows that the impedance equals the ratio E_0/I_0. This is somewhat analogous to $E/I = R$ (Ohm's law).

A general solution of the homogeneous equation corresponding to (1) is

$$I_h = c_1 e^{\lambda_1 t} + c_2 e^{\lambda_2 t}$$

where λ_1 and λ_2 are the roots of the characteristic equation

$$\lambda^2 + \frac{R}{L}\lambda + \frac{1}{LC} = 0.$$

We can write these roots in the form $\lambda_1 = -\alpha + \beta$ and $\lambda_2 = -\alpha - \beta$, where

$$\alpha = \frac{R}{2L}, \qquad \beta = \sqrt{\frac{R^2}{4L^2} - \frac{1}{LC}} = \frac{1}{2L}\sqrt{R^2 - \frac{4L}{C}}.$$

Now in an actual circuit, R is never zero (hence $R > 0$). From this it follows that I_h approaches zero, theoretically as $t \to \infty$, but practically after a relatively short time. (This is as for the motion in the previous section.) Hence the transient current $I = I_h + I_p$ tends

to the steady-state current I_p, and after some time the output will practically be a harmonic oscillation, which is given by (5) and whose frequency is that of the input (of the electromotive force).

EXAMPLE 1 ***RLC*-Circuit**

Find the current $I(t)$ in an *RLC*-circuit with $R = 11\ \Omega$ (ohms), $L = 0.1$ H (henry), $C = 10^{-2}$F (farad), which is connected to a source of voltage $E(t) = 100 \sin 400t$ (hence $63\frac{2}{3}$ Hz $= 63\frac{2}{3}$ cycles/sec, because $400 = 63\frac{2}{3} \cdot 2\pi$). Assume that current and charge are zero when $t = 0$.

Solution. ***Step 1. General solution of the homogeneous ODE.*** Substituting R, L, C, and the derivative $E'(t)$ into (1), we obtain

$$0.1I'' + 11I' + 100I = 100 \cdot 400 \cos 400t.$$

Hence the homogeneous ODE is $0.1I'' + 11I' + 100I = 0$. Its characteristic equation is

$$0.1\lambda^2 + 11\lambda + 100 = 0.$$

The roots are $\lambda_1 = -10$ and $\lambda_2 = -100$. The corresponding general solution of the homogeneous ODE is

$$I_h(t) = c_1e^{-10t} + c_2e^{-100t}.$$

***Step 2. Particular solution I_p of* (1).** We calculate the reactance $S = 40 - 1/4 = 39.75$ and the steady-state current

$$I_p(t) = a \cos 400t + b \sin 400t$$

with coefficients obtained from (4)

$$a = \frac{-100 \cdot 39.75}{11^2 + 39.75^2} = -2.3368, \qquad b = \frac{100 \cdot 11}{11^2 + 39.75^2} = 0.6467.$$

Hence in our present case, a general solution of the nonhomogeneous ODE (1) is

$$I(t) = c_1e^{-10t} + c_2e^{-100t} - 2.3368 \cos 400t + 0.6467 \sin 400t. \tag{6}$$

Step 3. Particular solution satisfying the initial conditions. How to use $Q(0) = 0$? We finally determine c_1 and c_2 from the initial conditions $I(0) = 0$ and $Q(0) = 0$. From the first condition and (6) we have

$$I(0) = c_1 + c_2 - 2.3368 = 0, \qquad \text{hence} \qquad c_2 = 2.3368 - c_1. \tag{7}$$

Furthermore, using (1′) with $t = 0$ and noting that the integral equals $Q(t)$ (see the formula before (1′)), we obtain

$$LI'(0) + R \cdot 0 + \frac{1}{C} \cdot 0 = 0, \qquad \text{hence} \qquad I'(0) = 0.$$

Differentiating (6) and setting $t = 0$, we thus obtain

$$I'(0) = -10c_1 - 100c_2 + 0 + 0.6467 \cdot 400 = 0, \qquad \text{hence} \qquad -10c_1 = 100(2.3368 - c_1) - 258.68.$$

The solution of this and (7) is $c_1 = -0.2776$, $c_2 = 2.6144$. Hence the answer is

$$I(t) = -0.2776e^{-10t} + 2.6144e^{-100t} - 2.3368 \cos 400t + 0.6467 \sin 400t.$$

Figure 62 on p. 96 shows $I(t)$ as well as $I_p(t)$, which practically coincide, except for a very short time near $t = 0$ because the exponential terms go to zero very rapidly. Thus after a very short time the current will practically execute harmonic oscillations of the input frequency $63\frac{2}{3}$ Hz $= 63\frac{2}{3}$ cycles/sec. Its maximum amplitude and phase lag can be seen from (5), which here takes the form

$$I_p(t) = 2.4246 \sin (400t - 1.3008).$$

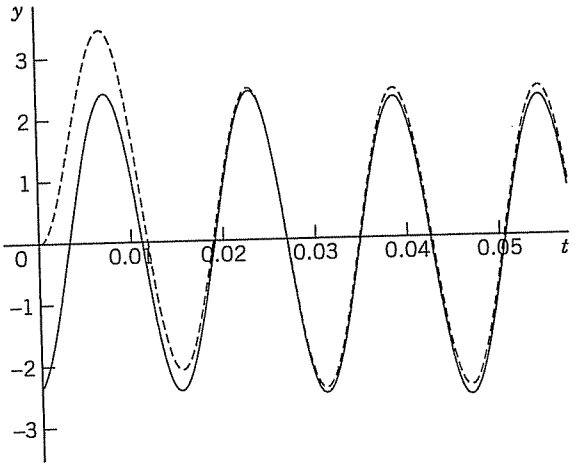


Fig. 62. Transient and steady-state currents in Example 1

Analogy of Electrical and Mechanical Quantities

Entirely different physical or other systems may have the same mathematical model. For instance, we have seen this from the various applications of the ODE $y' = ky$ in Chap. 1. Another impressive demonstration of this ***unifying power of mathematics*** is given by the ODE (1) for an electric *RLC*-circuit and the ODE (2) in the last section for a mass–spring system. Both equations

$$LI'' + RI' + \frac{1}{C}I = E_0\omega \cos \omega t \qquad \text{and} \qquad my'' + cy' + ky = F_0 \cos \omega t$$

are of the same form. Table 2.2 shows the analogy between the various quantities involved. The inductance L corresponds to the mass m and, indeed, an inductor opposes a change in current, having an "inertia effect" similar to that of a mass. The resistance R corresponds to the damping constant c, and a resistor causes loss of energy, just as a damping dashpot does. And so on.

This analogy is ***strictly quantitative*** in the sense that to a given mechanical system we can construct an electric circuit whose current will give the exact values of the displacement in the mechanical system when suitable scale factors are introduced.

The ***practical importance*** of this analogy is almost obvious. The analogy may be used for constructing an "electrical model" of a given mechanical model, resulting in substantial savings of time and money because electric circuits are easy to assemble, and electric quantities can be measured much more quickly and accurately than mechanical ones.

Table 2.2 Analogy of Electrical and Mechanical Quantities

Electrical System	Mechanical System
Inductance L	Mass m
Resistance R	Damping constant c
Reciprocal $1/C$ of capacitance	Spring modulus k
Derivative $E_0\omega \cos \omega t$ of electromotive force	Driving force $F_0 \cos \omega t$
Current $I(t)$	Displacement $y(t)$

PROBLEM SET 2.9

1. **(*RL*-circuit)** Model the *RL*-circuit in Fig. 63. Find the general solution when R, L, E are any constants. Graph or sketch solutions when $L = 0.1$ H, $R = 5\ \Omega$, $E = 12$ V.
2. **(*RL*-circuit)** Solve Prob. 1 when $E = E_0 \sin \omega t$ and R, L, E_0, ω are arbitrary. Sketch a typical solution.
3. **(*RC*-circuit)** Model the *RC*-circuit in Fig. 66. Find the current due to a constant E.
4. **(*RC*-circuit)** Find the current in the *RC*-circuit in Fig. 66 with $E = E_0 \sin \omega t$ and arbitrary R, C, E_0, and ω.

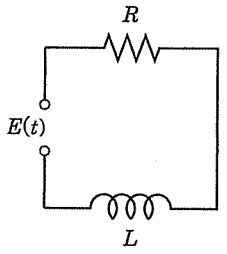


Fig. 63. *RL*-circuit

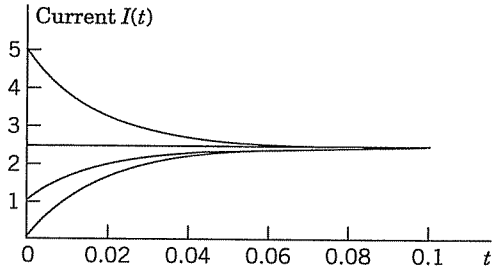


Fig. 64. Currents in Problem 1

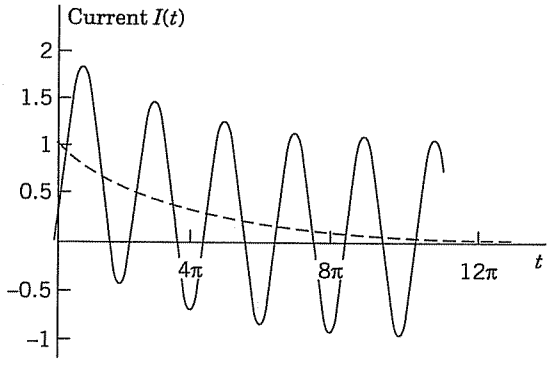


Fig. 65. Typical current $I = e^{-0.1t} + \sin(t - \frac{1}{4}\pi)$ in Problem 2

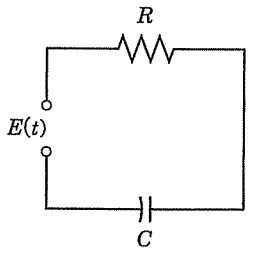


Fig. 66. *RC*-circuit

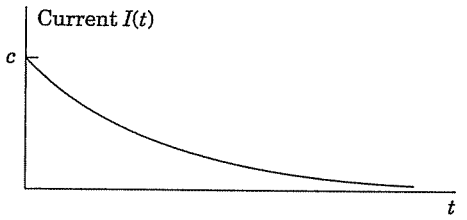


Fig. 67. Current I in Problem 3

5. **(*LC*-circuit)** This is an *RLC*-circuit with negligibly small R (analog of an undamped mass–spring system). Find the current when $L = 0.2$ H, $C = 0.05$ F, and $E = \sin t$ V, assuming zero initial current and charge.
6. **(*LC*-circuit)** Find the current when $L = 0.5$ H, $C = 8 \cdot 10^{-4}$ F, $E = t^2$ V and initial current and charge zero.

7–9 *RLC*-CIRCUITS (FIG. 60, P. 92)

7. **(Tuning)** In tuning a stereo system to a radio station, we adjust the tuning control (turn a knob) that changes C (or perhaps L) in an *RLC*-circuit so that the amplitude of the steady-state current (5) becomes maximum. For what C will this happen?
8. **(Transient current)** Prove the claim in the text that if $R \neq 0$ (hence $R > 0$), then the transient current approaches I_p as $t \to \infty$.
9. **(Cases of damping)** What are the conditions for an *RLC*-circuit to be (I) overdamped, (II) critically damped, (III) underdamped? What is the critical resistance R_{crit} (the analog of the critical damping constant $2\sqrt{mk}$)?

10–12 Find the **steady-state current** in the *RLC*-circuit in Fig. 60 on p. 92 for the given data. (Show the details of your work.)

10. $R = 8\ \Omega$, $L = 0.5$ H, $C = 0.1$ F, $E = 100 \sin 2t$ V
11. $R = 1\ \Omega$, $L = 0.25$ H, $C = 5 \cdot 10^{-5}$ F, $E = 110$ V
12. $R = 2\ \Omega$, $L = 1$ H, $C = 0.05$ F, $E = \frac{157}{9} \sin 3t$ V

13–15 Find the **transient current** (a general solution) in the *RLC*-circuit in Fig. 60 for the given data. (Show the details of your work.)

13. $R = 6\ \Omega$, $L = 0.2$ H, $C = 0.025$ F, $E = 110 \sin 10t$ V
14. $R = 0.2\ \Omega$, $L = 0.1$ H, $C = 2$ F, $E = 754 \sin 0.5t$ V
15. $R = 1/10\ \Omega$, $L = 1/2$ H, $C = 100/13$ F, $E = e^{-4t}(1.932 \cos \frac{1}{2}t + 0.246 \sin \frac{1}{2}t)$ V

16–18 Solve the **initial value problem** for the *RLC*-circuit in Fig. 60 with the given data, assuming zero initial current and charge. Graph or sketch the solution. (Show the details of your work.)

16. $R = 4\ \Omega$, $L = 0.1$ H, $C = 0.025$ F, $E = 10 \sin 10t$ V
17. $R = 6\ \Omega$, $L = 1$ H, $C = 0.04$ F, $E = 600(\cos t + 4 \sin t)$ V
18. $R = 3.6\ \Omega$, $L = 0.2$ H, $C = 0.0625$ F, $E = 164 \cos 10t$ V

19. **WRITING PROJECT. Analogy of *RLC*-Circuits and Damped Mass–Spring Systems.** **(a)** Write an essay of 2–3 pages based on Table 2.2. Describe the analogy in more detail and indicate its practical significance.

(b) What *RLC*-circuit with $L = 1$ H is the analog of the mass–spring system with mass 5 kg, damping constant 10 kg/sec, spring constant 60 kg/sec^2, and driving force 220 cos $10t$?

(c) Illustrate the analogy with another example of your own choice.

20. **TEAM PROJECT. Complex Method for Particular Solutions.** **(a)** Find a particular solution of the **complex ODE**

$$\text{(8)} \qquad L\tilde{I}'' + R\tilde{I}' + \frac{1}{C}\tilde{I} = E_0\omega e^{i\omega t} \qquad (i = \sqrt{-1})$$

by substituting $\tilde{I}_p = Ke^{i\omega t}$ (K unknown) and its derivatives into (8), and then take the real part I_p of $\tilde{I}_p$, showing that I_p agrees with (2), (4). *Hint.* Use the Euler formula $e^{i\omega t} = \cos \omega t + i \sin \omega t$ [(11) in Sec. 2.2 with ωt instead of t]. Note that $E_0\omega \cos \omega t$ in (1) is the real part of $E_0\omega e^{i\omega t}$ in (8). Use $i^2 = -1$.

(b) The **complex impedance** Z is defined by

$$Z = R + iS = R + i\left(\omega L - \frac{1}{\omega C}\right).$$

Show that K obtained in (a) can be written as

$$K = \frac{E_0}{iZ}.$$

Note that the real part of Z is R, the imaginary part is the reactance S, and the absolute value is the impedance $|Z| = \sqrt{R^2 + S^2}$ as defined before. See Fig. 68.

(c) Find the steady-state solution of the ODE $I'' + 2I' + 3I = 20 \cos t$, first by the real method and then by the complex method, and compare. (Show the details of your work.)

(d) Apply the complex method to an *RLC*-circuit of your choice.

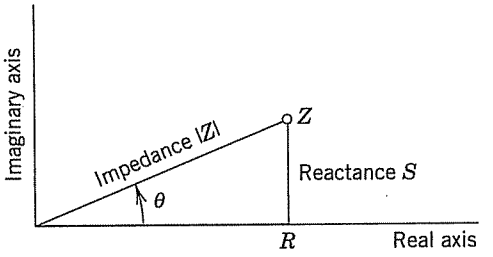


Fig. 68. Complex impedance Z

2.10 Solution by Variation of Parameters

We continue our discussion of nonhomogeneous linear ODEs

$$\text{(1)} \qquad y'' + p(x)y' + q(x)y = r(x).$$

In Sec. 2.6 we have seen that a general solution of (1) is the sum of a general solution y_h of the corresponding homogeneous ODE and any particular solution y_p of (1). To obtain y_p when $r(x)$ is not too complicated, we can often use the *method of undetermined coefficients,* as we have shown in Sec. 2.7 and applied to basic engineering models in Secs. 2.8 and 2.9.

However, since this method is restricted to functions $r(x)$ whose derivatives are of a form similar to $r(x)$ itself (powers, exponential functions, etc.), it is desirable to have a method valid for more general ODEs (1), which we shall now develop. It is called the **method of variation of parameters** and is credited to Lagrange (Sec. 2.1). Here p, q, r in (1) may be variable (given functions of x), but we assume that they are continuous on some open interval I.

Lagrange's method gives a particular solution y_p of (1) on I in the form

$$\text{(2)} \qquad y_p(x) = -y_1\int \frac{y_2 r}{W}\,dx + y_2\int \frac{y_1 r}{W}\,dx$$

where y_1, y_2 form a basis of solutions of the corresponding homogeneous ODE

$$y'' + p(x)y' + q(x)y = 0 \tag{3}$$

on I, and W is the Wronskian of y_1, y_2,

$$W = y_1y_2' - y_2y_1' \tag{4}$$

(see Sec. 2.6).

CAUTION! The solution formula (2) is obtained under the assumption that the ODE is written in standard form, with y'' as the first term as shown in (1). If it starts with $f(x)y''$, divide first by $f(x)$.

The integration in (2) may often cause difficulties, and so may the determination of y_1, y_2 if (1) has variable coefficients. If you have a choice, use the previous method. It is simpler. Before deriving (2) let us work an example for which you *do need* the new method. (Try otherwise.)

EXAMPLE 1 **Method of Variation of Parameters**

Solve the nonhomogeneous ODE

$$y'' + y = \sec x = \frac{1}{\cos x}.$$

Solution. A basis of solutions of the homogeneous ODE on any interval is $y_1 = \cos x$, $y_2 = \sin x$. This gives the Wronskian

$$W(y_1, y_2) = \cos x \cos x - \sin x\,(-\sin x) = 1.$$

From (2), choosing zero constants of integration, we get the particular solution of the given ODE

$$\begin{aligned} y_p &= -\cos x \int \sin x \sec x\, dx + \sin x \int \cos x \sec x\, dx \\ &= \cos x \ln |\cos x| + x \sin x \end{aligned} \qquad \text{(Fig. 69).}$$

Figure 69 shows y_p and its first term, which is small, so that $x \sin x$ essentially determines the shape of the curve of y_p. (Recall from Sec. 2.8 that we have seen $x \sin x$ in connection with resonance, except for notation.) From y_p and the general solution $y_h = c_1y_1 + c_2y_2$ of the homogeneous ODE we obtain the *answer*

$$y = y_h + y_p = (c_1 + \ln |\cos x|) \cos x + (c_2 + x) \sin x.$$

Had we included integration constants $-c_1$, c_2 in (2), then (2) would have given the additional $c_1 \cos x + c_2 \sin x = c_1y_1 + c_2y_2$, that is, a general solution of the given ODE directly from (2). This will always be the case. ■

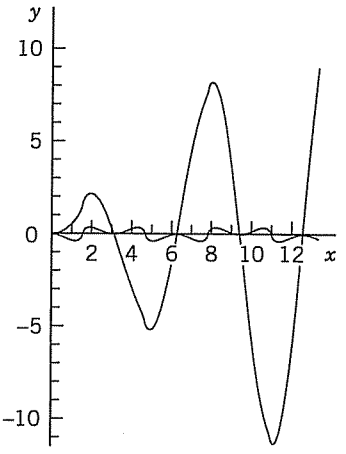


Fig. 69. Particular solution y_p and its first term in Example 1

Idea of the Method. Derivation of (2)

What idea did Lagrange have? What gave the method the name? Where do we use the continuity assumptions?

The idea is to start from a general solution

$$y_h(x) = c_1y_1(x) + c_2y_2(x)$$

of the homogeneous ODE (3) on an open interval I and to replace the constants ("the parameters") c_1 and c_2 by functions $u(x)$ and $v(x)$; this suggests the name of the method. We shall determine u and v so that the resulting function

$$y_p(x) = u(x)y_1(x) + v(x)y_2(x) \tag{5}$$

is a particular solution of the nonhomogeneous ODE (1). Note that y_h exists by Theorem 3 in Sec. 2.6 because of the continuity of p and q on I. (The continuity of r will be used later.)

We determine u and v by substituting (5) and its derivatives into (1). Differentiating (5), we obtain

$$y_p' = u'y_1 + uy_1' + v'y_2 + vy_2'.$$

Now y_p must satisfy (1). This is *one* condition for *two* functions u and v. It seems plausible that we may impose a *second* condition. Indeed, our calculation will show that we can determine u and v such that y_p satisfies (1) and u and v satisfy as a second condition the equation

$$u'y_1 + v'y_2 = 0. \tag{6}$$

This reduces the first derivative y_p' to the simpler form

$$y_p' = uy_1' + vy_2'. \tag{7}$$

Differentiating (7), we obtain

$$y_p'' = u'y_1' + uy_1'' + v'y_2' + vy_2''. \tag{8}$$

We now substitute y_p and its derivatives according to (5), (7), (8) into (1). Collecting terms in u and terms in v, we obtain

$$u(y_1'' + py_1' + qy_1) + v(y_2'' + py_2' + qy_2) + u'y_1' + v'y_2' = r.$$

Since y_1 and y_2 are solutions of the homogeneous ODE (3), this reduces to

$$u'y_1' + v'y_2' = r. \tag{9a}$$

Equation (6) is

$$u'y_1 + v'y_2 = 0. \tag{9b}$$

This is a linear system of two algebraic equations for the unknown functions u' and v'. We can solve it by elimination as follows (or by Cramer's rule in Sec. 7.6). To eliminate v', we multiply (9a) by $-y_2$ and (9b) by y_2' and add, obtaining

$$u'(y_1y_2' - y_2y_1') = -y_2r, \qquad \text{thus} \qquad u'W = -y_2r.$$

Here, W is the Wronskian (4) of y_1, y_2. To eliminate u' we multiply (9a) by y_1, and (9b) by $-y_1'$ and add, obtaining

$$v'(y_1y_2' - y_2y_1') = y_1r, \qquad \text{thus} \qquad v'W = y_1r.$$

Since y_1, y_2 form a basis, we have $W \neq 0$ (by Theorem 2 in Sec. 2.6) and can divide by W,

$$u' = -\frac{y_2r}{W}, \qquad v' = \frac{y_1r}{W}. \tag{10}$$

By integration,

$$u = -\int \frac{y_2r}{W}\,dx, \qquad v = \int \frac{y_1r}{W}\,dx.$$

These integrals exist because $r(x)$ is continuous. Inserting them into (5) gives (2) and completes the derivation. ∎

PROBLEM SET 2.10

1–17 GENERAL SOLUTION

Solve the given nonhomogeneous ODE by variation of parameters or undetermined coefficients. Give a general solution. (Show the details of your work.)

1. $y'' + y = \csc x$
2. $y'' - 4y' + 4y = x^2e^x$
3. $x^2y'' - 2xy' + 2y = x^3\cos x$
4. $y'' - 2y' + y = e^x \sin x$
5. $y'' + y = \tan x$
6. $x^2y'' - xy' + y = x \ln |x|$
7. $y'' + y = \cos x + \sec x$
8. $y'' - 4y' + 4y = 12e^{2x}/x^4$
9. $(D^2 - 2D + I)y = x^2 + x^{-2}e^x$
10. $(D^2 - I)y = 1/\cosh x$
11. $(D^2 + 4I)y = \cosh 2x$
12. $(x^2D^2 + xD - \frac{1}{4}I)y = 3x^{-1} + 3x$
13. $(x^2D^2 - 2xD + 2I)y = x^3 \sin x$
14. $(x^2D^2 + xD - 4I)y = 1/x^2$
15. $(D^2 + I)y = \sec x - 10 \sin 5x$
16. $(x^2D^2 + xD + (x^2 - \frac{1}{4})I)y = x^{3/2}\cos x$.
    Hint. To find y_1, y_2 set $y = ux^{-1/2}$.
17. $(x^2D^2 + xD + (x^2 - \frac{1}{4})I)y = x^{3/2}\sin x$.
    Hint: As in Prob. 16.
18. **TEAM PROJECT. Comparison of Methods.** The undetermined-coefficient method should be used whenever possible because it is simpler. Compare it with the present method as follows.

    (a) Solve $y'' + 2y' - 15y = 17 \sin 5x$ by both methods, showing all details, and compare.

    (b) Solve $y'' + 9y = r_1 + r_2$, $r_1 = \sec 3x$, $r_2 = \sin 3x$ by applying each method to a suitable function on the right.

    (c) Invent an undetermined-coefficient method for nonhomogeneous Euler–Cauchy equations by experimenting.

CHAPTER 2 REVIEW QUESTIONS AND PROBLEMS

1. What general properties make *linear* ODEs particularly attractive?
2. What is a general solution of a linear ODE? A basis of solutions?
3. How would you obtain a general solution of a nonhomogeneous linear ODE if you knew a general solution of the corresponding homogeneous ODE?
4. What does an initial value problem for a second-order ODE look like?
5. What is a particular solution and why is it more common than a general solution as the answer to practical problems?
6. Why are second-order ODEs more important in modeling than ODEs of higher order?
7. Describe the applications of ODEs in mechanical vibrating systems. What are the electrical analogs of those systems?
8. If a construction, such as a bridge, shows undesirable resonance, what could you do?

9–18 GENERAL SOLUTION

Find a general solution. Indicate the method you are using and show the details of your calculation.

9. $y'' - 2y' - 8y = 52 \cos 6x$
10. $y'' + 6y' + 9y = e^{-3x} - 27x^2$
11. $y'' + 8y' + 25y = 26 \sin 3x$
12. $yy'' = 2y'^2$
13. $(x^2D^2 + 2xD - 12I)y = 1/x^3$
14. $(x^2D^2 + 6xD + 6I)y = x^2$
15. $(D^2 - 2D + I)y = x^{-3}e^x$
16. $(D^2 - 4D + 5I)y = e^{2x} \csc x$
17. $(D^2 - 2D + 2I)y = e^x \csc x$
18. $(4x^2D^2 - 24xD + 49I)y = 36x^5$

19–25 INITIAL VALUE PROBLEMS

Solve the following initial value problems. Sketch or graph the solution. (Show the details of your work.)

19. $y'' + 5y' - 14y = 0, \quad y(0) = 6, \quad y'(0) = -6$
20. $y'' + 6y' + 18y = 0, \quad y(0) = 5, \quad y'(0) = -21$
21. $x^2y'' - xy' - 24y = 0, \quad y(1) = 15, \quad y'(1) = 0$
22. $x^2y'' + 15xy' + 49y = 0, \quad y(1) = 2, \quad y'(1) = -11$
23. $y'' + 5y' + 6y = 108x^2, \quad y(0) = 18, \quad y'(0) = -26$
24. $y'' + y' + 2.5y = 13 \cos x, \; y(0) = 8.0, \; y'(0) = 4.5$
25. $(x^2D^2 + xD - 4I)y = x^3, \; y(1) = -4/5, \; y'(1) = 93/5$

26–34 APPLICATIONS

26. Find the steady-state solution of the system in Fig. 70 when $m = 4$, $c = 4$, $k = 17$ and the driving force is $202 \cos 3t$.
27. Find the motion of the system in Fig. 70 with mass 0.25 kg, no damping, spring constant 1 kg/sec^2, and driving force $15 \cos 0.5t - 7 \sin 1.5t$ nt, assuming zero initial displacement and velocity. For what frequency of the driving force would you get resonance?
28. In Prob. 26 find the solution corresponding to initial displacement 10 and initial velocity 0.
29. Show that the system in Fig. 70 with $m = 4$, $c = 0$, $k = 36$, and driving force $61 \cos 3.1t$ exhibits beats. *Hint:* Choose zero initial conditions.
30. In Fig. 70 let $m = 2$, $c = 6$, $k = 27$, and $r(t) = 10 \cos \omega t$. For what ω will you obtain the steady-state vibration of maximum possible amplitude? Determine this amplitude. Then use this ω and the undetermined-coefficient method to see whether you obtain the same amplitude.
31. Find an electrical analog of the mass–spring system in Fig. 70 with mass 0.5 kg, spring constant 40 kg/sec^2, damping constant 9 kg/sec, and driving force $102 \cos 6t$ nt. Solve the analog, assuming zero initial current and charge.
32. Find the current in the *RLC*-circuit in Fig. 71 when $L = 0.1$ H, $R = 20\ \Omega$, $C = 2 \cdot 10^{-4}$ F, and $E(t) = 110 \sin 415t$ V (66 cycles/sec).
33. Find the current in the *RLC*-circuit when $L = 0.4$ H, $R = 40\ \Omega$, $C = 10^{-4}$ F, and $E(t) = 220 \sin 314t$ V (50 cycles/sec).
34. Find a particular solution in Prob. 33 by the complex method. (See Team Project 20 in Sec. 2.9.)

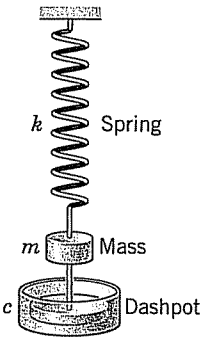


Fig. 70. Mass–spring system

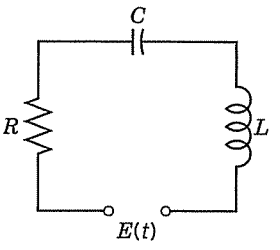


Fig. 71. *RLC*-circuit

SUMMARY OF CHAPTER 2
Second-Order Linear ODEs

Second-order linear ODEs are particularly important in applications, for instance, in mechanics (Secs. 2.4, 2.8) and electrical engineering (Sec. 2.9). A second-order ODE is called **linear** if it can be written

(1) $$y'' + p(x)y' + q(x)y = r(x) \qquad \text{(Sec. 2.1).}$$

(If the first term is, say, $f(x)y''$, divide by $f(x)$ to get the **"standard form"** (1) with y'' as the first term.) Equation (1) is called **homogeneous** if $r(x)$ is zero for all x considered, usually in some open interval; this is written $r(x) \equiv 0$. Then

(2) $$y'' + p(x)y' + q(x)y = 0.$$

Equation (1) is called **nonhomogeneous** if $r(x) \not\equiv 0$ (meaning $r(x)$ is not zero for some x considered).

For the homogeneous ODE (2) we have the important **superposition principle** (Sec. 2.1) that a linear combination $y = ky_1 + ly_2$ of two solutions y_1, y_2 is again a solution.

Two *linearly independent* solutions y_1, y_2 of (2) on an open interval I form a **basis** (or **fundamental system**) of solutions on I, and $y = c_1y_1 + c_2y_2$ with arbitrary constants c_1, c_2 is a **general solution** of (2) on I. From it we obtain a **particular solution** if we specify numeric values (numbers) for c_1 and c_2, usually by prescribing two **initial conditions**

(3) $$y(x_0) = K_0, \qquad y'(x_0) = K_1 \qquad (x_0, K_0, K_1 \text{ given numbers; Sec. 2.1}).$$

(2) and (3) together form an **initial value problem.** Similarly for (1) and (3).

For a nonhomogeneous ODE (1) a **general solution** is of the form

(4) $$y = y_h + y_p \qquad \text{(Sec. 2.7).}$$

Here y_h is a general solution of (2) and y_p is a particular solution of (1). Such a y_p can be determined by a general method (**variation of parameters,** Sec. 2.10) or in many practical cases by the **method of undetermined coefficients.** The latter applies when (1) has constant coefficients p and q, and $r(x)$ is a power of x, sine, cosine, etc. (Sec. 2.7). Then we write (1) as

(5) $$y'' + ay' + by = r(x) \qquad \text{(Sec. 2.7).}$$

The corresponding homogeneous ODE $y' + ay' + by = 0$ has solutions $y = e^{\lambda x}$, where λ is a root of

(6) $$\lambda^2 + a\lambda + b = 0.$$

Hence there are three cases (Sec. 2.2):

Case	Type of Roots	General Solution
I	Distinct real λ_1, λ_2	$y = c_1 e^{\lambda_1 x} + c_2 e^{\lambda_2 x}$
II	Double $-\frac{1}{2}a$	$y = (c_1 + c_2 x) e^{-ax/2}$
III	Complex $-\frac{1}{2}a \pm i\omega^*$	$y = e^{-ax/2}(A \cos \omega^* x + B \sin \omega^* x)$

Important applications of (5) in mechanical and electrical engineering in connection with **vibrations** and **resonance** are discussed in Secs. 2.4, 2.7, and 2.8.

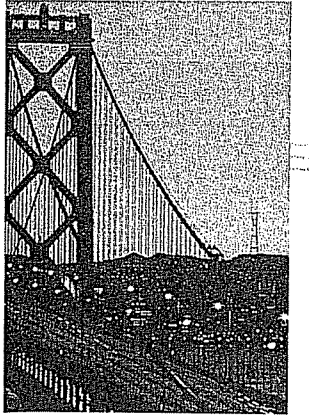
Another large class of ODEs solvable "algebraically" consists of the **Euler–Cauchy equations**

(7) $$x^2y'' + axy' + by = 0 \qquad \text{(Sec. 2.5).}$$

These have solutions of the form $y = x^m$, where m is a solution of the auxiliary equation

(8) $$m^2 + (a - 1)m + b = 0.$$

Existence and uniqueness of solutions of (1) and (2) is discussed in Secs. 2.6 and 2.7, and **reduction of order** in Sec. 2.1.



CHAPTER 3

Higher Order Linear ODEs

In this chapter we extend the concepts and methods of Chap. 2 for linear ODEs from order $n = 2$ to arbitrary order n. This will be straightforward and needs no new ideas. However, the formulas become more involved, the variety of roots of the characteristic equation (in Sec. 3.2) becomes much larger with increasing n, and the Wronskian plays a more prominent role.

Prerequisite: Secs. 2.1, 2.2, 2.6, 2.7, 2.10.
References and Answers to Problems: App. 1 Part A, and App. 2.

3.1 Homogeneous Linear ODEs

Recall from Sec. 1.1 that an ODE is of **nth order** if the nth derivative $y^{(n)} = d^n y/dx^n$ of the unknown function $y(x)$ is the highest occurring derivative. Thus the ODE is of the form

$$F(x, y, y', \cdots, y^{(n)}) = 0 \qquad \left(y^{(n)} = \frac{d^n y}{dx^n}\right)$$

where lower order derivatives and y itself may or may not occur. Such an ODE is called **linear** if it can be written

$$\textbf{(1)} \qquad y^{(n)} + p_{n-1}(x)y^{(n-1)} + \cdots + p_1(x)y' + p_0(x)y = r(x).$$

(For $n = 2$ this is (1) in Sec. 2.1 with $p_1 = p$ and $p_0 = q$). The **coefficients** $p_0, \cdots, p_{n-1}$ and the function r on the right are any given functions of x, and y is unknown. $y^{(n)}$ has coefficient 1. This is practical. We call this the **standard form.** (If you have $p_n(x)y^{(n)}$, divide by $p_n(x)$ to get this form.) An nth-order ODE that cannot be written in the form (1) is called **nonlinear.**

If $r(x)$ is identically zero, $r(x) \equiv 0$ (zero for all x considered, usually in some open interval I), then (1) becomes

$$\textbf{(2)} \qquad y^{(n)} + p_{n-1}(x)y^{(n-1)} + \cdots + p_1(x)y' + p_0(x)y = 0$$

and is called **homogeneous.** If $r(x)$ is not identically zero, then the ODE is called **nonhomogeneous.** This is as in Sec. 2.1.

A **solution** of an nth-order (linear or nonlinear) ODE on some open interval I is a function $y = h(x)$ that is defined and n times differentiable on I and is such that the ODE becomes an identity if we replace the unknown function y and its derivatives by h and its corresponding derivatives.

Homogeneous Linear ODE: Superposition Principle, General Solution

Sections 3.1–3.2 will be devoted to homogeneous linear ODEs and Sec. 3.3 to nonhomogeneous linear ODEs. The basic **superposition or linearity principle** in Sec. 2.1 extends to nth order homogeneous linear ODEs as follows.

THEOREM 1

Fundamental Theorem for the Homogeneous Linear ODE (2)

For a homogeneous linear ODE (2), *sums and constant multiples of solutions on some open interval I are again solutions on I.* (This does ***not*** hold for a nonhomogeneous or nonlinear ODE!)

The proof is a simple generalization of that in Sec. 2.1 and we leave it to the student.

Our further discussion parallels and extends that for second-order ODEs in Sec. 2.1. So we define next a general solution of (2), which will require an extension of linear independence from 2 to n functions.

DEFINITION

General Solution, Basis, Particular Solution

A **general solution** of (2) on an open interval I is a solution of (2) on I of the form

$$y(x) = c_1y_1(x) + \cdots + c_ny_n(x) \qquad (c_1, \cdots, c_n \text{ arbitrary}) \tag{3}$$

where $y_1, \cdots, y_n$ is a **basis** (or **fundamental system**) of solutions of (2) on I; that is, these solutions are linearly independent on I, as defined below.

A **particular solution** of (2) on I is obtained if we assign specific values to the n constants $c_1, \cdots, c_n$ in (3).

DEFINITION

Linear Independence and Dependence

n functions $y_1(x), \cdots, y_n(x)$ are called **linearly independent** *on some interval I* where they are defined if the equation

$$k_1y_1(x) + \cdots + k_ny_n(x) = 0 \qquad \text{on } I \tag{4}$$

implies that all $k_1, \cdots, k_n$ are zero. These functions are called **linearly dependent** on I if this equation also holds on I for some $k_1, \cdots, k_n$ not all zero.

(As in Secs. 1.1 and 2.1, the arbitrary constants $c_1, \cdots, c_n$ must sometimes be restricted to some interval.)

If and only if $y_1, \cdots, y_n$ are linearly dependent on I, we can express (at least) one of these functions on I as a **"linear combination"** of the other $n - 1$ functions, that is, as a sum of those functions, each multiplied by a constant (zero or not). This motivates the term "linearly dependent." For instance, if (4) holds with $k_1 \neq 0$, we can divide by k_1 and express y_1 as the linear combination

$$y_1 = -\frac{1}{k_1}(k_2y_2 + \cdots + k_ny_n).$$

Note that when $n = 2$, these concepts reduce to those defined in Sec. 2.1.

EXAMPLE 1 **Linear Dependence**

Show that the functions $y_1 = x^2$, $y_2 = 5x$, $y_3 = 2x$ are linearly dependent on any interval.

Solution. $y_2 = 0y_1 + 2.5y_3$. This proves linear dependence on any interval. ▩

EXAMPLE 2 **Linear Independence**

Show that $y_1 = x$, $y_2 = x^2$, $y_3 = x^3$ are linearly independent on any interval, for instance, on $-1 \leqq x \leqq 2$.

Solution. Equation (4) is $k_1x + k_2x^2 + k_3x^3 = 0$. Taking (a) $x = -1$, (b) $x = 1$, (c) $x = 2$, we get

$$\text{(a) } -k_1 + k_2 - k_3 = 0, \qquad \text{(b) } k_1 + k_2 + k_3 = 0, \qquad \text{(c) } 2k_1 + 4k_2 + 8k_3 = 0.$$

$k_2 = 0$ from (a) + (b). Then $k_3 = 0$ from (c) -2(b). Then $k_1 = 0$ from (b). This proves linear independence.
A better method for testing linear independence of solutions of ODEs will soon be explained. ▩

EXAMPLE 3 **General Solution. Basis**

Solve the fourth-order ODE

$$y^{\text{iv}} - 5y'' + 4y = 0 \qquad (\text{where } y^{\text{iv}} = d^4y/dx^4).$$

Solution. As in Sec. 2.2 we try and substitute $y = e^{\lambda x}$. Omitting the common factor $e^{\lambda x}$, we obtain the characteristic equation

$$\lambda^4 - 5\lambda^2 + 4 = 0.$$

This is a quadratic equation in $\mu = \lambda^2$, namely,

$$\mu^2 - 5\mu + 4 = (\mu - 1)(\mu - 4) = 0.$$

The roots are $\mu = 1$ and 4. Hence $\lambda = -2, -1, 1, 2$. This gives four solutions. A general solution on any interval is

$$y = c_1e^{-2x} + c_2e^{-x} + c_3e^{x} + c_4e^{2x}$$

provided those four solutions are linearly independent. This is true but will be shown later. ▩

Initial Value Problem. Existence and Uniqueness

An **initial value problem** for the ODE (2) consists of (2) and n **initial conditions**

$$\text{(5)} \qquad y(x_0) = K_0, \quad y'(x_0) = K_1, \quad \cdots, \quad y^{(n-1)}(x_0) = K_{n-1}$$

with given x_0 in the open interval I considered, and given $K_0, \cdots, K_{n-1}$.
In extension of the existence and uniqueness theorem in Sec. 2.6 we now have the following.

THEOREM 2

Existence and Uniqueness Theorem for Initial Value Problems

If the coefficients $p_0(x), \cdots, p_{n-1}(x)$ of (2) are continuous on some open interval I and x_0 is in I, then the initial value problem (2), (5) has a unique solution $y(x)$ on I.

Existence is proved in Ref. [A11] in App. 1. Uniqueness can be proved by a slight generalization of the uniqueness proof at the beginning of App. 4.

EXAMPLE 4 Initial Value Problem for a Third-Order Euler–Cauchy Equation

Solve the following initial value problem on any open interval I on the positive x-axis containing $x = 1$.

$$x^3y''' - 3x^2y'' + 6xy' - 6y = 0, \qquad y(1) = 2, \qquad y'(1) = 1, \qquad y''(1) = -4.$$

Solution. ***Step 1. General solution.*** As in Sec. 2.5 we try $y = x^m$. By differentiation and substitution,

$$m(m-1)(m-2)x^m - 3m(m-1)x^m + 6mx^m - 6x^m = 0.$$

Dropping x^m and ordering gives $m^3 - 6m^2 + 11m - 6 = 0$. If we can guess the root $m = 1$, we can divide by $m - 1$ and find the other roots 2 and 3, thus obtaining the solutions x, x^2, x^3, which are linearly independent on I (see Example 2). [In general one shall need a root-finding method, such as Newton's (Sec. 19.2), also available in a CAS (Computer Algebra System).] Hence a general solution is

$$y = c_1x + c_2x^2 + c_3x^3$$

valid on any interval I, even when it includes $x = 0$ where the coefficients of the ODE divided by x^3 (to have the standard form) are not continuous.

Step 2. Particular solution. The derivatives are $y' = c_1 + 2c_2x + 3c_3x^2$ and $y'' = 2c_2 + 6c_3x$. From this and y and the initial conditions we get by setting $x = 1$

$$\begin{aligned} \text{(a)}\ y(1) &= c_1 + c_2 + c_3 = 2 \\ \text{(b)}\ y'(1) &= c_1 + 2c_2 + 3c_3 = 1 \\ \text{(c)}\ y''(1) &= 2c_2 + 6c_3 = -4. \end{aligned}$$

This is solved by Cramer's rule (Sec. 7.6), or by elimination, which is simple, as follows. (b) − (a) gives (d) $c_2 + 2c_3 = -1$. Then (c) − 2(d) gives $c_3 = -1$. Then (c) gives $c_2 = 1$. Finally $c_1 = 2$ from (a).
Answer: $y = 2x + x^2 - x^3$. ■

Linear Independence of Solutions. Wronskian

Linear independence of solutions is crucial for obtaining general solutions. Although it can often be seen by inspection, it would be good to have a criterion for it. Now Theorem 2 in Sec. 2.6 extends from order $n = 2$ to any n. This extended criterion uses the **Wronskian** W of n solutions $y_1, \cdots, y_n$ defined as the nth order determinant

$$\textbf{(6)} \qquad W(y_1, \cdots, y_n) = \begin{vmatrix} y_1 & y_2 & \cdots & y_n \\ y_1' & y_2' & \cdots & y_n' \\ \cdot & \cdot & \cdots & \cdot \\ y_1^{(n-1)} & y_2^{(n-1)} & \cdots & y_n^{(n-1)} \end{vmatrix}.$$

Note that W depends on x since $y_1, \cdots, y_n$ does. The criterion states that these solutions form a basis if and only if W is not zero; more precisely:

THEOREM 3 Linear Dependence and Independence of Solutions

Let the ODE (2) *have continuous coefficients* $p_0(x), \cdots, p_{n-1}(x)$ *on an open interval* I. *Then* n *solutions* $y_1, \cdots, y_n$ *of* (2) *on* I *are linearly dependent on* I *if and only if their Wronskian is zero for some* $x = x_0$ *in* I. *Furthermore, if* W *is zero for* $x = x_0$, *then* W *is identically zero on* I. *Hence if there is an* x_1 *in* I *at which* W *is not zero, then* $y_1, \cdots, y_n$ *are linearly independent on* I, *so that they form a basis of solutions of* (2) *on* I.

PROOF **(a)** Let $y_1, \cdots, y_n$ be linearly dependent solutions of (2) on I. Then, by definition, there are constants $k_1, \cdots, k_n$ *not all zero*, such that for all x in I,

$$k_1 y_1 + \cdots + k_n y_n = 0. \tag{7}$$

By $n - 1$ differentiations of (7) we obtain for all x in I

$$\begin{aligned} k_1 y_1' + \cdots + k_n y_n' &= 0 \\ &\vdots \\ k_1 y_1^{(n-1)} + \cdots + k_n y_n^{(n-1)} &= 0. \end{aligned} \tag{8}$$

(7), (8) is a homogeneous linear system of algebraic equations with a nontrivial solution $k_1, \cdots, k_n$. Hence its coefficient determinant must be zero for every x on I, by Cramer's theorem (Sec. 7.7). But that determinant is the Wronskian W, as we see from (6). Hence W is zero for every x on I.

(b) Conversely, if W is zero at an x_0 in I, then the system (7), (8) with $x = x_0$ has a solution $k_1^*, \cdots, k_n^*$, not all zero, by the same theorem. With these constants we define the solution $y^* = k_1^* y_1 + \cdots + k_n^* y_n$ of (2) on I. By (7), (8) this solution satisfies the initial conditions $y^*(x_0) = 0, \cdots, y^{*(n-1)}(x_0) = 0$. But another solution satisfying the same conditions is $y \equiv 0$. Hence $y^* \equiv y$ by Theorem 2, which applies since the coefficients of (2) are continuous. Together, $y^* = k_1^* y_1 + \cdots + k_n^* y_n \equiv 0$ on I. This means linear dependence of $y_1, \cdots, y_n$ on I.

(c) If W is zero at an x_0 in I, we have linear dependence by (b) and then $W \equiv 0$ by (a). Hence if W is not zero at an x_1 in I, the solutions $y_1, \cdots, y_n$ must be linearly independent on I. ∎

EXAMPLE 5 **Basis, Wronskian**

We can now prove that in Example 3 we do have a basis. In evaluating W, pull out the exponential functions columnwise. In the result, subtract Column 1 from Columns 2, 3, 4 (without changing Column 1). Then expand by Row 1. In the resulting third-order determinant, subtract Column 1 from Column 2 and expand the result by Row 2:

$$W = \begin{vmatrix} e^{-2x} & e^{-x} & e^{x} & e^{2x} \\ -2e^{-2x} & -e^{-x} & e^{x} & 2e^{2x} \\ 4e^{-2x} & e^{-x} & e^{x} & 4e^{2x} \\ -8e^{-2x} & -e^{-x} & e^{x} & 8e^{2x} \end{vmatrix} = \begin{vmatrix} 1 & 1 & 1 & 1 \\ -2 & -1 & 1 & 2 \\ 4 & 1 & 1 & 4 \\ -8 & -1 & 1 & 8 \end{vmatrix} = \begin{vmatrix} 1 & 3 & 4 \\ -3 & -3 & 0 \\ 7 & 9 & 16 \end{vmatrix} = 72. \quad ∎$$

A General Solution of (2) Includes All Solutions

Let us first show that general solutions always exist. Indeed, Theorem 3 in Sec. 2.6 extends as follows.

THEOREM 4

Existence of a General Solution

If the coefficients $p_0(x), \cdots, p_{n-1}(x)$ of (2) are continuous on some open interval I, then (2) has a general solution on I.

PROOF We choose any fixed x_0 in I. By Theorem 2 the ODE (2) has n solutions $y_1, \cdots, y_n$, where y_j satisfies initial conditions (5) with $K_{j-1} = 1$ and all other K's equal to zero. Their Wronskian at x_0 equals 1. For instance, when $n = 3$, then $y_1(x_0) = 1$, $y_2'(x_0) = 1$, $y_3''(x_0) = 1$, and the other initial values are zero. Thus, as claimed,

$$W(y_1(x_0), y_2(x_0), y_3(x_0)) = \begin{vmatrix} y_1(x_0) & y_2(x_0) & y_3(x_0) \\ y_1'(x_0) & y_2'(x_0) & y_3'(x_0) \\ y_1''(x_0) & y_2''(x_0) & y_3''(x_0) \end{vmatrix} = \begin{vmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix} = 1.$$

Hence for any n those solutions $y_1, \cdots, y_n$ are linearly independent on I, by Theorem 3. They form a basis on I, and $y = c_1y_1 + \cdots + c_ny_n$ is a general solution of (2) on I. ▩

We can now prove the basic property that from a general solution of (2) every solution of (2) can be obtained by choosing suitable values of the arbitrary constants. Hence an nth order ***linear*** ODE has no **singular solutions,** that is, solutions that cannot be obtained from a general solution.

THEOREM 5

General Solution Includes All Solutions

If the ODE (2) *has continuous coefficients* $p_0(x), \cdots, p_{n-1}(x)$ *on some open interval* I, *then every solution* $y = Y(x)$ *of* (2) *on* I *is of the form*

$$Y(x) = C_1y_1(x) + \cdots + C_ny_n(x) \tag{9}$$

where $y_1, \cdots, y_n$ *is a basis of solutions of* (2) *on* I *and* $C_1, \cdots, C_n$ *are suitable constants.*

PROOF Let Y be a given solution and $y = c_1y_1 + \cdots + c_ny_n$ a general solution of (2) on I. We choose any fixed x_0 in I and show that we can find constants $c_1, \cdots, c_n$ for which y and its first $n - 1$ derivatives agree with Y and its corresponding derivatives at x_0. That is, we should have at $x = x_0$

$$\begin{aligned} c_1y_1 + \cdots + c_ny_n &= Y \\ c_1y_1' + \cdots + c_ny_n' &= Y' \\ &\vdots \\ c_1y_1^{(n-1)} + \cdots + c_ny_n^{(n-1)} &= Y^{(n-1)}. \end{aligned} \tag{10}$$

But this is a linear system of equations in the unknowns $c_1, \cdots, c_n$. Its coefficient determinant is the Wronskian W of $y_1, \cdots, y_n$ at x_0. Since $y_1, \cdots, y_n$ form a basis, they are linearly independent, so that W is not zero by Theorem 3. Hence (10) has a unique solution $c_1 = C_1, \cdots, c_n = C_n$ (by Cramer's theorem in Sec. 7.7). With these values we obtain the particular solution

$$y^*(x) = C_1y_1(x) + \cdots + C_ny_n(x)$$

on I. Equation (10) shows that y^* and its first $n - 1$ derivatives agree at x_0 with Y and its corresponding derivatives. That is, y^* and Y satisfy at x_0 the same initial conditions.

The uniqueness theorem (Theorem 2) now implies that $y^* \equiv Y$ on I. This proves the theorem. ∎

This completes our theory of the homogeneous linear ODE (2). Note that for $n = 2$ it is identical with that in Sec. 2.6. This had to be expected.

PROBLEM SET 3.1

1–5 TYPICAL EXAMPLES OF BASES

To get a feel for higher order ODEs, show that the given functions are solutions and form a basis on any interval. Use Wronskians. (In Prob. 2, $x > 0$.)

1. $1, x, x^2, x^3, \quad y^{iv} = 0$
2. $1, x^2, x^4, \quad x^2y''' - 3xy'' + 3y' = 0$
3. $e^x, xe^x, x^2e^x, \quad y''' - 3y'' + 3y' - y = 0$
4. $e^{2x}\cos x, e^{2x}\sin x, e^{-2x}\cos x, e^{-2x}\sin x, \quad y^{iv} - 6y'' + 25y = 0.$
5. $1, x, \cos 3x, \sin 3x, \quad y^{iv} + 9y'' = 0$

6. **TEAM PROJECT. General Properties of Solutions of Linear ODEs.** These properties are important in obtaining new solutions from given ones. Therefore extend Team Project 34 in Sec. 2.2 to nth-order ODEs. Explore statements on sums and multiples of solutions of (1) and (2) systematically and with proofs. Recognize clearly that no new ideas are needed in this extension from $n = 2$ to general n.

7–19 LINEAR INDEPENDENCE AND DEPENDENCE

Are the given functions linearly independent or dependent on the positive x-axis? (Give a reason.)

7. $1, e^x, e^{-x}$
8. $x + 1, x + 2, x$
9. $\ln x, \ln x^2, (\ln x)^2$
10. $e^x, e^{-x}, \sinh 2x$
11. $x^2, x|x|, x$
12. $x, 1/x, 0$
13. $\sin 2x, \sin x, \cos x$
14. $\cos^2 x, \sin^2 x, \cos 2x$
15. $\tan x, \cot x, 1$
16. $(x - 1)^2, (x + 1)^2, x$
17. $\sin x, \sin \frac{1}{2}x$
18. $\cosh x, \sinh x, \cosh^2 x$
19. $\cos^2 x, \sin^2 x, 2\pi$

20. **TEAM PROJECT. Linear Independence and Dependence. (a)** Investigate the given question about a set S of functions on an interval I. Give an example. Prove your answer.

(1) If S contains the zero function, can S be linearly independent?

(2) If S is linearly independent on a subinterval J of I, is it linearly independent on I?

(3) If S is linearly dependent on a subinterval J of I, is it linearly dependent on I?

(4) If S is linearly independent on I, is it linearly independent on a subinterval J?

(5) If S is linearly dependent on I, is it linearly independent on a subinterval J?

(6) If S is linearly dependent on I, and if T contains S, is T linearly dependent on I?

(b) In what cases can you use the Wronskian for testing linear independence? By what other means can you perform such a test?

3.2 Homogeneous Linear ODEs with Constant Coefficients

In this section we consider nth-order homogeneous linear ODEs with constant coefficients, which we write in the form

$$y^{(n)} + a_{n-1}y^{(n-1)} + \cdots + a_1y' + a_0y = 0 \tag{1}$$

where $y^{(n)} = d^ny/dx^n$, etc. We shall see that this extends the case $n = 2$ discussed in Sec. 2.2. Substituting $y = e^{\lambda x}$ (as in Sec. 2.2), we obtain the characteristic equation

$$\lambda^n + a_{n-1}\lambda^{(n-1)} + \cdots + a_1\lambda + a_0 = 0 \tag{2}$$

of (1). If λ is a root of (2), then $y = e^{\lambda x}$ is a solution of (1). To find these roots, you may need a numeric method, such as Newton's in Sec. 19.2, also available on the usual CASs. For general n there are more cases than for $n = 2$. We shall discuss all of them and illustrate them with typical examples.

Distinct Real Roots

If all the n roots $\lambda_1, \cdots, \lambda_n$ of (2) are real and different, then the n solutions

$$y_1 = e^{\lambda_1 x}, \quad \cdots, \quad y_n = e^{\lambda_n x} \tag{3}$$

constitute a basis for all x. The corresponding general solution of (1) is

$$y = c_1 e^{\lambda_1 x} + \cdots + c_n e^{\lambda_n x}. \tag{4}$$

Indeed, the solutions in (3) are linearly independent, as we shall see after the example.

EXAMPLE 1 **Distinct Real Roots**

Solve the ODE $y''' - 2y'' - y' + 2y = 0$.

Solution. The characteristic equation is $\lambda^3 - 2\lambda^2 - \lambda + 2 = 0$. It has the roots $-1, 1, 2$; if you find one of them by inspection, you can obtain the other two roots by solving a quadratic equation (explain!). The corresponding general solution (4) is $y = c_1 e^{-x} + c_2 e^{x} + c_3 e^{2x}$. ■

Linear Independence of (3). Students familiar with nth-order determinants may verify that by pulling out all exponential functions from the columns and denoting their product by E, thus $E = \exp[(\lambda_1 + \cdots + \lambda_n)x]$, the Wronskian of the solutions in (3) becomes

$$W = \begin{vmatrix} e^{\lambda_1 x} & e^{\lambda_2 x} & \cdots & e^{\lambda_n x} \\ \lambda_1 e^{\lambda_1 x} & \lambda_2 e^{\lambda_2 x} & \cdots & \lambda_n e^{\lambda_n x} \\ \lambda_1^2 e^{\lambda_1 x} & \lambda_2^2 e^{\lambda_2 x} & \cdots & \lambda_n^2 e^{\lambda_n x} \\ \cdot & \cdot & \cdots & \cdot \\ \lambda_1^{n-1} e^{\lambda_1 x} & \lambda_2^{n-1} e^{\lambda_2 x} & \cdots & \lambda_n^{n-1} e^{\lambda_n x} \end{vmatrix}$$

$$= E \begin{vmatrix} 1 & 1 & \cdots & 1 \\ \lambda_1 & \lambda_2 & \cdots & \lambda_n \\ \lambda_1^2 & \lambda_2^2 & \cdots & \lambda_n^2 \\ \cdot & \cdot & \cdots & \cdot \\ \lambda_1^{n-1} & \lambda_2^{n-1} & \cdots & \lambda_n^{n-1} \end{vmatrix}. \tag{5}$$

The exponential function E is never zero. Hence $W = 0$ if and only if the determinant on the right is zero. This is a so-called **Vandermonde** or **Cauchy determinant**[1]. It can be shown that it equals

[1]ALEXANDRE THÉOPHILE VANDERMONDE (1735–1796), French mathematician, who worked on solution of equations by determinants. For CAUCHY see footnote 4, in Sec. 2.5.

(6) $$(-1)^{n(n-1)/2}V$$

where V is the product of all factors $\lambda_j - \lambda_k$ with $j < k$ ($\leqq n$); for instance, when $n = 3$ we get $-V = -(\lambda_1 - \lambda_2)(\lambda_1 - \lambda_3)(\lambda_2 - \lambda_3)$. This shows that the Wronskian is not zero if and only if all the n roots of (2) are different and thus gives the following.

THEOREM 1

Basis

Solutions $y_1 = e^{\lambda_1 x}, \cdots, y_n = e^{\lambda_n x}$ of (1) (with any real or complex λ_j's) form a basis of solutions of (1) on any open interval if and only if all n roots of (2) are different.

Actually, Theorem 1 is an important special case of our more general result obtained from (5) and (6):

THEOREM 2

Linear Independence

Any number of solutions of (1) of the form $e^{\lambda x}$ are linearly independent on an open interval I if and only if the corresponding λ are all different.

Simple Complex Roots

If complex roots occur, they must occur in conjugate pairs since the coefficients of (1) are real. Thus, if $\lambda = \gamma + i\omega$ is a simple root of (2), so is the conjugate $\bar{\lambda} = \gamma - i\omega$, and two corresponding linearly independent solutions are (as in Sec. 2.2, except for notation)

$$y_1 = e^{\gamma x} \cos \omega x, \qquad y_2 = e^{\gamma x} \sin \omega x.$$

EXAMPLE 2 **Simple Complex Roots. Initial Value Problem**

Solve the initial value problem

$$y''' - y'' + 100y' - 100y = 0, \qquad y(0) = 4, \qquad y'(0) = 11, \qquad y''(0) = -299.$$

Solution. The characteristic equation is $\lambda^3 - \lambda^2 + 100\lambda - 100 = 0$. It has the root 1, as can perhaps be seen by inspection. Then division by $\lambda - 1$ shows that the other roots are $\pm 10i$. Hence a general solution and its derivatives (obtained by differentiation) are

$$y = c_1 e^x + A \cos 10x + B \sin 10x,$$
$$y' = c_1 e^x - 10A \sin 10x + 10B \cos 10x,$$
$$y'' = c_1 e^x - 100A \cos 10x - 100B \sin 10x.$$

From this and the initial conditions we obtain by setting $x = 0$

$$\text{(a)} \quad c_1 + A = 4, \qquad \text{(b)} \quad c_1 + 10B = 11, \qquad \text{(c)} \quad c_1 - 100A = -299.$$

We solve this system for the unknowns A, B, c_1. Equation (a) minus Equation (c) gives $101A = 303$, $A = 3$. Then $c_1 = 1$ from (a) and $B = 1$ from (b). The solution is (Fig. 72)

$$y = e^x + 3 \cos 10x + \sin 10x.$$

This gives the solution curve, which oscillates about e^x (dashed in Fig. 72 on p. 114). ■

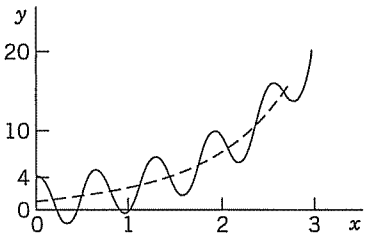


Fig. 72. Solution in Example 2

Multiple Real Roots

If a real double root occurs, say, $\lambda_1 = \lambda_2$, then $y_1 = y_2$ in (3), and we take y_1 and xy_1 as corresponding linearly independent solutions. This is as in Sec. 2.2.

More generally, if λ is a real root of order m, then m corresponding linearly independent solutions are

$$e^{\lambda x},\quad xe^{\lambda x},\quad x^2e^{\lambda x},\quad \cdots,\quad x^{m-1}e^{\lambda x}. \tag{7}$$

We derive these solutions after the next example and indicate how to prove their linear independence.

EXAMPLE 3 Real Double and Triple Roots

Solve the ODE $y^{v} - 3y^{iv} + 3y''' - y'' = 0$.

Solution. The characteristic equation $\lambda^5 - 3\lambda^4 + 3\lambda^3 - \lambda^2 = 0$ has the roots $\lambda_1 = \lambda_2 = 0$ and $\lambda_3 = \lambda_4 = \lambda_5 = 1$, and the answer is

$$y = c_1 + c_2x + (c_3 + c_4x + c_5x^2)e^x. \tag{8}$$

■

***Derivation of* (7).** We write the left side of (1) as

$$L[y] = y^{(n)} + a_{n-1}y^{(n-1)} + \cdots + a_0y.$$

Let $y = e^{\lambda x}$. Then by performing the differentiations we have

$$L[e^{\lambda x}] = (\lambda^n + a_{n-1}\lambda^{n-1} + \cdots + a_0)e^{\lambda x}.$$

Now let λ_1 be a root of mth order of the polynomial on the right, where $m \leqq n$. For $m < n$ let $\lambda_{m+1}, \cdots, \lambda_n$ be the other roots, all different from λ_1. Writing the polynomial in product form, we then have

$$L[e^{\lambda x}] = (\lambda - \lambda_1)^m h(\lambda)e^{\lambda x}$$

with $h(\lambda) = 1$ if $m = n$, and $h(\lambda) = (\lambda - \lambda_{m+1}) \cdots (\lambda - \lambda_n)$ if $m < n$. Now comes the key idea: We differentiate on both sides with respect to λ,

$$\frac{\partial}{\partial\lambda} L[e^{\lambda x}] = m(\lambda - \lambda_1)^{m-1}h(\lambda)e^{\lambda x} + (\lambda - \lambda_1)^m \frac{\partial}{\partial\lambda}[h(\lambda)e^{\lambda x}]. \tag{9}$$

The differentiations with respect to x and λ are independent and the occurring derivatives are continuous, so that we can interchange their order on the left:

$$\frac{\partial}{\partial \lambda} L[e^{\lambda x}] = L\left[\frac{\partial}{\partial \lambda} e^{\lambda x}\right] = L[xe^{\lambda x}]. \tag{10}$$

The right side of (9) is zero for $\lambda = \lambda_1$ because of the factors $\lambda - \lambda_1$ (and $m \geqq 2$ since we have a multiple root!). Hence $L[xe^{\lambda_1 x}] = 0$ by (9) and (10). This proves that $xe^{\lambda_1 x}$ is a solution of (1).

We can repeat this step and produce $x^2 e^{\lambda_1 x}, \cdots, x^{m-1} e^{\lambda_1 x}$ by another $m - 2$ such differentiations with respect to λ. Going one step further would no longer give zero on the right because the lowest power of $\lambda - \lambda_1$ would then be $(\lambda - \lambda_1)^0$, multiplied by $m!h(\lambda)$ and $h(\lambda_1) \neq 0$ because $h(\lambda)$ has no factors $\lambda - \lambda_1$; so we get *precisely* the solutions in (7).

We finally show that the solutions (7) are linearly independent. For a specific n this can be seen by calculating their Wronskian, which turns out to be nonzero. For arbitrary m we can pull out the exponential functions from the Wronskian. This gives $(e^{\lambda x})^m = e^{\lambda m x}$ times a determinant which by "row operations" can be reduced to the Wronskian of $1, x, \cdots, x^{m-1}$. The latter is constant and different from zero (equal to $1!2! \cdots (m-1)!$). These functions are solutions of the ODE $y^{(m)} = 0$, so that linear independence follows from Theroem 3 in Sec. 3.1. ▩

Multiple Complex Roots

In this case, real solutions are obtained as for complex simple roots above. Consequently, if $\lambda = \gamma + i\omega$ is a **complex double root,** so is the conjugate $\bar{\lambda} = \gamma - i\omega$. Corresponding linearly independent solutions are

$$e^{\gamma x} \cos \omega x, \quad e^{\gamma x} \sin \omega x, \quad xe^{\gamma x} \cos \omega x, \quad xe^{\gamma x} \sin \omega x. \tag{11}$$

The first two of these result from $e^{\lambda x}$ and $e^{\bar{\lambda} x}$ as before, and the second two from $xe^{\lambda x}$ and $xe^{\bar{\lambda} x}$ in the same fashion. Obviously, the corresponding general solution is

$$y = e^{\gamma x}[(A_1 + A_2 x) \cos \omega x + (B_1 + B_2 x) \sin \omega x]. \tag{12}$$

For *complex triple roots* (which hardly ever occur in applications), one would obtain two more solutions $x^2 e^{\lambda x} \cos \omega x$, $x^2 e^{\gamma x} \sin \omega x$, and so on.

PROBLEM SET 3.2

1–6 ODE FOR GIVEN BASIS

Find an ODE (1) for which the given functions form a basis of solutions.

1. e^x, e^{2x}, e^{3x}
2. $e^{-2x}, xe^{-2x}, x^2 e^{-2x}$
3. $e^x, e^{-x}, \cos x, \sin x$
4. $\cos x, \sin x, x \cos x, x \sin x$
5. $1, x, \cos 2x, \sin 2x$
6. $e^{-2x}, e^{-x}, e^x, e^{2x}, 1$

7–12 GENERAL SOLUTION

Solve the given ODE. (Show the details of your work.)

7. $y''' + y' = 0$
8. $y^{iv} - 29y'' + 100y = 0$
9. $y''' + y'' - y' - y = 0$
10. $16y^{iv} - 8y'' + y = 0$
11. $y''' - 3y'' - 4y' + 6y = 0$
12. $y^{iv} + 3y'' - 4y = 0$

13–18 INITIAL VALUE PROBLEMS

Solve by a CAS, giving a general solution and the particular solution and its graph.

13. $y^{iv} + 0.45y''' - 0.165y'' + 0.0045y' - 0.00175y = 0$, $y(0) = 17.4$, $y'(0) = -2.82$, $y''(0) = 2.0485$, $y'''(0) = -1.458675$

14. $4y''' + 8y'' + 41y' + 37y = 0$, $y(0) = 9$, $y'(0) = -6.5$, $y''(0) = -39.75$

15. $y''' + 3.2y'' + 4.81y' = 0$, $y(0) = 3.4$, $y'(0) = -4.6$, $y''(0) = 9.91$

16. $y^{iv} + 4y = 0$, $y(0) = \frac{1}{2}$, $y'(0) = -\frac{3}{2}$, $y''(0) = \frac{5}{2}$, $y'''(0) = -\frac{7}{2}$

17. $y^{iv} - 9y'' - 400y = 0$, $y(0) = 0$, $y'(0) = 0$, $y''(0) = 41$, $y'''(0) = 0$

18. $y''' + 7.5y'' + 14.25y' - 9.125y = 0$, $y(0) = 10.05$, $y'(0) = -54.975$, $y''(0) = 257.5125$

19. **CAS PROJECT. Wronskians. Euler–Cauchy Equations of Higher Order.** Although Euler–Cauchy equations have *variable* coefficients (powers of x), we include them here because they fit quite well into the present methods.

(a) Write a program for calculating Wronskians.

(b) Apply the program to some bases of third-order and fourth-order constant-coefficient ODEs. Compare the results with those obtained by the program most likely available for Wronskians in your CAS.

(c) Extend the solution method in Sec. 2.5 to any order n. Solve $x^3y''' + 2x^2y'' - 4xy' + 4y = 0$ and another ODE of your choice. In each case calculate the Wronskian.

20. **PROJECT. Reduction of Order.** This is of practical interest since a single solution of an ODE can often be guessed. For second order, see Example 7 in Sec. 2.1.

(a) How could you reduce the order of a linear constant-coefficient ODE if a solution is known?

(b) Extend the method to a variable-coefficient ODE

$$y''' + p_2(x)y'' + p_1(x)y' + p_0(x)y = 0.$$

Assuming a solution y_1 to be known, show that another solution is $y_2(x) = u(x)y_1(x)$ with $u(x) = \int z(x)\,dx$ and z obtained by solving

$$y_1z'' + (3y_1' + p_2y_1)z' + (3y_1'' + 2p_2y_1' + p_1y_1)z = 0.$$

(c) Reduce

$$x^3y''' - 3x^2y'' + (6 - x^2)xy' - (6 - x^2)y = 0,$$

using $y_1 = x$ (perhaps obtainable by inspection).

21. **CAS EXPERIMENT. Reduction of Order.** Starting with a basis, find third-order ODEs with variable coefficients for which the reduction to second order turns out to be relatively simple.

3.3 Nonhomogeneous Linear ODEs

We now turn from homogeneous to nonhomogeneous linear ODEs of nth order. We write them in standard form

$$y^{(n)} + p_{n-1}(x)y^{(n-1)} + \cdots + p_1(x)y' + p_0(x)y = r(x) \tag{1}$$

with $y^{(n)} = d^ny/dx^n$ as the first term, which is practical, and $r(x) \not\equiv 0$. As for second-order ODEs, a general solution of (1) on an open interval I of the x-axis is of the form

$$y(x) = y_h(x) + y_p(x). \tag{2}$$

Here $y_h(x) = c_1y_1(x) + \cdots + c_ny_n(x)$ is a general solution of the corresponding homogeneous ODE

$$y^{(n)} + p_{n-1}(x)y^{(n-1)} + \cdots + p_1(x)y' + p_0(x)y = 0 \tag{3}$$

on I. Also, y_p is any solution of (1) on I containing no arbitrary constants. If (1) has continuous coefficients and a continuous $r(x)$ on I, then a general solution of (1) exists and includes all solutions. Thus (1) has no singular solutions.

An **initial value problem** for (1) consists of (1) and n **initial conditions**

$$y(x_0) = K_0, \quad y'(x_0) = K_1, \quad \cdots, \quad y^{(n-1)}(x_0) = K_{n-1} \tag{4}$$

with x_0 in I. Under those continuity assumptions it has a unique solution. The ideas of proof are the same as those for $n = 2$ in Sec. 2.7.

Method of Undetermined Coefficients

Equation (2) shows that for solving (1) we have to determine a particular solution of (1). For a constant-coefficient equation

$$y^{(n)} + a_{n-1}y^{(n-1)} + \cdots + a_1y' + a_0y = r(x) \tag{5}$$

($a_0, \cdots, a_{n-1}$ constant) and special $r(x)$ as in Sec. 2.7, such a $y_p(x)$ can be determined by the **method of undetermined coefficients,** as in Sec. 2.7, using the following rules.

(A) Basic Rule as in Sec. 2.7.

(B) Modification Rule. *If a term in your choice for $y_p(x)$ is a solution of the homogeneous equation* (3)*, then multiply $y_p(x)$ by x^k, where k is the smallest positive integer such that no term of $x^k y_p(x)$ is a solution of* (3).

(C) Sum Rule as in Sec. 2.7.

The practical application of the method is the same as that in Sec. 2.7. It suffices to illustrate the typical steps of solving an initial value problem and, in particular, the new Modification Rule, which includes the old Modification Rule as a particular case (with $k = 1$ or 2). We shall see that the technicalities are the same as for $n = 2$, perhaps except for the more involved determination of the constants.

EXAMPLE 1 **Initial Value Problem. Modification Rule**

Solve the initial value problem

$$y''' + 3y'' + 3y' + y = 30e^{-x}, \quad y(0) = 3, \quad y'(0) = -3, \quad y''(0) = -47. \tag{6}$$

Solution. ***Step 1.*** The characteristic equation is $\lambda^3 + 3\lambda^2 + 3\lambda + 1 = (\lambda + 1)^3 = 0$. It has the triple root $\lambda = -1$. Hence a general solution of the homogeneous ODE is

$$\begin{aligned} y_h &= c_1e^{-x} + c_2xe^{-x} + c_3x^2e^{-x} \\ &= (c_1 + c_2x + c_3x^2)e^{-x}. \end{aligned}$$

Step 2. If we try $y_p = Ce^{-x}$, we get $-C + 3C - 3C + C = 30$, which has no solution. Try Cxe^{-x} and Cx^2e^{-x}. The Modification Rule calls for

$$y_p = Cx^3e^{-x}.$$

Then

$$\begin{aligned} y_p' &= C(3x^2 - x^3)e^{-x}, \\ y_p'' &= C(6x - 6x^2 + x^3)e^{-x}, \\ y_p''' &= C(6 - 18x + 9x^2 - x^3)e^{-x}. \end{aligned}$$

Substitution of these expressions into (6) and omission of the common factor e^{-x} gives

$$C(6 - 18x + 9x^2 - x^3) + 3C(6x - 6x^2 + x^3) + 3C(3x^2 - x^3) + Cx^3 = 30.$$

The linear, quadratic, and cubic terms drop out, and $6C = 30$. Hence $C = 5$. This gives $y_p = 5x^3e^{-x}$.

Step 3. We now write down $y = y_h + y_p$, the general solution of the given ODE. From it we find c_1 by the first initial condition. We insert the value, differentiate, and determine c_2 from the second initial condition, insert the value, and finally determine c_3 from $y''(0)$ and the third initial condition:

$$y = y_h + y_p = (c_1 + c_2x + c_3x^2)e^{-x} + 5x^3e^{-x}, \qquad y(0) = c_1 = 3$$

$$y' = [-3 + c_2 + (-c_2 + 2c_3)x + (15 - c_3)x^2 - 5x^3]e^{-x}, \qquad y'(0) = -3 + c_2 = -3, \qquad c_2 = 0$$

$$y'' = [3 + 2c_3 + (30 - 4c_3)x + (-30 + c_3)x^2 + 5x^3]e^{-x}, \qquad y''(0) = 3 + 2c_3 = -47, \qquad c_3 = -25.$$

Hence the *answer* to our problem is (Fig. 73)

$$y = (3 - 25x^2)e^{-x} + 5x^3e^{-x}.$$

The curve of y begins at (0, 3) with a negative slope, as expected from the initial values, and approaches zero as $x \to \infty$. The dashed curve in Fig. 73 is y_p. ▩

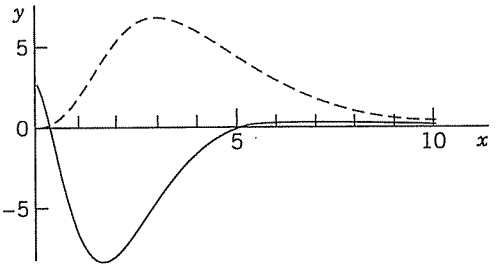


Fig. 73. y and y_p (dashed) in Example 1

Method of Variation of Parameters

The method of variation of parameters (see Sec. 2.10) also extends to arbitrary order n. It gives a particular solution y_p for the nonhomogeneous equation (1) (in standard form with $y^{(n)}$ as the first term!) by the formula

$$(7) \qquad y_p(x) = \sum_{k=1}^{n} y_k(x) \int \frac{W_k(x)}{W(x)}\, r(x)\, dx$$

$$= y_1(x) \int \frac{W_1(x)}{W(x)}\, r(x)\, dx + \cdots + y_n(x) \int \frac{W_n(x)}{W(x)}\, r(x)\, dx$$

on an open interval I on which the coefficients of (1) and $r(x)$ are continuous. In (7) the functions $y_1, \cdots, y_n$ form a basis of the homogeneous ODE (3), with Wronskian W, and W_j ($j = 1, \cdots, n$) is obtained from W by replacing the jth column of W by the column $[0 \quad 0 \quad \cdots \quad 0 \quad 1]^T$. Thus, when $n = 2$, this becomes identical with (2) in Sec. 2.10,

$$W = \begin{vmatrix} y_1 & y_2 \\ y_1' & y_2' \end{vmatrix}, \qquad W_1 = \begin{vmatrix} 0 & y_2 \\ 1 & y_2' \end{vmatrix} = -y_2, \qquad W_2 = \begin{vmatrix} y_1 & 0 \\ y_1' & 1 \end{vmatrix} = y_1.$$

The proof of (7) uses an extension of the idea of the proof of (2) in Sec. 2.10 and can be found in Ref [A11] listed in App. 1.

EXAMPLE 2 Variation of Parameters. Nonhomogeneous Euler–Cauchy Equation

Solve the nonhomogeneous Euler–Cauchy equation

$$x^3y''' - 3x^2y'' + 6xy' - 6y = x^4 \ln x \qquad (x > 0).$$

Solution. ***Step 1. General solution of the homogeneous ODE.*** Substitution of $y = x^m$ and the derivatives into the homogeneous ODE and deletion of the factor x^m gives

$$m(m-1)(m-2) - 3m(m-1) + 6m - 6 = 0.$$

The roots are 1, 2, 3 and give as a basis

$$y_1 = x, \qquad y_2 = x^2, \qquad y_3 = x^3.$$

Hence the corresponding general solution of the homogeneous ODE is

$$y_h = c_1x + c_2x^2 + c_3x^3.$$

Step 2. Determinants needed in (7). These are

$$W = \begin{vmatrix} x & x^2 & x^3 \\ 1 & 2x & 3x^2 \\ 0 & 2 & 6x \end{vmatrix} = 2x^3$$

$$W_1 = \begin{vmatrix} 0 & x^2 & x^3 \\ 0 & 2x & 3x^2 \\ 1 & 2 & 6x \end{vmatrix} = x^4$$

$$W_2 = \begin{vmatrix} x & 0 & x^3 \\ 1 & 0 & 3x^2 \\ 0 & 1 & 6x \end{vmatrix} = -2x^3$$

$$W_3 = \begin{vmatrix} x & x^2 & 0 \\ 1 & 2x & 0 \\ 0 & 2 & 1 \end{vmatrix} = x^2.$$

Step 3. Integration. In (7) we also need the right side $r(x)$ of our ODE in standard form, obtained by division of the given equation by the coefficient x^3 of y'''; thus, $r(x) = (x^4 \ln x)/x^3 = x \ln x$. In (7) we have the simple quotients $W_1/W = x/2$, $W_2/W = -1$, $W_3/W = 1/(2x)$. Hence (7) becomes

$$y_p = x\int \frac{x}{2}\, x \ln x\, dx - x^2 \int x \ln x\, dx + x^3 \int \frac{1}{2x}\, x \ln x\, dx$$

$$= \frac{x}{2}\left(\frac{x^3}{3}\ln x - \frac{x^3}{9}\right) - x^2\left(\frac{x^2}{2}\ln x - \frac{x^2}{4}\right) + \frac{x^3}{2}(x \ln x - x).$$

Simplification gives $y_p = \frac{1}{6}x^4(\ln x - \frac{11}{6})$. Hence the answer is

$$y = y_h + y_p = c_1x + c_2x^2 + c_3x^3 + \tfrac{1}{6}x^4(\ln x - \tfrac{11}{6}).$$

Figure 74 shows y_p. Can you explain the shape of this curve? Its behavior near $x = 0$? The occurrence of a minimum? Its rapid increase? Why would the method of undetermined coefficients not have given the solution? ■

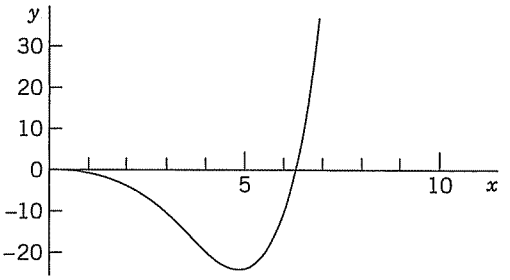


Fig. 74. Particular solution y_p of the nonhomogeneous Euler–Cauchy equation in Example 2

Application: Elastic Beams

Whereas second-order ODEs have various applications, some of the more important ones we have seen, higher order ODEs occur much more rarely in engineering work. An important fourth-order ODE governs the bending of elastic beams, such as wooden or iron girders in a building or a bridge.

Vibrations of beams will be considered in Sec. 12.3.

EXAMPLE 3 **Bending of an Elastic Beam under a Load**

We consider a beam B of length L and constant (e.g., **rectangular**) cross section and homogeneous elastic material (e.g., steel); see Fig. 75. We assume that under its own weight the beam is bent so little that it is practically straight. If we apply a load to B in a vertical plane through the axis of symmetry (the x-axis in Fig. 75), B is bent. Its axis is curved into the so-called **elastic curve** C (or **deflection curve**). It is shown in elasticity theory that the bending moment $M(x)$ is proportional to the curvature $k(x)$ of C. We assume the bending to be small, so that the deflection $y(x)$ and its derivative $y'(x)$ (determining the tangent direction of C) are small. Then, by calculus, $k = y''/(1 + y'^2)^{3/2} \approx y''$. Hence

$$M(x) = EIy''(x).$$

EI is the constant of proportionality. E is *Young's modulus of elasticity* of the material of the beam. I is the moment of inertia of the cross section about the (horizontal) z-axis in Fig. 75.

Elasticity theory shows further that $M''(x) = f(x)$, where $f(x)$ is the load per unit length. Together,

$$EIy^{\text{iv}} = f(x). \tag{8}$$

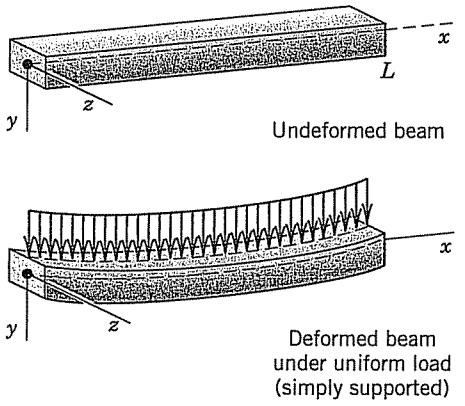


Fig. 75. Elastic Beam

The practically most important supports and corresponding boundary conditions are as follows (see Fig. 76).

(A)	Simply supported	$y = y'' = 0$ at $x = 0$ and L
(B)	Clamped at both ends	$y = y' = 0$ at $x = 0$ and L
(C)	Clamped at $x = 0$, free at $x = L$	$y(0) = y'(0) = 0, y''(L) = y'''(L) = 0.$

The boundary condition $y = 0$ means no displacement at that point, $y' = 0$ means a horizontal tangent, $y'' = 0$ means no bending moment, and $y''' = 0$ means no shear force.

Let us apply this to the uniformly loaded simply supported beam in Fig. 75. The load is $f(x) \equiv f_0 = const.$ Then (8) is

$$y^{\mathrm{iv}} = k, \qquad k = \frac{f_0}{EI}. \tag{9}$$

This can be solved simply by calculus. Two integrations give

$$y'' = \frac{k}{2}x^2 + c_1x + c_2.$$

$y''(0) = 0$ gives $c_2 = 0$. Then $y''(L) = L(\frac{1}{2}kL + c_1) = 0$, $c_1 = -kL/2$ (since $L \neq 0$). Hence

$$y'' = \frac{k}{2}(x^2 - Lx).$$

Integrating this twice, we obtain

$$y = \frac{k}{2}\left(\frac{1}{12}x^4 - \frac{L}{6}x^3 + c_3x + c_4\right)$$

with $c_4 = 0$ from $y(0) = 0$. Then

$$y(L) = \frac{kL}{2}\left(\frac{L^3}{12} - \frac{L^3}{6} + c_3\right) = 0, \qquad c_3 = \frac{L^3}{12}.$$

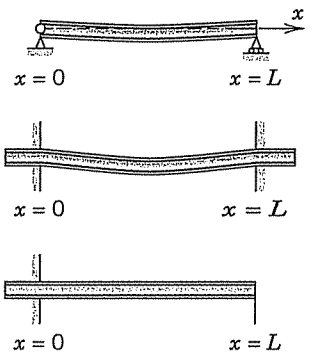
Inserting the expression for k, we obtain as our solution

$$y = \frac{f_0}{24EI}(x^4 - 2Lx^3 + L^3x).$$

Since the boundary conditions at both ends are the same, we expect the deflection $y(x)$ to be "symmetric" with respect to $L/2$, that is, $y(x) = y(L - x)$. Verify this directly or set $x = u + L/2$ and show that y becomes an even function of u,

$$y = \frac{f_0}{24EI}\left(u^2 - \frac{1}{4}L^2\right)\left(u^2 - \frac{5}{4}L^2\right).$$

From this we can see that the maximum deflection in the middle at $u = 0$ ($x = L/2$) is $5f_0L^4/(16 \cdot 24EI)$. Recall that the positive direction points downward. ■



(A) Simply supported

(B) Clamped at both ends

(C) Clamped at the left end, free at the right end

Fig. 76. Supports of a Beam

PROBLEM SET 3.5

1–8 GENERAL SOLUTION

Solve the following ODEs. (Show the details of your work.)

1. $y''' - 2y'' - 4y' + 8y = e^{-3x} + 8x^2$
2. $y''' + 3y'' - 5y' - 39y = 30 \cos x$
3. $y^{iv} + 0.5y'' + 0.0625y = e^{-x} \cos 0.5x$
4. $y''' + 2y'' - 5y' - 6y = 100e^{-3x} + 18e^{-x}$
5. $x^3y''' + 0.75xy' - 0.75y = 9x^{5.5}$
6. $(xD^3 + 4D^2)y = 8e^x$
7. $(D^4 + 10D^2 + 9I)y = 13 \cosh 2x$
8. $(D^3 - 2D^2 - 9D + 18I)y = e^{2x}$

9–14 INITIAL VALUE PROBLEMS

Solve the following initial value problems. (Show the details.)

9. $y''' - 9y'' + 27y' - 27y = 54 \sin 3x$, $y(0) = 3.5$, $y'(0) = 13.5$, $y''(0) = 38.5$
10. $y^{iv} - 16y = 128 \cosh 2x$, $y(0) = 1$, $y'(0) = 24$, $y''(0) = 20$, $y'''(0) = -160$
11. $(x^3D^3 - x^2D^2 - 7xD + 16I)y = 9x \ln x$, $y(1) = 6$, $Dy(1) = 18$, $D^2y(1) = 65$
12. $(D^4 - 26D^2 + 25I)y = 50(x + 1)^2$, $y(0) = 12.16$, $Dy(0) = -6$, $D^2y(0) = 34$, $D^3y(0) = -130$
13. $(D^3 + 4D^2 + 85D)y = 135xe^x$, $y(0) = 10.4$, $Dy(0) = -18.1$, $D^2y(0) = -691.6$
14. $(2D^3 - D^2 - 8D + 4I)y = \sin x$, $y(0) = 1$, $Dy(0) = 0$, $D^2y(0) = 0$

15. **WRITING PROJECT. Comparison of Methods.** Write a report on the method of undetermined coefficients and the method of variation of parameters, discussing and comparing the advantages and disadvantages of each method. Illustrate your findings with typical examples. Try to show that the method of undetermined coefficients, say, for a third-order ODE with constant coefficients and an exponential function on the right, can be derived from the method of variation of parameters.

16. **CAS EXPERIMENT. Undetermined Coefficients.** Since variation of parameters is generally complicated, it seems worthwhile to try to extend the other method. Find out experimentally for what ODEs this is possible and for what not. *Hint:* Work backward, solving ODEs with a CAS and then looking whether the solution could be obtained by undetermined coefficients. For example, consider

$$y''' - 12y'' + 48y' - 64y = x^{1/2}e^{4x} \quad \text{and}$$

$$x^3y''' + x^2y'' - 6xy' + 6y = x \ln x.$$

CHAPTER 3 REVIEW QUESTIONS AND PROBLEMS

1. What is the superposition or linearity principle? For what nth-order ODEs does it hold?
2. List some other basic theorems that extend from second-order to nth-order ODEs.
3. If you know a general solution of a homogeneous linear ODE, what do you need to obtain from it a general solution of a corresponding nonhomogeneous linear ODE?
4. What is an initial value problem for an nth-order linear ODE?
5. What is the Wronskian? What is it used for?

6–15 GENERAL SOLUTION

Solve the given ODE. (Show the details of your work.)

6. $y''' + 6y'' + 18y' + 40y = 0$
7. $4x^2y''' + 12xy'' + 3y' = 0$
8. $y^{iv} + 10y'' + 9y = 0$
9. $8y''' + 12y'' - 2y' - 3y = 0$
10. $(D^3 + 3D^2 + 3D + I)y = x^2$
11. $(xD^4 + D^3)y = 150x^4$
12. $(D^4 - 2D^3 - 8D^2)y = 16 \cos 2x$
13. $(D^3 + I)y = 9e^{x/2}$
14. $(x^3D^3 - 3x^2D^2 + 6xD - 6I)y = 30x^{-2}$
15. $(D^3 - D^2 - D + I)y = e^x$

16–20 INITIAL VALUE PROBLEMS

Solve the given problem. (Show the details.)

16. $y''' - 2y'' + 4y' - 8y = 0$, $y(0) = -1$, $y'(0) = 30$, $y''(0) = 28$
17. $x^3y''' + 7x^2y'' - 2xy' - 10y = 0$, $y(1) = 1$, $y'(1) = -7$, $y''(1) = 44$
18. $(D^3 + 25D)y = 32 \cos^2 4x$, $y(0) = 0$, $Dy(0) = 0$, $D^2y(0) = 0$
19. $(D^4 + 40D^2 - 441I)y = 8 \cosh x$, $y(0) = 1.98$, $Dy(0) = 3$, $D^2y(0) = -40.02$, $D^3y(0) = 27$
20. $(x^3D^3 + 5x^2D^2 + 2xD - 2I)y = 7x^{3/2}$, $y(1) = 10.6$, $Dy(1) = -3.6$, $D^2y(1) = 31.2$

SUMMARY OF CHAPTER 3
Higher Order Linear ODEs

Compare with the similar Summary of Chap. 2 (the case $n = 2$).
Chapter 3 extends Chap. 2 from order $n = 2$ to arbitrary order n. An **nth-order linear ODE** is an ODE that can be written

$$(1) \qquad y^{(n)} + p_{n-1}(x)y^{(n-1)} + \cdots + p_1(x)y' + p_0(x)y = r(x)$$

with $y^{(n)} = d^n y/dx^n$ as the first term; we again call this the **standard form.** Equation (1) is called **homogeneous** if $r(x) \equiv 0$ on a given open interval I considered, **nonhomogeneous** if $r(x) \not\equiv 0$ on I. For the homogeneous ODE

$$(2) \qquad y^{(n)} + p_{n-1}(x)y^{(n-1)} + \cdots + p_1(x)y' + p_0(x)y = 0$$

the **superposition principle** (Sec. 3.1) holds, just as in the case $n = 2$. A **basis** or **fundamental system** of solutions of (2) on I consists of n linearly independent solutions $y_1, \cdots, y_n$ of (2) on I. A **general solution** of (2) on I is a linear combination of these,

$$(3) \qquad y = c_1y_1 + \cdots + c_ny_n \qquad (c_1, \cdots, c_n \text{ arbitrary constants}).$$

A **general solution** of the nonhomogeneous ODE (1) on I is of the form

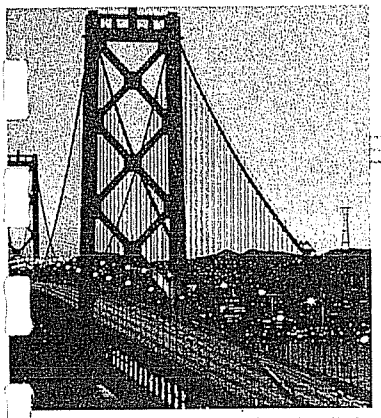
$$(4) \qquad y = y_h + y_p \qquad \text{(Sec. 3.3)}.$$

Here, y_p is a particular solution of (1) and is obtained by two methods (**undetermined coefficients** or **variation of parameters**) explained in Sec. 3.3.

An **initial value problem** for (1) or (2) consists of one of these ODEs and n initial conditions (Secs. 3.1, 3.3)

$$(5) \qquad y(x_0) = K_0, \quad y'(x_0) = K_1, \quad \cdots, \quad y^{(n-1)}(x_0) = K_{n-1}$$

with given x_0 in I and given $K_0, \cdots, K_{n-1}$. If $p_0, \cdots, p_{n-1}, r$ are continuous on I, then general solutions of (1) and (2) on I exist, and initial value problems (1), (5) or (2), (5) have a unique solution.



CHAPTER 4

Systems of ODEs. Phase Plane. Qualitative Methods

Systems of ODEs have various applications (see, for instance, Secs. 4.1 and 4.5). Their theory is outlined in Sec. 4.2 and includes that of a single ODE. The practically important conversion of a single nth-order ODE to a system is shown in Sec. 4.1.

Linear systems (Secs. 4.3, 4.4, 4.6) are best treated by the use of vectors and matrices, of which, however, only a few elementary facts will be needed here, as given in Sec. 4.0 and probably familiar to most students.

Qualitative methods. In addition to actually solving systems (Sec. 4.3, 4.6), which is often difficult or even impossible, we shall explain a totally different method, namely, the powerful method of investigating the general behavior of whole families of solutions in the **phase plane** (Sec. 4.3). This approach to systems of ODEs is called a **qualitative method** because it does not need actual solutions (in contrats to a *"quantitative method"* of actually solving a system).

This *phase plane method,* as it is called, also gives information on **stability** of solutions, which is of general importance in control theory, circuit theory, population dynamics, and so on. Here, *stability of a physical system* means that, roughly speaking, a small change at some instant causes only small changes in the behavior of the system at all later times.

Phase plane methods can be extended to nonlinear systems, for which they are particularly useful. We will show this in Sec. 4.5, which includes a discussion of the pendulum equation and the Lotka-Volterra population model. We finally discuss nonhomogeneous linear systems in Sec. 4.6.

NOTATION. Analogous to Chaps. 1–3, we continue to denote unknown functions by y; thus, $y_1(t)$, $y_2(t)$. This seems preferable to suddenly using x for functions, $x_1(t)$, $x_2(t)$, as is sometimes done in systems of ODEs.

Prerequisite: Chap. 2.
References and Answers to Problems: App. 1 Part A, and App. 2.

4.0 Basics of Matrices and Vectors

In discussing *linear* systems of ODEs we shall use matrices and vectors. This simplifies formulas and clarifies ideas. But we shall need only a few elementary facts (by no means the bulk of material in Chaps. 7 and 8). These facts will very likely be at the disposal of most students. Hence ***this section is for reference only. Begin with Sec.* 4.1 *and consult* 4.0 *as needed.***

Most of our linear systems will consist of two ODEs in two unknown functions $y_1(t)$, $y_2(t)$,

$$(1)\qquad \begin{aligned} y_1' &= a_{11}y_1 + a_{12}y_2, \\ y_2' &= a_{21}y_1 + a_{22}y_2, \end{aligned} \qquad \text{for example,} \qquad \begin{aligned} y_1' &= -5y_1 + 2y_2 \\ y_2' &= 13y_1 + \tfrac{1}{2}y_2 \end{aligned}$$

(perhaps with additional *given* functions $g_1(t)$, $g_2(t)$ in the two ODEs on the right).

Similarly, a linear system of n first-order ODEs in n unknown functions $y_1(t), \cdots, y_n(t)$ is of the form

$$(2)\qquad \begin{aligned} y_1' &= a_{11}y_1 + a_{12}y_2 + \cdots + a_{1n}y_n \\ y_2' &= a_{21}y_1 + a_{22}y_2 + \cdots + a_{2n}y_n \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ y_n' &= a_{n1}y_1 + a_{n2}y_2 + \cdots + a_{nn}y_n \end{aligned}$$

(perhaps with an additional given function in each ODE on the right).

Some Definitions and Terms

Matrices. In (1) the (constant or variable) coefficients form a **2 × 2 matrix A,** that is, an array

$$(3)\qquad \mathbf{A} = [a_{jk}] = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \qquad \text{for example,} \qquad \mathbf{A} = \begin{bmatrix} -5 & 2 \\ 13 & \frac{1}{2} \end{bmatrix}.$$

Similarly, the coefficients in (2) form an ***n* × *n* matrix**

$$(4)\qquad \mathbf{A} = [a_{jk}] = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}.$$

The a_{11}, a_{12}, $\cdots$ are called **entries,** the horizontal lines **rows,** and the vertical lines **columns.** Thus, in (3) the first row is $[a_{11} \quad a_{12}]$, the second row is $[a_{21} \quad a_{22}]$, and the first and second columns are

$$\begin{bmatrix} a_{11} \\ a_{21} \end{bmatrix} \qquad \text{and} \qquad \begin{bmatrix} a_{12} \\ a_{22} \end{bmatrix}.$$

In the ***"double subscript notation"*** for entries, the first subscript denotes the *row* and the second the *column* in which the entry stands. Similarly in (4). The **main diagonal** is the diagonal $a_{11} \quad a_{22} \quad \cdots \quad a_{nn}$ in (4), hence $a_{11} \quad a_{22}$ in (3).

We shall need only **square matrices,** that is, matrices with the same number of rows and columns, as in (3) and (4).

Vectors. A **column vector x** with n **components** $x_1, \cdots, x_n$ is of the form

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \qquad \text{thus if } n = 2, \qquad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

Similarly, a **row vector v** is of the form

$$\mathbf{v} = [v_1 \quad \cdots \quad v_n], \qquad \text{thus if } n = 2, \text{ then} \qquad \mathbf{v} = [v_1, \quad v_2].$$

Calculations with Matrices and Vectors

Equality. Two $n \times n$ matrices are *equal* if and only if corresponding entries are equal. Thus for $n = 2$, let

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \qquad \text{and} \qquad \mathbf{B} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}.$$

Then $\mathbf{A} = \mathbf{B}$ if and only if

$$a_{11} = b_{11}, \qquad a_{12} = b_{12}$$
$$a_{21} = b_{21}, \qquad a_{22} = b_{22}.$$

Two column vectors (or two row vectors) are *equal* if and only if they both have n components and corresponding components are equal. Thus, let

$$\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \quad \text{and} \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}. \qquad \text{Then} \qquad \mathbf{v} = \mathbf{x} \quad \text{if and only if} \quad \begin{matrix} v_1 = x_1 \\ v_2 = x_2. \end{matrix}$$

Addition is performed by adding corresponding entries (or components); here, matrices must both be $n \times n$, and vectors must both have the same number of components. Thus for $n = 2$,

$$\text{(5)} \qquad \mathbf{A} + \mathbf{B} = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} \\ a_{21} + b_{21} & a_{22} + b_{22} \end{bmatrix}, \qquad \mathbf{v} + \mathbf{x} = \begin{bmatrix} v_1 + x_1 \\ v_2 + x_2 \end{bmatrix}.$$

Scalar multiplication (multiplication by a number c) is performed by multiplying each entry (or component) by c. For example, if

$$\mathbf{A} = \begin{bmatrix} 9 & 3 \\ -2 & 0 \end{bmatrix}, \qquad \text{then} \qquad -7\mathbf{A} = \begin{bmatrix} -63 & -21 \\ 14 & 0 \end{bmatrix}.$$

If

$$\mathbf{v} = \begin{bmatrix} 0.4 \\ -13 \end{bmatrix}, \qquad \text{then} \qquad 10\mathbf{v} = \begin{bmatrix} 4 \\ -130 \end{bmatrix}.$$

Matrix Multiplication. The product $\mathbf{C} = \mathbf{AB}$ (in this order) of two $n \times n$ matrices $\mathbf{A} = [a_{jk}]$ and $\mathbf{B} = [b_{jk}]$ is the $n \times n$ matrix $\mathbf{C} = [c_{jk}]$ with entries

$$(6) \qquad c_{jk} = \sum_{m=1}^{n} a_{jm} b_{mk} \qquad \begin{matrix} j = 1, \cdots, n \\ k = 1, \cdots, n, \end{matrix}$$

that is, multiply each entry in the *j*th *row* of **A** by the corresponding entry in the *k*th *column* of **B** and then add these n products. One says briefly that this is a "multiplication of rows into columns." For example,

$$\begin{bmatrix} 9 & 3 \\ -2 & 0 \end{bmatrix} \begin{bmatrix} 1 & -4 \\ 2 & 5 \end{bmatrix} = \begin{bmatrix} 9 \cdot 1 + 3 \cdot 2 & 9 \cdot (-4) + 3 \cdot 5 \\ -2 \cdot 1 + 0 \cdot 2 & (-2) \cdot (-4) + 0 \cdot 5 \end{bmatrix} = \begin{bmatrix} 15 & -21 \\ -2 & 8 \end{bmatrix}.$$

CAUTION! Matrix multiplication is ***not commutative,*** $\mathbf{AB} \neq \mathbf{BA}$ in general. In our example,

$$\begin{bmatrix} 1 & -4 \\ 2 & 5 \end{bmatrix} \begin{bmatrix} 9 & 3 \\ -2 & 0 \end{bmatrix} = \begin{bmatrix} 1 \cdot 9 + (-4) \cdot (-2) & 1 \cdot 3 + (-4) \cdot 0 \\ 2 \cdot 9 + 5 \cdot (-2) & 2 \cdot 3 + 5 \cdot 0 \end{bmatrix} = \begin{bmatrix} 17 & 3 \\ 8 & 6 \end{bmatrix}.$$

Multiplication of an $n \times n$ matrix **A** by a vector **x** with n components is defined by the same rule: $\mathbf{v} = \mathbf{Ax}$ is the vector with the n components

$$v_j = \sum_{m=1}^{n} a_{jm} x_m \qquad j = 1, \cdots, n.$$

For example,

$$\begin{bmatrix} 12 & 7 \\ -8 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 12x_1 + 7x_2 \\ -8x_1 + 3x_2 \end{bmatrix}.$$

Systems of ODEs as Vector Equations

Differentiation. The *derivative* of a matrix (or vector) with variable entries (or components) is obtained by differentiating each entry (or component). Thus, if

$$\mathbf{y}(t) = \begin{bmatrix} y_1(t) \\ y_2(t) \end{bmatrix} = \begin{bmatrix} e^{-2t} \\ \sin t \end{bmatrix}, \quad \text{then} \quad \mathbf{y}'(t) = \begin{bmatrix} y_1'(t) \\ y_2'(t) \end{bmatrix} = \begin{bmatrix} -2e^{-2t} \\ \cos t \end{bmatrix}.$$

Using matrix multiplication and differentiation, we can now write (1) as

$$(7) \quad \mathbf{y}' = \begin{bmatrix} y_1' \\ y_2' \end{bmatrix} = \mathbf{Ay} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \quad \text{e.g.,} \quad \mathbf{y}' = \begin{bmatrix} -5 & 2 \\ 13 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}.$$

Similarly for (2) by means of an $n \times n$ matrix $\mathbf{A}$ and a column vector $\mathbf{y}$ with n components, namely, $\mathbf{y}' = \mathbf{Ay}$. The vector equation (7) is equivalent to two equations for the components, and these are precisely the two ODEs in (1).

Some Further Operations and Terms

Transposition is the operation of writing columns as rows and conversely and is indicated by T. Thus the transpose $\mathbf{A}^T$ of the 2×2 matrix

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} -5 & 2 \\ 13 & \frac{1}{2} \end{bmatrix} \quad \text{is} \quad \mathbf{A}^T = \begin{bmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{bmatrix} = \begin{bmatrix} -5 & 13 \\ 2 & \frac{1}{2} \end{bmatrix}.$$

The transpose of a column vector, say,

$$\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}, \qquad \text{is a row vector,} \qquad \mathbf{v}^T = [v_1 \quad v_2],$$

and conversely.

Inverse of a Matrix. The $n \times n$ **unit matrix I** is the $n \times n$ matrix with main diagonal $1, 1, \cdots, 1$ and all other entries zero. If for a given $n \times n$ matrix $\mathbf{A}$ there is an $n \times n$ matrix $\mathbf{B}$ such that $\mathbf{AB} = \mathbf{BA} = \mathbf{I}$, then $\mathbf{A}$ is called **nonsingular** and $\mathbf{B}$ is called the **inverse** of $\mathbf{A}$ and is denoted by $\mathbf{A}^{-1}$; thus

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}. \tag{8}$$

If $\mathbf{A}$ has no inverse, it is called **singular.** For $n = 2$,

$$\mathbf{A}^{-1} = \frac{1}{\det \mathbf{A}} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}, \tag{9}$$

where the **determinant** of $\mathbf{A}$ is

$$\det \mathbf{A} = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}. \tag{10}$$

(For general n, see Sec. 7.7, but this will not be needed in this chapter.)

Linear Independence. r given vectors $\mathbf{v}^{(1)}, \cdots, \mathbf{v}^{(r)}$ with n components are called a *linearly independent set* or, more briefly, **linearly independent,** if

$$c_1\mathbf{v}^{(1)} + \cdots + c_r\mathbf{v}^{(r)} = \mathbf{0} \tag{11}$$

implies that all scalars $c_1, \cdots, c_r$ must be zero; here, $\mathbf{0}$ denotes the **zero vector,** whose n components are all zero. If (11) also holds for scalars not all zero (so that at least one of these scalars is not zero), then these vectors are called a *linearly dependent set* or, briefly, **linearly dependent,** because then at least one of them can be expressed as

a **linear combination** of the others; that is, if, for instance, $c_1 \neq 0$ in (11), then we can obtain

$$\mathbf{v}^{(1)} = -\frac{1}{c_1}(c_2\mathbf{v}^{(2)} + \cdots + c_r\mathbf{v}^{(r)}).$$

Eigenvalues, Eigenvectors

Eigenvalues and eigenvectors will be very important in this chapter (and, as a matter of fact, throughout mathematics).

Let $\mathbf{A} = [a_{jk}]$ be an $n \times n$ matrix. Consider the equation

$$\mathbf{Ax} = \lambda\mathbf{x} \tag{12}$$

where λ is a scalar (a real or complex number) to be determined and $\mathbf{x}$ is a vector to be determined. Now for every λ a solution is $\mathbf{x} = \mathbf{0}$. A scalar λ such that (12) holds for some vector $\mathbf{x} \neq \mathbf{0}$ is called an **eigenvalue** of $\mathbf{A}$, and this vector is called an **eigenvector** of $\mathbf{A}$ corresponding to this eigenvalue λ.

We can write (12) as $\mathbf{Ax} - \lambda\mathbf{x} = \mathbf{0}$ or

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}. \tag{13}$$

These are n linear algebraic equations in the n unknowns $x_1, \cdots, x_n$ (the components of $\mathbf{x}$). For these equations to have a solution $\mathbf{x} \neq \mathbf{0}$, the determinant of the coefficient matrix $\mathbf{A} - \lambda\mathbf{I}$ must be zero. This is proved as a basic fact in linear algebra (Theorem 4 in Sec. 7.7). In this chapter we need this only for $n = 2$. Then (13) is

$$\begin{bmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}; \tag{14}$$

in components,

$$\begin{aligned} (a_{11} - \lambda)x_1 + \quad a_{12}x_2 \quad &= 0 \\ a_{21}x_1 \quad + (a_{22} - \lambda)x_2 &= 0. \end{aligned} \tag{14*}$$

Now $\mathbf{A} - \lambda\mathbf{I}$ is singular if and only if its determinant $\det(\mathbf{A} - \lambda\mathbf{I})$, called the **characteristic determinant** of $\mathbf{A}$ (also for general n), is zero. This gives

$$\begin{aligned} \det(\mathbf{A} - \lambda\mathbf{I}) &= \begin{vmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{vmatrix} \\ &= (a_{11} - \lambda)(a_{22} - \lambda) - a_{12}a_{21} \\ &= \lambda^2 - (a_{11} + a_{22})\lambda + a_{11}a_{22} - a_{12}a_{21} = 0. \end{aligned} \tag{15}$$

This quadratic equation in λ is called the **characteristic equation** of $\mathbf{A}$. Its solutions are the eigenvalues λ_1 and λ_2 of $\mathbf{A}$. First determine these. Then use (14*) with $\lambda = \lambda_1$ to determine an eigenvector $\mathbf{x}^{(1)}$ of $\mathbf{A}$ corresponding to λ_1. Finally use (14*) with $\lambda = \lambda_2$ to find an eigenvector $\mathbf{x}^{(2)}$ of $\mathbf{A}$ corresponding to λ_2. Note that if $\mathbf{x}$ is an eigenvector of $\mathbf{A}$, so is $k\mathbf{x}$ for any $k \neq 0$.

EXAMPLE 1 Eigenvalue Problem

Find the eigenvalues and eigenvectors of the matrix

$$\mathbf{A} = \begin{bmatrix} -4.0 & 4.0 \\ -1.6 & 1.2 \end{bmatrix}. \tag{16}$$

Solution. The characteristic equation is the quadratic equation

$$\det [\mathbf{A} - \lambda \mathbf{I}] = \begin{vmatrix} -4 - \lambda & 4 \\ -1.6 & 1.2 - \lambda \end{vmatrix} = \lambda^2 + 2.8\lambda + 1.6 = 0.$$

It has the solutions $\lambda_1 = -2$ and $\lambda_2 = -0.8$. These are the eigenvalues of $\mathbf{A}$.

Eigenvectors are obtained from (14*). For $\lambda = \lambda_1 = -2$ we have from (14*)

$$\begin{aligned} (-4.0 + 2.0)x_1 + \quad 4.0x_2 \quad &= 0 \\ -1.6x_1 \quad + (1.2 + 2.0)x_2 &= 0. \end{aligned}$$

A solution of the first equation is $x_1 = 2$, $x_2 = 1$. This also satisfies the second equation. (Why?). Hence an eigenvector of $\mathbf{A}$ corresponding to $\lambda_1 = -2.0$ is

$$\mathbf{x}^{(1)} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}. \qquad \text{Similarly,} \qquad \mathbf{x}^{(2)} = \begin{bmatrix} 1 \\ 0.8 \end{bmatrix} \tag{17}$$

is an eigenvector of $\mathbf{A}$ corresponding to $\lambda_2 = -0.8$, as obtained from (14*) with $\lambda = \lambda_2$. Verify this. ▩

4.1 Systems of ODEs as Models

We first illustrate with a few typical examples that systems of ODEs can serve as models in various applications. We further show that a higher order ODE (with the highest derivative standing alone on one side) can be reduced to a first-order system. Both facts account for the practical importance of these systems.

EXAMPLE 1 Mixing Problem Involving Two Tanks

A mixing problem involving a single tank is modeled by a single ODE, and you may first review the corresponding Example 3 in Sec. 1.3 because the principle of modeling will be the same for two tanks. The model will be a system of two first-order ODEs.

Tank T_1 and T_2 in Fig. 77 contain initially 100 gal of water each. In T_1 the water is pure, whereas 150 lb of fertilizer are dissolved in T_2. By circulating liquid at a rate of 2 gal/min and stirring (to keep the mixture uniform) the amounts of fertilizer $y_1(t)$ in T_1 and $y_2(t)$ in T_2 change with time t. How long should we let the liquid circulate so that T_1 will contain at least half as much fertilizer as there will be left in T_2?

Solution. *Step 1.* **Setting up the model.** As for a single tank, the time rate of change $y_1'(t)$ of $y_1(t)$ equals inflow minus outflow. Similarly for tank T_2. From Fig. 77 we see that

$$y_1' = \text{Inflow/min} - \text{Outflow/min} = \frac{2}{100}y_2 - \frac{2}{100}y_1 \qquad \text{(Tank } T_1\text{)}$$

$$y_2' = \text{Inflow/min} - \text{Outflow/min} = \frac{2}{100}y_1 - \frac{2}{100}y_2 \qquad \text{(Tank } T_2\text{).}$$

Hence the mathematical model of our mixture problem is the system of first-order ODEs

$$\begin{aligned} y_1' &= -0.02y_1 + 0.02y_2 \qquad \text{(Tank } T_1\text{)} \\ y_2' &= \quad 0.02y_1 - 0.02y_2 \qquad \text{(Tank } T_2\text{).} \end{aligned}$$

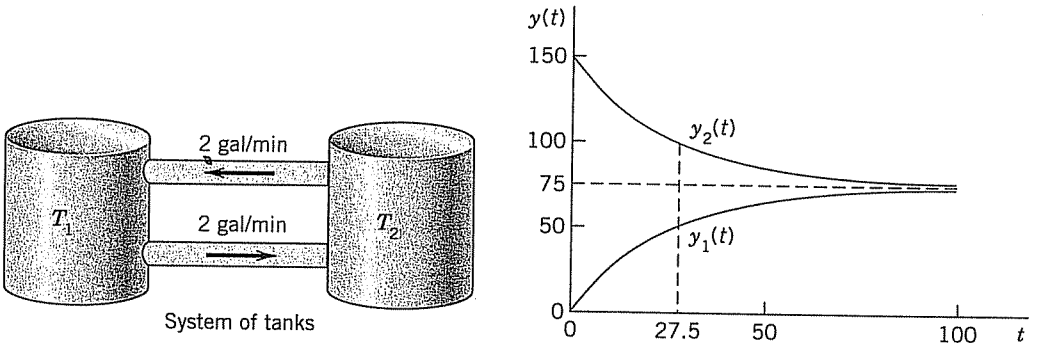


Fig. 77. Fertilizer content in Tanks T_1 (lower curve) and T_2

As a vector equation with column vector $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$ and matrix $\mathbf{A}$ this becomes

$$\mathbf{y}' = \mathbf{A}\mathbf{y}, \qquad \text{where} \qquad \mathbf{A} = \begin{bmatrix} -0.02 & 0.02 \\ 0.02 & -0.02 \end{bmatrix}.$$

Step 2. **General solution.** As for a single equation, we try an exponential function of t,

(1) $$\mathbf{y} = \mathbf{x}e^{\lambda t}. \qquad \text{Then} \qquad \mathbf{y}' = \lambda \mathbf{x}e^{\lambda t} = \mathbf{A}\mathbf{x}e^{\lambda t}.$$

Dividing the last equation $\lambda \mathbf{x}e^{\lambda t} = \mathbf{A}\mathbf{x}e^{\lambda t}$ by $e^{\lambda t}$ and interchanging the left and right sides, we obtain

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

We need nontrivial solutions (solutions that are not identically zero). Hence we have to look for eigenvalues and eigenvectors of $\mathbf{A}$. The eigenvalues are the solutions of the characteristic equation

(2) $$\det (\mathbf{A} - \lambda \mathbf{I}) = \begin{vmatrix} -0.02 - \lambda & 0.02 \\ 0.02 & -0.02 - \lambda \end{vmatrix} = (-0.02 - \lambda)^2 - 0.02^2 = \lambda(\lambda + 0.04) = 0.$$

We see that $\lambda_1 = 0$ (which can very well happen—don't get mixed up—it is eigen*vectors* that must not be zero) and $\lambda_2 = -0.04$. Eigenvectors are obtained from (14*) in Sec. 4.0 with $\lambda = 0$ and $\lambda = -0.04$. For our present $\mathbf{A}$ this gives [we need only the first equation in (14*)]

$$-0.02x_1 + 0.02x_2 = 0 \qquad \text{and} \qquad (-0.02 + 0.04)x_1 + 0.02x_2 = 0,$$

respectively. Hence $x_1 = x_2$ and $x_1 = -x_2$, respectively, and we can take $x_1 = x_2 = 1$ and $x_1 = -x_2 = 1$. This gives two eigenvectors corresponding to $\lambda_1 = 0$ and $\lambda_2 = -0.04$, respectively, namely,

$$\mathbf{x}^{(1)} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad \text{and} \qquad \mathbf{x}^{(2)} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

From (1) and the superposition principle (which continues to hold for systems of homogeneous linear ODEs) we thus obtain a solution

(3) $$\mathbf{y} = c_1\mathbf{x}^{(1)}e^{\lambda_1 t} + c_2\mathbf{x}^{(2)}e^{\lambda_2 t} = c_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} + c_2 \begin{bmatrix} 1 \\ -1 \end{bmatrix} e^{-0.04t}$$

where c_1 and c_2 are arbitrary constants. Later we shall call this a **general solution.**

Step 3. **Use of initial conditions.** The initial conditions are $y_1(0) = 0$ (no fertilizer in tank T_1) and $y_2(0) = 150$. From this and (3) with $t = 0$ we obtain

$$\mathbf{y}(0) = c_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} + c_2 \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} c_1 + c_2 \\ c_1 - c_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 150 \end{bmatrix}.$$

In components this is $c_1 + c_2 = 0$, $c_1 - c_2 = 150$. The solution is $c_1 = 75$, $c_2 = -75$. This gives the answer

$$\mathbf{y} = 75\mathbf{x}^{(1)} - 75\mathbf{x}^{(2)}e^{-0.04t} = 75\begin{bmatrix}1\\1\end{bmatrix} - 75\begin{bmatrix}1\\-1\end{bmatrix}e^{-0.04t}.$$

In components,

$$y_1 = 75 - 75e^{-0.04t} \qquad \text{(Tank } T_1\text{, lower curve)}$$

$$y_2 = 75 + 75e^{-0.04t} \qquad \text{(Tank } T_2\text{, upper curve).}$$

Figure 77 shows the exponential increase of y_1 and the exponential decrease of y_2 to the common limit 75 lb. Did you expect this for physical reasons? Can you physically explain why the curves look "symmetric"? Would the limit change if T_1 initially contained 100 lb of fertilizer and T_2 contained 50 lb?

Step 4. **Answer.** T_1 contains half the fertilizer amount of T_2 if it contains 1/3 of the total amount, that is, 50 lb. Thus

$$y_1 = 75 - 75e^{-0.04t} = 50, \qquad e^{-0.04t} = \tfrac{1}{3}, \qquad t = (\ln 3)/0.04 = 27.5.$$

Hence the fluid should circulate for at least about half an hour. ▩

EXAMPLE 2 Electrical Network

Find the currents $I_1(t)$ and $I_2(t)$ in the network in Fig. 78. Assume all currents and charges to be zero at $t = 0$, the instant when the switch is closed.

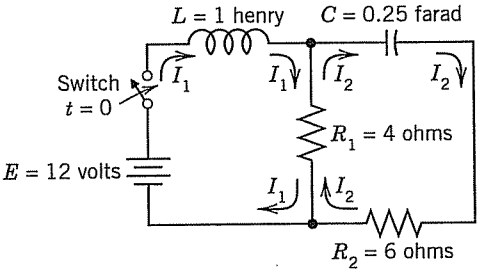


Fig. 78. Electrical network in Example 2

Solution. *Step 1.* **Setting up the mathematical model.** The model of this network is obtained from Kirchhoff's voltage law, as in Sec. 2.9 (where we considered single circuits). Let $I_1(t)$ and $I_2(t)$ be the currents in the left and right loops, respectively. In the left loop the voltage drops are $LI_1' = I_1'$ [V] over the inductor and $R_1(I_1 - I_2) = 4(I_1 - I_2)$ [V] over the resistor, the difference because I_1 and I_2 flow through the resistor in opposite directions. By Kirchhoff's voltage law the sum of these drops equals the voltage of the battery; that is, $I_1' + 4(I_1 - I_2) = 12$, hence

$$I_1' = -4I_1 + 4I_2 + 12. \tag{4a}$$

In the right loop the voltage drops are $R_2I_2 = 6I_2$ [V] and $R_1(I_2 - I_1) = 4(I_2 - I_1)$ [V] over the resistors and $(1/C)\int I_2\,dt = 4\int I_2\,dt$ [V] over the capacitor, and their sum is zero,

$$6I_2 + 4(I_2 - I_1) + 4\int I_2\,dt = 0 \qquad \text{or} \qquad 10I_2 - 4I_1 + 4\int I_2\,dt = 0.$$

Division by 10 and differentiation gives $I_2' - 0.4I_1' + 0.4I_2 = 0$.

To simplify the solution process, we first get rid of $0.4I_1'$, which by (4a) equals $0.4(-4I_1 + 4I_2 + 12)$. Substitution into the present ODE gives

$$I_2' = 0.4I_1' - 0.4I_2 = 0.4(-4I_1 + 4I_2 + 12) - 0.4I_2$$

and by simplification

(4b) $$I_2' = -1.6I_1 + 1.2I_2 + 4.8.$$

In matrix form, (4) is (we write $\mathbf{J}$ since $\mathbf{I}$ is the unit matrix)

(5) $$\mathbf{J}' = \mathbf{AJ} + \mathbf{g}, \qquad \text{where} \qquad \mathbf{J} = \begin{bmatrix} I_1 \\ I_2 \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} -4.0 & 4.0 \\ -1.6 & 1.2 \end{bmatrix}, \quad \mathbf{g} = \begin{bmatrix} 12.0 \\ 4.8 \end{bmatrix}.$$

Step 2. **Solving (5).** Because of the vector $\mathbf{g}$ this is a *nonhomogeneous* system, and we try to proceed as for a single ODE, solving first the *homogeneous* system $\mathbf{J}' = \mathbf{AJ}$ (thus $\mathbf{J}' - \mathbf{AJ} = \mathbf{0}$) by substituting $\mathbf{J} = \mathbf{x}e^{\lambda t}$. This gives

$$\mathbf{J}' = \lambda \mathbf{x}e^{\lambda t} = \mathbf{A}\mathbf{x}e^{\lambda t}, \qquad \text{hence} \qquad \mathbf{Ax} = \lambda \mathbf{x}.$$

Hence to obtain a nontrivial solution, we again need the eigenvalues and eigenvectors. For the present matrix $\mathbf{A}$ they are derived in Example 1 in Sec. 4.0:

$$\lambda_1 = -2, \quad \mathbf{x}^{(1)} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}; \qquad \lambda_2 = -0.8, \quad \mathbf{x}^{(2)} = \begin{bmatrix} 1 \\ 0.8 \end{bmatrix}.$$

Hence a "general solution" of the homogeneous system is

$$\mathbf{J}_h = c_1\mathbf{x}^{(1)}e^{-2t} + c_2\mathbf{x}^{(2)}e^{-0.8t}.$$

For a particular solution of the nonhomogeneous system (5), since $\mathbf{g}$ is constant, we try a constant column vector $\mathbf{J}_p = \mathbf{a}$ with components a_1, a_2. Then $\mathbf{J}_p' = \mathbf{0}$, and substitution into (5) gives $\mathbf{Aa} + \mathbf{g} = \mathbf{0}$; in components,

$$\begin{aligned} -4.0a_1 + 4.0a_2 + 12.0 &= 0 \\ -1.6a_1 + 1.2a_2 + 4.8 &= 0. \end{aligned}$$

The solution is $a_1 = 3$, $a_2 = 0$; thus $\mathbf{a} = \begin{bmatrix} 3 \\ 0 \end{bmatrix}$. Hence

(6) $$\mathbf{J} = \mathbf{J}_h + \mathbf{J}_p = c_1\mathbf{x}^{(1)}e^{-2t} + c_2\mathbf{x}^{(2)}e^{-0.8t} + \mathbf{a};$$

in components,

$$\begin{aligned} I_1 &= 2c_1e^{-2t} + c_2e^{-0.8t} + 3 \\ I_2 &= c_1e^{-2t} + 0.8c_2e^{-0.8t}. \end{aligned}$$

The initial conditions give

$$\begin{aligned} I_1(0) &= 2c_1 + c_2 + 3 = 0 \\ I_2(0) &= c_1 + 0.8c_2 = 0. \end{aligned}$$

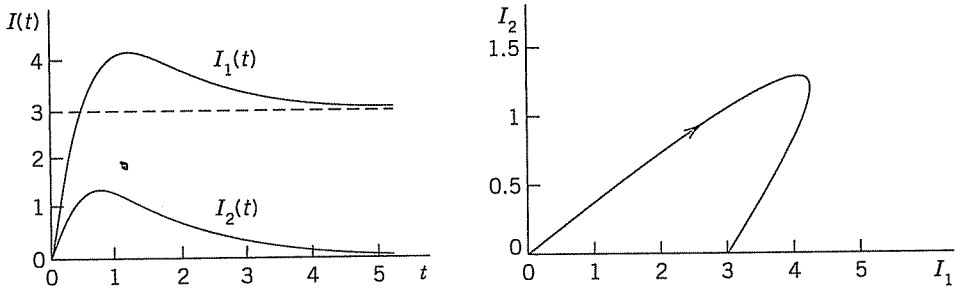
Hence $c_1 = -4$ and $c_2 = 5$. As the solution of our problem we thus obtain

(7) $$\mathbf{J} = -4\mathbf{x}^{(1)}e^{-2t} + 5\mathbf{x}^{(2)}e^{-0.8t} + \mathbf{a}.$$

In components (Fig. 79b),

$$\begin{aligned} I_1 &= -8e^{-2t} + 5e^{-0.8t} + 3 \\ I_2 &= -4e^{-2t} + 4e^{-0.8t}. \end{aligned}$$

Now comes an important idea, on which we shall elaborate further, beginning in Sec. 4.3. Figure 79a shows $I_1(t)$ and $I_2(t)$ as two separate curves. Figure 79b shows these two currents as a single curve $[I_1(t), I_2(t)]$ in the I_1I_2-plane. This is a parametric representation with time t as the parameter. It is often important to know in which sense such a curve is traced. This can be indicated by an arrow in the sense of increasing t, as is shown. The I_1I_2-plane is called the **phase plane** of our system (5), and the curve in Fig. 79b is called a **trajectory.** We shall see that such **"phase plane representations"** are far more important than graphs as in Fig. 79a because they will give a much better qualitative overall impression of the general behavior of whole families of solutions, not merely of one solution as in the present case. ▩



(a) Currents I_1 (upper curve) and I_2

(b) Trajectory $[I_1(t), I_2(t)]^T$ in the I_1I_2-plane (the "phase plane")

Fig. 79. Currents in Example 2

Conversion of an *n*th-Order ODE to a System

We show that an *n*th-order ODE of the general form (8) (see Theorem 1) can be converted to a system of n first-order ODEs. This is practically and theoretically important—practically because it permits the study and solution of single ODEs by methods for systems, and theoretically because it opens a way of including the theory of higher order ODEs into that of first-order systems. This conversion is another reason for the importance of systems, in addition to their use as models in various basic applications. The idea of the conversion is simple and straightforward, as follows.

THEOREM 1

Conversion of an ODE

An nth-order ODE

$$y^{(n)} = F(t, y, y', \cdots, y^{(n-1)}) \tag{8}$$

can be converted to a system of n first-order ODEs by setting

$$y_1 = y, \quad y_2 = y', \quad y_3 = y'', \cdots, y_n = y^{(n-1)}. \tag{9}$$

This system is of the form

$$\begin{aligned} y_1' &= y_2 \\ y_2' &= y_3 \\ &\vdots \\ y_{n-1}' &= y_n \\ y_n' &= F(t, y_1, y_2, \cdots, y_n). \end{aligned} \tag{10}$$

PROOF The first $n - 1$ of these n ODEs follow immediately from (9) by differentiation. Also, $y_n' = y^{(n)}$ by (9), so that the last equation in (10) results from the given ODE (8). ∎

EXAMPLE 3 Mass on a Spring

To gain confidence in the conversion method, let us apply it to an old friend of ours, modeling the free motions of a mass on a spring (see Sec. 2.4)

$$my'' + cy' + ky = 0 \qquad \text{or} \qquad y'' = -\frac{c}{m}y' - \frac{k}{m}y.$$

For this ODE (8) the system (10) is linear and homogeneous,

$$y_1' = y_2$$

$$y_2' = -\frac{k}{m}y_1 - \frac{c}{m}y_2.$$

Setting $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$, we get in matrix form

$$\mathbf{y}' = \mathbf{A}\mathbf{y} = \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{c}{m} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}.$$

The characteristic equation is

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} -\lambda & 1 \\ -\frac{k}{m} & -\frac{c}{m} - \lambda \end{vmatrix} = \lambda^2 + \frac{c}{m}\lambda + \frac{k}{m} = 0.$$

It agrees with that in Sec. 2.4. For an illustrative computation, let $m = 1$, $c = 2$, and $k = 0.75$. Then

$$\lambda^2 + 2\lambda + 0.75 = (\lambda + 0.5)(\lambda + 1.5) = 0.$$

This gives the eigenvalues $\lambda_1 = -0.5$ and $\lambda_2 = -1.5$. Eigenvectors follow from the first equation in $\mathbf{A} - \lambda\mathbf{I} = \mathbf{0}$, which is $-\lambda x_1 + x_2 = 0$. For λ_1 this gives $0.5x_1 + x_2 = 0$, say, $x_1 = 2$, $x_2 = -1$. For $\lambda_2 = -1.5$ it gives $1.5x_1 + x_2 = 0$, say, $x_1 = 1$, $x_2 = -1.5$. These eigenvectors

$$\mathbf{x}^{(1)} = \begin{bmatrix} 2 \\ -1 \end{bmatrix}, \quad \mathbf{x}^{(2)} = \begin{bmatrix} 1 \\ -1.5 \end{bmatrix} \qquad \text{give} \qquad \mathbf{y} = c_1 \begin{bmatrix} 2 \\ -1 \end{bmatrix} e^{-0.5t} + c_2 \begin{bmatrix} 1 \\ -1.5 \end{bmatrix} e^{-1.5t}.$$

This vector solution has the first component

$$y = y_1 = 2c_1 e^{-0.5t} + c_2 e^{-1.5t}$$

which is the expected solution. The second component is its derivative

$$y_2 = y_1' = y' = -c_1 e^{-0.5t} - 1.5c_2 e^{-1.5t}.$$ ■

PROBLEM SET 4.1

1–6 MIXING PROBLEMS

1. Find out without calculation whether doubling the flow rate in Example 1 has the same effect as halfing the tank sizes. (Give a reason.)
2. What happens in Example 1 if we replace T_2 by a tank containing 500 gal of water and 150 lb of fertilizer dissolved in it?
3. Derive the eigenvectors in Example 1 without consulting this book.
4. In Example 1 find a "general solution" for any ratio $a = (\textit{flow rate})/(\textit{tank size})$, tank sizes being equal. Comment on the result.
5. If you extend Example 1 by a tank T_3 of the same size as the others and connected to T_2 by two tubes with

flow rates as between T_1 and T_2, what system of ODEs will you get?

6. Find a "general solution" of the system in Prob. 5.

7–10 ELECTRICAL NETWORKS

7. Find the currents in Example 2 if the initial currents are 0 and −3 A (minus meaning that $I_2(0)$ flows against the direction of the arrow).
8. Find the currents in Example 2 if the resistance of R_1 and R_2 is doubled (general solution only). First, guess.
9. What are the limits of the currents in Example 2? Explain them in terms of physics.
10. Find the currents in Example 2 if the capacitance is changed to $C = 1/5.4$ F (farad).

11–15 CONVERSION TO SYSTEMS

Find a general solution of the given ODE (a) by first converting it to a system, (b), as given. (Show the details of your work.)

11. $y'' - 4y = 0$
12. $y'' + 2y' - 24y = 0$
13. $y'' - y' = 0$
14. $y'' + 15y' + 50y = 0$
15. $64y'' - 48y' - 7y = 0$

16. **TEAM PROJECT. Two Masses on Springs.** (a) Set up the model for the (undamped) system in Fig. 80.

(b) Solve the system of ODEs obtained. *Hint.* Try $\mathbf{y} = \mathbf{x}e^{\omega t}$ and set $\omega^2 = \lambda$. Proceed as in Example 1 or 2.

(c) Describe the influence of initial conditions on the possible kind of motions.

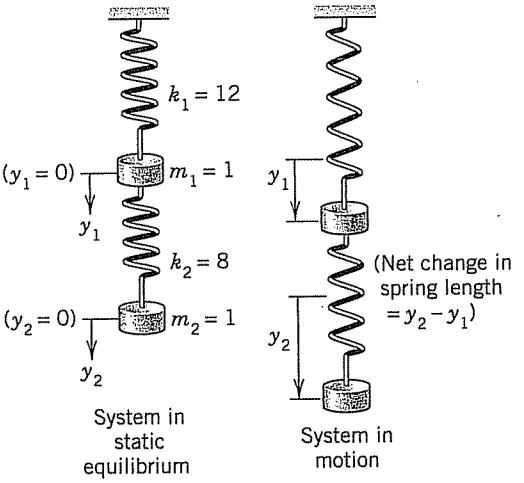


Fig. 80. Mechanical system in Team Project 16

4.2 Basic Theory of Systems of ODEs

In this section we discuss some basic concepts and facts about systems of ODEs that are quite similar to those for single ODEs.

The first-order systems in the last section were special cases of the more general system

$$\begin{aligned} y_1' &= f_1(t, y_1, \cdots, y_n) \\ y_2' &= f_2(t, y_1, \cdots, y_n) \\ &\cdots \\ y_n' &= f_n(t, y_1, \cdots, y_n). \end{aligned} \tag{1}$$

We can write the system (1) as a vector equation by introducing the column vectors $\mathbf{y} = [y_1 \quad \cdots \quad y_n]^T$ and $\mathbf{f} = [f_1 \quad \cdots \quad f_n]^T$ (where T means *transposition* and saves us the space that would be needed for writing $\mathbf{y}$ and $\mathbf{f}$ as columns). This gives

$$\mathbf{y}' = \mathbf{f}(t, \mathbf{y}). \tag{1}$$

This system (1) includes almost all cases of practical interest. For $n = 1$ it becomes $y_1' = f_1(t, y_1)$ or, simply, $y' = f(t, y)$, well known to us from Chap. 1.

A **solution** of (1) on some interval $a < t < b$ is a set of n differentiable functions

$$y_1 = h_1(t), \quad \cdots, \quad y_n = h_n(t)$$

on $a < t < b$ that satisfy (1) throughout this interval. In vector form, introducing the *"solution vector"* $\mathbf{h} = [h_1 \quad \cdots \quad h_n]^T$ (a column vector!) we can write

$$\mathbf{y} = \mathbf{h}(t).$$

An **initial value problem** for (1) consists of (1) and n given **initial conditions**

$$(2) \qquad y_1(t_0) = K_1, \qquad y_2(t_0) = K_2, \qquad \cdots, \qquad y_n(t_0) = K_n,$$

in vector form, $\mathbf{y}(t_0) = \mathbf{K}$, where t_0 is a specified value of t in the interval considered and the components of $\mathbf{K} = [K_1 \quad \cdots \quad K_n]^T$ are given numbers. Sufficient conditions for the existence and uniqueness of a solution of an initial value problem (1), (2) are stated in the following theorem, which extends the theorems in Sec. 1.7 for a single equation. (For a proof, see Ref. [A7].)

THEOREM 1

Existence and Uniqueness Theorem

Let $f_1, \cdots, f_n$ in (1) be continuous functions having continuous partial derivatives $\partial f_1/\partial y_1, \cdots, \partial f_1/\partial y_n, \cdots, \partial f_n/\partial y_n$ in some domain R of $ty_1y_2 \cdots y_n$-space containing the point $(t_0, K_1, \cdots, K_n)$. Then (1) has a solution on some interval $t_0 - \alpha < t < t_0 + \alpha$ satisfying (2), and this solution is unique.

Linear Systems

Extending the notion of a *linear* ODE, we call (1) a **linear system** if it is linear in $y_1 \cdots, y_n$; that is, if it can be written

$$(3) \qquad \begin{aligned} y_1' &= a_{11}(t)y_1 + \cdots + a_{1n}(t)y_n + g_1(t) \\ &\vdots \\ y_n' &= a_{n1}(t)y_1 + \cdots + a_{nn}(t)y_n + g_n(t). \end{aligned}$$

In vector form, this becomes

$$(3) \qquad \mathbf{y}' = \mathbf{A}\mathbf{y} + \mathbf{g}$$

where $$\mathbf{A} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \cdot & \cdots & \cdot \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \qquad \mathbf{g} = \begin{bmatrix} g_1 \\ \vdots \\ g_n \end{bmatrix}.$$

This system is called **homogeneous** if $\mathbf{g} = \mathbf{0}$, so that it is

$$(4) \qquad \mathbf{y}' = \mathbf{A}\mathbf{y}.$$

If $\mathbf{g} \neq \mathbf{0}$, then (3) is called **nonhomogeneous.** The system in Example 1 in the last section is homogeneous and in Example 2 nonhomogeneous. The system in Example 3 is homogeneous.

For a linear system (3) we have $\partial f_1/\partial y_1 = a_{11}(t), \cdots, \partial f_n/\partial y_n = a_{nn}(t)$ in Theorem 1 Hence for a linear system we simply obtain the following.

THEOREM 2

Existence and Uniqueness in the Linear Case

Let the a_{jk}'s and g_j's in (3) *be continuous functions of t on an open interval $\alpha < t < \beta$ containing the point $t = t_0$. Then* (3) *has a solution* $\mathbf{y}(t)$ *on this interval satisfying* (2)*, and this solution is unique.*

As for a single homogeneous linear ODE we have

THEOREM 3

Superposition Principle or Linearity Principle

If $\mathbf{y}^{(1)}$ *and* $\mathbf{y}^{(2)}$ *are solutions of the* ***homogeneous linear*** *system* (4) *on some interval, so is any linear combination* $\mathbf{y} = c_1\mathbf{y}^{(1)} + c_2\mathbf{y}^{(2)}$.

PROOF Differentiating and using (4), we obtain

$$\begin{aligned}\mathbf{y}' &= [c_1\mathbf{y}^{(1)} + c_2\mathbf{y}^{(2)}]' \\ &= c_1\mathbf{y}^{(1)\prime} + c_2\mathbf{y}^{(2)\prime} \\ &= c_1\mathbf{A}\mathbf{y}^{(1)} + c_2\mathbf{A}\mathbf{y}^{(2)} \\ &= \mathbf{A}(c_1\mathbf{y}^{(1)} + c_2\mathbf{y}^{(2)}) = \mathbf{A}\mathbf{y}.\end{aligned}$$

∎

The general theory of linear systems of ODEs is quite similar to that of a single linea ODE in Secs. 2.6 and 2.7. To see this, we explain the most basic concepts and facts. Fo proofs we refer to more advanced texts, such as [A7].

Basis. General Solution. Wronskian

By a **basis** or a **fundamental system** of solutions of the homogeneous system (4) on som interval J we mean a linearly independent set of n solutions $\mathbf{y}^{(1)}, \cdots, \mathbf{y}^{(n)}$ of (4) on th interval. (We write J because we need $\mathbf{I}$ to denote the unit matrix.) We call a correspondin linear combination

$$\mathbf{y} = c_1\mathbf{y}^{(1)} \cdots + c_n\mathbf{y}^{(n)} \qquad (c_1, \cdots, c_n \text{ arbitrar} \tag{5}$$

a **general solution** of (4) on J. It can be shown that if the $a_{jk}(t)$ in (4) are continuous c J, then (4) has a basis of solutions on J, hence a general solution, which includes eve solution of (4) on J.

We can write n solutions $\mathbf{y}^{(1)}, \cdots, \mathbf{y}^{(n)}$ of (4) on some interval J as columns of a $n \times n$ matrix

$$\mathbf{Y} = \begin{bmatrix}\mathbf{y}^{(1)} & \cdots & \mathbf{y}^{(n)}\end{bmatrix}. \tag{6}$$

The determinant of $\mathbf{Y}$ is called the **Wronskian** of $\mathbf{y}^{(1)}, \cdots, \mathbf{y}^{(n)}$, written

$$(7) \quad W(\mathbf{y}^{(1)}, \cdots, \mathbf{y}^{(n)}) = \begin{vmatrix} y_1^{(1)} & y_1^{(2)} & \cdots & y_1^{(n)} \\ y_2^{(1)} & y_2^{(2)} & \cdots & y_2^{(n)} \\ \cdot & \cdot & \cdots & \cdot \\ y_n^{(1)} & y_n^{(2)} & \cdots & y_n^{(n)} \end{vmatrix}.$$

The columns are these solutions, each in terms of components. These solutions form a basis on J if and only if W is not zero at any t_1 in this interval. W either is identically zero or is nowhere zero in J. (This is similar to Secs. 2.6 and 3.1.)

If the solutions $\mathbf{y}^{(1)}, \cdots, \mathbf{y}^{(n)}$ in (5) form a basis (a fundamental system), then (6) is often called a **fundamental matrix.** Introducing a column vector $\mathbf{c} = [c_1 \quad c_2 \quad \cdots \quad c_n]^T$, we can now write (5) simply as

$$(8) \quad \mathbf{y} = \mathbf{Yc}.$$

Furthermore, we can relate (7) to Sec. 2.6, as follows. If y and z are solutions of a second-order homogeneous linear ODE, their Wronskian is

$$W(y, z) = \begin{vmatrix} y & z \\ y' & z' \end{vmatrix}.$$

To write this ODE as a system, we have to set $y = y_1$, $y' = y_1' = y_2$ and similarly for z (see Sec. 4.1). But then $W(y, z)$ becomes (7), except for notation.

4.3 Constant-Coefficient Systems. Phase Plane Method

Continuing, we now assume that our homogeneous linear system

$$(1) \quad \mathbf{y}' = \mathbf{Ay}$$

under discussion has ***constant coefficients,*** so that the $n \times n$ matrix $\mathbf{A} = [a_{jk}]$ has entries not depending on t. We want to solve (1). Now a single ODE $y' = ky$ has the solution $y = Ce^{kt}$. So let us try

$$(2) \quad \mathbf{y} = \mathbf{x}e^{\lambda t}.$$

Substitution into (1) gives $\mathbf{y}' = \lambda \mathbf{x}e^{\lambda t} = \mathbf{Ay} = \mathbf{A}\mathbf{x}e^{\lambda t}$. Dividing by $e^{\lambda t}$, we obtain the **eigenvalue problem**

$$(3) \quad \mathbf{Ax} = \lambda \mathbf{x}.$$

Thus the nontrivial solutions of (1) (solutions that are not zero vectors) are of the form (
where λ is an eigenvalue of $\mathbf{A}$ and $\mathbf{x}$ is a corresponding eigenvector.

We assume that $\mathbf{A}$ has a linearly independent set of n eigenvectors. This holds in mc
applications, in particular if $\mathbf{A}$ is symmetric ($a_{kj} = a_{jk}$) or skew-symmetric ($a_{kj} = -a_{jk}$)
or has *n different* eigenvalues.

Let those eigenvectors be $\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(n)}$ and let them correspond to eigenvalu
$\lambda_1, \cdots, \lambda_n$ (which may be all different, or some—or even all—may be equal). Then t
corresponding solutions (2) are

$$\mathbf{y}^{(1)} = \mathbf{x}^{(1)}e^{\lambda_1 t}, \quad \cdots, \quad \mathbf{y}^{(n)} = \mathbf{x}^{(n)}e^{\lambda_n t}. \tag{4}$$

Their Wronskian $W = W(\mathbf{y}^{(1)}, \cdots, \mathbf{y}^{(n)})$ [(7) in Sec. 4.2] is given by

$$W = (\mathbf{y}^{(1)}, \cdots, \mathbf{y}^{(n)}) = \begin{vmatrix} x_1^{(1)}e^{\lambda_1 t} & \cdots & x_1^{(n)}e^{\lambda_n t} \\ x_2^{(1)}e^{\lambda_1 t} & \cdots & x_2^{(n)}e^{\lambda_n t} \\ \cdot & \cdots & \cdot \\ x_n^{(1)}e^{\lambda_1 t} & \cdots & x_n^{(n)}e^{\lambda_n t} \end{vmatrix} = e^{\lambda_1 t + \cdots + \lambda_n t} \begin{vmatrix} x_1^{(1)} & \cdots & x_1^{(n)} \\ x_2^{(1)} & \cdots & x_2^{(n)} \\ \cdot & \cdots & \cdot \\ x_n^{(1)} & \cdots & x_n^{(n)} \end{vmatrix}$$

On the right, the exponential function is never zero, and the determinant is not zero eith
because its columns are the n linearly independent eigenvectors. This proves the followin
theorem, whose assumption is true if the matrix $\mathbf{A}$ is symmetric or skew-symmetric, or
the n eigenvalues of $\mathbf{A}$ are all different.

THEOREM 1

General Solution

If the constant matrix $\mathbf{A}$ *in the system* (1) *has a linearly independent set of* n *eigenvectors, then the corresponding solutions* $\mathbf{y}^{(1)}, \cdots, \mathbf{y}^{(n)}$ *in* (4) *form a basis of solutions of* (1), *and the corresponding general solution is*

$$\mathbf{y} = c_1\mathbf{x}^{(1)}e^{\lambda_1 t} + \cdots + c_n\mathbf{x}^{(n)}e^{\lambda_n t}. \tag{5}$$

How to Graph Solutions in the Phase Plane

We shall now concentrate on systems (1) with constant coefficients consisting of tw
ODEs

$$\mathbf{y}' = \mathbf{A}\mathbf{y}; \quad \text{in components,} \quad \begin{aligned} y_1' &= a_{11}y_1 + a_{12}y_2 \\ y_2' &= a_{21}y_1 + a_{22}y_2. \end{aligned} \tag{6}$$

Of course, we can graph solutions of (6),

$$\mathbf{y}(t) = \begin{bmatrix} y_1(t) \\ y_2(t) \end{bmatrix}, \tag{7}$$

as two curves over the t-axis, one for each component of $\mathbf{y}(t)$. (Figure 79a in Sec. 4.1 shows an example.) But we can also graph (7) as a single curve in the y_1y_2-plane. This is a *parametric representation* (*parametric equation*) with parameter t. (See Fig. 79b for an example. Many more follow. Parametric equations also occur in calculus.) Such a curve is called a **trajectory** (or sometimes an *orbit* or *path*) of (6). The y_1y_2-plane is called the **phase plane.**[1] If we fill the phase plane with trajectories of (6), we obtain the so-called **phase portrait** of (6).

EXAMPLE 1 **Trajectories in the Phase Plane (Phase Portrait)**

In order to see what is going on, let us find and graph solutions of the system

$$\mathbf{y}' = \mathbf{A}\mathbf{y} = \begin{bmatrix} -3 & 1 \\ 1 & -3 \end{bmatrix} \mathbf{y}, \quad \text{thus} \quad \begin{aligned} y_1' &= -3y_1 + y_2 \\ y_2' &= y_1 - 3y_2. \end{aligned} \tag{8}$$

Solution. By substituting $\mathbf{y} = \mathbf{x}e^{\lambda t}$ and $\mathbf{y}' = \lambda \mathbf{x}e^{\lambda t}$ and dropping the exponential function we get $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$. The characteristic equation is

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} -3 - \lambda & 1 \\ 1 & -3 - \lambda \end{vmatrix} = \lambda^2 + 6\lambda + 8 = 0.$$

This gives the eigenvalues $\lambda_1 = -2$ and $\lambda_2 = -4$. Eigenvectors are then obtained from

$$(-3 - \lambda)x_1 + x_2 = 0.$$

For $\lambda_1 = -2$ this is $-x_1 + x_2 = 0$. Hence we can take $\mathbf{x}^{(1)} = [1 \quad 1]^\mathsf{T}$. For $\lambda_2 = -4$ this becomes $x_1 + x_2 = 0$, and an eigenvector is $\mathbf{x}^{(2)} = [1 \quad -1]^\mathsf{T}$. This gives the general solution

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = c_1\mathbf{y}^{(1)} + c_2\mathbf{y}^{(2)} = c_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} e^{-2t} + c_2 \begin{bmatrix} 1 \\ -1 \end{bmatrix} e^{-4t}.$$

Figure 81 on p. 142 shows a phase portrait of some of the trajectories (to which more trajectories could be added if so desired). The two straight trajectories correspond to $c_1 = 0$ and $c_2 = 0$ and the others to other choices of c_1, c_2. ■

Studies of solutions in the phase plane have recently become quite important, along with advances in computer graphics, because a phase portrait gives a good general qualitative impression of the entire family of solutions. This method becomes particularly valuable in the frequent cases when solving an ODE or a system is inconvenient or impossible.

Critical Points of the System (6)

The point $\mathbf{y} = \mathbf{0}$ in Fig. 81 seems to be a common point of all trajectories, and we want to explore the reason for this remarkable observation. The answer will follow by calculus. Indeed, from (6) we obtain

$$\frac{dy_2}{dy_1} = \frac{y_2'\,dt}{y_1'\,dt} = \frac{y_2'}{y_1'} = \frac{a_{21}y_1 + a_{22}y_2}{a_{11}y_1 + a_{12}y_2}. \tag{9}$$

[1] A name that comes from physics, where it is the y-(mv)-plane, used to plot a motion in terms of position y and velocity $y' = v$ (m = mass); but the name is now used quite generally for the y_1y_2-plane.

The use of the phase plane is a **qualitative method,** a method of obtaining general qualitative information on solutions without actually solving an ODE or a system. This method was created by HENRI POINCARÉ (1854–1912), a great French mathematician, whose work was also fundamental in complex analysis, divergent series, topology, and astronomy.

This associates with every point P: (y_1, y_2) a unique tangent direction dy_2/dy_1 of t trajectory passing through P, except for the point $P = P_0$: (0, 0), where the right side (9) becomes 0/0. This point P_0, at which dy_2/dy_1 becomes undetermined, is called a **critic point** of (6).

Five Types of Critical Points

There are five types of critical points depending on the geometric shape of the trajectori near them. They are called **improper nodes, proper nodes, saddle points, centers,** a **spiral points.** We define and illustrate them in Examples 1–5.

EXAMPLE 1 **(*Continued*) Improper Node (Fig. 81)**

An **improper node** is a critical point P_0 at which all the trajectories, except for two of them, have the sa limiting direction of the tangent. The two exceptional trajectories also have a limiting direction of the tang at P_0 which, however, is different.

The system (8) has an improper node at **0,** as its phase portrait Fig. 81 shows. The common limiting directi at **0** is that of the eigenvector $\mathbf{x}^{(1)} = [1 \quad 1]^T$ because e^{-4t} goes to zero faster than e^{-2t} as t increases. The t exceptional limiting tangent directions are those of $\mathbf{x}^{(2)} = [1 \quad -1]^T$ and $-\mathbf{x}^{(2)} = [-1 \quad 1]^T$.

EXAMPLE 2 **Proper Node (Fig. 82)**

A **proper node** is a critical point P_0 at which every trajectory has a definite limiting direction and for any giv direction $\mathbf{d}$ at P_0 there is a trajectory having $\mathbf{d}$ as its limiting direction.

The system

$$\text{(10)} \qquad \mathbf{y}' = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \mathbf{y}, \qquad \text{thus} \qquad \begin{aligned} y_1' &= y_1 \\ y_2' &= y_2 \end{aligned}$$

has a proper node at the origin (see Fig. 82). Indeed, the matrix is the unit matrix. Its characteristic equati $(1 - \lambda)^2 = 0$ has the root $\lambda = 1$. Any $\mathbf{x} \neq \mathbf{0}$ is an eigenvector, and we can take $[1 \quad 0]^T$ and $[0 \quad 1]^T$. Her a general solution is

$$\mathbf{y} = c_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} e^t + c_2 \begin{bmatrix} 0 \\ 1 \end{bmatrix} e^t \qquad \text{or} \qquad \begin{aligned} y_1 &= c_1 e^t \\ y_2 &= c_2 e^t \end{aligned} \qquad \text{or} \qquad c_1 y_2 = c_2 y_1.$$

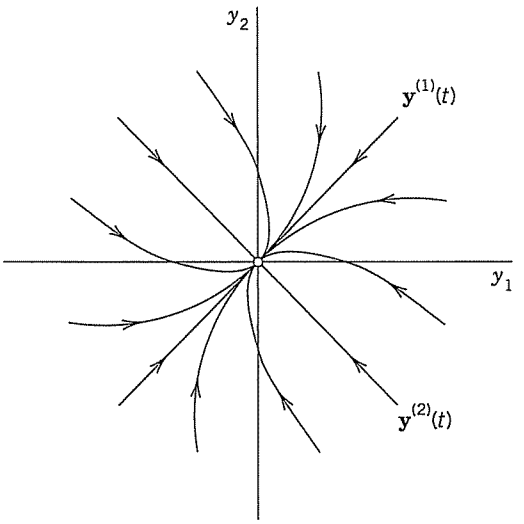


Fig. 81. Trajectories of the system (8) (Improper node)

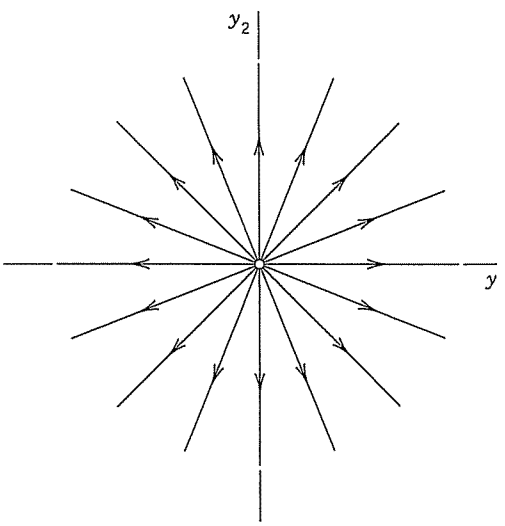


Fig. 82. Trajectories of the system (10) (Proper node)

EXAMPLE 3 **Saddle Point (Fig. 83)**

A **saddle point** is a critical point P_0 at which there are two incoming trajectories, two outgoing trajectories, and all the other trajectories in a neighborhood of P_0 bypass P_0.

The system

$$\mathbf{y}' = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \mathbf{y}, \qquad \text{thus} \qquad \begin{aligned} y_1' &= y_1 \\ y_2' &= -y_2 \end{aligned} \tag{11}$$

has a saddle point at the origin. Its characteristic equation $(1 - \lambda)(-1 - \lambda) = 0$ has the roots $\lambda_1 = 1$ and $\lambda_2 = -1$. For $\lambda = 1$ an eigenvector $[1 \quad 0]^T$ is obtained from the second row of $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$, that is, $0x_1 + (-1 - 1)x_2 = 0$. For $\lambda_2 = -1$ the first row gives $[0 \quad 1]^T$. Hence a general solution is

$$\mathbf{y} = c_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} e^{t} + c_2 \begin{bmatrix} 0 \\ 1 \end{bmatrix} e^{-t} \quad \text{or} \quad \begin{aligned} y_1 &= c_1 e^{t} \\ y_2 &= c_2 e^{-t} \end{aligned} \quad \text{or} \quad y_1 y_2 = const.$$

This is a family of hyperbolas (and the coordinate axes); see Fig. 83. ■

EXAMPLE 4 **Center (Fig. 84)**

A **center** is a critical point that is enclosed by infinitely many closed trajectories.

The system

$$\mathbf{y}' = \begin{bmatrix} 0 & 1 \\ -4 & 0 \end{bmatrix} \mathbf{y}, \qquad \text{thus} \qquad \begin{aligned} y_1' &= y_2 \\ y_2' &= -4y_1 \end{aligned} \tag{12}$$

has a center at the origin. The characteristic equation $\lambda^2 + 4 = 0$ gives the eigenvalues $2i$ and $-2i$. For $2i$ an eigenvector follows from the first equation $-2ix_1 + x_2 = 0$ of $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$, say, $[1 \quad 2i]^T$. For $\lambda = -2i$ that equation is $-(-2i)x_1 + x_2 = 0$ and gives, say, $[1 \quad -2i]^T$. Hence a complex general solution is

$$\mathbf{y} = c_1 \begin{bmatrix} 1 \\ 2i \end{bmatrix} e^{2it} + c_2 \begin{bmatrix} 1 \\ -2i \end{bmatrix} e^{-2it}, \qquad \text{thus} \qquad \begin{aligned} y_1 &= c_1 e^{2it} + c_2 e^{-2it} \\ y_2 &= 2ic_1 e^{2it} - 2ic_2 e^{-2it}. \end{aligned} \tag{12*}$$

The next step would be the transformation of this solution to real form by the Euler formula (Sec. 2.2). But we were just curious to see what kind of eigenvalues we obtain in the case of a center. Accordingly, we do not continue, but start again from the beginning and use a shortcut. We rewrite the given equations in the form $y_1' = y_2$, $4y_1 = -y_2'$; then the product of the left sides must equal the product of the right sides,

$$4y_1 y_1' = -y_2 y_2'. \qquad \text{By integration,} \qquad 2y_1^2 + \tfrac{1}{2}y_2^2 = const.$$

This is a family of ellipses (see Fig. 84) enclosing the center at the origin. ■

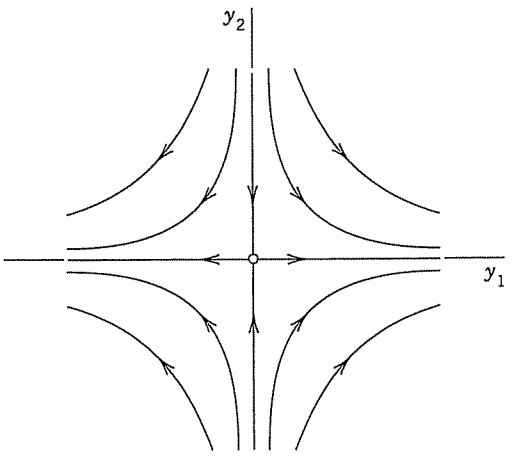


Fig. 83. Trajectories of the system (11) (Saddle point)

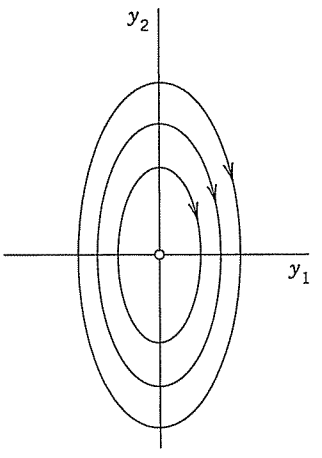


Fig. 84. Trajectories of the system (12) (Center)

EXAMPLE 5 Spiral Point (Fig. 85)

A **spiral point** is a critical point P_0 about which the trajectories spiral, approaching P_0 as $t \to \infty$ (or tracing these spirals in the opposite sense, away from P_0).

The system

$$\mathbf{y}' = \begin{bmatrix} -1 & 1 \\ -1 & -1 \end{bmatrix} \mathbf{y}, \qquad \text{thus} \qquad \begin{aligned} y_1' &= -y_1 + y_2 \\ y_2' &= -y_1 - y_2 \end{aligned} \tag{13}$$

has a spiral point at the origin, as we shall see. The characteristic equation is $\lambda^2 + 2\lambda + 2 = 0$. It gives the eigenvalues $-1 + i$ and $-1 - i$. Corresponding eigenvectors are obtained from $(-1 - \lambda)x_1 + x_2 = 0$. For $\lambda = -1 + i$ this becomes $-ix_1 + x_2 = 0$ and we can take $[1 \quad i]^T$ as an eigenvector. Similarly, an eigenvector corresponding to $-1 - i$ is $[1 \quad -i]^T$. This gives the complex general solution

$$\mathbf{y} = c_1 \begin{bmatrix} 1 \\ i \end{bmatrix} e^{(-1+i)t} + c_2 \begin{bmatrix} 1 \\ -i \end{bmatrix} e^{(-1-i)t}.$$

The next step would be the transformation of this complex solution to a real general solution by the Euler formula. But, as in the last example, we just wanted to see what eigenvalues to expect in the case of a spiral point. Accordingly, we start again from the beginning and instead of that rather lengthy systematic calculation we use a shortcut. We multiply the first equation in (13) by y_1, the second by y_2, and add, obtaining

$$y_1 y_1' + y_2 y_2' = -(y_1^2 + y_2^2).$$

We now introduce polar coordinates r, t, where $r^2 = y_1^2 + y_2^2$. Differentiating this with respect to t gives $2rr' = 2y_1 y_1' + 2y_2 y_2'$. Hence the previous equation can be written

$$rr' = -r^2, \qquad \text{Thus,} \qquad r' = -r, \qquad dr/r = -dt, \qquad \ln|r| = -t + c^*, \qquad r = ce^{-t}.$$

For each real c this is a spiral, as claimed. (see Fig. 85). ∎

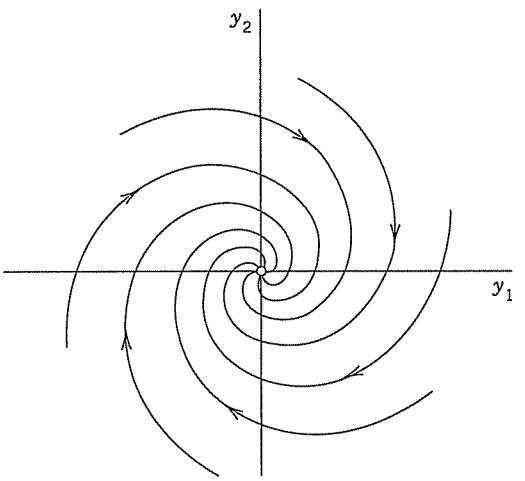


Fig. 85. Trajectories of the system (13) (Spiral point)

EXAMPLE 6 No Basis of Eigenvectors Available. Degenerate Node (Fig. 86)

This cannot happen if A in (1) is symmetric ($a_{kj} = a_{jk}$, as in Examples 1–3) or skew-symmetric ($a_{kj} = -a_{jk}$ thus $a_{jj} = 0$). And it does not happen in many other cases (see Examples 4 and 5). Hence it suffices to explain the method to be used by an example.

Find and graph a general solution of

$$\mathbf{y}' = \mathbf{A}\mathbf{y} = \begin{bmatrix} 4 & 1 \\ -1 & 2 \end{bmatrix} \mathbf{y}. \tag{14}$$

Solution. $\mathbf{A}$ is not skew-symmetric! Its characteristic equation is

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} 4-\lambda & 1 \\ -1 & 2-\lambda \end{vmatrix} = \lambda^2 - 6\lambda + 9 = (\lambda - 3)^2 = 0.$$

It has a double root $\lambda = 3$. Hence eigenvectors are obtained from $(4 - \lambda)x_1 + x_2 = 0$, thus from $x_1 + x_2 = 0$, say, $\mathbf{x}^{(1)} = [1 \quad -1]^T$ and nonzero multiples of it (which do not help). The method now is to substitute

$$\mathbf{y}^{(2)} = \mathbf{x}te^{\lambda t} + \mathbf{u}e^{\lambda t}$$

with constant $\mathbf{u} = [u_1 \quad u_2]^T$ into (14). (The $\mathbf{x}t$-term alone, the analog of what we did in Sec. 2.2 in the case of a double root, would not be enough. Try it.) This gives

$$\mathbf{y}^{(2)\prime} = \mathbf{x}e^{\lambda t} + \lambda\mathbf{x}te^{\lambda t} + \lambda\mathbf{u}e^{\lambda t} = \mathbf{A}\mathbf{y}^{(2)} = \mathbf{A}\mathbf{x}te^{\lambda t} + \mathbf{A}\mathbf{u}e^{\lambda t}.$$

On the right, $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$. Hence the terms $\lambda\mathbf{x}te^{\lambda t}$ cancel, and then division by $e^{\lambda t}$ gives

$$\mathbf{x} + \lambda\mathbf{u} = \mathbf{A}\mathbf{u}, \qquad \text{thus} \qquad (\mathbf{A} - \lambda\mathbf{I})\mathbf{u} = \mathbf{x}.$$

Here $\lambda = 3$ and $\mathbf{x} = [1 \quad -1]^T$, so that

$$(\mathbf{A} - 3\mathbf{I})\mathbf{u} = \begin{bmatrix} 4-3 & 1 \\ -1 & 2-3 \end{bmatrix} \mathbf{u} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \qquad \text{thus} \qquad \begin{aligned} u_1 + u_2 &= 1 \\ -u_1 - u_2 &= -1. \end{aligned}$$

A solution, linearly independent of $\mathbf{x} = [1 \quad -1]^T$, is $\mathbf{u} = [0 \quad 1]^T$. This yields the answer (Fig. 86)

$$\mathbf{y} = c_1\mathbf{y}^{(1)} + c_2\mathbf{y}^{(2)} = c_1 \begin{bmatrix} 1 \\ -1 \end{bmatrix} e^{3t} + c_2\left(\begin{bmatrix} 1 \\ -1 \end{bmatrix} t + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right) e^{3t}.$$

The critical point at the origin is often called a **degenerate node.** $c_1\mathbf{y}^{(1)}$ gives the heavy straight line, with $c_1 > 0$ the lower part and $c_1 < 0$ the upper part of it. $\mathbf{y}^{(2)}$ gives the right part of the heavy curve from 0 through the second, first, and—finally—fourth quadrants. $-\mathbf{y}^{(2)}$ gives the other part of that curve. ■

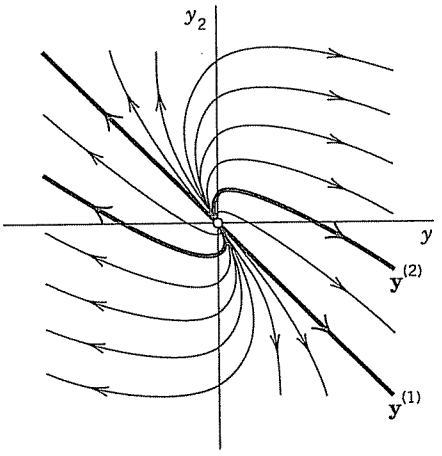


Fig. 86. Degenerate node in Example 6

We mention that for a system (1) with three or more equations and a triple eigenvalu with only one linearly independent eigenvector, one will get two solutions, as jus discussed, and a third linearly independent one from

$$\mathbf{y}^{(3)} = \tfrac{1}{2}\mathbf{x}t^2e^{\lambda t} + \mathbf{u}te^{\lambda t} + \mathbf{v}e^{\lambda t} \qquad \text{with } \mathbf{v} \text{ from} \qquad (\mathbf{u} + \lambda\mathbf{v} = \mathbf{A}\mathbf{v}.$$

PROBLEM SET 4.5

1–9 GENERAL SOLUTION

Find a real general solution of the following systems. (Show the details.)

1. $y_1' = 3y_2$
$y_2' = 12y_1$

2. $y_1' = 5y_2$
$y_2' = 5y_1$

3. $y_1' = y_1 + y_2$
$y_2' = y_1 + y_2$

4. $y_1' = 9y_1 + 13.5y_2$
$y_2' = 1.5y_1 + 9y_2$

5. $y_1' = 4y_2$
$y_2' = -4y_1$

6. $y_1' = 2y_1 - 2y_2$
$y_2' = 2y_1 + 2y_2$

7. $y_1' = 2y_1 + 8y_2 - 4y_3$
$y_2' = -4y_1 - 10y_2 + 2y_3$
$y_3' = -4y_1 - 4y_2 - 4y_3$

8. $y_1' = 8y_1 - y_2$
$y_2' = y_1 + 10y_2$

9. $y_1' = -y_1 + y_2 + 0.4y_3$
$y_2' = y_1 - 0.1y_2 + 1.4y_3$
$y_3' = 0.4y_1 + 1.4y_2 + 0.2y_3$

10–15 INITIAL VALUE PROBLEMS

Solve the following initial value problems. (Show the details.)

10. $y_1' = y_1 + y_2$
$y_2' = 4y_1 + y_2$
$y_1(0) = 1,\ y_2(0) = 6$

11. $y_1' = y_1 + 2y_2$
$y_2' = \frac{1}{2}y_1 + y_2$
$y_1(0) = 16,\ y_2(0) = -2$

12. $y_1' = 3y_1 + 2y_2$
$y_2' = 2y_1 + 3y_2$
$y_1(0) = 7,\ y_2(0) = 7$

13. $y_1' = \frac{1}{2}y_1 - 2y_2$
$y_2' = -\frac{3}{2}y_1 + y_2$
$y_1(0) = 0.4,\ y_2(0) = 3.8$

14. $y_1' = -y_1 + 5y_2$
$y_2' = -y_1 + 3y_2$
$y_1(0) = 7,\ y_2(0) = 2$

15. $y_1' = 2y_1 + 5y_2$
$y_2' = 5y_1 + 12.5y_2$
$y_1(0) = 12,\ y_2(0) = 1$

16–17 CONVERSION

Find a general solution by conversion to a single ODE.

16. The system in Prob. 8.

17. The system in Example 5.

18. (Mixing problem, Fig. 87) Each of the two tank contains 400 gal of water, in which initially 100 l (Tank T_1) and 40 lb (Tank T_2) of fertilizer ar dissolved. The inflow, circulation, and outflow ar shown in Fig. 87. The mixture is kept uniform b stirring. Find the fertilizer contents $y_1(t)$ in T_1 and $y_2($ in T_2.

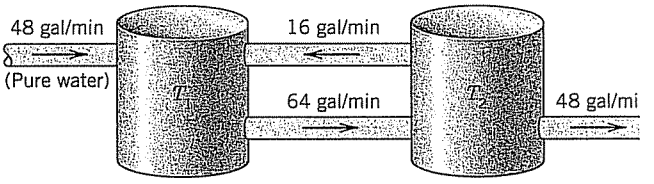


Fig. 87. Tanks in Problem 18

19. (Network) Show that a model for the currents $I_1(t)$ an $I_2(t)$ in Fig. 88 is

$$\frac{1}{C}\int I_1\,dt + R(I_1 - I_2) = 0, \qquad LI_2' + R(I_2 - I_1) =$$

Find a general solution, assuming that $R = 20$ $L = 0.5$ H, $C = 2 \cdot 10^{-4}$ F.

20. CAS PROJECT. Phase Portraits. Graph some of th figures in this section, in particular Fig. 86 on th degenerate node, in which the vector $\mathbf{y}^{(2)}$ depends c t. In each figure highlight a trajectory that satisfies a initial condition of your choice.

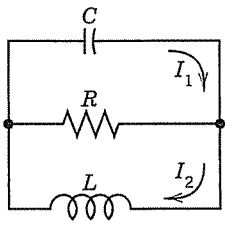


Fig. 88. Network in Problem 19

4.4 Criteria for Critical Points. Stability

We continue our discussion of homogeneous linear systems with constant coefficients

$$(1)\quad \mathbf{y}' = \mathbf{A}\mathbf{y} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \mathbf{y}, \qquad \text{in components,} \qquad \begin{aligned} y_1' &= a_{11}y_1 + a_{12}y_2 \\ y_2' &= a_{21}y_1 + a_{22}y_2. \end{aligned}$$

From the examples in the last section we have seen that we can obtain an overview of families of solution curves if we represent them parametrically as $\mathbf{y}(t) = [y_1(t) \quad y_2(t)]^T$ and graph them as curves in the y_1y_2-plane, called the **phase plane.** Such a curve is called a **trajectory** of (1), and their totality is known as the **phase portrait** of (1).

Now we have seen that solutions are of the form

$$\mathbf{y}(t) = \mathbf{x}e^{\lambda t}. \qquad \text{Substitution into (1) gives} \qquad \mathbf{y}'(t) = \lambda\mathbf{x}e^{\lambda t} = \mathbf{A}\mathbf{y} = \mathbf{A}\mathbf{x}e^{\lambda t}.$$

Dropping the common factor $e^{\lambda t}$, we have

$$(2)\qquad \mathbf{A}\mathbf{x} = \lambda\mathbf{x}.$$

Hence $\mathbf{y}(t)$ is a (nonzero) solution of (1) if λ is an eigenvalue of $\mathbf{A}$ and $\mathbf{x}$ a corresponding eigenvector.

Our examples in the last section show that the general form of the phase portrait is determined to a large extent by the type of **critical point** of the system (1) defined as a point at which dy_2/dy_1 becomes undetermined, 0/0; here [see (9) in Sec. 4.3]

$$(3)\qquad \frac{dy_2}{dy_1} = \frac{y_2'\,dt}{y_1'\,dt} = \frac{a_{21}y_1 + a_{22}y_2}{a_{11}y_1 + a_{12}y_2}.$$

We also recall from Sec. 4.3 that there are various types of critical points, and we shall now see how these types are related to the eigenvalues. The latter are solutions $\lambda = \lambda_1$ and λ_2 of the characteristic equation

$$(4)\qquad \det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{vmatrix} = \lambda^2 - (a_{11} + a_{22})\lambda + \det\mathbf{A} = 0.$$

This is a quadratic equation $\lambda^2 - p\lambda + q = 0$ with coefficients p, q and discriminant Δ given by

$$(5)\qquad p = a_{11} + a_{22}, \qquad q = \det\mathbf{A} = a_{11}a_{22} - a_{12}a_{21}, \qquad \Delta = p^2 - 4q.$$

From calculus we know that the solutions of this equation are

$$(6)\qquad \lambda_1 = \tfrac{1}{2}(p + \sqrt{\Delta}), \qquad \lambda_2 = \tfrac{1}{2}(p - \sqrt{\Delta}).$$

Furthermore, the product representation of the equation gives

$$\lambda^2 - p\lambda + q = (\lambda - \lambda_1)(\lambda - \lambda_2) = \lambda^2 - (\lambda_1 + \lambda_2)\lambda + \lambda_1\lambda_2.$$

Hence p is the sum and q the product of the eigenvalues. Also $\lambda_1 - \lambda_2 = \sqrt{\Delta}$ from Together,

$$p = \lambda_1 + \lambda_2, \qquad q = \lambda_1\lambda_2, \qquad \Delta = (\lambda_1 - \lambda_2)^2. \tag{7}$$

This gives the criteria in Table 4.1 for classifying critical points. A derivation will indicated later in this section.

Table 4.1 Eigenvalue Criteria for Critical Points (Derivation after Table 4.2)

Name	$p = \lambda_1 + \lambda_2$	$q = \lambda_1\lambda_2$	$\Delta = (\lambda_1 - \lambda_2)^2$	Comments on λ_1,
(a) Node		$q > 0$	$\Delta \geqq 0$	Real, same sign
(b) Saddle point		$q < 0$		Real, opposite sign
(c) Center	$p = 0$	$q > 0$		Pure imaginary
(d) Spiral point	$p \neq 0$		$\Delta < 0$	Complex, not pure imaginary

Stability

Critical points may also be classified in terms of their stability. Stability concepts are ba in engineering and other applications. They are suggested by physics, where **stabili** means, roughly speaking, that a small change (a small disturbance) of a physical syste at some instant changes the behavior of the system only slightly at all future times t. F critical points, the following concepts are appropriate.

DEFINITIONS

Stable, Unstable, Stable and Attractive

A critical point P_0 of (1) is called **stable**[2] if, roughly, all trajectories of (1) that at some instant are close to P_0 remain close to P_0 at all future times; precisely: if for every disk D_ϵ of radius $\epsilon > 0$ with center P_0 there is a disk D_δ of radius $\delta > 0$ with center P_0 such that every trajectory of (1) that has a point P_1 (corresponding to $t = t_1$, say) in D_δ has all its points corresponding to $t \geqq t_1$ in D_ϵ. See Fig. 89.

P_0 is called **unstable** if P_0 is not stable.

P_0 is called **stable and attractive** (or *asymptotically stable*) if P_0 is stable and every trajectory that has a point in D_δ approaches P_0 as $t \to \infty$. See Fig. 90.

Classification criteria for critical points in terms of stability are given in Table 4.2. Bot tables are summarized in the **stability chart** in Fig. 91. In this chart the region of instabilit is dark blue.

[2]In the sense of the Russian mathematician ALEXANDER MICHAILOVICH LJAPUNOV (1857–1918 whose work was fundamental in stability theory for ODEs. This is perhaps the most appropriate definition c stability (and the only we shall use), but there are others, too.

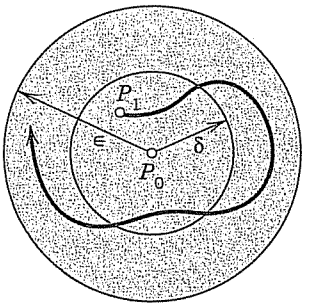


Fig. 89. Stable critical point P_0 of (1) (The trajectory initiating at P_1 stays in the disk of radius ϵ.)

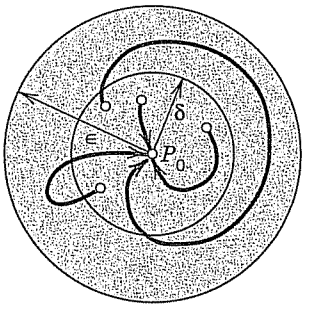


Fig. 90. Stable and attractive critical point P_0 of (1)

Table 4.2 Stability Criteria for Critical Points

Type of Stability	$p = \lambda_1 + \lambda_2$		$q = \lambda_1\lambda_2$
(a) Stable and attractive	$p < 0$		$q > 0$
(b) Stable	$p \leqq 0$		$q > 0$
(c) Unstable	$p > 0$	OR	$q < 0$

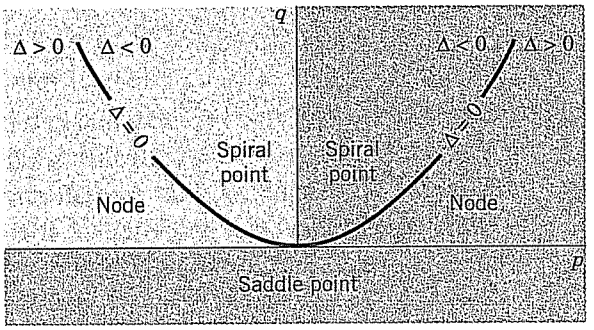


Fig. 91. Stability chart of the system (1) with p, q, Δ defined in (5).
Stable and attractive: The second quadrant without the q-axis.
Stability also on the positive q-axis (which corresponds to centers).
Unstable: Dark blue region

We indicate how the criteria in Tables 4.1 and 4.2 are obtained. If $q = \lambda_1\lambda_2 > 0$, both eigenvalues are positive or both are negative or complex conjugates. If also $p = \lambda_1 + \lambda_2 < 0$, both are negative or have a negative real part. Hence P_0 is stable and attractive. The reasoning for the other two lines in Table 4.2 is similar.

If $\Delta < 0$, the eigenvalues are complex conjugates, say, $\lambda_1 = \alpha + i\beta$ and $\lambda_2 = \alpha - i\beta$. If also $p = \lambda_1 + \lambda_2 = 2\alpha < 0$, this gives a spiral point that is stable and attractive. If $p = 2\alpha > 0$, this gives an unstable spiral point.

If $p = 0$, then $\lambda_2 = -\lambda_1$ and $q = \lambda_1\lambda_2 = -\lambda_1^2$. If also $q > 0$, then $\lambda_1^2 = -q < 0$, so that λ_1, and thus λ_2, must be pure imaginary. This gives periodic solutions, their trajectories being closed curves around P_0, which is a center.

EXAMPLE 1 Application of the Criteria in Tables 4.1 and 4.2

In Example 1, Sec. 4.3, we have $\mathbf{y}' = \begin{bmatrix} -3 & 1 \\ 1 & -3 \end{bmatrix} \mathbf{y}$, $p = -6$, $q = 8$, $\Delta = 4$, a node by Table 4.1(a), which is stable and attractive by Table 4.2(a). ▩

EXAMPLE 2 Free Motions of a Mass on a Spring

What kind of critical point does $my'' + cy' + ky = 0$ in Sec. 2.4 have?

Solution. Division by m gives $y'' = -(k/m)y - (c/m)y'$. To get a system, set $y_1 = y$, $y_2 = y'$ (see 4.1). Then $y_2' = y'' = -(k/m)y_1 - (c/m)y_2$. Hence

$$\mathbf{y}' = \begin{bmatrix} 0 & 1 \\ -k/m & -c/m \end{bmatrix} \mathbf{y}, \qquad \det(\mathbf{A} - \lambda \mathbf{I}) = \begin{vmatrix} -\lambda & 1 \\ -k/m & -c/m - \lambda \end{vmatrix} = \lambda^2 + \frac{c}{m}\lambda + \frac{k}{m} = 0.$$

We see that $p = -c/m$, $q = k/m$, $\Delta = (c/m)^2 - 4k/m$. From this and Tables 4.1 and 4.2 we obtain the follc results. Note that in the last three cases the discriminant Δ plays an essential role.

No damping. $c = 0$, $p = 0$, $q > 0$, a center.
Underdamping. $c^2 < 4mk$, $p < 0$, $q > 0$, $\Delta < 0$, a stable and attractive spiral point.
Critical damping. $c^2 = 4mk$, $p < 0$, $q > 0$, $\Delta = 0$, a stable and attractive node.
Overdamping. $c^2 > 4mk$, $p < 0$, $q > 0$, $\Delta > 0$, a stable and attractive node.

PROBLEM SET 4.4

1–9 TYPE AND STABILITY OF CRITICAL POINT

Determine the type and stability of the critical point. Then find a real general solution and sketch or graph some of the trajectories in the phase plane. (Show the details of your work.)

1. $y_1' = 2y_2$
   $y_2' = 8y_1$

2. $y_1' = 4y_1$
   $y_2' = 3y_2$

3. $y_1' = 2y_1 + y_2$
   $y_2' = y_1 + 2y_2$

4. $y_1' = y_2$
   $y_2' = -5y_1 - 2y_2$

5. $y_1' = -4y_1 + y_2$
   $y_2' = y_1 - 4y_2$

6. $y_1' = y_1 + 10y_2$
   $y_2' = 7y_1 - 8y_2$

7. $y_1' = -2y_2$
   $y_2' = 8y_1$

8. $y_1' = 3y_1 + 5y_2$
   $y_2' = -5y_1 - 3y_2$

9. $y_1' = y_1 + 2y_2$
   $y_2' = 2y_1 + y_2$

10–12 FORM OF TRAJECTORIES

What kind of curves are the trajectories of the following ODEs in the phase plane?

10. $y'' + 5y' = 0$
11. $y'' - k^2y = 0$
12. $y'' + \frac{1}{16}y = 0$

13. **(Damped oscillation)** Solve $y'' + 4y' + 5y = 0$. What kind of curves do you get as trajectories?

14. **(Transformation of variable)** What happens to system (1) and its critical point if you introduce $\tau =$ as a new independent variable?

15. **(Types of critical points)** Discuss the critical poini (10)–(14) in Sec. 4.3 by applying the criteria in Tal 4.1 and 4.2 in this section.

16. **(Perturbation of center)** If a system has a cente its critical point, what happens if you replace the ma $\mathbf{A}$ by $\tilde{\mathbf{A}} = \mathbf{A} + k\mathbf{I}$ with any real number $k \neq$ (representing measurement errors in the diagc entries)?

17. **(Perturbation)** The system in Example 4 in Sec. has a center as its critical point. Replace each a_{jk} Example 4, Sec. 4.3, by $a_{jk} + b$. Find values of b s that you get (a) a saddle point, (b) a stable and attract node, (c) a stable and attractive spiral, (d) an unsta spiral, (e) an unstable node.

18. **CAS EXPERIMENT. Phase Portraits.** Graph ph portraits for the systems in Prob. 17 with the values b suggested in the answer. Try to illustrate how the ph portrait changes "continuously" under a continuc change of b.

19. **WRITING EXPERIMENT. Stability.** Stabil concepts are basic in physics and engineering. Write two-part report of 3 pages each (A) on gene applications in which stability plays a role (be precise as you can), and (B) on material related stability in this section. Use your own formulations a examples; do not copy.

20. **(Stability chart)** Locate the critical points of t systems (10)–(14) in Sec. 4.3 and of Probs. 1, 3, 5 this problem set on the stability chart.

4.5 Qualitative Methods for Nonlinear Systems

Qualitative methods are methods of obtaining qualitative information on solutions without actually solving a system. These methods are particularly valuable for systems whose solution by analytic methods is difficult or impossible. This is the case for many practically important **nonlinear systems**

$$\mathbf{y}' = \mathbf{f}(\mathbf{y}), \quad \text{thus} \quad \begin{aligned} y_1' &= f_1(y_1, y_2) \\ y_2' &= f_2(y_1, y_2). \end{aligned} \tag{1}$$

In this section we extend **phase plane methods,** as just discussed, from linear systems to nonlinear systems (1). We assume that (1) is **autonomous,** that is, the independent variable t does not occur explicitly. (All examples in the last section are autonomous.) We shall again exhibit entire families of solutions. This is an advantage over numeric methods, which give only one (approximate) solution at a time.

Concepts needed from the last section are the **phase plane** (the y_1y_2-plane), **trajectories** (solution curves of (1) in the phase plane), the **phase portrait** of (1) (the totality of these trajectories), and **critical points** of (1) (points (y_1, y_2) at which both $f_1(y_1, y_2)$ and $f_2(y_1, y_2)$ are zero).

Now (1) may have several critical points. Then we discuss one after another. As a technical convenience, each time we first move the critical point P_0: (a, b) to be considered to the origin (0, 0). This can be done by a translation

$$\tilde{y}_1 = y_1 - a, \qquad \tilde{y}_2 = y_2 - b$$

which moves P_0 to (0, 0). Thus we can assume P_0 to be the origin (0, 0), and for simplicity we continue to write y_1, y_2 (instead of $\tilde{y}_1$, $\tilde{y}_2$). We also assume that P_0 is **isolated,** that is, it is the only critical point of (1) within a (sufficiently small) disk with center at the origin. If (1) has only finitely many critical points, this is automatically true. (Explain!)

Linearization of Nonlinear Systems

How can we determine the kind and stability property of a critical point P_0: (0, 0) of (1)? In most cases this can be done by **linearization** of (1) near P_0, writing (1) as $\mathbf{y}' = \mathbf{f}(\mathbf{y}) = \mathbf{A}\mathbf{y} + \mathbf{h}(\mathbf{y})$ and dropping $\mathbf{h}(\mathbf{y})$, as follows.

Since P_0 is critical, $f_1(0, 0) = 0$, $f_2(0, 0) = 0$, so that f_1 and f_2 have no constant terms and we can write

$$\mathbf{y}' = \mathbf{A}\mathbf{y} + \mathbf{h}(\mathbf{y}), \quad \text{thus} \quad \begin{aligned} y_1' &= a_{11}y_1 + a_{12}y_2 + h_1(y_1, y_2) \\ y_2' &= a_{21}y_1 + a_{22}y_2 + h_2(y_1, y_2). \end{aligned} \tag{2}$$

$\mathbf{A}$ is constant (independent of t) since (1) is autonomous. One can prove the following (proof in Ref. [A7], pp. 375–388, listed in App. 1).

THEOREM 1

Linearization

If f_1 and f_2 in (1) *are continuous and have continuous partial derivatives in a neighborhood of the critical point P_0:* (0, 0), *and if* det $\mathbf{A} \neq 0$ *in* (2), *then the kind and stability of the critical point of* (1) *are the same as those of the* **linearized system**

$$\text{(3)} \qquad \mathbf{y}' = \mathbf{A}\mathbf{y}, \qquad \text{thus} \qquad \begin{aligned} y_1' &= a_{11}y_1 + a_{12}y_2 \\ y_2' &= a_{21}y_1 + a_{22}y_2. \end{aligned}$$

Exceptions occur if $\mathbf{A}$ *has equal or pure imaginary eigenvalues; then* (1) *may have the same kind of critical point as* (3) *or a spiral point.*

EXAMPLE 1 **Free Undamped Pendulum. Linearization**

Figure 92a shows a pendulum consisting of a body of mass m (the bob) and a rod of length L. Determine locations and types of the critical points. Assume that the mass of the rod and air resistance are negligible

Solution. ***Step 1. Setting up the mathematical model.*** Let θ denote the angular displacement, measu counterclockwise from the equilibrium position. The weight of the bob is mg (g the acceleration of gravity causes a restoring force $mg \sin\theta$ tangent to the curve of motion (circular arc) of the bob. By Newton's sec law, at each instant this force is balanced by the force of acceleration $mL\theta''$, where $L\theta''$ is the accelerat hence the resultant of these two forces is zero, and we obtain as the mathematical model

$$mL\theta'' + mg \sin\theta = 0.$$

Dividing this by mL, we have

$$\text{(4)} \qquad \theta'' + k \sin\theta = 0 \qquad \left(k = \frac{g}{L}\right.$$

When θ is very small, we can approximate $\sin\theta$ rather accurately by θ and obtain as an *approximate* solu $A\cos\sqrt{k}t + B\sin\sqrt{k}t$, but the *exact* solution for any θ is not an elementary function.

Step 2. Critical points **(0, 0), ±(2π, 0), ±(4π, 0), · · ·** ***, Linearization.*** To obtain a system of ODEs, we $\theta = y_1$, $\theta' = y_2$. Then from (4) we obtain a nonlinear system (1) of the form

$$\text{(4*)} \qquad \begin{aligned} y_1' &= f_1(y_1, y_2) = y_2 \\ y_2' &= f_2(y_1, y_2) = -k \sin y_1. \end{aligned}$$

The right sides are both zero when $y_2 = 0$ and $\sin y_1 = 0$. This gives infinitely many critical points ($n\pi$, where $n = 0, \pm1, \pm2, \cdots$. We consider (0, 0). Since the Maclaurin series is

$$\sin y_1 = y_1 - \tfrac{1}{6}y_1^3 + - \cdots \approx y_1,$$

the linearized system at (0, 0) is

$$\mathbf{y}' = \mathbf{A}\mathbf{y} = \begin{bmatrix} 0 & 1 \\ -k & 0 \end{bmatrix}\mathbf{y}, \qquad \text{thus} \qquad \begin{aligned} y_1' &= y_2 \\ y_2' &= -ky_1. \end{aligned}$$

To apply our criteria in Sec. 4.4 we calculate $p = a_{11} + a_{22} = 0$, $q = \det\mathbf{A} = k = g/L$ (> 0), $\Delta = p^2 - 4q = -4k$. From this and Table 4.1(c) in Sec. 4.4 we conclude that (0, 0) is a center, which is alw stable. Since $\sin\theta = \sin y_1$ is periodic with period 2π, the critical points ($n\pi$, 0), $n = \pm2, \pm4, \cdots$, are centers.

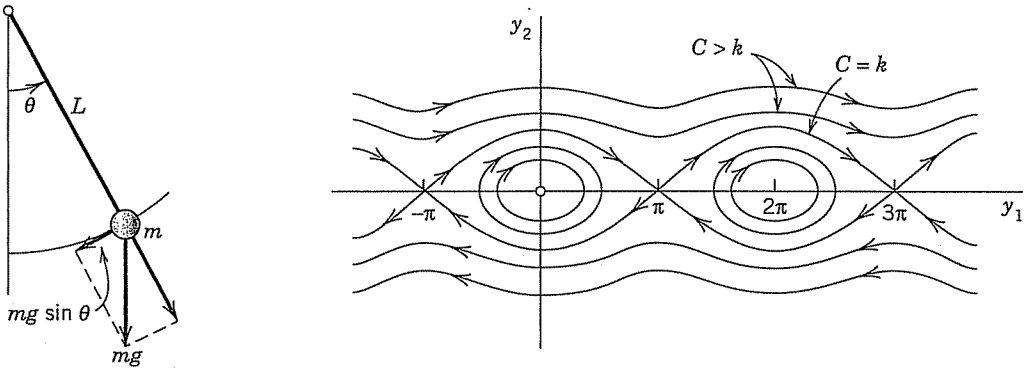
Step 3. Critical points **±(π, 0), ±(3π, 0), ±(5π, 0), · · ·** ***, Linearization.*** We now consider the critical p (π, 0), setting $\theta - \pi = y_1$ and $(\theta - \pi)' = \theta' = y_2$. Then in (4),

$$\sin\theta = \sin(y_1 + \pi) = -\sin y_1 = -y_1 + \tfrac{1}{6}y_1^3 - + \cdots \approx -y_1$$

and the linearized system at $(\pi, 0)$ is now

$$\mathbf{y}' = \mathbf{A}\mathbf{y} = \begin{bmatrix} 0 & 1 \\ k & 0 \end{bmatrix} \mathbf{y}, \qquad \text{thus} \qquad \begin{aligned} y_1' &= y_2 \\ y_2' &= k y_1. \end{aligned}$$

We see that $p = 0$, $q = -k\ (< 0)$, and $\Delta = -4q = 4k$. Hence, by Table 4.1(b), this gives a saddle point, which is always unstable. Because of periodicity, the critical points $(n\pi, 0)$, $n = \pm 1, \pm 3, \cdots$, are all saddle points. These results agree with the impression we get from Fig. 92b. ▩



(a) Pendulum

(b) Solution curves $y_2(y_1)$ of (4) in the phase plane

Fig. 92. Example 1 (C will be explained in Example 4.)

EXAMPLE 2 Linearization of the Damped Pendulum Equation

To gain further experience in investigating critical points, as another practically important case, let us see how Example 1 changes when we add a damping term $c\theta'$ (damping proportional to the angular velocity) to equation (4), so that it becomes

$$\theta'' + c\theta' + k \sin \theta = 0 \tag{5}$$

where $k > 0$ and $c \geqq 0$ (which includes our previous case of no damping, $c = 0$). Setting $\theta = y_1$, $\theta' = y_2$, as before, we obtain the nonlinear system (use $\theta'' = y_2'$)

$$\begin{aligned} y_1' &= y_2 \\ y_2' &= -k \sin y_1 - c y_2. \end{aligned}$$

We see that the critical points have the same locations as before, namely, $(0, 0)$, $(\pm\pi, 0)$, $(\pm 2\pi, 0)$, $\cdots$. We consider $(0, 0)$. Linearizing $\sin y_1 \approx y_1$ as in Example 1, we get the linearized system at $(0, 0)$

$$\mathbf{y}' = \mathbf{A}\mathbf{y} = \begin{bmatrix} 0 & 1 \\ -k & -c \end{bmatrix} \mathbf{y}, \qquad \text{thus} \qquad \begin{aligned} y_1' &= y_2 \\ y_2' &= -k y_1 - c y_2. \end{aligned} \tag{6}$$

This is identical with the system in Example 2 of Sec 4.4, except for the (positive!) factor m (and except for the physical meaning of y_1). Hence for $c = 0$ (no damping) we have a center (see Fig. 92b), for small damping we have a spiral point (see Fig. 93), and so on.

We now consider the critical point $(\pi, 0)$. We set $\theta - \pi = y_1$, $(\theta - \pi)' = \theta' = y_2$ and linearize

$$\sin \theta = \sin (y_1 + \pi) = -\sin y_1 \approx -y_1.$$

This gives the new linearized system at $(\pi, 0)$

$$\mathbf{y}' = \mathbf{A}\mathbf{y} = \begin{bmatrix} 0 & 1 \\ k & -c \end{bmatrix} \mathbf{y}, \qquad \text{thus} \qquad \begin{aligned} y_1' &= y_2 \\ y_2' &= k y_1 - c y_2. \end{aligned} \tag{6*}$$

For our criteria in Sec 4.4 we calculate $p = a_{11} + a_{22} = -c$, $q = \det \mathbf{A} = -k$, and $\Delta = p^2 - 4q = c^2 + 4k$. This gives the following results for the critical point at $(\pi, 0)$.

No damping. $c = 0$, $p = 0$, $q < 0$, $\Delta > 0$, a saddle point. See Fig. 92b.
Damping. $c > 0$, $p < 0$, $q < 0$, $\Delta > 0$, a saddle point. See Fig. 93.

Since $\sin y_1$ is periodic with period 2π, the critical points $(\pm 2\pi, 0)$, $(\pm 4\pi, 0)$, $\cdots$ are of the same type as $(0, 0)$, and the critical points $(-\pi, 0)$, $(\pm 3\pi, 0)$, $\cdots$ are of the same type as $(\pi, 0)$, so that our task is finished.

Figure 93 shows the trajectories in the case of damping. What we see agrees with our physical intuition. Indeed, damping means loss of energy. Hence instead of the closed trajectories of periodic solutions in Fig. 92b we now have trajectories spiraling around one of the critical points $(0, 0)$, $(\pm 2\pi, 0)$, $\cdots$. Even the wavy trajectories corresponding to whirly motions eventually spiral around one of these points. Furthermore, there are no more trajectories that connect critical points (as there were in the undamped case for the saddle points). ▩

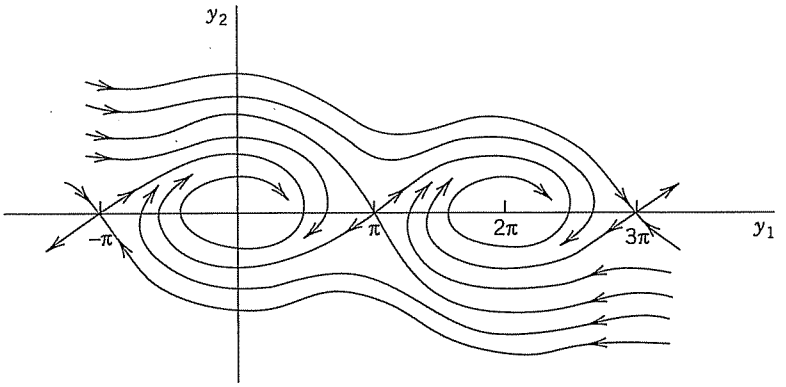


Fig. 93. Trajectories in the phase plane for the damped pendulum in Example 2

Lotka–Volterra Population Model

EXAMPLE 3 **Predator–Prey Population Model**[3]

This model concerns two species, say, rabbits and foxes, and the foxes prey on the rabbits.

Step 1. Setting up the model. We assume the following.

1. Rabbits have unlimited food supply. Hence if there were no foxes, their number $y_1(t)$ would grow exponentially, $y_1' = ay_1$.
2. Actually, y_1 is decreased because of the kill by foxes, say, at a rate proportional to $y_1 y_2$, where $y_2(t)$ is the number of foxes. Hence $y_1' = ay_1 - by_1y_2$, where $a > 0$ and $b > 0$.
3. If there were no rabbits, then $y_2(t)$ would exponentially decrease to zero, $y_2' = -ly_2$. However, y_2 is increased by a rate proportional to the number of encounters between predator and prey; together we have $y_2' = -ly_2 + ky_1y_2$, where $k > 0$ and $l > 0$.

This gives the (nonlinear!) Lotka–Volterra system

$$\begin{aligned} y_1' &= f_1(y_1, y_2) = ay_1 - by_1y_2 \\ y_2' &= f_2(y_1, y_2) = ky_1y_2 - ly_2 . \end{aligned} \tag{7}$$

[3]Introduced by ALFRED J. LOTKA (1880–1949), American biophysicist, and VITO VOLTERRA (1860–1940), Italian mathematician, the initiator of functional analysis (see [GR7] in App. 1).

Step 2. Critical point (0, 0), Linearization. We see from (7) that the critical points are the solutions of

$$(7^*) \qquad f_1(y_1, y_2) = y_1(a - by_2) = 0, \qquad f_2(y_1, y_2) = y_2(ky_1 - l) = 0.$$

The solutions are $(y_1, y_2) = (0, 0)$ and $(\frac{l}{k}, \frac{a}{b})$. We consider (0, 0). Dropping $-by_1y_2$ and ky_1y_2 from (7) gives the linearized system

$$\mathbf{y}' = \begin{bmatrix} a & 0 \\ 0 & -l \end{bmatrix} \mathbf{y}.$$

Its eigenvalues are $\lambda_1 = a > 0$ and $\lambda_2 = -l < 0$. They have opposite signs, so that we get a saddle point.

Step 3. Critical point (l/k, a/b), Linearization. We set $y_1 = \tilde{y}_1 + l/k$, $y_2 = \tilde{y}_2 + a/b$. Then the critical point $(l/k, a/b)$ corresponds to $(\tilde{y}_1, \tilde{y}_2) = (0, 0)$. Since $\tilde{y}_1' = y_1'$, $\tilde{y}_2' = y_2'$, we obtain from (7) [factorized as in (8)]

$$\tilde{y}_1' = \left(\tilde{y}_1 + \frac{l}{k}\right)\left[a - b\left(\tilde{y}_2 + \frac{a}{b}\right)\right] = \left(\tilde{y}_1 + \frac{l}{k}\right)(-b\tilde{y}_2)$$

$$\tilde{y}_2' = \left(\tilde{y}_2 + \frac{a}{b}\right)\left[k\left(\tilde{y}_1 + \frac{l}{k}\right) - l\right] = \left(\tilde{y}_2 + \frac{a}{b}\right)k\tilde{y}_1.$$

Dropping the two nonlinear terms $-b\tilde{y}_1\tilde{y}_2$ and $k\tilde{y}_1\tilde{y}_2$, we have the linearized system

$$(7^{**}) \qquad \text{(a)} \quad \tilde{y}_1' = -\frac{lb}{k}\tilde{y}_2$$

$$\text{(b)} \quad \tilde{y}_2' = \frac{ak}{b}\tilde{y}_1.$$

The left side of (a) times the right side of (b) must equal the right side of (a) times the left side of (b),

$$\frac{ak}{b}\tilde{y}_1\tilde{y}_1' = -\frac{lb}{k}\tilde{y}_2\tilde{y}_2' \qquad \text{By integration,} \qquad \frac{ak}{b}\tilde{y}_1^2 + \frac{lb}{k}\tilde{y}_2^2 = const.$$

This is a family ellipses, so that the critical point $(l/k, a/b)$ of the linearized system (7**) is a center (Fig. 94). It can be shown by a complicated analysis that the nonlinear system (7) also has a center (rather than a spiral point) at $(l/k, a/b)$ surrounded by closed trajectories (not ellipses).

We see that the predators and prey have a cyclic variation about the critical point. Let us move counterclockwise around the ellipse, beginning at the right vertex, where the rabbits have a maximum number. Foxes are sharply increasing in number until they reach a maximum at the upper vertex, and the number of rabbits is then sharply decreasing until it reaches a minimum at the left vertex, and so on. Cyclic variations of this kind have been observed in nature, for example, for lynx and snowshoe hare near the Hudson Bay, with a cycle of about 10 years.

For models of more complicated situations and a systematic discussion, see C. W. Clark, *Mathematical Bioeconomics* (Wiley, 1976). ▩

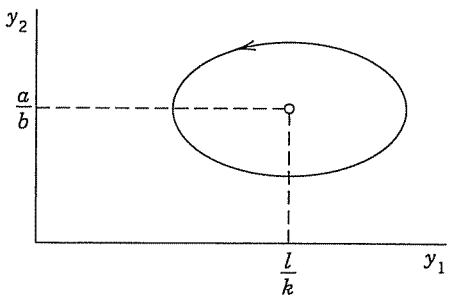


Fig. 94. Ecological equilibrium point and trajectory of the linearized Lotka–Volterra system (7**)

Transformation to a First-Order Equation in the Phase Plane

Another phase plane method is based on the idea of transforming a second-order **autonomous ODE** (an ODE in which t does not occur explicitly)

$$F(y, y', y'') = 0$$

to first order by taking $y = y_1$ as the independent variable, setting $y' = y_2$ and transforming y'' by the chain rule,

$$y'' = y_2' = \frac{dy_2}{dt} = \frac{dy_2}{dy_1}\frac{dy_1}{dt} = \frac{dy_2}{dy_1}\, y_2.$$

Then the ODE becomes of first order,

$$F\left(y_1, y_2, \frac{dy_2}{dy_1}\, y_2\right) = 0 \tag{8}$$

and can sometimes be solved or treated by direction fields. We illustrate this for the equation in Example 1 and shall gain much more insight into the behavior of solutions.

EXAMPLE 4 **An ODE (8) for the Free Undamped Pendulum**

If in (4) $\theta'' + k \sin\theta = 0$ we set $\theta = y_1$, $\theta' = y_2$ (the angular velocity) and use

$$\theta'' = \frac{dy_2}{dt} = \frac{dy_2}{dy_1}\frac{dy_1}{dt} = \frac{dy_2}{dy_1} y_2, \qquad \text{we get} \qquad \frac{dy_2}{dy_1}\, y_2 = -k \sin y_1.$$

Separation of variables gives $y_2\, dy_2 = -k \sin y_1\, dy_1$. By integration,

$$\tfrac{1}{2}{y_2}^2 = k \cos y_1 + C \qquad (C \text{ constant}) \tag{9}$$

Multiplying this by mL^2, we get

$$\tfrac{1}{2}m(Ly_2)^2 - mL^2 k \cos y_1 = mL^2 C.$$

We see that these three terms are **energies.** Indeed, y_2 is the angular velocity, so that Ly_2 is the velocity and th first term is the kinetic energy. The second term (including the minus sign) is the potential energy of the pendulum and mL^2C is its total energy, which is constant, as expected from the law of conservation of energy, becaus there is no damping (no loss of energy). The type of motion depends on the total energy, hence on C, as follows

Figure 92b on p. 153 shows trajectories for various values of C. These graphs continue periodically wit period 2π to the left and to the right. We see that some of them are ellipse-like and closed, others are wavy and there are two trajectories (passing through the saddle points $(n\pi, 0)$, $n = \pm 1, \pm 3, \cdots$) that separat those two types of trajectories. From (9) we see that the smallest possible C is $C = -k$; then $y_2 = 0$, an $\cos y_1 = 1$, so that the pendulum is at rest. The pendulum will change its direction of motion if there are point at which $y_2 = \theta' = 0$. Then $k \cos y_1 + C = 0$ by (9). If $y_1 = \pi$, then $\cos y_1 = -1$ and $C = k$. Hence $-k < C < k$, then the pendulum reverses its direction for a $|y_1| = |\theta| < \pi$, and for these values of C wit $|C| < k$ the pendulum oscillates. This corresponds to the closed trajectories in the figure. However, if $C >$ then $y_2 = 0$ is impossible and the pendulum makes a whirly motion that appears as a wavy trajectory in th y_1y_2-plane. Finally, the value $C = k$ corresponds to the two "separating trajectories" in Fig. 92b connecting th saddle points. ∎

The phase plane method of deriving a single first-order equation (8) may be of practical interes not only when (8) can be solved (as in Example 4) but also when solution is not possible an we have to utilize direction fields (Sec. 1.2). We illustrate this with a very famous example

EXAMPLE 5 **Self-Sustained Oscillations. Van der Pol Equation**

There are physical systems such that for small oscillations, energy is fed into the system, whereas for large oscillations, energy is taken from the system. In other words, large oscillations will be damped, whereas for small oscillations there is "negative damping" (feeding of energy into the system). For physical reasons we expect such a system to approach a periodic behavior, which will thus appear as a closed trajectory in the phase plane, called a **limit cycle.** A differential equation describing such vibrations is the famous **van der Pol equation**[4]

$$y'' - \mu(1 - y^2)y' + y = 0 \qquad (\mu > 0, \text{ constant}). \tag{10}$$

It first occurred in the study of electrical circuits containing vacuum tubes. For $\mu = 0$ this equation becomes $y'' + y = 0$ and we obtain harmonic oscillations. Let $\mu > 0$. The damping term has the factor $-\mu(1 - y^2)$. This is negative for small oscillations, when $y^2 < 1$, so that we have "negative damping," is zero for $y^2 = 1$ (no damping), and is positive if $y^2 > 1$ (positive damping, loss of energy). If μ is small, we expect a limit cycle that is almost a circle because then our equation differs but little from $y'' + y = 0$. If μ is large, the limit cycle will probably look different.

Setting $y = y_1$, $y' = y_2$ and using $y'' = (dy_2/dy_1)y_2$ as in (8), we have from (10)

$$\frac{dy_2}{dy_1}\, y_2 - \mu(1 - y_1^2)y_2 + y_1 = 0. \tag{11}$$

The isoclines in the y_1y_2-plane (the phase plane) are the curves $dy_2/dy_1 = K = const$, that is,

$$\frac{dy_2}{dy_1} = \mu(1 - y_1^2) - \frac{y_1}{y_2} = K.$$

Solving algebraically for y_2, we see that the isoclines are given by

$$y_2 = \frac{y_1}{\mu(1 - y_1^2) - K} \qquad \text{(Figs. 95, 96).}$$

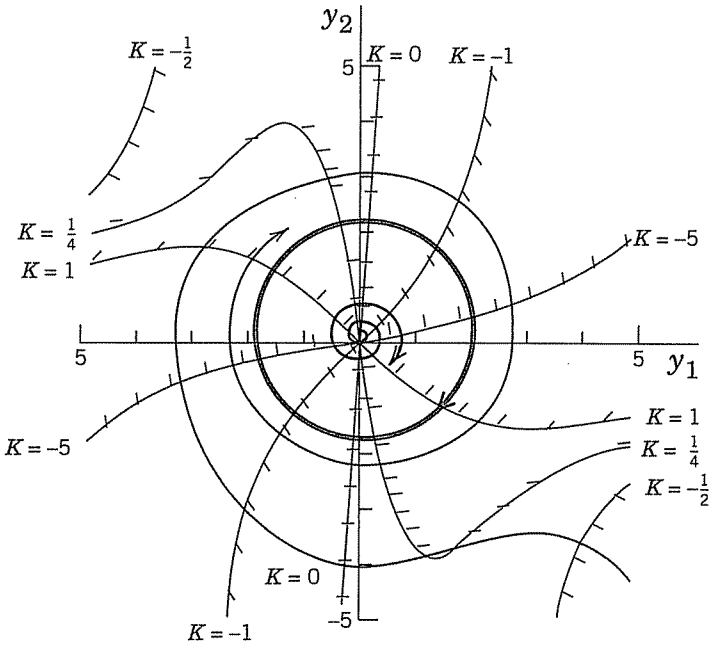


Fig. 95. Direction field for the van der Pol equation with $\mu = 0.1$ in the phase plane, showing also the limit cycle and two trajectories. See also Fig. 8 in Sec. 1.2.

[4]BALTHASAR VAN DER POL (1889–1959), Dutch physicist and engineer.

Figure 95 shows some isoclines when μ is small, $\mu = 0.1$, the limit cycle (almost a circle), and two (blu trajectories approaching it, one from the outside and the other from the inside, of which only the initial portic a small spiral, is shown. Due to this approach by trajectories, a limit cycle differs conceptually from a clos curve (a trajectory) surrounding a center, which is not approached by trajectories. For larger μ the limit cyc no longer resembles a circle, and the trajectories approach it more rapidly than for smaller μ. Figure 96 illustrat this for $\mu = 1$.

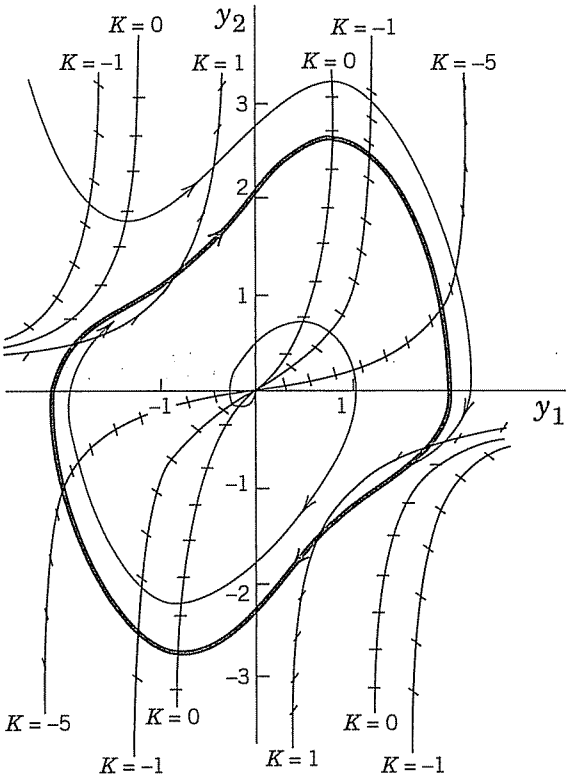


Fig. 96. Direction field for the van der Pol equation with $\mu = 1$ in the phase plane, showing also the limit cycle and two trajectories approaching it

PROBLEM SET 4.5

1–12 CRITICAL POINTS, LINEARIZATION

Determine the location and type of all critical points by linearization. In Probs. 7–12 first transform the ODE to a system. (Show the details of your work.)

1. $y_1' = y_2 + {y_2}^2$
$y_2' = 3y_1$

2. $y_1' = 4y_1 - {y_1}^2$
$y_2' = y_2$

3. $y_1' = 4y_2$
$y_2' = 2y_1 - {y_1}^2$

4. $y_1' = -3y_1 + y_2 - {y_2}^2$
$y_2' = y_1 - 3y_2$

5. $y_1' = -y_1 + y_2 - {y_2}^2$
$y_2' = -y_1 - y_2$

6. $y_1' = y_2 - {y_2}^2$
$y_2' = y_1 - {y_1}^2$

7. $y'' + y - 4y^2 = 0$

8. $y'' + 9y + y^2 = 0$

9. $y'' + \cos y = 0$

10. $y'' + \sin y = 0$

11. $y'' + 4y - y^3 = 0$

12. $y'' + y' + 2y - y^2 =$

13. (Trajectories) What kind of curves are the trajectori of $yy'' + 2y'^2 = 0$?

14. (Trajectories) Write the ODE $y'' - 4y + y^3 = 0$ as system, solve it for y_2 as a function of y_1, and sketc or graph some of the trajectories in the phase plane.

15. (Trajectories) What is the radius of a real gener solution of $y'' + y = 0$ in the phase plane?

16. (Trajectories) In Prob. 14 add a linear damping ter to get $y'' + 2y' - 4y + y^3 = 0$. Using arguments fro mechanics and a comparison with Prob. 14, as well a with Examples 1 and 2, guess the type of each critic point. Then determine these types by linearizatio (Show all details of your work.)

17. **(Pendulum)** To what state (position, speed, direction of motion) do the four points of intersection of a closed trajectory with the axes in Fig. 92b correspond? The point of intersection of a wavy curve with the y_2-axis?

18. **(Limit cycle)** What is the essential difference between a limit cycle and a closed trajectory surrounding a center?

19. **CAS EXPERIMENT. Deformation of Limit Cycle.** Convert the van der Pol equation to a system. Graph the limit cycle and some approaching trajectories for $\mu = 0.2, 0.4, 0.6, 0.8, 1.0, 1.5, 2.0$. Try to observe how the limit cycle changes its form continuously if you vary μ continuously. Describe in words how the limit cycle is deformed with growing μ.

20. **TEAM PROJECT. Self-sustained oscillations.** **(a) Van der Pol Equation.** Determine the type of the critical point at (0, 0) when $\mu > 0$, $\mu = 0$, $\mu < 0$. Show that if $\mu \to 0$, the isoclines approach straight lines through the origin. Why is this to be expected?

(b) Rayleigh equation. Show that the so-called Rayleigh equation[5]

$$Y'' - \mu(1 - \tfrac{1}{3}Y'^2)Y' + Y = 0 \quad (\mu > 0)$$

also describes self-sustained oscillations and that by differentiating it and setting $y = Y'$ one obtains the van der Pol equation.

(c) Duffing equation. The Duffing equation is

$$y'' + \omega_0^2 y + \beta y^3 = 0$$

where usually $|\beta|$ is small, thus characterizing a small deviation of the restoring force from linearity. $\beta > 0$ and $\beta < 0$ are called the cases of a *hard spring* and a *soft spring*, respectively. Find the equation of the trajectories in the phase plane. (Note that for $\beta > 0$ all these curves are closed.)

4.6 Nonhomogeneous Linear Systems of ODEs

In this last section of Chap. 4 we discuss methods for solving nonhomogeneous linear systems of ODEs

$$\mathbf{y}' = \mathbf{A}\mathbf{y} + \mathbf{g} \tag{1}$$

(see Sec. 4.2)

where the vector $\mathbf{g}(t)$ is not identically zero. We assume $\mathbf{g}(t)$ and the entries of the $n \times n$ matrix $\mathbf{A}(t)$ to be continuous on some interval J of the t-axis. From a general solution $\mathbf{y}^{(h)}(t)$ of the homogeneous system $\mathbf{y}' = \mathbf{A}\mathbf{y}$ on J and a **particular solution** $\mathbf{y}^{(p)}(t)$ of (1) on J [i.e., a solution of (1) containing no arbitrary constants], we get a solution of (1),

$$\mathbf{y} = \mathbf{y}^{(h)} + \mathbf{y}^{(p)}. \tag{2}$$

$\mathbf{y}$ is called a **general solution** of (1) on J because it includes every solution of (1) on J. This follows from Theorem 2 in Sec. 4.2 (see Prob. 1 of this section).

Having studied homogeneous linear systems in Secs. 4.1–4.4, our present task will be to explain methods for obtaining particular solutions of (1). We discuss the method of undetermined coefficients and the method of the variation of parameters; these have counterparts for a single ODE, as we know from Secs. 2.7 and 2.10.

---

[5]LORD RAYLEIGH (JOHN WILLIAM STRUTT) (1842–1919), great English physicist and mathematician, professor at Cambridge and London, known by his important contributions to the theory of waves, elasticity theory, hydrodynamics, and various other branches of applied mathematics and theoretical physics. In 1904 he received the Nobel Prize in physics.

Method of Undetermined Coefficients

As for a single ODE, this method is suitable if the entries of $\mathbf{A}$ are constants and components of $\mathbf{g}$ are constants, positive integer powers of t, exponential functions, cosines and sines. In such a case a particular solution $\mathbf{y}^{(p)}$ is assumed in a form similar to $\mathbf{g}$; for instance, $\mathbf{y}^{(p)} = \mathbf{u} + \mathbf{v}t + \mathbf{w}t^2$ if $\mathbf{g}$ has components quadratic in t, with $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$ to be determined by substitution into (1). This is similar to Sec. 2.7, except for the Modification Rule. It suffices to show this by an example.

EXAMPLE 1 **Method of Undetermined Coefficients. Modification Rule**

Find a general solution of

$$\mathbf{y}' = \mathbf{A}\mathbf{y} + \mathbf{g} = \begin{bmatrix} -3 & 1 \\ 1 & -3 \end{bmatrix} \mathbf{y} + \begin{bmatrix} -6 \\ 2 \end{bmatrix} e^{-2t}. \tag{3}$$

Solution. A general equation of the homogeneous system is (see Example 1 in Sec. 4.3)

$$\mathbf{y}^{(h)} = c_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} e^{-2t} + c_2 \begin{bmatrix} 1 \\ -1 \end{bmatrix} e^{-4t}. \tag{4}$$

Since $\lambda = -2$ is an eigenvalue of $\mathbf{A}$, the function e^{-2t} on the right also appears in $\mathbf{y}^{(h)}$, and we must apply the Modification Rule by setting

$$\mathbf{y}^{(p)} = \mathbf{u}te^{-2t} + \mathbf{v}e^{-2t} \qquad \text{(rather than } \mathbf{u}e$$

Note that the first of these two terms is the analog of the modification in Sec. 2.7, but it would not be sufficient here. (Try it.) By substitution,

$$\mathbf{y}^{(p)\prime} = \mathbf{u}e^{-2t} - 2\mathbf{u}te^{-2t} - 2\mathbf{v}e^{-2t} = \mathbf{A}\mathbf{u}te^{-2t} + \mathbf{A}\mathbf{v}e^{-2t} + \mathbf{g}.$$

Equating the te^{-2t}-terms on both sides, we have $-2\mathbf{u} = \mathbf{A}\mathbf{u}$. Hence $\mathbf{u}$ is an eigenvector of $\mathbf{A}$ corresponding to $\lambda = -2$; thus [see (5)] $\mathbf{u} = a[1 \quad 1]^T$ with any $a \neq 0$. Equating the other terms gives

$$\mathbf{u} - 2\mathbf{v} = \mathbf{A}\mathbf{v} + \begin{bmatrix} -6 \\ 2 \end{bmatrix} \quad \text{thus} \quad \begin{bmatrix} a \\ a \end{bmatrix} - \begin{bmatrix} 2v_1 \\ 2v_2 \end{bmatrix} = \begin{bmatrix} -3v_1 + v_2 \\ v_1 - 3v_2 \end{bmatrix} + \begin{bmatrix} -6 \\ 2 \end{bmatrix}.$$

Collecting terms and reshuffling gives

$$v_1 - v_2 = -a - 6$$
$$-v_1 + v_2 = -a + 2.$$

By addition, $0 = -2a - 4$, $a = -2$, and then $v_2 = v_1 + 4$, say, $v_1 = k$, $v_2 = k + 4$, thus, $\mathbf{v} = [k \quad k + 4]^T$. We can simply choose $k = 0$. This gives the *answer*

$$\mathbf{y} = \mathbf{y}^{(h)} + \mathbf{y}^{(p)} = c_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} e^{-2t} + c_2 \begin{bmatrix} 1 \\ -1 \end{bmatrix} e^{-4t} - 2 \begin{bmatrix} 1 \\ 1 \end{bmatrix} te^{-2t} + \begin{bmatrix} 0 \\ 4 \end{bmatrix} e^{-2t}. \tag{5}$$

For other k we get other $\mathbf{v}$; for instance, $k = -2$ gives $\mathbf{v} = [-2 \quad 2]^T$, so that the *answer* becomes

$$\mathbf{y} = c_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} e^{-2t} + c_2 \begin{bmatrix} 1 \\ -1 \end{bmatrix} e^{-4t} - 2 \begin{bmatrix} 1 \\ 1 \end{bmatrix} te^{-2t} + \begin{bmatrix} -2 \\ 2 \end{bmatrix} e^{-2t}, \qquad \text{etc.} \tag{5*}$$

Method of Variation of Parameters

This method can be applied to nonhomogeneous linear systems

$$\mathbf{y}' = \mathbf{A}(t)\mathbf{y} + \mathbf{g}(t) \tag{6}$$

with variable $\mathbf{A} = \mathbf{A}(t)$ and general $\mathbf{g}(t)$. It yields a particular solution $\mathbf{y}^{(p)}$ of (6) on some open interval J on the t-axis if a general solution of the homogeneous system $\mathbf{y}' = \mathbf{A}(t)\mathbf{y}$ on J is known. We explain the method in terms of the previous example.

EXAMPLE 2 Solution by the Method of Variation of Parameters

Solve (3) in Example 1.

Solution. A basis of solutions of the homogeneous system is $[e^{-2t} \quad e^{-2t}]^T$ and $[e^{-4t} \quad -e^{-4t}]^T$. Hence the general solution (4) of the homogenous system may be written

$$\mathbf{y}^{(h)} = \begin{bmatrix} e^{-2t} & e^{-4t} \\ e^{-2t} & -e^{-4t} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \mathbf{Y}(t)\mathbf{c}. \tag{7}$$

Here, $\mathbf{Y}(t) = [\mathbf{y}^{(1)} \quad \mathbf{y}^{(2)}]^T$ is the fundamental matrix (see Sec. 4.2). As in Sec. 2.10 we replace the constant vector $\mathbf{c}$ by a variable vector $\mathbf{u}(t)$ to obtain a particular solution

$$\mathbf{y}^{(p)} = \mathbf{Y}(t)\mathbf{u}(t).$$

Substitution into (3) $\mathbf{y}' = \mathbf{A}\mathbf{y} + \mathbf{g}$ gives

$$\mathbf{Y}'\mathbf{u} + \mathbf{Y}\mathbf{u}' = \mathbf{A}\mathbf{Y}\mathbf{u} + \mathbf{g}. \tag{8}$$

Now since $\mathbf{y}^{(1)}$ and $\mathbf{y}^{(2)}$ are solutions of the homogeneous system, we have

$$\mathbf{y}^{(1)\prime} = \mathbf{A}\mathbf{y}^{(1)}, \qquad \mathbf{y}^{(2)\prime} = \mathbf{A}\mathbf{y}^{(2)}, \qquad \text{thus} \qquad \mathbf{Y}' = \mathbf{A}\mathbf{Y}.$$

Hence $\mathbf{Y}'\mathbf{u} = \mathbf{A}\mathbf{Y}\mathbf{u}$, so that (8) reduces to

$$\mathbf{Y}\mathbf{u}' = \mathbf{g}. \qquad \text{The solution is} \qquad \mathbf{u}' = \mathbf{Y}^{-1}\mathbf{g};$$

here we use that the inverse $\mathbf{Y}^{-1}$ of $\mathbf{Y}$ (Sec. 4.0) exists because the determinant of $\mathbf{Y}$ is the Wronskian W, which is not zero for a basis. Equation (9) in Sec. 4.0 gives the form of $\mathbf{Y}^{-1}$,

$$\mathbf{Y}^{-1} = \frac{1}{-2e^{-6t}} \begin{bmatrix} -e^{-4t} & -e^{-4t} \\ -e^{-2t} & e^{-2t} \end{bmatrix} = \frac{1}{2} \begin{bmatrix} e^{2t} & e^{2t} \\ e^{4t} & -e^{4t} \end{bmatrix}.$$

We multiply this by $\mathbf{g}$, obtaining

$$\mathbf{u}' = \mathbf{Y}^{-1}\mathbf{g} = \frac{1}{2} \begin{bmatrix} e^{2t} & e^{2t} \\ e^{4t} & -e^{4t} \end{bmatrix} \begin{bmatrix} -6e^{-2t} \\ 2e^{-2t} \end{bmatrix} = \frac{1}{2} \begin{bmatrix} -4 \\ -8e^{2t} \end{bmatrix} = \begin{bmatrix} -2 \\ -4e^{2t} \end{bmatrix}.$$

Integration is done componentwise (just as differentiation) and gives

$$\mathbf{u}(t) = \int_0^t \begin{bmatrix} -2 \\ -4e^{2\tilde{t}} \end{bmatrix} d\tilde{t} = \begin{bmatrix} -2t \\ -2e^{2t} + 2 \end{bmatrix}$$

(where $+ 2$ comes from the lower limit of integration). From this and $\mathbf{Y}$ in (7) we obtain

$$\mathbf{Y}\mathbf{u} = \begin{bmatrix} e^{-2t} & e^{-4t} \\ e^{-2t} & -e^{-4t} \end{bmatrix} \begin{bmatrix} -2t \\ -2e^{2t} + 2 \end{bmatrix} = \begin{bmatrix} -2te^{-2t} - 2e^{-2t} + 2e^{-4t} \\ -2te^{-2t} + 2e^{-2t} - 2e^{-4t} \end{bmatrix} = \begin{bmatrix} -2t - 2 \\ -2t + 2 \end{bmatrix} e^{-2t} + \begin{bmatrix} 2 \\ -2 \end{bmatrix} e^{-4t}.$$

The last term on the right is a solution of the homogeneous system. Hence we can absorb it into $\mathbf{y}^{(h)}$. We thus obtain as a general solution of the system (3), in agreement with (5*),

$$\mathbf{y} = c_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} e^{-2t} + c_2 \begin{bmatrix} 1 \\ -1 \end{bmatrix} e^{-4t} - 2 \begin{bmatrix} 1 \\ 1 \end{bmatrix} te^{-2t} + \begin{bmatrix} -2 \\ 2 \end{bmatrix} e^{-2t}. \tag{9}$$

■

PROBLEM SET 4.6

1. (General solution) Prove that (2) includes every solution of (1).

2–9 GENERAL SOLUTION

Find a general solution. (Show the details of your work.)

2. $y_1' = y_2 + t$

$y_2' = y_1 - 3t$

3. $y_1' = 4y_2 + 9t$

$y_2' = -4y_1 + 5$

4. $y_1' = y_1 + y_2 + 5\cos t$

$y_2' = 3y_1 - y_2 - 5\sin t$

5. $y_1' = 2y_1 + 2y_2 + 12$

$y_2' = 5y_1 - y_2 - 30$

6. $y_1' = -y_1 + y_2 + e^{-2t}$

$y_2' = -y_1 - y_2 - e^{-2t}$

7. $y_1' = -14y_1 + 10y_2 + 162$

$y_2' = -5y_1 + y_2 - 324t$

8. $y_1' = 10y_1 - 6y_2 + 10(1 - t - t^2)$

$y_2' = 6y_1 - 10y_2 + 4 - 20t - 6t^2$

9. $y_1' = -3y_1 - 4y_2 + 11t + 15$

$y_2' = 5y_1 + 6y_2 + 3e^{-t} - 15t - 20$

10. CAS EXPERIMENT. Undetermined Coefficients. Find out experimentally how general you must choose $\mathbf{y}^{(p)}$, in particular when the components of $\mathbf{g}$ have a different form (e.g., as in Prob. 9). Write a short report, covering also the situation in the case of the modification rule.

11–16 INITIAL VALUE PROBLEM

Solve (showing details):

11. $y_1' = -2y_2 + 4t$

$y_2' = 2y_1 - 2t$

$y_1(0) = 4,\ y_2(0) = \frac{1}{2}$

12. $y_1' = 4y_2 + 5e^{t}$

$y_2' = -y_1 - 20e^{-t}$

$y_1(0) = 1,\ y_2(0) = 0$

13. $y_1' = y_1 + 2y_2 + e^{2t} - 2t$

$y_2' = -y_2 + 1 + t$

$y_1(0) = 1,\ y_2(0) = -4$

14. $y_1' = 3y_1 - 4y_2 + 20\cos t$

$y_2' = y_1 - 2y_2$

$y_1(0) = 0,\ y_2(0) = 8$

15. $y_1' = 4y_2 + 3e^{3t}$

$y_2' = 2y_2 - 15e^{-3t}$

$y_1(0) = 2,\ y_2(0) = 2$

16. $y_1' = 4y_1 + 8y_2 + 2\cos t - 16\sin t$

$y_2' = 6y_1 + 2y_2 + \cos t - 14\sin t$

$y_1(0) = 15,\ y_2(0) = 13$

17. (Network) Find the currents in Fig. 97 when $R = 2.5\ \Omega$, $L = 1$ H, $C = 0.04$ F, $E(t) = 845\sin t$ V, and $I_1(0) = 0$, $I_2(0) = 0$. (Show the details.)

18. (Network) Find the currents in Fig. 97 when $R = 1\ \Omega$, $L = 10$ H, $C = 1.25$ F, $E(t) = 10$ kV, and $I_1(0) = 0$, $I_2(0) = 0$. (Show the details.)

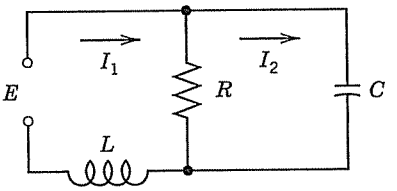


Fig. 97. Network in Probs. 17, 18

19. (Network) Find the currents in Fig. 98 when $R_1 = 2\ \Omega$, $R_2 = 8\ \Omega$, $L = 1$ H, $C = 0.5$ F, $E = 200$ V. (Show the details.)

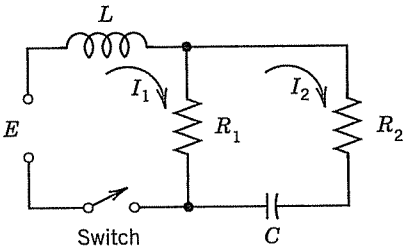


Fig. 98. Network in Prob. 19

20. WRITING PROJECT. Undetermined Coefficients. Write a short report in which you compare the application of the method of undetermined coefficients to a single ODE and to a system of two ODEs, using ODEs and systems of your choice.

CHAPTER 4 REVIEW QUESTIONS AND PROBLEMS

1. State some applications that can be modeled by systems of ODEs.
2. What is population dynamics? Give examples.
3. How can you transform an ODE into a system of ODEs?
4. What are qualitative methods for systems? Why are they important?
5. What is the phase plane? The phase plane method? The phase portrait of a system of ODEs?
6. What is a critical point of a system of ODEs? How did we classify these points?
7. What are eigenvalues? What role did they play in this chapter?
8. What does stability mean in general? In connection with critical points?
9. What does linearization of a system mean? Give an example.
10. What is a limit cycle? When may it occur in mechanics?

11–19 GENERAL SOLUTION. CRITICAL POINTS

Find a general solution. Determine the kind and stability of the critical point. (Show the details of your work.)

11. $y_1' = 4y_2$
$y_2' = 16y_1$

12. $y_1' = 9y_1$
$y_2' = y_2$

13. $y_1' = y_2$
$y_2' = 6y_1 - 5y_2$

14. $y_1' = 3y_1 - 3y_2$
$y_2' = 3y_1 + 3y_2$

15. $y_1' = 1.5y_1 - 6y_2$
$y_2' = -4.5y_1 + 3y_2$

16. $y_1' = -3y_1 - 2y_2$
$y_2' = -2y_1 - 3y_2$

17. $y_1' = 3y_1 + 2y_2$
$y_2' = 2y_1 + 3y_2$

18. $y_1' = 3y_1 + 5y_2$
$y_2' = -5y_1 - 3y_2$

19. $y_1' = -y_1 + 2y_2$
$y_2' = -2y_1 - y_2$

20–25 NONHOMOGENEOUS SYSTEMS

Find a general solution. (Show the details.)

20. $y_1' = 3y_2 + 6t$
$y_2' = 12y_1 + 1$

21. $y_1' = y_1 + 2y_2 + e^{2t}$
$y_2' = -y_2 + 1.5e^{-2t}$

22. $y_1' = y_1 + y_2 + \sin t$
$y_2' = 4y_1 + y_2$

23. $y_1' = 4y_1 + 3y_2 + 2$
$y_2' = -6y_1 - 5y_2 + 4e^{-t}$

24. $y_1' = y_1 - 2y_2 - \sin t$
$y_2' = 3y_1 - 4y_2 - \cos t$

25. $y_1' = y_1 + 2y_2 + t^2$
$y_2' = 2y_1 + y_2 - t^2$

26. (Mixing problem) Tank T_1 in Fig. 99 contains initially 200 gal of water in which 160 lb of salt are dissolved. Tank T_2 contains initially 100 gal of pure water. Liquid is pumped through the system as indicated, and the mixtures are kept uniform by stirring. Find the amounts of salt $y_1(t)$ and $y_2(t)$ in T_1 and T_2, respectively.

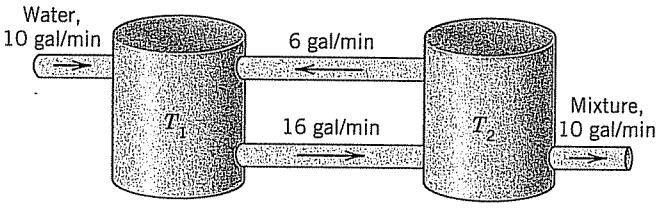


Fig. 99. Tanks in Problem 26

27. (Critical point) What kind of critical point does $\mathbf{y}' = \mathbf{A}\mathbf{y}$ have if $\mathbf{A}$ has the eigenvalues -6 and 1?

28. (Network) Find the currents in Fig. 100, where $R_1 = 0.5\ \Omega$, $R_2 = 0.7\ \Omega$, $L_1 = 0.4$ H, $L_2 = 0.5$ H, $E = 1$ kV $= 1000$ V, and $I_1(0) = 0$, $I_2(0) = 0$.

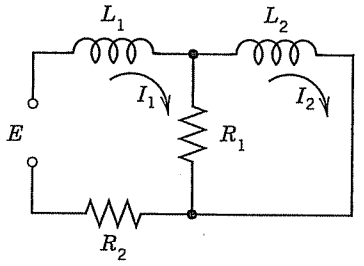


Fig. 100. Network in Problem 28

29. (Network) Find the currents in Fig. 101 when $R = 10\ \Omega$, $L = 1.25$ H, $C = 0.002$ F, and $I_1(0) = I_2(0) = 3$ A.

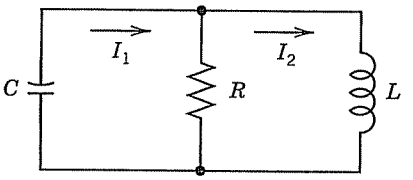


Fig. 101. Network in Problem 29

30–33 LINEARIZATION

Determine the location and kind of all critical points of the given nonlinear system by linearization.

30. $y_1' = y_2$
$y_2' = 4y_1 - y_1^3$

31. $y_1' = -9y_2$
$y_2' = \sin y_1$

32. $y_1' = \cos y_2$
$y_2' = 3y_1$

33. $y_1' = y_2 - 2y_2^2$
$y_2' = y_1 - 2y_1^2$

SUMMARY OF CHAPTER 4
Systems of ODEs. Phase Plane. Qualitative Methods

Whereas single electric circuits or single mass–spring systems are modeled by single ODEs (Chap. 2), networks of several circuits, systems of several masses and springs, and other engineering problems lead to **systems of ODEs,** involving several unknown functions $y_1(t), \cdots, y_n(t)$. Of central interest are **first-order systems** (Sec. 4.2):

$$\mathbf{y}' = \mathbf{f}(t, \mathbf{y}), \qquad \text{in components,} \qquad \begin{aligned} y_1' &= f_1(t, y_1, \cdots, y_n) \\ &\vdots \\ y_n' &= f_n(t, y_1, \cdots, y_n), \end{aligned}$$

to which higher order ODEs and systems of ODEs can be reduced (Sec. 4.1). In this summary we let $n = 2$, so that

$$(1) \qquad \mathbf{y}' = \mathbf{f}(t, \mathbf{y}), \qquad \text{in components,} \qquad \begin{aligned} y_1' &= f_1(t, y_1, y_2) \\ y_2' &= f_2(t, y_1, y_2) \end{aligned}$$

Then we can represent solution curves as **trajectories** in the **phase plane** (the y_1y_2-plane), investigate their totality [the **"phase portrait"** of (1)], and study the kind and **stability** of the **critical points** (points at which both f_1 and f_2 are zero), and classify them as **nodes, saddle points, centers,** or **spiral points** (Secs. 4.3, 4.4). These phase plane methods are **qualitative;** with their use we can discover various general properties of solutions without actually solving the system. They are primarily used for **autonomous systems,** that is, systems in which t does not occur explicitly.

A **linear system** is of the form

$$(2) \quad \mathbf{y}' = \mathbf{A}\mathbf{y} + \mathbf{g}, \quad \text{where} \quad \mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \quad \mathbf{g} = \begin{bmatrix} g_1 \\ g_2 \end{bmatrix}.$$

If $\mathbf{g} = \mathbf{0}$, the system is called **homogeneous** and is of the form

$$(3) \qquad \mathbf{y}' = \mathbf{A}\mathbf{y}.$$

If $a_{11}, \cdots, a_{22}$ are constants, it has solutions $\mathbf{y} = \mathbf{x}e^{\lambda t}$, where λ is a solution of the quadratic equation

$$\begin{vmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{vmatrix} = (a_{11} - \lambda)(a_{22} - \lambda) - a_{12}a_{21} = 0$$

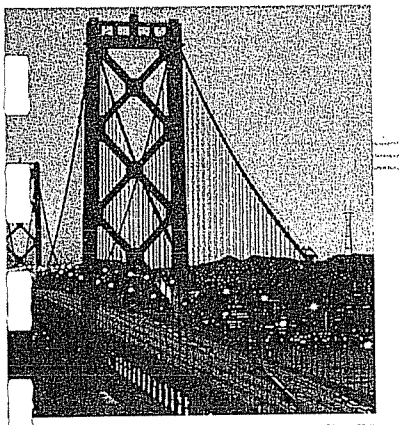
and $\mathbf{x} \neq \mathbf{0}$ has components x_1, x_2 determined up to a multiplicative constant by

$$(a_{11} - \lambda)x_1 + a_{12}x_2 = 0.$$

(These λ's are called the **eigenvalues** and these vectors $\mathbf{x}$ **eigenvectors** of the matrix $\mathbf{A}$. Further explanation is given in Sec. 4.0.)

A system (2) with $\mathbf{g} \neq \mathbf{0}$ is called **nonhomogeneous.** Its general solution is of the form $\mathbf{y} = \mathbf{y}_h + \mathbf{y}_p$, where $\mathbf{y}_h$ is a general solution of (3) and $\mathbf{y}_p$ a particular solution of (2). Methods of determining the latter are discussed in Sec. 4.6.

The discussion of critical points of linear systems based on eigenvalues is summarized in Tables 4.1 and 4.2 in Sec. 4.4. It also applies to nonlinear systems if the latter are first linearized. The key theorem for this is Theorem 1 in Sec. 4.5, which also includes three famous applications, namely the pendulum and van der Pol equations and the Lotka–Volterra predator–prey population model.



CHAPTER 5

Series Solutions of ODEs. Special Functions

In Chaps. 2 and 3 we have seen that linear ODEs with *constant* coefficients can be solved by functions known from calculus. However, if a linear ODE has *variable* coefficients (functions of x), it must usually be solved by other methods, as we shall see in this chapter.

Legendre polynomials, Bessel functions, and eigenfunction expansions are the three main topics in this chapter. These are of greatest importance to the applied mathematician.

Legendre's ODE and **Legendre polynomials** (Sec. 5.3) are likely to occur in problems showing *spherical symmetry*. They are obtained by the **power series method** (Secs. 5.1, 5.2), which gives solutions of ODEs in power series.

Bessel's ODE and **Bessel functions** (Secs. 5.5, 5.6) are likely to occur in problems showing *cylindrical symmetry*. They are obtained by the **Frobenius method** (Sec. 5.4), an extension of the power series method which gives solutions of ODEs in power series, possibly multiplied by a logarithmic term or by a fractional power.

Eigenfunction expansions (Sec. 5.8) are infinite series obtained by the **Sturm–Liouville theory** (Sec. 5.7). The terms of these series may be Legendre polynomials or other functions, and their coefficients are obtained by the **orthogonality** of those functions. These expansions include **Fourier series** in terms of cosine and sine, which are so important that we shall devote a whole chapter (Chap. 11) to them.

Special functions (also called **higher functions**) is a name for more advanced functions not considered in calculus. If a function occurs in many applications, it gets a name, and its properties and values are investigated in all details, resulting in hundreds of formulas which together with the underlying theory often fill whole books. This is what has happened to the gamma, Legendre, Bessel, and several other functions (take a look into Refs. [GR1], [GR10], [A11] in App. 1).

Your CAS knows most of the special functions and corresponding formulas that you will ever need in your later work in industry, and this chapter will give you a feel for the basics of their theory and their application in modeling.

COMMENT. You can study this chapter directly after Chap. 2 because it needs no material from Chaps. 3 or 4.

Prerequisite: Chap. 2.
Sections that may be omitted in a shorter course: 5.2, 5.6–5.8.
References and Answers to Problems: App. 1 Part A, and App. 2.

5.1 Power Series Method

The **power series method** is the standard method for solving linear ODEs with ***variable*** coefficients. It gives solutions in the form of power series. These series can be used for computing values, graphing curves, proving formulas, and exploring properties of solutions, as we shall see. In this section we begin by explaining the idea of the power series method.

Power Series

From calculus we recall that a **power series** (in powers of $x - x_0$) is an infinite series of the form

$$\text{(1)} \qquad \sum_{m=0}^{\infty} a_m(x - x_0)^m = a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + \cdots.$$

Here, x is a variable. a_0, a_1, a_2, $\cdots$ are constants, called the **coefficients** of the series. x_0 is a constant, called the **center** of the series. In particular, if $x_0 = 0$, we obtain a **power series in powers of x**

$$\text{(2)} \qquad \sum_{m=0}^{\infty} a_m x^m = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots.$$

We shall assume that all variables and constants are real.

Familiar examples of power series are the Maclaurin series

$$\frac{1}{1-x} = \sum_{m=0}^{\infty} x^m = 1 + x + x^2 + \cdots \qquad (|x| < 1, \textbf{ geometric series})$$

$$e^x = \sum_{m=0}^{\infty} \frac{x^m}{m!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$$

$$\cos x = \sum_{m=0}^{\infty} \frac{(-1)^m x^{2m}}{(2m)!} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - + \cdots$$

$$\sin x = \sum_{m=0}^{\infty} \frac{(-1)^m x^{2m+1}}{(2m+1)!} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - + \cdots.$$

We note that the term "power series" usually refers to a series of the form (1) [or (2)] but ***does not include*** series of negative or fractional powers of x. We use m as the summation letter, reserving n as a standard notation in the Legendre and Bessel equations for integer values of the parameter.

Idea of the Power Series Method

The idea of the power series method for solving ODEs is simple and natural. We describe the practical procedure and illustrate it for two ODEs whose solution we know, so that

we can see what is going on. The mathematical justification of the method follows in the next section.

For a given ODE

$$y'' + p(x)y' + q(x)y = 0$$

we first represent $p(x)$ and $q(x)$ by power series in powers of x (or of $x - x_0$ if solutions in powers of $x - x_0$ are wanted). Often $p(x)$ and $q(x)$ are polynomials, and then nothing needs to be done in this first step. Next we assume a solution in the form of a power series with unknown coefficients,

$$y = \sum_{m=0}^{\infty} a_m x^m = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots \tag{3}$$

and insert this series and the series obtained by termwise differentiation,

$$\text{(a)} \quad y' = \sum_{m=1}^{\infty} m a_m x^{m-1} = a_1 + 2a_2 x + 3a_3 x^2 + \cdots$$

$$\text{(b)} \quad y'' = \sum_{m=2}^{\infty} m(m-1) a_m x^{m-2} = 2a_2 + 3 \cdot 2a_3 x + 4 \cdot 3a_4 x^2 + \cdots \tag{4}$$

into the ODE. Then we collect like powers of x and equate the sum of the coefficients of each occurring power of x to zero, starting with the constant terms, then taking the terms containing x, then the terms in x^2, and so on. This gives equations from which we can determine the unknown coefficients of (3) successively.

Let us show this for two simple ODEs that can also be solved by elementary methods, so that we would not need power series.

EXAMPLE 1 Solve the following ODE by power series. To grasp the idea, do this by hand; do not use your CAS (for which you could program the whole process).

$$y' = 2xy.$$

Solution. We insert (3) and (4a) into the given ODE, obtaining

$$a_1 + 2a_2 x + 3a_3 x^2 + \cdots = 2x(a_0 + a_1 x + a_2 x^2 + \cdots).$$

We must perform the multiplication by $2x$ on the right and can write the resulting equation conveniently as

$$\begin{aligned} a_1 + 2a_2 x + 3a_3 x^2 + 4a_4 x^3 + 5a_5 x^4 + 6a_6 x^5 + \cdots \\ = \quad 2a_0 x + 2a_1 x^2 + 2a_2 x^3 + 2a_3 x^4 + 2a_4 x^5 + \cdots. \end{aligned}$$

For this equation to hold, the two coefficients of every power of x on both sides must be equal, that is,

$$a_1 = 0, \quad 2a_2 = 2a_0, \quad 3a_3 = 2a_1, \quad 4a_4 = 2a_2, \quad 5a_5 = 2a_3, \quad 6a_6 = 2a_4, \cdots.$$

Hence $a_3 = 0$, $a_5 = 0$, $\cdots$ and for the coefficients with even subscripts,

$$a_2 = a_0, \qquad a_4 = \frac{a_2}{2} = \frac{a_0}{2!}, \qquad a_6 = \frac{a_4}{3} = \frac{a_0}{3!}, \cdots;$$

a_0 remains arbitrary. With these coefficients the series (3) gives the following solution, which you should confirm by the method of separating variables.

$$y = a_0\left(1 + x^2 + \frac{x^4}{2!} + \frac{x^6}{3!} + \frac{x^8}{4!} + \cdots\right) = a_0 e^{x^2}.$$

More rapidly, (3) and (4) give for the ODE $y' = 2xy$

$$1 \cdot a_1 x^0 + \sum_{m=2}^{\infty} m a_m x^{m-1} = 2x \sum_{m=0}^{\infty} a_m x^m = \sum_{m=0}^{\infty} 2a_m x^{m+1}.$$

Now, to get the same general power on both sides, we make a **"shift of index"** on the left by setting $m = s + 2$, thus $m - 1 = s + 1$. Then a_m becomes a_{s+2} and x^{m-1} becomes x^{s+1}. Also the summation, which started with $m = 2$, now starts with $s = 0$ because $s = m - 2$. On the right we simply make a change of notation $m = s$, hence $a_m = a_s$ and $x^{m+1} = x^{s+1}$; also the summation now starts with $s = 0$. This altogether gives

$$a_1 + \sum_{s=0}^{\infty} (s + 2)a_{s+2}x^{s+1} = \sum_{s=0}^{\infty} 2a_s x^{s+1}.$$

Every occurring power of x must have the same coefficient on both sides; hence

$$a_1 = 0 \qquad \text{and} \qquad (s + 2)a_{s+2} = 2a_s \qquad \text{or} \qquad a_{s+2} = \frac{2}{s + 2}a_s.$$

For $s = 0, 1, 2, \cdots$ we thus have $a_2 = (2/2)a_0$, $a_3 = (2/3)a_1 = 0$, $a_4 = (2/4)a_2$, $\cdots$ as before. ▩

EXAMPLE 2 Solve

$$y'' + y = 0.$$

Solution. By inserting (3) and (4b) into the ODE we have

$$\sum_{m=2}^{\infty} m(m - 1)a_m x^{m-2} + \sum_{m=0}^{\infty} a_m x^m = 0.$$

To obtain the same general power on both series, we set $m = s + 2$ in the first series and $m = s$ in the second, and then we take the latter to the right side. This gives

$$\sum_{s=0}^{\infty} (s + 2)(s + 1)a_{s+2}x^s = -\sum_{s=0}^{\infty} a_s x^s.$$

Each power x^s must have the same coefficient on both sides. Hence $(s + 2)(s + 1)a_{s+2} = -a_s$. This gives the **recursion formula**

$$a_{s+2} = -\frac{a_s}{(s + 2)(s + 1)} \qquad (s = 0, 1, \cdots).$$

We thus obtain successively

$$a_2 = -\frac{a_0}{2 \cdot 1} = -\frac{a_0}{2!}, \qquad a_3 = -\frac{a_1}{3 \cdot 2} = -\frac{a_1}{3!}$$

$$a_4 = -\frac{a_2}{4 \cdot 3} = \frac{a_0}{4!}, \qquad a_5 = -\frac{a_3}{5 \cdot 4} = \frac{a_1}{5!}.$$

and so on. a_0 and a_1 remain arbitrary. With these coefficients the series (3) becomes

$$y = a_0 + a_1 x - \frac{a_0}{2!}x^2 - \frac{a_1}{3!}x^3 + \frac{a_0}{4!}x^4 + \frac{a_1}{5!}x^5 + \cdots.$$

Reordering terms (which is permissible for a power series), we can write this in the form

$$y = a_0\left(1 - \frac{x^2}{2!} + \frac{x^4}{4!} - + \cdots\right) + a_1\left(x - \frac{x^3}{3!} + \frac{x^5}{5!} - + \cdots\right)$$

and we recognize the familiar general solution

$$y = a_0 \cos x + a_1 \sin x.$$

Do we need the power series method for these or similar ODEs? Of course not; we used them just for explaining the idea of the method. What happens if we apply the method to an ODE not of the kind considered so far, even to an innocent-looking one such as $y'' + xy = 0$ ("Airy's equation")? We most likely end up with new special functions given by power series. And if such an ODE and its solutions are of practical (or theoretical) interest, we name and investigate them in terms of formulas and graphs and by numeric methods.

We shall discuss Legendre's, Bessel's, and the hypergeometric equations and their solutions, to mention just the most prominent of these ODEs. To do this with a good understanding, also in the light of your CAS, we first explain the power series method (and later an extension, the Frobenius method) in more detail.

PROBLEM SET 5.1

1–10 POWER SERIES METHOD: TECHNIQUE, FEATURES

Apply the power series method. Do this by hand, not by a CAS, so that you get a feel for the method, e.g., why a series may terminate, or has even powers only, or has no constant or linear terms, etc. Show the details of your work.

1. $y' - y = 0$
2. $y' + xy = 0$
3. $y'' + 4y = 0$
4. $y'' - y = 0$
5. $(2 + x)y' = y$
6. $y' + 3(1 + x^2)y = 0$
7. $y' = y + x$
8. $(x^5 + 4x^3)y' = (5x^4 + 12x^2)y$
9. $y'' - y' = 0$
10. $y'' - xy' + y = 0$

11–16 CAS PROBLEMS. INITIAL VALUE PROBLEMS

Solve the initial value problems by a power series. Graph the partial sum s of the powers up to and including x^5. Find the value of s (5 digits) at x_1.

11. $y' + 4y = 1$, $y(0) = 1.25$, $x_1 = 0.2$
12. $y' = 1 + y^2$, $y(0) = 0$, $x_1 = \frac{1}{4}\pi$
13. $y' = y - y^2$, $y(0) = \frac{1}{2}$, $x_1 = 1$
14. $(x - 2)y' = xy$, $y(0) = 4$, $x_1 = 2$
15. $y'' + 3xy' + 2y = 0$, $y(0) = 1$, $y'(0) = 1$, $x_1 = 0.5$
16. $(1 - x^2)y'' - 2xy' + 30y = 0$, $y(0) = 0$, $y'(0) = 1.875$, $x_1 = 0.5$

17. **WRITING PROJECT. Power Series.** Write a review (2–3 pages) on power series as they are discussed in calculus, using your own formulation and examples—do not just copy passages from calculus texts.
18. **LITERATURE PROJECT. Maclaurin Series.** Collect Maclaurin series of the functions known from calculus and arrange them systematically in a list that you can use for your work.

5.2 Theory of the Power Series Method

In the last section we saw that the power series method gives solutions of ODEs in the form of power series. In this section we justify the method mathematically as follows. We first review relevant facts on power series from calculus. Then we list the operations on power series needed in the method (differentiation, addition, multiplication, etc.). Near the end we state the basic existence theorem for power series solutions of ODEs.

Basic Concepts

Recall from calculus that a power series is an infinite series of the form

$$\sum_{m=0}^{\infty} a_m(x - x_0)^m = a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + \cdots \tag{1}$$

As before, we assume the variable x, the **center** x_0, and the **coefficients** $a_0, a_1, \cdots$ to be real. The ***n*th partial sum** of (1) is

$$s_n(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + \cdots + a_n(x - x_0)^n \tag{2}$$

where $n = 0, 1, \cdots$. Clearly, if we omit the terms of s_n from (1), the remaining expression is

$$R_n(x) = a_{n+1}(x - x_0)^{n+1} + a_{n+2}(x - x_0)^{n+2} + \cdots. \tag{3}$$

This expression is called the **remainder** *of* (1) *after the term* $a_n(x - x_0)^n$.

For example, in the case of the geometric series

$$1 + x + x^2 + \cdots + x^n + \cdots$$

we have

$$\begin{aligned} s_0 &= 1, & R_0 &= x + x^2 + x^3 + \cdots, \\ s_1 &= 1 + x, & R_1 &= x^2 + x^3 + x^4 + \cdots, \\ s_2 &= 1 + x + x^2, & R_2 &= x^3 + x^4 + x^5 + \cdots, \qquad \text{etc.} \end{aligned}$$

In this way we have now associated with (1) the sequence of the partial sums $s_0(x), s_1(x), s_2(x), \cdots$. If for some $x = x_1$ this sequence converges, say,

$$\lim_{n\to\infty} s_n(x_1) = s(x_1),$$

then the series (1) is called **convergent** *at* $x = x_1$, the number $s(x_1)$ is called the **value** or *sum* of (1) at x_1, and we write

$$s(x_1) = \sum_{m=0}^{\infty} a_m(x_1 - x_0)^m.$$

Then we have for every n,

$$s(x_1) = s_n(x_1) + R_n(x_1). \tag{4}$$

If that sequence diverges at $x = x_1$, the series (1) is called **divergent** at $x = x_1$.

In the case of convergence, for any positive ϵ there is an N (depending on ϵ) such that, by (4),

$$|R_n(x_1)| = |s(x_1) - s_n(x_1)| < \epsilon \qquad \text{for all } n > N. \tag{5}$$

Geometrically, this means that all $s_n(x_1)$ with $n > N$ lie between $s(x_1) - \epsilon$ and $s(x_1) + \epsilon$ (Fig. 102). Practically, this means that in the case of convergence we can approximate the sum $s(x_1)$ of (1) at x_1 by $s_n(x_1)$ as accurately as we please, by taking n large enough.

Convergence Interval. Radius of Convergence

With respect to the convergence of the power series (1) there are three cases, the useless Case 1, the usual Case 2, and the best Case 3, as follows.

Case 1. The series (1) always converges at $x = x_0$, because for $x = x_0$ all its terms are zero, perhaps except for the first one, a_0. In exceptional cases $x = x_0$ may be the only x for which (1) converges. Such a series is of no practical interest.

Case 2. If there are further values of x for which the series converges, these values form an interval, called the **convergence interval.** If this interval is finite, it has the midpoint x_0, so that it is of the form

$$|x - x_0| < R \tag{6}$$

(Fig. 103)

and the series (1) converges for all x such that $|x - x_0| < R$ and diverges for all x such that $|x - x_0| > R$. (No general statement about convergence or divergence can be made for $x - x_0 = R$ or $-R$.) The number R is called the **radius of convergence** of (1). (R is called "radius" because for a *complex* power series it is the radius of a disk of convergence.) R can be obtained from either of the formulas

$$\text{(7)} \qquad \text{(a)} \quad R = 1\Big/\lim_{m\to\infty} \sqrt[m]{|a_m|} \qquad \text{(b)} \quad R = 1\Big/\lim_{m\to\infty} \left|\frac{a_{m+1}}{a_m}\right| \tag{7}$$

provided these limits exist and are not zero. [If these limits are infinite, then (1) converges only at the center x_0.]

Case 3. The convergence interval may sometimes be infinite, that is, (1) converges for all x. For instance, if the limit in (7a) or (7b) is zero, this case occurs. One then writes $R = \infty$, for convenience. (Proofs of all these facts can be found in Sec. 15.2.)

For each x for which (1) converges, it has a certain value $s(x)$. We say that (1) **represents** the function $s(x)$ in the convergence interval and write

$$s(x) = \sum_{m=0}^{\infty} a_m(x - x_0)^m \qquad (|x - x_0| < R).$$

Let us illustrate these three possible cases with typical examples.

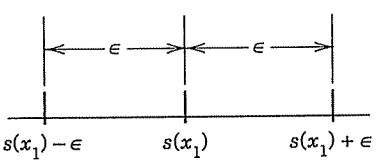


Fig. 102. Inequality (5)

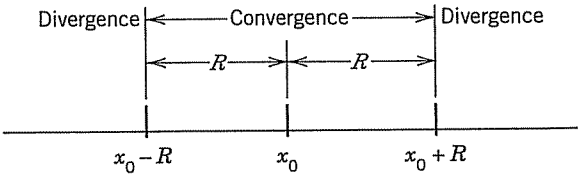


Fig. 103. Convergence interval (6) of a power series with center x_0

EXAMPLE 1 **The Useless Case 1 of Convergence Only at the Center**

In the case of the series

$$\sum_{m=0}^{\infty} m!x^m = 1 + x + 2x^2 + 6x^3 + \cdots$$

we have $a_m = m!$, and in (7b),

$$\frac{a_{m+1}}{a_m} = \frac{(m+1)!}{m!} = m + 1 \to \infty \qquad \text{as } m \to \infty.$$

Thus this series converges only at the center $x = 0$. Such a series is useless. ■

EXAMPLE 2 **The Usual Case 2 of Convergence in a Finite Interval. Geometric Series**

For the **geometric series** we have

$$\frac{1}{1-x} = \sum_{m=0}^{\infty} x^m = 1 + x + x^2 + \cdots \qquad (|x| < 1).$$

In fact, $a_m = 1$ for all m, and from (7) we obtain $R = 1$, that is, the geometric series converges and represents $1/(1 - x)$ when $|x| < 1$. ■

EXAMPLE 3 **The Best Case 3 of Convergence for All x**

In the case of the series

$$e^x = \sum_{m=0}^{\infty} \frac{x^m}{m!} = 1 + x + \frac{x^2}{2!} + \cdots$$

we have $a_m = 1/m!$. Hence in (7b),

$$\frac{a_{m+1}}{a_m} = \frac{1/(m+1)!}{1/m!} = \frac{1}{m+1} \to 0 \qquad \text{as } m \to \infty,$$

so that the series converges for all x. ■

EXAMPLE 4 **Hint for Some of the Problems**

Find the radius of convergence of the series

$$\sum_{m=0}^{\infty} \frac{(-1)^m}{8^m} x^{3m} = 1 - \frac{x^3}{8} + \frac{x^6}{64} - \frac{x^9}{512} + - \cdots.$$

Solution. This is a series in powers of $t = x^3$ with coefficients $a_m = (-1)^m/8^m$, so that in (7b),

$$\left| \frac{a_{m+1}}{a_m} \right| = \frac{8^m}{8^{m+1}} = \frac{1}{8}.$$

Thus $R = 8$. Hence the series converges for $|t| = |x^3| < 8$, that is, $|x| < 2$. ■

Operations on Power Series

In the power series method we differentiate, add, and multiply power series. These three operations are permissible, in the sense explained in what follows. We also list a condition about the vanishing of all coefficients of a power series, which is a basic tool of the power series method. (Proofs can be found in Sec. 15.3.)

Termwise Differentiation

A power series may be differentiated term by term. More precisely: if

$$y(x) = \sum_{m=0}^{\infty} a_m(x - x_0)^m$$

converges for $|x - x_0| < R$, where $R > 0$, then the series obtained by differentiating term by term also converges for those x and represents the derivative y' of y for those x, that is,

$$y'(x) = \sum_{m=1}^{\infty} m a_m(x - x_0)^{m-1} \qquad (|x - x_0| < R).$$

Similarly,

$$y''(x) = \sum_{m=2}^{\infty} m(m-1) a_m(x - x_0)^{m-2} \qquad (|x - x_0| < R), \text{ etc.}$$

Termwise Addition

Two power series may be added term by term. More precisely: if the series

$$(8) \qquad \sum_{m=0}^{\infty} a_m(x - x_0)^m \qquad \text{and} \qquad \sum_{m=0}^{\infty} b_m(x - x_0)^m$$

have positive radii of convergence and their sums are $f(x)$ and $g(x)$, then the series

$$\sum_{m=0}^{\infty} (a_m + b_m)(x - x_0)^m$$

converges and represents $f(x) + g(x)$ for each x that lies in the interior of the convergence interval of each of the two given series.

Termwise Multiplication

Two power series may be multiplied term by term. More precisely: Suppose that the series (8) have positive radii of convergence and let $f(x)$ and $g(x)$ be their sums. Then the series obtained by multiplying each term of the first series by each term of the second series and collecting like powers of $x - x_0$, that is,

$$\sum_{m=0}^{\infty} (a_0 b_m + a_1 b_{m-1} + \cdots + a_m b_0)(x - x_0)^m$$
$$= a_0 b_0 + (a_0 b_1 + a_1 b_0)(x - x_0) + (a_0 b_2 + a_1 b_1 + a_2 b_0)(x - x_0)^2 + \cdots$$

converges and represents $f(x)g(x)$ for each x in the interior of the convergence interval of each of the two given series.

Vanishing of All Coefficients

If a power series has a positive radius of convergence and a sum that is identically zero throughout its interval of convergence, then each coefficient of the series must be zero.

Existence of Power Series Solutions of ODEs. Real Analytic Functions

The properties of power series just discussed form the foundation of the power series method. The remaining question is whether an ODE has power series solutions at all. An answer is simple: If the coefficients p and q and the function r on the right side of

$$y'' + p(x)y' + q(x)y = r(x) \tag{9}$$

have power series representations, then (9) has power series solutions. The same is true if $\tilde{h}$, $\tilde{p}$, $\tilde{q}$, and $\tilde{r}$ in

$$\tilde{h}(x)y'' + \tilde{p}(x)y' + \tilde{q}(x)y = \tilde{r}(x) \tag{10}$$

have power series representations and $\tilde{h}(x_0) \neq 0$ (x_0 the center of the series). Almost all ODEs in practice have polynomials as coefficients (thus terminating power series), so that (when $r(x) \equiv 0$ or is a power series, too) those conditions are satisfied, except perhaps the condition $\tilde{h}(x_0) \neq 0$. If $\tilde{h}(x_0) \neq 0$, division of (10) by $\tilde{h}(x)$ gives (9) with $p = \tilde{p}/\tilde{h}$, $q = \tilde{q}/\tilde{h}$, $r = \tilde{r}/\tilde{h}$. This motivates our notation in (10).

To formulate all this in a precise and simple way, we use the following concept (which is of general interest).

DEFINITION

Real Analytic Function

A real function $f(x)$ is called **analytic** *at a point* $x = x_0$ if it can be represented by a power series in powers of $x - x_0$ with radius of convergence $R > 0$.

Using this concept, we can state the following basic theorem.

THEOREM 1

Existence of Power Series Solutions

If p, q, and r in (9) *are analytic at $x = x_0$, then every solution of* (9) *is analytic at $x = x_0$ and can thus be represented by a power series in powers of $x - x_0$ with radius of convergence $R > 0$. Hence the same is true if $\tilde{h}$, $\tilde{p}$, $\tilde{q}$, and $\tilde{r}$ in* (10) *are analytic at $x = x_0$ and $\tilde{h}(x_0) \neq 0$.*

The proof of this theorem requires advanced methods of complex analysis and can be found in Ref. [A11] listed in App. 1.

We mention that the radius of convergence R in Theorem 1 is at least equal to the distance from the point $x = x_0$ to the point (or points) closest to x_0 at which one of the functions p, q, r, as functions of a *complex variable,* is not analytic. (Note that that point may not lie on the x-axis but somewhere in the complex plane.)

PROBLEM SET 5.2

1–12 RADIUS OF CONVERGENCE

Determine the radius of convergence. (Show the details.)

1. $\sum_{m=0}^{\infty} \frac{x^m}{c^m} \quad (c \neq 0)$

2. $\sum_{m=0}^{\infty} \frac{(-1)^m}{3^m(m+1)^2} (x+1)^{2m}$

3. $\sum_{m=1}^{\infty} \frac{(m+1)m}{2^m} (x-3)^{2m}$

4. $\sum_{m=0}^{\infty} (-1)^m x^{4m}$

5. $\sum_{m=0}^{\infty} \frac{(2m)!}{(2m+2)(2m+4)} x^m$

6. $\sum_{m=0}^{\infty} \frac{(-1)^m}{(m!)^2} x^{2m+10}$

7. $\sum_{m=2}^{\infty} \frac{(-1)^m}{4^m} (x-1)^{2m}$

8. $\sum_{m=1}^{\infty} \frac{(4m)!}{(m!)^4} x^m$

9. $\sum_{m=4}^{\infty} \frac{(m+3)^2}{(m-3)^4} x^m$

10. $\sum_{m=1}^{\infty} \frac{(2m)!}{m^2} x^m$

11. $\sum_{m=1}^{\infty} \frac{1}{\pi^m} (x - \frac{1}{2}\pi)^m$

12. $\sum_{m=1}^{\infty} \frac{(m+1)m}{(2m+1)!} x^{2m+1}$

13–15 SHIFTING SUMMATION INDICES (CF. SEC. 5.1)

This is often convenient or necessary in the power series method. Shift the index so that the power under the summation sign is x^s. Check by writing the first few terms explicitly. Also determine the radius of convergence R.

13. $\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{5n} x^{n+2}$

14. $\sum_{m=3}^{\infty} \frac{(-1)^{m+1}}{4^m} x^{m-3}$

15. $\sum_{p=1}^{\infty} \frac{p^2}{(p+1)!} x^{p+4}$

16–23 POWER SERIES SOLUTIONS

Find a power series solution in powers of x. (Show the details of your work.)

16. $y'' + xy = 0$

17. $y'' - y' + x^2y = 0$

18. $y'' - y' + xy = 0$

19. $y'' + 4xy' = 0$

20. $y'' + 2xy' + y = 0$

21. $y'' + (1 + x^2)y = 0$

22. $y'' - 4xy' + (4x^2 - 2)y = 0$

23. $(2x^2 - 3x + 1)y'' + 2xy' - 2y = 0$

24. **TEAM PROJECT. Properties from Power Series.** In the next sections we shall define new functions (Legendre functions, etc.) by power series, deriving properties of the functions directly from the series. To understand this idea, do the same for functions familiar from calculus, using Maclaurin series.

(a) Show that $\cosh x + \sinh x = e^x$. Show that $\cosh x > 0$ for all x. Show that $e^x \geqq e^{-x}$ for all $x \geqq 0$.

(b) Derive the differentiation formulas for e^x, $\cos x$, $\sin x$, $1/(1-x)$ and other functions of your choice. Show that $(\cos x)'' = -\cos x$, $(\cosh x)'' = \cosh x$. Consider integration similarly.

(c) What can you conclude if a series contains only odd powers? Only even powers? No constant term? If all its coefficients are positive? Give examples.

(d) What properties of $\cos x$ and $\sin x$ are *not* obvious from the Maclaurin series? What properties of other functions?

25. **CAS EXPERIMENT. Information from Graphs of Partial Sums.** In connection with power series in numerics we use partial sums. To get a feel for the accuracy for various x, experiment with $\sin x$ and graphs of partial sums of the Maclaurin series of an increasing number of terms, describing qualitatively the "breakaway points" of these graphs from the graph of $\sin x$. Consider other examples of your own choice.

5.3 Legendre's Equation. Legendre Polynomials $P_n(x)$

In order to first gain skill, we have applied the power series method to ODEs that can also be solved by other methods. We now turn to the first "big" equation of physics, for which we do need the power series method. This is **Legendre's equation**[1]

$$(1) \qquad (1 - x^2)y'' - 2xy' + n(n + 1)y = 0$$

where n is a given constant. Legendre's equation arises in numerous problems, particularly in boundary value problems for spheres (take a quick look at Example 1 in Sec. 12.10). The **parameter** n in (1) is a given real number. Any solution of (1) is called a **Legendre function.** The study of these and other "higher" functions not occurring in calculus is called the **theory of special functions.** Further special functions will occur in the next sections.

Dividing (1) by the coefficient $1 - x^2$ of y'', we see that the coefficients $-2x/(1 - x^2)$ and $n(n + 1)/(1 - x^2)$ of the new equation are analytic at $x = 0$. Hence by Theorem 1, in Sec. 5.2, Legendre's equation has power series solutions of the form

$$(2) \qquad y = \sum_{m=0}^{\infty} a_m x^m.$$

Substituting (2) and its derivatives into (1), and denoting the constant $n(n + 1)$ simply by k, we obtain

$$(1 - x^2) \sum_{m=2}^{\infty} m(m - 1)a_m x^{m-2} - 2x \sum_{m=1}^{\infty} m a_m x^{m-1} + k \sum_{m=0}^{\infty} a_m x^m = 0.$$

By writing the first expression as two separate series we have the equation

$$\sum_{m=2}^{\infty} m(m - 1)a_m x^{m-2} - \sum_{m=2}^{\infty} m(m - 1)a_m x^m - \sum_{m=1}^{\infty} 2m a_m x^m + \sum_{m=0}^{\infty} k a_m x^m = 0.$$

To obtain the same general power x^s in all four series, we set $m - 2 = s$ (thus $m = s + 2$) in the first series and simply write s instead of m in the other three series. This gives

$$\sum_{s=0}^{\infty} (s + 2)(s + 1)a_{s+2} x^s - \sum_{s=2}^{\infty} s(s - 1)a_s x^s - \sum_{s=1}^{\infty} 2s a_s x^s + \sum_{s=0}^{\infty} k a_s x^s = 0.$$

---

[1]ADRIEN-MARIE LEGENDRE (1752–1833), French mathematician, who became a professor in Paris in 1775 and made important contributions to special functions, elliptic integrals, number theory, and the calculus of variations, His book *Éléments de géométrie* (1794) became very famous and had 12 editions in less than 30 years.

Formulas on Legendre functions may be found in Refs. [GR1] and [GR10].

(Note that in the first series the summation begins with $s = 0$.) Since this equation with right side 0 must be an identity in x if (2) is to be a solution of (1), the sum of the coefficients of each power of x on the left must be zero. Now x^0 occurs in the first and fourth series and gives [remember that $k = n(n + 1)$]

$$2 \cdot 1a_2 + n(n + 1)a_0 = 0. \tag{3a}$$

x^1 occurs in the first, third, and fourth series and gives

$$3 \cdot 2a_3 + [-2 + n(n + 1)]a_1 = 0. \tag{3b}$$

The higher powers $x^2, x^3, \cdots$ occur in all four series and give

$$(s + 2)(s + 1)a_{s+2} + [-s(s - 1) - 2s + n(n + 1)]a_s = 0. \tag{3c}$$

The expression in the brackets $[\cdots]$ can be written $(n - s)(n + s + 1)$, as you may readily verify. Solving (3a) for a_2 and (3b) for a_3 as well as (3c) for a_{s+2}, we obtain the general formula

$$a_{s+2} = -\frac{(n - s)(n + s + 1)}{(s + 2)(s + 1)}a_s \qquad (s = 0, 1, \cdots). \tag{4}$$

This is called a **recurrence relation** or **recursion formula.** (Its derivation you may verify with your CAS.) It gives each coefficient in terms of the second one preceding it, except for a_0 and a_1, which are left as arbitrary constants. We find successively

$$a_2 = -\frac{n(n + 1)}{2!}a_0 \qquad a_3 = -\frac{(n - 1)(n + 2)}{3!}a_1$$

$$a_4 = -\frac{(n - 2)(n + 3)}{4 \cdot 3}a_2 \qquad a_5 = -\frac{(n - 3)(n + 4)}{5 \cdot 4}a_3$$

$$= \frac{(n - 2)n(n + 1)(n + 3)}{4!}a_0 \qquad = \frac{(n - 3)(n - 1)(n + 2)(n + 4)}{5!}a_1$$

and so on. By inserting these expressions for the coefficients into (2) we obtain

$$y(x) = a_0y_1(x) + a_1y_2(x) \tag{5}$$

where

$$y_1(x) = 1 - \frac{n(n + 1)}{2!}x^2 + \frac{(n - 2)n(n + 1)(n + 3)}{4!}x^4 - + \cdots \tag{6}$$

$$y_2(x) = x - \frac{(n - 1)(n + 2)}{3!}x^3 + \frac{(n - 3)(n - 1)(n + 2)(n + 4)}{5!}x^5 - + \cdots. \tag{7}$$

These series converge for $|x| < 1$ (see Prob. 4; or they may terminate, see below). Since (6) contains even powers of x only, while (7) contains odd powers of x only, the ratio y_1/y_2 is not a constant, so that y_1 and y_2 are not proportional and are thus linearly independent solutions. Hence (5) is a general solution of (1) on the interval $-1 < x < 1$.

Legendre Polynomials $P_n(x)$

In various applications, power series solutions of ODEs reduce to polynomials, that is, they terminate after finitely many terms. This is a great advantage and is quite common for special functions, leading to various important families of polynomials (see Refs. [GR1] or [GR10] in App. 1). For Legendre's equation this happens when the parameter n is a nonnegative integer because then the right side of (4) is zero for $s = n$, so that $a_{n+2} = 0$, $a_{n+4} = 0$, $a_{n+6} = 0, \cdots$. Hence if n is even, $y_1(x)$ reduces to a polynomial of degree n. If n is odd, the same is true for $y_2(x)$. These polynomials, multiplied by some constants, are called **Legendre polynomials** and are denoted by $P_n(x)$. The standard choice of a constant is done as follows. We choose the coefficient a_n of the highest power x^n as

$$a_n = \frac{(2n)!}{2^n(n!)^2} = \frac{1 \cdot 3 \cdot 5 \cdots (2n-1)}{n!} \qquad (n \text{ a positive integer}) \tag{8}$$

(and $a_n = 1$ if $n = 0$). Then we calculate the other coefficients from (4), solved for a_s in terms of a_{s+2}, that is,

$$a_s = -\frac{(s+2)(s+1)}{(n-s)(n+s+1)}\, a_{s+2} \qquad (s \leqq n-2). \tag{9}$$

The choice (8) makes $P_n(1) = 1$ for every n (see Fig. 104 on p. 180); this motivates (8). From (9) with $s = n - 2$ and (8) we obtain

$$a_{n-2} = -\frac{n(n-1)}{2(2n-1)}\, a_n = -\frac{n(n-1)(2n)!}{2(2n-1)2^n(n!)^2}.$$

Using $(2n)! = 2n(2n-1)(2n-2)!$, $n! = n(n-1)!$, and $n! = n(n-1)(n-2)!$, we obtain

$$a_{n-2} = -\frac{n(n-1)2n(2n-1)(2n-2)!}{2(2n-1)2^n n(n-1)!\, n(n-1)(n-2)!}.$$

$n(n-1)2n(2n-1)$ cancels, so that we get

$$a_{n-2} = -\frac{(2n-2)!}{2^n(n-1)!\,(n-2)!}.$$

Similarly,

$$a_{n-4} = -\frac{(n-2)(n-3)}{4(2n-3)}\, a_{n-2}$$

$$= \frac{(2n-4)!}{2^n 2!\,(n-2)!\,(n-4)!}$$

and so on, and in general, when $n - 2m \geqq 0$,

$$a_{n-2m} = (-1)^m \frac{(2n-2m)!}{2^n m!\,(n-m)!\,(n-2m)!}. \tag{10}$$

The resulting solution of Legendre's differential equation (1) is called the **Legendre polynomial** *of degree n* and is denoted by $P_n(x)$.

From (10) we obtain

$$\textbf{(11)} \qquad P_n(x) = \sum_{m=0}^{M} (-1)^m \frac{(2n-2m)!}{2^n m!\,(n-m)!\,(n-2m)!}\, x^{n-2m}$$

$$= \frac{(2n)!}{2^n (n!)^2}\, x^n - \frac{(2n-2)!}{2^n 1!\,(n-1)!\,(n-2)!}\, x^{n-2} + - \cdots$$

where $M = n/2$ or $(n-1)/2$, whichever is an integer. The first few of these functions are (Fig. 104)

$$\textbf{(11')} \qquad \begin{aligned} &P_0(x) = 1, && P_1(x) = x \\ &P_2(x) = \tfrac{1}{2}(3x^2 - 1), && P_3(x) = \tfrac{1}{2}(5x^3 - 3x) \\ &P_4(x) = \tfrac{1}{8}(35x^4 - 30x^2 + 3), && P_5(x) = \tfrac{1}{8}(63x^5 - 70x^3 + 15x) \end{aligned}$$

and so on. You may now program (11) on your CAS and calculate $P_n(x)$ as needed.

The so-called **orthogonality** of the Legendre polynomials will be considered in Secs. 5.7 and 5.8.

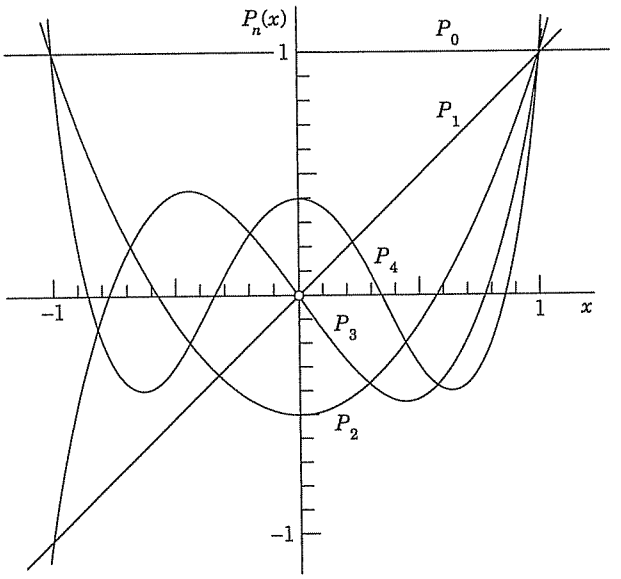


Fig. 104. Legendre polynomials

PROBLEM SET 5.3

1. Verify that the polynomials in (11′) satisfy Legendre's equation.
2. Derive (11′) from (11).
3. Obtain P_6 and P_7 from (11).
4. **(Convergence)** Show that for any n for which (6) or (7) does not reduce to a polynomial, the series has radius of convergence 1.
5. **(Legendre function $Q_0(x)$ for $n = 0$)** Show that (6) with $n = 0$ gives $y_1(x) = P_0(x) = 1$ and (7) gives

$$y_2(x) = x + \frac{2}{3!}x^3 + \frac{(-3)(-1)\cdot 2\cdot 4}{5!}x^5 + \cdots$$

$$= x + \frac{x^3}{3} + \frac{x^5}{5} + \cdots = \frac{1}{2}\ln\frac{1+x}{1-x}.$$

Verify this by solving (1) with $n = 0$, setting $z = y'$ and separating variables.

6. **(Legendre function $-Q_1(x)$ for $n = 1$)** Show that (7) with $n = 1$ gives $y_2(x) = P_1(x) = x$ and (6) gives $y_1(x) = -Q_1(x)$ (the minus sign in the notation being conventional),

$$y_1(x) = 1 - \frac{x^2}{1} - \frac{x^4}{3} - \frac{x^6}{5} - \cdots$$
$$= 1 - x\left(x + \frac{x^3}{3} + \frac{x^5}{5} + \cdots\right)$$
$$= 1 - \frac{1}{2}\,x \ln \frac{1+x}{1-x}.$$

7. **(ODE)** Find a solution of $(a^2 - x^2)y'' - 2xy' + n(n+1)y = 0$, $a \neq 0$, by reduction to the Legendre equation.

8. **[Rodrigues's formula (12)][2]** Applying the binomial theorem to $(x^2 - 1)^n$, differentiating it n times term by term, and comparing the result with (11), show that

$$\text{(12)} \qquad P_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}\left[(x^2-1)^n\right].$$

9. **(Rodrigues's formula)** Obtain (11′) from (12).

10–13 CAS PROBLEMS

10. Graph $P_2(x), \cdots, P_{10}(x)$ on common axes. For what x (approximately) and $n = 2, \cdots, 10$ is $|P_n(x)| < \frac{1}{2}$?

11. From what n on will your CAS no longer produce faithful graphs of $P_n(x)$? Why?

12. Graph $Q_0(x)$, $Q_1(x)$, and some further Legendre functions.

13. Substitute $a_s x^s + a_{s+1}x^{s+1} + a_{s+2}x^{s+2}$ into Legendre's equation and obtain the coefficient recursion (4).

14. **TEAM PROJECT. Generating Functions.** Generating functions play a significant role in modern applied mathematics (see [GR5]). The idea is simple. If we want to study a certain sequence $(f_n(x))$ and can find a function

$$G(u, x) = \sum_{n=0}^{\infty} f_n(x)u^n,$$

we may obtain properties of $(f_n(x))$ from those of G, which "generates" this sequence and is called a **generating function** of the sequence.

(a) Legendre polynomials. Show that

$$\text{(13)} \qquad G(u, x) = \frac{1}{\sqrt{1 - 2xu + u^2}} = \sum_{n=0}^{\infty} P_n(x)u^n$$

is a generating function of the Legendre polynomials. *Hint:* Start from the binomial expansion of $1/\sqrt{1-v}$, then set $v = 2xu - u^2$, multiply the powers of $2xu - u^2$ out, collect all the terms involving u^n, and verify that the sum of these terms is $P_n(x)u^n$.

(b) Potential theory. Let A_1 and A_2 be two points in space (Fig. 105, $r_2 > 0$). Using (13), show that

$$\frac{1}{r} = \frac{1}{\sqrt{r_1^2 + r_2^2 - 2r_1r_2\cos\theta}} = \frac{1}{r_2}\sum_{m=0}^{\infty} P_m(\cos\theta)\left(\frac{r_1}{r_2}\right)^m.$$

This formula has applications in potential theory. (Q/r is the electrostatic potential at A_2 due to a charge Q located at A_1. And the series expresses $1/r$ in terms of the distances of A_1 and A_2 from any origin O and the angle θ between the segments OA_1 and OA_2.)

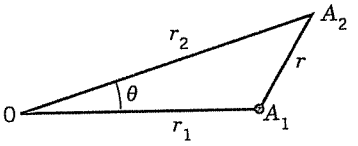


Fig. 105. Team Project 14

(c) Further applications of (13). Show that $P_n(1) = 1$, $P_n(-1) = (-1)^n$, $P_{2n+1}(0) = 0$, and

$$P_{2n}(0) = (-1)^n \cdot 1 \cdot 3 \cdots (2n-1)/[2 \cdot 4 \cdots (2n)].$$

(d) Bonnet's recursion.[3] Differentiating (13) with respect to u, using (13) in the resulting formula, and comparing coefficients of u^n, obtain the *Bonnet recursion*

$$\text{(14)} \qquad (n+1)P_{n+1}(x) = (2n+1)xP_n(x) - nP_{n-1}(x),$$

where $n = 1, 2, \cdots$. This formula is useful for computations, the loss of significant digits being small (except near zeros). Try (14) out for a few computations of your own choice.

---

[2] OLINDE RODRIGUES (1794–1851), French mathematician and economist.

[3] OSSIAN BONNET (1819–1892), French mathematician, whose main work was in differential geometry.

15. (Associated Legendre functions) The associated Legendre functions $P_n^k(x)$ play a role in quantum physics. They are defined by

$$P_n^k(x) = (1 - x^2)^{k/2} \frac{d^k P_n}{dx^k} \tag{15}$$

and are solutions of the ODE

$$(1 - x^2)y'' - 2xy' + \left[n(n+1) - \frac{k^2}{1 - x^2}\right] y = 0. \tag{16}$$

Find $P_1^1(x)$, $P_2^1(x)$, $P_2^2(x)$, and $P_4^2(x)$ and verify that they satisfy (16).

5.4 Frobenius Method

Several second-order ODEs of considerable practical importance—the famous Bessel equation among them—have coefficients that are not analytic (definition in Sec. 5.2), but are "not too bad," so that these ODEs can still be solved by series (power series times a logarithm or times a fractional power of x, etc.). Indeed, the following theorem permits an extension of the power series method that is called the **Frobenius method.** The latter—as well as the power series method itself—has gained in significance due to the use of software in the actual calculations.

THEOREM 1

Frobenius Method

Let $b(x)$ and $c(x)$ be any functions that are analytic at $x = 0$. Then the ODE

$$y'' + \frac{b(x)}{x} y' + \frac{c(x)}{x^2} y = 0 \tag{1}$$

has at least one solution that can be represented in the form

$$y(x) = x^r \sum_{m=0}^{\infty} a_m x^m = x^r(a_0 + a_1 x + a_2 x^2 + \cdots) \qquad (a_0 \neq 0) \tag{2}$$

where the exponent r may be any (real or complex) number (and r is chosen so that $a_0 \neq 0$).

The ODE (1) *also has a second solution (such that these two solutions are linearly independent) that may be similar to* (2) *(with a different r and different coefficients) or may contain a logarithmic term.* (Details in Theorem 2 below.)[4]

For example, Bessel's equation (to be discussed in the next section)

$$y'' + \frac{1}{x} y' + \left(\frac{x^2 - \nu^2}{x^2}\right) y = 0 \qquad (\nu \text{ a parameter})$$

[4]GEORG FROBENIUS (1849–1917), German mathematician, also known for his work on matrices and in group theory.

In this theorem we may replace x by $x - x_0$ with any number x_0. The condition $a_0 \neq 0$ is no restriction; it simply means that we factor out the highest possible power of x.

The singular point of (1) at $x = 0$ is sometimes called a **regular singular point,** a term confusing to the student, which we shall not use.

is of the form (1) with $b(x) = 1$ and $c(x) = x^2 - \nu^2$ analytic at $x = 0$, so that the theorem applies. This ODE could not be handled in full generality by the power series method.

Similarly, the so-called hypergeometric differential equation (see Problem Set 5.4) also requires the Frobenius method.

The point is that in (2) we have a power series times a single power of x whose exponent r is not restricted to be a nonnegative integer. (The latter restriction would make the whole expression a power series, by definition; see Sec. 5.1.)

The proof of the theorem requires advanced methods of complex analysis and can be found in Ref. [A11] listed in App. 1.

Regular and Singular Points

The following commonly used terms are practical. A **regular point** of

$$y'' + p(x)y' + q(x)y = 0$$

is a point x_0 at which the coefficients p and q are analytic. Then the power series method can be applied. If x_0 is not regular, it is called **singular.** Similarly, a **regular point** of the ODE

$$\tilde{h}(x)y'' + \tilde{p}(x)y'(x) + \tilde{q}(x)y = 0$$

is an x_0 at which $\tilde{h}$, $\tilde{p}$, $\tilde{q}$ are analytic and $\tilde{h}(x_0) \neq 0$ (so what we can divide by $\tilde{h}$ and get the previous standard form). If x_0 is not regular, it is called **singular.**

Indicial Equation, Indicating the Form of Solutions

We shall now explain the Frobenius method for solving (1). Multiplication of (1) by x^2 gives the more convenient form

$$x^2y'' + xb(x)y' + c(x)y = 0. \tag{1'}$$

We first expand $b(x)$ and $c(x)$ in power series,

$$b(x) = b_0 + b_1x + b_2x^2 + \cdots, \qquad c(x) = c_0 + c_1x + c_2x^2 + \cdots$$

or we do nothing if $b(x)$ and $c(x)$ are polynomials. Then we differentiate (2) term by term, finding

$$y'(x) = \sum_{m=0}^{\infty} (m + r)a_m x^{m+r-1} = x^{r-1}\,[ra_0 + (r + 1)a_1x + \cdots]$$

$$y''(x) = \sum_{m=0}^{\infty} (m + r)(m + r - 1)a_m x^{m+r-2} \tag{2*}$$

$$= x^{r-2}[r(r - 1)a_0 + (r + 1)ra_1x + \cdots].$$

By inserting all these series into (1') we readily obtain

$$x^r[r(r - 1)a_0 + \cdots] + (b_0 + b_1x + \cdots)x^r(ra_0 + \cdots) + (c_0 + c_1x + \cdots)x^r(a_0 + a_1x + \cdots) = 0. \tag{3}$$

We now equate the sum of the coefficients of each power $x^r, x^{r+1}, x^{r+2}, \cdots$ to zero. This yields a system of equations involving the unknown coefficients a_m. The equation corresponding to the power x^r is

$$[r(r-1) + b_0 r + c_0]a_0 = 0.$$

Since by assumption $a_0 \neq 0$, the expression in the brackets $[\cdots]$ must be zero. This gives

$$r(r-1) + b_0 r + c_0 = 0. \tag{4}$$

This important quadratic equation is called the **indicial equation** of the ODE (1). Its role is as follows.

The Frobenius method yields a basis of solutions. One of the two solutions will always be of the form (2), where r is a root of (4). The other solution will be of a form indicated by the indicial equation. There are three cases:

Case 1. Distinct roots not differing by an integer $1, 2, 3, \cdots$.

Case 2. A double root.

Case 3. Roots differing by an integer $1, 2, 3, \cdots$.

Cases 1 and 2 are not unexpected because of the Euler–Cauchy equation (Sec. 2.5), the simplest ODE of the form (1). Case 1 includes complex conjugate roots r_1 and $r_2 = \bar{r}_1$ because $r_1 - r_2 = r_1 - \bar{r}_1 = 2i \operatorname{Im} r_1$ is imaginary, so it cannot be a *real* integer. The form of a basis will be given in Theorem 2 (which is proved in App. 4), without a general theory of convergence, but convergence of the occurring series can be tested in each individual case as usual. Note that in Case 2 we ***must*** have a logarithm, whereas in Case 3 we *may* or *may not.*

THEOREM 2

Frobenius Method. Basis of Solutions. Three Cases

Suppose that the ODE (1) *satisfies the assumptions in Theorem* 1. *Let* r_1 *and* r_2 *be the roots of the indicial equation* (4). *Then we have the following three cases.*

Case* 1. *Distinct Roots Not Differing by an Integer. *A basis is*

$$y_1(x) = x^{r_1}(a_0 + a_1 x + a_2 x^2 + \cdots) \tag{5}$$

and

$$y_2(x) = x^{r_2}(A_0 + A_1 x + A_2 x^2 + \cdots) \tag{6}$$

with coefficients obtained successively from (3) *with* $r = r_1$ *and* $r = r_2$, *respectively.*

Case* 2. *Double Root $r_1 = r_2 = r$. *A basis is*

$$y_1(x) = x^r(a_0 + a_1 x + a_2 x^2 + \cdots) \qquad [r = \tfrac{1}{2}(1 - b_0)] \tag{7}$$

(of the same general form as before) and

$$y_2(x) = y_1(x) \ln x + x^r(A_1 x + A_2 x^2 + \cdots) \qquad (x > 0). \tag{8}$$

Case 3. Roots Differing by an Integer. *A basis is*

$$y_1(x) = x^{r_1}(a_0 + a_1x + a_2x^2 + \cdorts) \tag{9}$$

(of the same general form as before) and

$$y_2(x) = ky_1(x) \ln x + x^{r_2}(A_0 + A_1x + A_2x^2 + \cdots), \tag{10}$$

where the roots are so denoted that $r_1 - r_2 > 0$ and k may turn out to be zero.

Typical Applications

Technically, the Frobenius method is similar to the power series method, once the roots of the indicial equation have been determined. However, (5)–(10) merely indicate the general form of a basis, and a second solution can often be obtained more rapidly by reduction of order (Sec. 2.1).

EXAMPLE 1 **Euler–Cauchy Equation, Illustrating Cases 1 and 2 and Case 3 without a Logarithm**

For the Euler–Cauchy equation (Sec. 2.5)

$$x^2y'' + b_0xy' + c_0y = 0 \qquad (b_0, c_0 \text{ constant})$$

substitution of $y = x^r$ gives the auxiliary equation

$$r(r - 1) + b_0r + c_0 = 0,$$

which is the indicial equation [and $y = x^r$ is a very special form of (2)!]. For different roots r_1, r_2 we get a basis $y_1 = x^{r_1}$, $y_2 = x^{r_2}$, and for a double root r we get a basis x^r, $x^r \ln x$. Accordingly, for this simple ODE, Case 3 plays no extra role. ■

EXAMPLE 2 **Illustration of Case 2 (Double Root)**

Solve the ODE

$$x(x - 1)y'' + (3x - 1)y' + y = 0. \tag{11}$$

(This is a special hypergeometric equation, as we shall see in the problem set.)

Solution. Writing (11) in the standard form (1), we see that it satisfies the assumptions in Theorem 1. [What are $b(x)$ and $c(x)$ in (11)?] By inserting (2) and its derivatives (2*) into (11) we obtain

$$\begin{aligned}&\sum_{m=0}^{\infty} (m + r)(m + r - 1)a_mx^{m+r} - \sum_{m=0}^{\infty} (m + r)(m + r - 1)a_mx^{m+r-1} \\ &+ 3\sum_{m=0}^{\infty} (m + r)a_mx^{m+r} - \sum_{m=0}^{\infty} (m + r)a_mx^{m+r-1} + \sum_{m=0}^{\infty} a_mx^{m+r} = 0.\end{aligned} \tag{12}$$

The smallest power is x^{r-1}, occurring in the second and the fourth series; by equating the sum of its coefficients to zero we have

$$[-r(r - 1) - r]a_0 = 0, \qquad \text{thus} \qquad r^2 = 0.$$

Hence this indicial equation has the double root $r = 0$.

First Solution. We insert this value $r = 0$ into (12) and equate the sum of the coefficients of the power x^s to zero, obtaining

$$s(s-1)a_s - (s+1)sa_{s+1} + 3sa_s - (s+1)a_{s+1} + a_s = 0$$

thus $a_{s+1} = a_s$. Hence $a_0 = a_1 = a_2 = \cdots$, and by choosing $a_0 = 1$ we obtain the solution

$$y_1(x) = \sum_{m=0}^{\infty} x^m = \frac{1}{1-x} \qquad (|x| < 1).$$

Second Solution. We get a second independent solution y_2 by the method of reduction of order (Sec. 2.1), substituting $y_2 = uy_1$ and its derivatives into the equation. This leads to (9), Sec. 2.1, which we shall use in this example, instead of starting reduction of order from scratch (as we shall do in the next example). In (9) of Sec. 2.1 we have $p = (3x-1)/(x^2-x)$, the coefficient of y' in (11) ***in standard form.*** By partial fractions,

$$-\int p\,dx = -\int \frac{3x-1}{x(x-1)}\,dx = -\int \left(\frac{2}{x-1} + \frac{1}{x}\right) dx = -2\ln(x-1) - \ln x.$$

Hence (9), Sec. 2.1, becomes

$$u' = U = y_1^{-2}e^{-\int p\,dx} = \frac{(x-1)^2}{(x-1)^2 x} = \frac{1}{x}, \qquad u = \ln x, \qquad y_2 = uy_1 = \frac{\ln x}{1-x}.$$

y_1 and y_2 are shown in Fig. 106. These functions are linearly independent and thus form a basis on the interval $0 < x < 1$ (as well as on $1 < x < \infty$). ▩

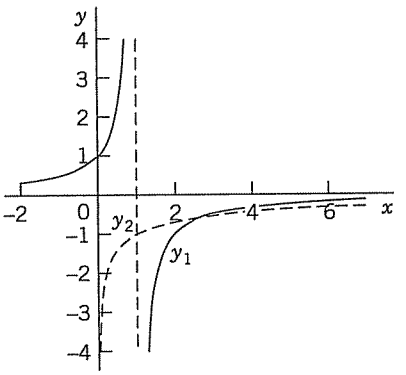


Fig. 106. Solutions in Example 2

EXAMPLE 3 **Case 3, Second Solution with Logarithmic Term**

Solve the ODE

$$(x^2 - x)y'' - xy' + y = 0. \tag{13}$$

Solution. Substituting (2) and (2*) into (13), we have

$$(x^2-x)\sum_{m=0}^{\infty}(m+r)(m+r-1)a_m x^{m+r-2} - x\sum_{m=0}^{\infty}(m+r)a_m x^{m+r-1} + \sum_{m=0}^{\infty} a_m x^{m+r} = 0.$$

We now take x^2, x, and x inside the summations and collect all terms with power x^{m+r} and simplify algebraically,

$$\sum_{m=0}^{\infty}(m+r-1)^2 a_m x^{m+r} - \sum_{m=0}^{\infty}(m+r)(m+r-1)a_m x^{m+r-1} = 0.$$

In the first series we set $m = s$ and in the second $m = s+1$, thus $s = m-1$. Then

$$\sum_{s=0}^{\infty}(s+r-1)^2 a_s x^{s+r} - \sum_{s=-1}^{\infty}(s+r+1)(s+r)a_{s+1}x^{s+r} = 0. \tag{14}$$

The lowest power is x^{r-1} (take $s = -1$ in the second series) and gives the indicial equation

$$r(r-1) = 0.$$

The roots are $r_1 = 1$ and $r_2 = 0$. They differ by an integer. This is Case 3.

First Solution. From (14) with $r = r_1 = 1$ we have

$$\sum_{s=0}^{\infty} [s^2 a_s - (s+2)(s+1)a_{s+1}]x^{s+1} = 0.$$

This gives the recurrence relation

$$a_{s+1} = \frac{s^2}{(s+2)(s+1)}\, a_s \qquad (s = 0, 1, \cdots).$$

Hence $a_1 = 0$, $a_2 = 0$, $\cdots$ successively. Taking $a_0 = 1$, we get as a first solution $y_1 = x^{r_1}a_0 = x$.

Second Solution. Applying reduction of order (Sec. 2.1), we substitute $y_2 = y_1 u = xu$, $y_2' = xu' + u$ and $y_2'' = xu'' + 2u'$ into the ODE, obtaining

$$(x^2 - x)(xu'' + 2u') - x(xu' + u) + xu = 0.$$

xu drops out. Division by x and simplification give

$$(x^2 - x)u'' + (x - 2)u' = 0.$$

From this, using partial fractions and integrating (taking the integration constant zero), we get

$$\frac{u''}{u'} = -\frac{x-2}{x^2-x} = -\frac{2}{x} + \frac{1}{x-1}, \qquad \ln u' = \ln\left|\frac{x-1}{x^2}\right|.$$

Taking exponents and integrating (again taking the integration constant zero), we obtain

$$u' = \frac{x-1}{x^2} = \frac{1}{x} - \frac{1}{x^2}, \qquad u = \ln x + \frac{1}{x}, \qquad y_2 = xu = x\ln x + 1.$$

y_1 and y_2 are linearly independent, and y_2 has a logarithmic term. Hence y_1 and y_2 constitute a basis of solutions for positive x. ▩

The Frobenius method solves the **hypergeometric equation,** whose solutions include many known functions as special cases (see the problem set). In the next section we use the method for solving Bessel's equation.

PROBLEM SET 5.4

1–17 BASIS OF SOLUTIONS BY THE FROBENIUS METHOD

Find a basis of solutions. Try to identify the series as expansions of known functions. (Show the details of your work.)

1. $xy'' + 2y' - xy = 0$

2. $(x+2)^2y'' - 2y = 0$

3. $xy'' + 5y' + xy = 0$

4. $2xy'' + (3 - 4x)y' + (2x - 3)y = 0$

5. $x^2y'' + 4xy' + (x^2 + 2)y = 0$

6. $4xy'' + 2y' + y = 0$

7. $(x+3)^2y'' - 9(x+3)y' + 25y = 0$

8. $xy'' - y = 0$

9. $xy'' + (2x+1)y' + (x+1)y = 0$

10. $x^2y'' + 2x^3y' + (x^2 - 2)y = 0$

11. $(x^2 + x)y'' + (4x+2)y' + 2y = 0$

12. $x^2y'' + 6xy' + (4x^2 + 6)y = 0$

13. $2xy'' - (8x - 1)y' + (8x - 2)y = 0$

14. $xy'' + y' - xy = 0$

15. $(x-4)^2y'' - (x-4)y' - 35y = 0$

16. $x^2y'' + 4xy' - (x^2 - 2)y = 0$

17. $y'' + (x-6)y = 0$

18. TEAM PROJECT. Hypergeometric Equation, Series, and Function. Gauss's **hypergeometric ODE**[5] is

$$x(1-x)y'' + [c - (a+b+1)x]y' - aby = 0. \tag{15}$$

Here, a, b, c are constants. This ODE is of the form $p_2y'' + p_1y' + p_0y = 0$, where p_2, p_1, p_0 are polynomials of degree 2, 1, 0, respectively. These polynomials are written so that the series solution takes a most practical form, namely,

$$y_1(x) = 1 + \frac{ab}{1!\,c}x + \frac{a(a+1)b(b+1)}{2!\,c(c+1)}x^2 + \frac{a(a+1)(a+2)b(b+1)(b+2)}{3!\,c(c+1)(c+2)}x^3 + \cdots. \tag{16}$$

This series is called the **hypergeometric series.** Its sum $y_1(x)$ is called the **hypergeometric function** and is denoted by $F(a, b, c; x)$. Here, $c \neq 0, -1, -2, \cdots$. By choosing specific values of a, b, c we can obtain an incredibly large number of special functions as solutions of (15) [see the small sample of elementary functions in part (c)]. This accounts for the importance of (15).

(a) Hypergeometric series and function. Show that the indicial equation of (15) has the roots $r_1 = 0$ and $r_2 = 1 - c$. Show that for $r_1 = 0$ the Frobenius method gives (16). Motivate the name for (16) by showing that

$$F(1, 1, 1; x) = F(1, b, b; x) = F(a, 1, a; x) = \frac{1}{1-x}.$$

(b) Convergence. For what a or b will (16) reduce to a polynomial? Show that for any other a, b, c ($c \neq 0, -1, -2, \cdots$) the series (16) converges when $|x| < 1$.

(c) Special cases. Show that

$$(1+x)^n = F(-n, b, b; -x),$$

$$(1-x)^n = 1 - nxF(1-n, 1, 2; x),$$

$$\arctan x = xF(\tfrac{1}{2}, 1, \tfrac{3}{2}; -x^2),$$

$$\arcsin x = xF(\tfrac{1}{2}, \tfrac{1}{2}, \tfrac{3}{2}; x^2),$$

$$\ln(1+x) = xF(1, 1, 2; -x),$$

$$\ln\frac{1+x}{1-x} = 2xF(\tfrac{1}{2}, 1, \tfrac{3}{2}; x^2).$$

Find more such relations from the literature on special functions.

(d) Second solution. Show that for $r_2 = 1 - c$ the Frobenius method yields the following solution (where $c \neq 2, 3, 4, \cdots$):

$$y_2(x) = x^{1-c}\left(1 + \frac{(a-c+1)(b-c+1)}{1!\,(-c+2)}x + \frac{(a-c+1)(a-c+2)(b-c+1)(b-c+2)}{2!\,(-c+2)(-c+3)}x^2 + \cdots\right). \tag{17}$$

Show that

$$y_2(x) = x^{1-c}F(a-c+1, b-c+1, 2-c; x).$$

(e) On the generality of the hypergeometric equation. Show that

$$(t^2 + At + B)\ddot{y} + (Ct + D)\dot{y} + Ky = 0 \tag{18}$$

with $\dot{y} = dy/dt$, etc., constant A, B, C, D, K, and $t^2 + At + B = (t - t_1)(t - t_2)$, $t_1 \neq t_2$, can be reduced to the hypergeometric equation with independent variable

$$x = \frac{t - t_1}{t_2 - t_1}$$

and parameters related by $Ct_1 + D = -c(t_2 - t_1)$, $C = a + b + 1$, $K = ab$. From this you see that (15) is a "normalized form" of the more general (18) and that various cases of (18) can thus be solved in terms of hypergeometric functions.

19–24 HYPERGEOMETRIC EQUATIONS

Find a general solution in terms of hypergeometric functions.

19. $x(1-x)y'' + (\frac{1}{2} - 2x)y' - \frac{1}{4}y = 0$

20. $2x(1-x)y'' - (1+6x)y' - 2y = 0$

21. $x(1-x)y'' + \frac{1}{2}y' + 2y = 0$

22. $3t(1+t)\ddot{y} + t\dot{y} - y = 0$

23. $2(t^2 - 5t + 6)\ddot{y} + (2t - 3)\dot{y} - 8y = 0$

24. $4(t^2 - 3t + 2)\ddot{y} - 2\dot{y} + y = 0$

---

[5]CARL FRIEDRICH GAUSS (1777–1855), great German mathematician. He already made the first of his great discoveries as a student at Helmstedt and Göttingen. In 1807 he became a professor and director of the Observatory at Göttingen. His work was of basic importance in algebra, number theory, differential equations, differential geometry, non-Euclidean geometry, complex analysis, numeric analysis, astronomy, geodesy, electromagnetism, and theoretical mechanics. He also paved the way for a general and systematic use of complex numbers.

5.5 Bessel's Equation. Bessel Functions $J_\nu(x)$

One of the most important ODEs in applied mathematics in **Bessel's equation**,[6]

$$x^2y'' + xy' + (x^2 - \nu^2)y = 0. \tag{1}$$

Its diverse applications range from electric fields to heat conduction and vibrations (see Sec. 12.9). It often appears when a problem shows *cylindrical symmetry* (just as Legendre's equation may appear in cases of *spherical* symmetry). The parameter ν in (1) is a given number. We assume that ν is real and nonnegative.

Bessel's equation can be solved by the Frobenius method, as we mentioned at the beginning of the preceding section, where the equation is written in standard form (obtained by dividing (1) by x^2). Accordingly, we substitute the series

$$y(x) = \sum_{m=0}^{\infty} a_m x^{m+r} \qquad (a_0 \neq 0) \tag{2}$$

with undetermined coefficients and its derivatives into (1). This gives

$$\sum_{m=0}^{\infty} (m + r)(m + r - 1)a_m x^{m+r} + \sum_{m=0}^{\infty} (m + r)a_m x^{m+r} + \sum_{m=0}^{\infty} a_m x^{m+r+2} - \nu^2 \sum_{m=0}^{\infty} a_m x^{m+r} = 0.$$

We equate the sum of the coefficients of x^{s+r} to zero. Note that this power x^{s+r} corresponds to $m = s$ in the first, second, and fourth series, and to $m = s - 2$ in the third series. Hence for $s = 0$ and $s = 1$, the third series does not contribute since $m \geqq 0$. For $s = 2, 3, \cdots$ all four series contribute, so that we get a general formula for all these s. We find

$$\begin{aligned}
&\text{(a)} && r(r - 1)a_0 + ra_0 - \nu^2 a_0 = 0 && (s = 0)\\
&\text{(b)} && (r + 1)ra_1 + (r + 1)a_1 - \nu^2 a_1 = 0 && (s = 1)\\
&\text{(c)} && (s + r)(s + r - 1)a_s + (s + r)a_s + a_{s-2} - \nu^2 a_s = 0 && (s = 2, 3, \cdots).
\end{aligned} \tag{3}$$

From (3a) we obtain the **indicial equation** by dropping a_0,

$$(r + \nu)(r - \nu) = 0. \tag{4}$$

The roots are $r_1 = \nu\ (\geqq 0)$ and $r_2 = -\nu$.

---

[6]FRIEDRICH WILHELM BESSEL (1784–1846), German astronomer and mathematician, studied astronomy on his own in his spare time as an apprentice of a trade company and finally became director of the new Königsberg Observatory.

Formulas on Bessel functions are contained in Ref. [GR1] and the standard treatise [A13].

Coefficient Recursion for $r = r_1 = \nu$. For $r = \nu$, Eq. (3b) reduces to $(2\nu + 1)a_1 = 0$. Hence $a_1 = 0$ since $\nu \geqq 0$. Substituting $r = \nu$ in (3c) and combining the three terms containing a_s gives simply

$$(s + 2\nu)sa_s + a_{s-2} = 0. \tag{5}$$

Since $a_1 = 0$ and $\nu \geqq 0$, it follows from (5) that $a_3 = 0$, $a_5 = 0, \cdots$. Hence we have to deal only with *even-numbered* coefficients a_s with $s = 2m$. For $s = 2m$, Eq. (5) becomes

$$(2m + 2\nu)2ma_{2m} + a_{2m-2} = 0.$$

Solving for a_{2m} gives the recursion formula

$$a_{2m} = -\frac{1}{2^2 m(\nu + m)} a_{2m-2}, \qquad m = 1, 2, \cdots. \tag{6}$$

From (6) we can now determine a_2, $a_4, \cdots$ successively. This gives

$$a_2 = -\frac{a_0}{2^2(\nu + 1)}$$

$$a_4 = -\frac{a_2}{2^2 2(\nu + 2)} = \frac{a_0}{2^4 2!\,(\nu + 1)(\nu + 2)}$$

and so on, and in general

$$a_{2m} = \frac{(-1)^m a_0}{2^{2m} m!\,(\nu + 1)(\nu + 2) \cdots (\nu + m)}, \qquad m = 1, 2, \cdots. \tag{7}$$

Bessel Functions $J_n(x)$ For Integer $\nu = n$

Integer values of ν are denoted by n. This is standard. For $\nu = n$ the relation (7) becomes

$$a_{2m} = \frac{(-1)^m a_0}{2^{2m} m!\,(n + 1)(n + 2) \cdots (n + m)}, \qquad m = 1, 2, \cdots. \tag{8}$$

a_0 is still arbitrary, so that the series (2) with these coefficients would contain this arbitrary factor a_0. This would be a highly impractical situation for developing formulas or computing values of this new function. Accordingly, we have to make a choice. $a_0 = 1$ would be possible, but more practical turns out to be

$$a_0 = \frac{1}{2^n n!}. \tag{9}$$

because then $n!(n + 1) \cdots (n + m) = (m + n)!$ in (8), so that (8) simply becomes

$$a_{2m} = \frac{(-1)^m}{2^{2m+n} m!\,(n + m)!}, \qquad m = 1, 2, \cdots. \tag{10}$$

This simplicity of the denominator of (10) partially motivates the choice (9). With these coefficients and $r_1 = \nu = n$ we get from (2) a particular solution of (1), denoted by $J_n(x)$ and given by

$$J_n(x) = x^n \sum_{m=0}^{\infty} \frac{(-1)^m x^{2m}}{2^{2m+n} m!\,(n+m)!}. \tag{11}$$

$J_n(x)$ is called the **Bessel function of the first kind** *of order n.* The series (11) converges for all x, as the ratio test shows. In fact, it converges very rapidly because of the factorials in the denominator.

EXAMPLE 1 **Bessel Functions $J_0(x)$ and $J_1(x)$**

For $n = 0$ we obtain from (11) the **Bessel function of order 0**

$$J_0(x) = \sum_{m=0}^{\infty} \frac{(-1)^m x^{2m}}{2^{2m}(m!)^2} = 1 - \frac{x^2}{2^2(1!)^2} + \frac{x^4}{2^4(2!)^2} - \frac{x^6}{2^6(3!)^2} + - \cdots \tag{12}$$

which looks similar to a cosine (Fig. 107). For $n = 1$ we obtain the **Bessel function of order 1**

$$J_1(x) = \sum_{m=0}^{\infty} \frac{(-1)^m x^{2m+1}}{2^{2m+1} m!\,(m+1)!} = \frac{x}{2} - \frac{x^3}{2^3 1!2!} + \frac{x^5}{2^5 2!3!} - \frac{x^7}{2^7 3!4!} + - \cdots, \tag{13}$$

which looks similar to a sine (Fig. 107). But the zeros of these functions are not completely regularly spaced (see also Table A1 in App. 5) and the height of the "waves" decreases with increasing x. Heuristically, n^2/x^2 in (1) in standard form [(1) divided by x^2] is zero (if $n = 0$) or small in absolute value for large x, and so is y'/x, so that then Bessel's equation comes close to $y'' + y = 0$, the equation of $\cos x$ and $\sin x$; also y'/x acts as a "damping term," in part responsible for the decrease in height. One can show that for large x,

$$J_n(x) \sim \sqrt{\frac{2}{\pi x}} \cos\left(x - \frac{n\pi}{2} - \frac{\pi}{4}\right) \tag{14}$$

where $\sim$ is read **"asymptotically equal"** and means that *for fixed n* the quotient of the two sides approaches 1 as $x \to \infty$.

Formula (14) is surprisingly accurate even for smaller x (> 0). For instance, it will give you good starting values in a computer program for the basic task of computing zeros. For example, for the first three zeros of J_0 you obtain the values 2.356 (2.405 exact to 3 decimals, error 0.049), 5.498 (5.520, error 0.022), 8.639 (8.654, error 0.015), etc. ■

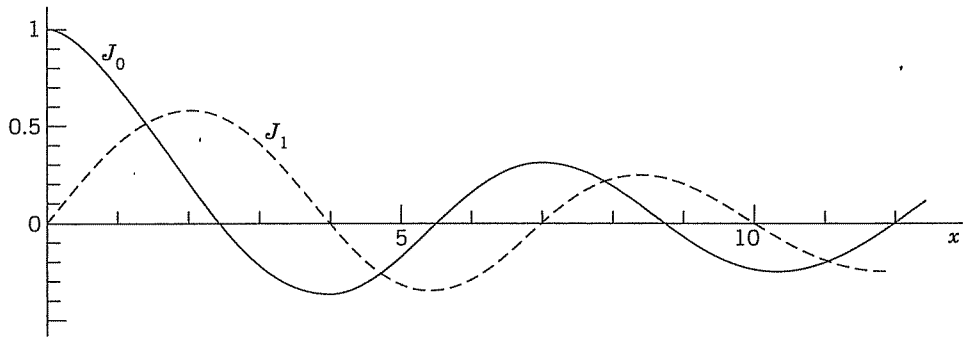


Fig. 107. Bessel functions of the first kind J_0 and J_1

Bessel Functions $J_\nu(x)$ for any $\nu \geqq 0$. Gamma Function

We now extend our discussion from integer $\nu = n$ to any $\nu \geqq 0$. All we need is an extension of the factorials in (9) and (11) to any ν. This is done by the **gamma function** $\Gamma(\nu)$ defined by the integral

(15) $$\Gamma(\nu) = \int_0^\infty e^{-t} t^{\nu-1}\, dt \qquad (\nu > 0).$$

By integration by parts we obtain

$$\Gamma(\nu + 1) = \int_0^\infty e^{-t} t^\nu\, dt = -e^{-t} t^\nu \Big|_0^\infty + \nu \int_0^\infty e^{-t} t^{\nu-1}\, dt.$$

The first expression on the right is zero. The integral on the right is $\Gamma(\nu)$. This yields the basic functional relation

(16) $$\Gamma(\nu + 1) = \nu\Gamma(\nu).$$

Now by (15)

$$\Gamma(1) = \int_0^\infty e^{-t}\, dt = -e^{-t} \Big|_0^\infty = 0 - (-1) = 1.$$

From this and (16) we obtain successively $\Gamma(2) = \Gamma(1) = 1!$, $\Gamma(3) = 2\Gamma(2) = 2!$, $\cdots$ and in general

(17) $$\Gamma(n + 1) = n! \qquad (n = 0, 1, \cdots).$$

This shows the ***the gamma function does in fact generalize the factorial function.***

Now in (9) we had $a_0 = 1/(2^n n!)$. This is $1/(2^n\Gamma(n + 1))$ by (17). It suggests to choose, for any ν,

(18) $$a_0 = \frac{1}{2^\nu \Gamma(\nu + 1)}.$$

Then (7) becomes

$$a_{2m} = \frac{(-1)^m}{2^{2m} m!\, (\nu + 1)(\nu + 2) \cdots (\nu + m) 2^\nu \Gamma(\nu + 1)}.$$

But (16) gives in the denominator

$$(\nu + 1)\Gamma(\nu + 1) = \Gamma(\nu + 2), \qquad (\nu + 2)\Gamma(\nu + 2) = \Gamma(\nu + 3)$$

and so on, so that

$$(\nu + 1)(\nu + 2) \cdots (\nu + m)\Gamma(\nu + 1) = \Gamma(\nu + m + 1).$$

Hence because of our (standard!) choice (18) of a_0 the coefficients (7) simply are

$$a_{2m} = \frac{(-1)^m}{2^{2m+\nu} m!\, \Gamma(\nu + m + 1)}. \tag{19}$$

With these coefficients and $r = r_1 = \nu$ we get from (2) a particular solution of (1), denoted by $J_\nu(x)$ and given by

$$J_\nu(x) = x^\nu \sum_{m=0}^{\infty} \frac{(-1)^m x^{2m}}{2^{2m+\nu} m!\, \Gamma(\nu + m + 1)}. \tag{20}$$

$J_\nu(x)$ is called the **Bessel function of the first kind of order ν.** The series (20) converges for all x, as one can verify by the ratio test.

General Solution for Noninteger ν. Solution $J_{-\nu}$

For a general solution, in addition to J_ν we need a second linearly independent solution. For ν not an integer this is easy. Replacing ν by $-\nu$ in (20), we have

$$J_{-\nu}(x) = x^{-\nu} \sum_{m=0}^{\infty} \frac{(-1)^m x^{2m}}{2^{2m-\nu} m!\, \Gamma(m - \nu + 1)}. \tag{21}$$

Since Bessel's equation involves ν^2, the functions J_ν and $J_{-\nu}$ are solutions of the equation for the same ν. If ν is not an integer, they are linearly independent, because the first term in (20) and the first term in (21) are finite nonzero multiples of x^ν and $x^{-\nu}$, respectively. $x = 0$ must be excluded in (21) because of the factor $x^{-\nu}$ (with $\nu > 0$). This gives

THEOREM 1

General Solution of Bessel's Equation

If ν is not an integer, a general solution of Bessel's equation for all $x \neq 0$ is

$$y(x) = c_1 J_\nu(x) + c_2 J_{-\nu}(x). \tag{22}$$

But if ν is an integer, then (22) is not a general solution because of linear dependence:

THEOREM 2

Linear Dependence of Bessel Functions J_n and J_{-n}

For integer $\nu = n$ the Bessel functions $J_n(x)$ and $J_{-n}(x)$ are linearly dependent, because

$$J_{-n}(x) = (-1)^n J_n(x) \qquad (n = 1, 2, \cdots). \tag{23}$$

PROOF We use (21) and let ν approach a positive integer n. Then the gamma functions in the coefficients of the first n terms become infinite (see Fig. 552 in App. A3.1), the coefficients become zero, and the summation starts with $m = n$. Since in this case $\Gamma(m - n + 1) = (m - n)!$ by (17), we obtain

$$J_{-n}(x) = \sum_{m=n}^{\infty} \frac{(-1)^m x^{2m-n}}{2^{2m-n} m!\,(m-n)!} = \sum_{s=0}^{\infty} \frac{(-1)^{n+s} x^{2s+n}}{2^{2s+n}(n+s)!\,s!} \qquad (m = n + s).$$

The last series represents $(-1)^n J_n(x)$, as you can see from (11) with m replaced by s. This completes the proof. ∎

A general solution for integer n will be given in the next section, based on some further interesting ideas.

Discovery of Properties From Series

Bessel functions are a model case for showing how to discover properties and relations of functions from series by which they are *defined.* Bessel functions satisfy an incredibly large number of relationships—look at Ref. [A13] in App. 1; also, find out what your CAS knows. In Theorem 3 we shall discuss four formulas that are backbones in applications.

THEOREM 3

Derivatives, Recursions

The derivative of $J_\nu(x)$ with respect to x can be expressed by $J_{\nu-1}(x)$ or $J_{\nu+1}(x)$ by the formulas

(24)

$$\textbf{(a)} \quad [x^\nu J_\nu(x)]' = x^\nu J_{\nu-1}(x)$$

$$\textbf{(b)} \quad [x^{-\nu} J_\nu(x)]' = -x^{-\nu} J_{\nu+1}(x).$$

Furthermore, $J_\nu(x)$ and its derivative satisfy the recurrence relations

(24)

$$\textbf{(c)} \quad J_{\nu-1}(x) + J_{\nu+1}(x) = \frac{2\nu}{x} J_\nu(x)$$

$$\textbf{(d)} \quad J_{\nu-1}(x) - J_{\nu+1}(x) = 2J_\nu'(x).$$

PROOF **(a)** We multiply (20) by x^ν and take $x^{2\nu}$ under the summation sign. Then we have

$$x^\nu J_\nu(x) = \sum_{m=0}^{\infty} \frac{(-1)^m x^{2m+2\nu}}{2^{2m+\nu} m!\,\Gamma(\nu + m + 1)}.$$

We now differentiate this, cancel a factor 2, pull $x^{2\nu-1}$ out, and use the functional relationship $\Gamma(\nu + m + 1) = (\nu + m)\Gamma(\nu + m)$ [see (16)]. Then (20) with $\nu - 1$ instead of ν shows that we obtain the right side of (24a). Indeed,

$$(x^\nu J_\nu)' = \sum_{m=0}^{\infty} \frac{(-1)^m 2(m+\nu) x^{2m+2\nu-1}}{2^{2m+\nu} m!\,\Gamma(\nu + m + 1)} = x^\nu x^{\nu-1} \sum_{m=0}^{\infty} \frac{(-1)^m x^{2m}}{2^{2m+\nu-1} m!\,\Gamma(\nu + m)}.$$

(b) Similarly, we multiply (20) by $x^{-\nu}$, so that x^ν in (20) cancels. Then we differentiate, cancel $2m$, and use $m! = m(m-1)!$. This gives, with $m = s + 1$,

$$(x^{-\nu}J_\nu)' = \sum_{m=1}^{\infty} \frac{(-1)^m x^{2m-1}}{2^{2m+\nu-1}(m-1)!\,\Gamma(\nu+m+1)} = \sum_{s=0}^{\infty} \frac{(-1)^{s+1}x^{2s+1}}{2^{2s+\nu+1}s!\,\Gamma(\nu+s+2)}.$$

Equation (20) with $\nu + 1$ instead of ν and s instead of m shows that the expression on the right is $-x^{-\nu}J_{\nu+1}(x)$. This proves (24b).

(c), (d) We perform the differentiation in (24a). Then we do the same in (24b) and multiply the result on both sides by $x^{2\nu}$. This gives

$$\text{(a*)} \qquad \nu x^{\nu-1}J_\nu + x^\nu J_\nu' = x^\nu J_{\nu-1}$$

$$\text{(b*)} \qquad -\nu x^{\nu-1}J_\nu + x^\nu J_\nu' = -x^\nu J_{\nu+1}.$$

Substracting (b*) from (a*) and dividing the result by x^ν gives (24c). Adding (a*) and (b*) and dividing the result by x^ν gives (24d). ∎

EXAMPLE 2 Application of Theorem 3 in Evaluation and Integration

Formula (24c) can be used recursively in the form

$$J_{\nu+1}(x) = \frac{2\nu}{x}J_\nu(x) - J_{\nu-1}(x)$$

for calculating Bessel functions of higher order from those of lower order. For instance, $J_2(x) = 2J_1(x)/x - J_0(x)$, so that J_2 can be obtained from tables of J_0 and J_1 (in App. 5 or, more accurately, in Ref. [GR1] in App. 1).

To illustrate how Theorem 3 helps in integration, we use (24b) with $\nu = 3$ integrated on both sides. This evaluates, for instance, the integral

$$I = \int_1^2 x^{-3}J_4(x)\,dx = -x^{-3}J_3(x)\Big|_1^2 = -\frac{1}{8}J_3(2) + J_3(1).$$

A table of J_3 (on p. 398 of Ref. [GR1]) or your CAS will give you

$$-\tfrac{1}{8}\cdot 0.128943 + 0.019563 = 0.003445.$$

Your CAS (or a human computer in precomputer times) obtains J_3 from (24), first using (24c) with $\nu = 2$, that is, $J_3 = 4x^{-1}J_2 - J_1$, then (24c) with $\nu = 1$, that is, $J_2 = 2x^{-1}J_1 - J_0$. Together,

$$\begin{aligned} I &= x^{-3}(4x^{-1}(2x^{-1}J_1 - J_0) - J_1)\Big|_1^2 \\ &= -\tfrac{1}{8}[2J_1(2) - 2J_0(2) - J_1(2)] + [8J_1(1) - 4J_0(1) - J_1(1)] \\ &= -\tfrac{1}{8}J_1(2) + \tfrac{1}{4}J_0(2) + 7J_1(1) - 4J_0(1). \end{aligned}$$

This is what you get, for instance, with Maple if you type int(· · ·). And if you type evalf(int(· · ·)), you obtain 0.003445448, in agreement with the result near the beginning of the example. ∎

In the theory of special functions it often happens that for certain values of a parameter a higher function becomes elementary. We have seen this in the last problem set, and we now show this for J_ν.

THEOREM 4

Elementary J_ν for Half-Integer Order ν

Bessel functions J_ν of orders $\pm\frac{1}{2}, \pm\frac{3}{2}, \pm\frac{5}{2}, \cdots$ are elementary; they can be expressed by finitely many cosines and sines and powers of x. In particular,

$$\textbf{(25)} \qquad \textbf{(a)}\quad J_{1/2}(x) = \sqrt{\frac{2}{\pi x}}\sin x, \qquad \textbf{(b)}\quad J_{-1/2}(x) = \sqrt{\frac{2}{\pi x}}\cos x.$$

PROOF When $\nu = \frac{1}{2}$, then (20) is

$$J_{1/2}(x) = \sqrt{x}\sum_{m=0}^{\infty}\frac{(-1)^m x^{2m}}{2^{2m+1/2}m!\,\Gamma(m+\frac{3}{2})} = \sqrt{\frac{2}{x}}\sum_{m=0}^{\infty}\frac{(-1)^m x^{2m+1}}{2^{2m+1}m!\,\Gamma(m+\frac{3}{2})}.$$

To simplify the denominator, we first write it out as a product AB, where

$$A = 2^m m! = 2m(2m-2)(2m-4)\cdots 4\cdot 2$$

and [use (16)]

$$B = 2^{m+1}\Gamma(m+\tfrac{3}{2}) = 2^{m+1}(m+\tfrac{1}{2})(m-\tfrac{1}{2})\cdots\tfrac{3}{2}\cdot\tfrac{1}{2}\Gamma(\tfrac{1}{2})$$
$$= (2m+1)(2m-1)\cdots 3\cdot 1\cdot\sqrt{\pi};$$

here we used

$$\textbf{(26)} \qquad \Gamma(\tfrac{1}{2}) = \sqrt{\pi}.$$

We see that the product of the two right sides of A and B is simply $(2m+1)!\sqrt{\pi}$, so that $J_{1/2}$ becomes

$$J_{1/2}(x) = \sqrt{\frac{2}{\pi x}}\sum_{m=0}^{\infty}\frac{(-1)^m x^{2m+1}}{(2m+1)!} = \sqrt{\frac{2}{\pi x}}\sin x,$$

as claimed. Differentiation and the use of (24a) with $\nu = \frac{1}{2}$ now gives

$$\left[\sqrt{x}J_{1/2}(x)\right]' = \sqrt{\frac{2}{\pi}}\cos x = x^{1/2}J_{-1/2}(x).$$

This proves (25b). From (25) follow further formulas successively by (24c), used as in Example 2. This completes the proof. ∎

EXAMPLE 3 **Further Elementary Bessel Functions**

From (24c) with $\nu = \frac{1}{2}$ and $\nu = -\frac{1}{2}$ and (25) we obtain

$$J_{3/2}(x) = \frac{1}{x}J_{1/2}(x) - J_{-1/2}(x) = \sqrt{\frac{2}{\pi x}}\left(\frac{\sin x}{x} - \cos x\right)$$

$$J_{-3/2}(x) = -\frac{1}{x}J_{-1/2}(x) - J_{1/2}(x) = -\sqrt{\frac{2}{\pi x}}\left(\frac{\cos x}{x} + \sin x\right)$$

respectively, and so on. ∎

We hope that our study has not only helped you to become acquainted with Bessel functions but has also convinced you that series can be quite useful in obtaining various properties of the corresponding functions.

PROBLEM SET 5.5

1. **(Convergence)** Show that the series in (11) converges for all x. Why is the convergence very rapid?
2. **(Approximation)** Show that for small $|x|$ we have $J_0 \approx 1 - 0.25x^2$. From this compute $J_0(x)$ for $x = 0, 0.1, 0.2, \cdots, 1.0$ and determine the error by using Table A1 in App. 5 or your CAS.
3. **("Large" values)** Using (14), compute $J_0(x)$ for $x = 1.0, 2.0, 3.0, \cdots, 8.0$, determine the error by Table A1 or your CAS, and comment.
4. **(Zeros)** Compute the first four positive zeros of $J_0(x)$ and $J_1(x)$ from (14). Determine the error and comment.

5–20 ODEs REDUCIBLE TO BESSEL'S EQUATION

Using the indicated substitutions, find a general solution in terms of J_ν and $J_{-\nu}$ or indicate when this is not possible. (This is just a sample of various ODEs reducible to Bessel's equation. Some more follow in the next problem set. Show the details of your work.)

5. **(ODE with two parameters)** $x^2y'' + xy' + (\lambda^2x^2 - \nu^2)y = 0 \quad (\lambda x = z)$
6. $x^2y'' + xy' + (x^2 - \frac{1}{16})y = 0$
7. $x^2y'' + xy' + \frac{1}{4}(x - \nu^2)y = 0 \quad (\sqrt{x} = z)$
8. $(2x + 1)^2y'' + 2(2x + 1)y' + 16x(x + 1)y = 0 \quad (2x + 1 = z)$
9. $xy'' - y' + 4xy = 0 \quad (y = xu, 2x = z)$
10. $x^2y'' + xy' + \frac{1}{4}(x^2 - 1)y = 0 \quad (x = 2z)$
11. $xy'' + (2\nu + 1)y' + xy = 0 \quad (y = x^{-\nu}u)$
12. $x^2y'' + xy' + 4(x^4 - \nu^2)y = 0 \quad (x^2 = z)$
13. $x^2y'' + xy' + 9(x^6 - \nu^2)y = 0 \quad (x^3 = z)$
14. $y'' + (e^{2x} - \frac{1}{9})y = 0 \quad (e^x = z)$
15. $xy'' + y = 0 \quad (y = \sqrt{x}u, 2\sqrt{x} = z)$
16. $16x^2y'' + 8xy' + (x^{1/2} + \frac{15}{16})y = 0 \quad (y = x^{1/4}u, x^{1/4} = z)$
17. $36x^2y'' + 18xy' + \sqrt{x}\,y = 0 \quad (y = x^{1/4}u, \frac{2}{3}x^{1/4} = z)$
18. $x^2y'' + xy' + \sqrt{x}\,y = 0 \quad (4x^{1/4} = z)$
19. $x^2y'' + \frac{1}{5}xy' + \sqrt{x}\,y = 0 \quad (y = x^{2/5}u, 4x^{1/4} = z)$
20. $x^2y'' + (1 - 2\nu)xy' + \nu^2(x^{2\nu} + 1 - \nu^2)y = 0 \quad (y = x^\nu u, x^\nu = z)$

21–28 APPLICATION OF (24): DERIVATIVES, INTEGRALS

Use the powerful formulas (24) to do Probs. 21–28. (Show the details of your work.)

21. **(Derivatives)** Show that $J_0'(x) = -J_1(x)$, $J_1'(x) = J_0(x) - J_1(x)/x$, $J_2'(x) = \frac{1}{2}[J_1(x) - J_3(x)]$.
22. **(Interlacing of zeros)** Using (24) and Rolle's theorem, show that between two consecutive zeros of $J_0(x)$ there is precisely one zero of $J_1(x)$.
23. **(Interlacing of zeros)** Using (24) and Rolle's theorem, show that between any two consecutive positive zeros of $J_n(x)$ there is precisely one zero of $J_{n+1}(x)$.
24. **(Bessel's equation)** Derive (1) from (24).
25. **(Basic integral formulas)** Show that

$$\int x^\nu J_{\nu-1}(x)\,dx = x^\nu J_\nu(x) + c,$$

$$\int x^{-\nu} J_{\nu+1}(x)\,dx = -x^{-\nu} J_\nu(x) + c,$$

$$\int J_{\nu+1}(x)\,dx = \int J_{\nu-1}(x)\,dx - 2J_\nu(x).$$

26. **(Integration)** Evaluate $\int x^{-1}J_4(x)\,dx$. (Use Prob. 25; integrate by parts.)
27. **(Integration)** Show that $\int x^2J_0(x)\,dx = x^2J_1(x) + xJ_0(x) - \int J_0(x)\,dx$. (The last integral is nonelementary; tables exist, e.g. in Ref. [A13] in App. 1.)
28. **(Integration)** Evaluate $\int J_5(x)\,dx$.
29. **(Elimination of first derivative)** Show that $y = uv$ with $v(x) = \exp(-\frac{1}{2}\int p(x)\,dx)$ gives from the ODE $y'' + p(x)y' + q(x)y = 0$ the ODE

$$u'' + \left[q(x) - \tfrac{1}{4}p(x)^2 - \tfrac{1}{2}p'(x)\right]u = 0$$

no longer containing the first derivative. Show that for the Bessel equation the substitution is $y = ux^{-1/2}$ and gives

(27) $$x^2u'' + (x^2 + \tfrac{1}{4} - \nu^2)u = 0.$$

30. **(Elementary Bessel functions)** Derive (25) in Theorem 4 from (27).

31. **CAS EXPERIMENT. Change of Coefficient.** Find and graph (on common axes) the solutions of

$$y'' + kx^{-1}y' + y = 0,\ y(0) = 1,\ y'(0) = 0,$$

for $k = 0, 1, 2, \cdots, 10$ (or as far as you get useful graphs). For what k do you get elementary functions? Why? Try for noninteger k, particularly between 0 and 2, to see the continuous change of the curve. Describe the change of the location of the zeros and of the extrema as k increases from 0. Can you interpret the ODE as a model in mechanics, thereby explaining your observations?

32. **TEAM PROJECT. Modeling a Vibrating Cable (Fig. 108).** A flexible cable, chain, or rope of length L and density (mass per unit length) ρ is fixed at the upper end ($x = 0$) and allowed to make small vibrations (small angles α in the horizontal displacement $u(x, t)$, t = time) in a vertical plane.

(a) Show the following. The weight of the cable below a point x is $W(x) = \rho g(L - x)$. The restoring force is $F(x) = W \sin \alpha \approx Wu_x$, $u_x = \partial u/\partial x$. The difference in force between x and $x + \Delta x$ is $\Delta x\ (Wu_x)_x$. Newton's second law now gives

$$\rho\ \Delta x\ u_{tt} = \Delta x\ \rho g[(L - x)u_x]_x.$$

For the expected periodic motion $u(x, t) = y(x) \cos(\omega t + \delta)$ the model of the problem is the ODE

$$(L - x)y'' - y' + \lambda^2 y = 0, \qquad \lambda^2 = \omega^2/g.$$

(b) Transform this ODE to $\ddot{y} + s^{-1}\dot{y} + y = 0$, $\dot{y} = dy/ds$, $s = 2\lambda z^{1/2}$, $z = L - x$, so that the solution is

$$y(x) = J_0(2\omega\sqrt{(L - x)/g}).$$

(c) Conclude that possible frequencies $\omega/2\pi$ are those for which $s = 2\omega\sqrt{L/g}$ is a zero of J_0. The corresponding solutions are called **normal modes.** Figure 108 shows the first of them. What does the second normal mode look like? The third? What is the frequency (cycles/min) of a cable of length 2 m? Of length 10 m?

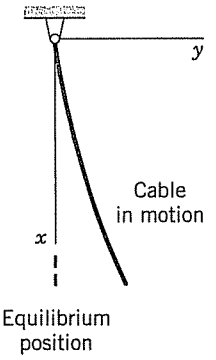


Fig. 108. Vibrating cable in Team Project 32

33. **CAS EXPERIMENT. Bessel Functions for Large x.**

(a) Graph $J_n(x)$ for $n = 0, \cdots, 5$ on common axes.

(b) Experiment with (14) for integer n. Using graphs, find out from which $x = x_n$ on the curves of (11) and (14) practically coincide. How does x_n change with n?

(c) What happens in (b) if $n = \pm\frac{1}{2}$? (Our usual notation in this case would be ν.)

(d) How does the error of (14) behave as function of x for fixed n? [Error = exact value minus approximation (14).]

(e) Show from the graphs that $J_0(x)$ has extrema where $J_1(x) = 0$. Which formula proves this? Find further relations between zeros and extrema.

(f) Raise and answer questions of your own, for instance, on the zeros of J_0 and J_1. How accurately are they obtained from (14)?

5.6 Bessel Functions of the Second Kind $Y_\nu(x)$

From the last section we know that J_ν and $J_{-\nu}$ form a basis of solutions of Bessel's equation, provided ν is not an integer. But when ν is an integer, these two solutions are linearly dependent on any interval (see Theorem 2 in Sec. 5.5). Hence to have a general solution also when $\nu = n$ is an integer, we need a second linearly independent solution besides J_n. This solution is called a **Bessel function of the second kind** and is denoted by Y_n. We shall now derive such a solution, beginning with the case $n = 0$.

$n = 0$: Bessel Function of the Second Kind $Y_0(x)$

When $n = 0$, Bessel's equation can be written

(1) $$xy'' + y' + xy = 0.$$

Then the indicial equation (4) in Sec. 5.5 has a double root $r = 0$. This is Case 2 in Sec. 5.4. In this case we first have only one solution, $J_0(x)$. From (8) in Sec. 5.4 we see that the desired second solution must be of the form

$$(2) \qquad y_2(x) = J_0(x) \ln x + \sum_{m=1}^{\infty} A_m x^m.$$

We substitute y_2 and its derivatives

$$y_2' = J_0' \ln x + \frac{J_0}{x} + \sum_{m=1}^{\infty} m A_m x^{m-1}$$

$$y_2'' = J_0'' \ln x + \frac{2J_0'}{x} - \frac{J_0}{x^2} + \sum_{m=1}^{\infty} m(m-1) A_m x^{m-2}$$

into (1). Then the sum of the three logarithmic terms $xJ_0'' \ln x$, $J_0' \ln x$, and $xJ_0 \ln x$ is zero because J_0 is a solution of (1). The terms $-J_0/x$ and J_0/x (from xy'' and y') cancel. Hence we are left with

$$2J_0' + \sum_{m=1}^{\infty} m(m-1) A_m x^{m-1} + \sum_{m=1}^{\infty} m A_m x^{m-1} + \sum_{m=1}^{\infty} A_m x^{m+1} = 0.$$

Addition of the first and second series gives $\Sigma m^2 A_m x^{m-1}$. The power series of $J_0'(x)$ is obtained from (12) in Sec. 5.5 and the use of $m!/m = (m-1)!$ in the form

$$J_0'(x) = \sum_{m=1}^{\infty} \frac{(-1)^m 2m x^{2m-1}}{2^{2m}(m!)^2} = \sum_{m=1}^{\infty} \frac{(-1)^m x^{2m-1}}{2^{2m-1} m!\,(m-1)!}.$$

Together with $\Sigma m^2 A_m x^{m-1}$ and $\Sigma A_m x^{m+1}$ this gives

$$(3^*) \qquad \sum_{m=1}^{\infty} \frac{(-1)^m x^{2m-1}}{2^{2m-2} m!\,(m-1)!} + \sum_{m=1}^{\infty} m^2 A_m x^{m-1} + \sum_{m=1}^{\infty} A_m x^{m+1} = 0.$$

First, we show that the A_m with odd subscripts are all zero. The power x^0 occurs only in the second series, with coefficient A_1. Hence $A_1 = 0$. Next, we consider the even powers x^{2s}. The first series contains none. In the second series, $m - 1 = 2s$ gives the term $(2s+1)^2 A_{2s+1} x^{2s}$. In the third series, $m + 1 = 2s$. Hence by equating the sum of the coefficients of x^{2s} to zero we have

$$(2s+1)^2 A_{2s+1} + A_{2s-1} = 0, \qquad s = 1, 2, \cdots.$$

Since $A_1 = 0$, we thus obtain $A_3 = 0$, $A_5 = 0$, $\cdots$, successively.

We now equate the sum of the coefficients of x^{2s+1} to zero. For $s = 0$ this gives

$$-1 + 4A_2 = 0, \qquad \text{thus} \qquad A_2 = \tfrac{1}{4}.$$

For the other values of s we have in the first series in (3*) $2m - 1 = 2s + 1$, hence $m = s + 1$, in the second $m - 1 = 2s + 1$, and in the third $m + 1 = 2s + 1$. We thus obtain

$$\frac{(-1)^{s+1}}{2^{2s}(s+1)!\,s!} + (2s+2)^2 A_{2s+2} + A_{2s} = 0.$$

For $s = 1$ this yields

$$\frac{1}{8} + 16A_4 + A_2 = 0, \qquad \text{thus} \qquad A_4 = -\frac{3}{128}$$

and in general

$$A_{2m} = \frac{(-1)^{m-1}}{2^{2m}(m!)^2}\left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{m}\right), \qquad m = 1, 2, \cdots. \tag{3}$$

Using the short notations

$$h_1 = 1 \qquad h_m = 1 + \frac{1}{2} + \cdots + \frac{1}{m} \qquad m = 2, 3, \cdots \tag{4}$$

and inserting (4) and $A_1 = A_3 = \cdots = 0$ into (2), we obtain the result

$$\begin{aligned} y_2(x) &= J_0(x) \ln x + \sum_{m=1}^{\infty} \frac{(-1)^{m-1} h_m}{2^{2m}(m!)^2} x^{2m} \\ &= J_0(x) \ln x + \frac{1}{4}x^2 - \frac{3}{128}x^4 + \frac{11}{13824}x^6 - + \cdots. \end{aligned} \tag{5}$$

Since J_0 and y_2 are linearly independent functions, they form a basis of (1) for $x > 0$. Of course, another basis is obtained if we replace y_2 by an independent particular solution of the form $a(y_2 + bJ_0)$, where a ($\neq 0$) and b are constants. It is customary to choose $a = 2/\pi$ and $b = \gamma - \ln 2$, where the number $\gamma = 0.577\,215\,664\,90 \cdots$ is the so-called **Euler constant,** which is defined as the limit of

$$1 + \frac{1}{2} + \cdots + \frac{1}{s} - \ln s$$

as s approaches infinity. The standard particular solution thus obtained is called the **Bessel function of the second kind** *of order zero* (Fig. 109) or **Neumann's function** *of order zero* and is denoted by $Y_0(x)$. Thus [see (4)]

$$Y_0(x) = \frac{2}{\pi}\left[J_0(x)\left(\ln\frac{x}{2} + \gamma\right) + \sum_{m=1}^{\infty} \frac{(-1)^{m-1} h_m}{2^{2m}(m!)^2} x^{2m}\right]. \tag{6}$$

For small $x > 0$ the function $Y_0(x)$ behaves about like $\ln x$ (see Fig. 109, why?), and $Y_0(x) \to -\infty$ as $x \to 0$.

Bessel Functions of the Second Kind $Y_n(x)$

For $\nu = n = 1, 2, \cdots$ a second solution can be obtained by manipulations similar to those for $n = 0$, starting from (10), Sec 5.4. It turns out that in these cases the solution also contains a logarithmic term.

The situation is not yet completely satisfactory, because the second solution is defined differently, depending on whether the order ν is an integer or not. To provide uniformity

of formalism, it is desirable to adopt a form of the second solution that is valid for all values of the order. For this reason we introduce a standard second solution $Y_\nu(x)$ defined for all ν by the formula

$$\text{(7)} \quad \text{(a)} \quad Y_\nu(x) = \frac{1}{\sin \nu\pi}\,[J_\nu(x)\cos \nu\pi - J_{-\nu}(x)]$$

$$\text{(b)} \quad Y_n(x) = \lim_{\nu \to n} Y_\nu(x).$$

This function is called the **Bessel function of the second kind** *of order* ν or **Neumann's function**[7] *of order* ν. Figure 109 shows $Y_0(x)$ and $Y_1(x)$.

Let us show that J_ν and Y_ν are indeed linearly independent for all ν (and $x > 0$).

For noninteger order ν, the function $Y_\nu(x)$ is evidently a solution of Bessel's equation because $J_\nu(x)$ and $J_{-\nu}(x)$ are solutions of that equation. Since for those ν the solutions J_ν and $J_{-\nu}$ are linearly independent and Y_ν involves $J_{-\nu}$, the functions J_ν and Y_ν are linearly independent. Furthermore, it can be shown that the limit in (7b) exists and Y_n is a solution of Bessel's equation for integer order; see Ref. [A13] in App. 1. We shall see that the series development of $Y_n(x)$ contains a logarithmic term. Hence $J_n(x)$ and $Y_n(x)$ are linearly independent solutions of Bessel's equation. The series development of $Y_n(x)$ can be obtained if we insert the series (20) and (21), Sec. 5.5, for $J_\nu(x)$ and $J_{-\nu}(x)$ into (7a) and then let ν approach n; for details see Ref. [A13]. The result is

$$\text{(8)} \quad Y_n(x) = \frac{2}{\pi} J_n(x)\left(\ln \frac{x}{2} + \gamma\right) + \frac{x^n}{\pi} \sum_{m=0}^{\infty} \frac{(-1)^{m-1}(h_m + h_{m+n})}{2^{2m+n} m!\,(m+n)!} x^{2m} - \frac{x^{-n}}{\pi} \sum_{m=0}^{n-1} \frac{(n-m-1)!}{2^{2m-n} m!} x^{2m}$$

where $x > 0$, $n = 0, 1, \cdots$, and [as in (4)] $h_0 = 0$, $h_1 = 1$,

$$h_m = 1 + \frac{1}{2} + \cdots + \frac{1}{m}, \qquad h_{m+n} = 1 + \frac{1}{2} + \cdots + \frac{1}{m+n}.$$

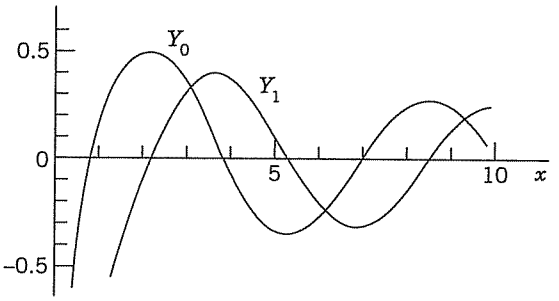


Fig. 109. Bessel functions of the second kind Y_0 and Y_1.
(For a small table, see App. 5.)

[7] CARL NEUMANN (1832–1925), German mathematician and physicist. His work on potential theory sparked the development in the field of integral equations by VITO VOLTERRA (1860–1940) of Rome, ERIC IVAR FREDHOLM (1866–1927) of Stockholm, and DAVID HILBERT (1862–1943) of Göttingen (see the footnote in Sec. 7.9).

The solutions $Y_\nu(x)$ are sometimes denoted by $N_\nu(x)$; in Ref. [A13] they are called **Weber's functions;** Euler's constant in (6) is often denoted by C or $\ln \gamma$.

For $n = 0$ the last sum in (8) is to be replaced by 0 [giving agreement with (6)].

Furthermore, it can be shown that

$$Y_{-n}(x) = (-1)^n Y_n(x).$$

Our main result may now be formulated as follows.

THEOREM 1

General Solution of Bessel's Equation

A general solution of Bessel's equation for all values of ν (and $x > 0$) is

$$\textbf{(9)} \qquad y(x) = C_1 J_\nu(x) + C_2 Y_\nu(x).$$

We finally mention that there is a practical need for solutions of Bessel's equation that are complex for real values of x. For this purpose the solutions

$$\textbf{(10)} \qquad \begin{aligned} H_\nu^{(1)}(x) &= J_\nu(x) + iY_\nu(x) \\ H_\nu^{(2)}(x) &= J_\nu(x) - iY_\nu(x) \end{aligned}$$

are frequently used. These linearly independent functions are called **Bessel functions of the third kind** *of order ν* or *first and second* **Hankel functions**[8] *of order ν.*

This finishes our discussion on Bessel functions, except for their "orthogonality," which we explain in Sec. 5.7. Applications to vibrations follow in Sec. 12.9.

PROBLEM SET 5.6

1–10 SOME FURTHER ODEs REDUCIBLE TO BESSEL'S EQUATIONS

(See also Sec. 5.5.)

Using the indicated substitutions, find a general solution in terms of J_ν and Y_ν. Indicate whether you could also use $J_{-\nu}$ instead of Y_ν. (Show the details of your work.)

1. $x^2y'' + xy' + (x^2 - 25)y = 0$
2. $x^2y'' + xy' + (9x^2 - \frac{1}{9})y = 0 \quad (3x = z)$
3. $4xy'' + 4y' + y = 0 \quad (\sqrt{x} = z)$
4. $xy'' + y' + 36y = 0 \quad (12\sqrt{x} = z)$
5. $x^2y'' + xy' + (4x^4 - 16)y = 0 \quad (x^2 = z)$
6. $x^2y'' + xy' + (x^6 - 1)y = 0 \quad (\frac{1}{3}x^3 = z)$
7. $xy'' + 11y' + xy = 0 \quad (y = x^{-5}u)$
8. $y'' + 4x^2y = 0 \quad (y = u\sqrt{x},\ x^2 = z)$
9. $x^2y'' - 5xy' + 9(x^6 - 8)y = 0 \quad (y = x^3u,\ x^3 = z)$
10. $xy'' + 7y' + 4xy = 0 \quad (y = x^{-3}u,\ 2x = z)$

11. **(Hankel functions)** Show that the Hankel functions (10) form a basis of solutions of Bessel's equation for any ν.

12. **CAS EXPERIMENT. Bessel Functions for Large x.** It can be shown that for large x,

$$\textbf{(11)} \qquad Y_n(x) \sim \sqrt{2/(\pi x)} \sin\left(x - \tfrac{1}{2}n\pi - \tfrac{1}{4}\pi\right)$$

with $\sim$ defined as in (14) of Sec. 5.5.

(a) Graph $Y_n(x)$ for $n = 0, \cdots, 5$ on common axes. Are there relations between zeros of one function and extrema of another? For what functions?

(b) Find out from graphs from which $x = x_n$ on the curves of (8) and (11) (both obtained from your CAS) practically coincide. How does x_n change with n?

(c) Calculate the first ten zeros x_m, $m = 1, \cdots, 10$, of $Y_0(x)$ from your CAS and from (11). How does the error behave as m increases?

(d) Do (c) for $Y_1(x)$ and $Y_2(x)$. How do the errors compare to those in (c)?

[8] HERMANN HANKEL (1839–1873), German mathematician.

13. Modified Bessel functions of the first kind of order ν are defined by $I_\nu(x) = i^{-\nu}J_\nu(ix)$, $i = \sqrt{-1}$. Show that I_ν satisfies the ODE

$$x^2y'' + xy' - (x^2 + \nu^2)y = 0 \tag{12}$$

and has the representation

$$I_\nu(x) = \sum_{m=0}^{\infty} \frac{x^{2m+\nu}}{2^{2m+\nu}m!\,\Gamma(m + \nu + 1)}. \tag{13}$$

14. (Modified Bessel functions I_ν) Show that $I_\nu(x)$ is real for all real x (and real ν), $I_\nu(x) \neq 0$ for all real $x \neq 0$, and $I_{-n}(x) = I_n(x)$, where n is any integer.

15. Modified Bessel functions of the third kind (sometimes called *of the second kind*) are defined by the formula (14) below. Show that they satisfy the ODE (12).

$$K_\nu(x) = \frac{\pi}{2 \sin \nu\pi}\,[I_{-\nu}(x) - I_\nu(x)] \tag{14}$$

5.7 Sturm–Liouville Problems. Orthogonal Functions

So far we have considered initial value problems. We recall from Sec. 2.1 that such a problem consists of an ODE, say, of second order, and initial conditions $y(x_0) = K_0$, $y'(x_0) = K_1$ referring to the *same point* (initial point) $x = x_0$. We now turn to boundary value problems. A **boundary value problem** consists of an ODE and given **boundary conditions** referring to the two boundary points (endpoints) $x = a$ and $x = b$ of a given interval $a \leqq x \leqq b$. To solve such a problem means to find a solution of the ODE on the interval $a \leqq x \leqq b$ satisfying the boundary conditions.

We shall see that Legendre's, Bessel's, and other ODEs of importance in engineering can be written as a **Sturm–Liouville equation**

$$[p(x)y']' + [q(x) + \lambda r(x)]y = 0 \tag{1}$$

involving a parameter λ. The boundary value problem consisting of an ODE (1) and given **Sturm–Liouville boundary conditions**

$$\begin{aligned} &\text{(a)} \quad k_1y(a) + k_2y'(a) = 0 \\ &\text{(b)} \quad l_1y(b) + l_2y'(b) = 0 \end{aligned} \tag{2}$$

is called a **Sturm–Liouville problem.**[9] We shall see further that these problems lead to useful series developments in terms of particular solutions of (1), (2). Crucial in this connection is *orthogonality* to be discussed later in this section.

In (1) we make the ***assumptions*** that p, q, r, and p' are continuous on $a \leqq x \leqq b$, and

$$r(x) > 0 \qquad (a \leqq x \leqq b).$$

In (2) we assume that k_1, k_2 are given constants, not both zero, and so are l_1, l_2, not both zero.

---

[9]JACQUES CHARLES FRANÇOIS STURM (1803–1855), was born and studied in Switzerland and then moved to Paris, where he later became the successor of Poisson in the chair of mechanics at the Sorbonne (the University of Paris).

JOSEPH LIOUVILLE (1809–1882), French mathematician and professor in Paris, contributed to various fields in mathematics and is particularly known by his important work in complex analysis (Liouville's theorem; Sec. 14.4), special functions, differential geometry, and number theory.

EXAMPLE 1 Legendre's and Bessel's Equations are Sturm–Liouville Equations

Legendre's equation $(1 - x^2)y'' - 2xy' + n(n + 1)y = 0$ may be written

$$[(1 - x^2)y']' + \lambda y = 0 \qquad \lambda = n(n + 1).$$

This is (1) with $p = 1 - x^2$, $q = 0$, and $r = 1$.

In Bessel's equation

$$\tilde{x}^2\ddot{y} + \tilde{x}\dot{y} + (\tilde{x}^2 - n^2)y = 0 \qquad \dot{y} = dy/d\tilde{x}, \quad \text{etc.}$$

as a model in physics or elsewhere, one often likes to have another parameter k in addition to n. For this reason we set $\tilde{x} = kx$. Then by the chain rule $\dot{y} = dy/d\tilde{x} = (dy/dx)\,dx/d\tilde{x} = y'/k$, $\ddot{y} = y''/k^2$. In the first two terms, k^2 and k drop out and we get

$$x^2y'' + xy' + (k^2x^2 - n^2)y = 0.$$

Division by x gives the Sturm–Liouville equation

$$[xy']' + \left(-\frac{n^2}{x} + \lambda x\right)y = 0 \qquad \lambda = k^2.$$

This is (1) with $p = x$, $q = -n^2/x$, and $r = x$. ▩

Eigenfunctions, Eigenvalues

Clearly, $y \equiv 0$ is a solution—the **"trivial solution"**—for any λ because (1) is homogeneous and (2) has zeros on the right. This is of no interest. We want to find **eigenfunctions** $y(x)$, that is, solutions of (1) satisfying (2) without being identically zero. We call a number λ for which an eigenfunction exists an **eigenvalue** of the Sturm–Liouville problem (1), (2).

EXAMPLE 2 Trigonometric Functions as Eigenfunctions. Vibrating String

Find the eigenvalues and eigenfunctions of the Sturm–Liouville problem

$$y'' + \lambda y = 0, \qquad y(0) = 0, \quad y(\pi) = 0. \tag{3}$$

This problem arises, for instance, if an elastic string (a violin string, for example) is stretched a little and then fixed at its ends $x = 0$ and $x = \pi$ and allowed to vibrate. Then $y(x)$ is the "space function" of the deflection $u(x, t)$ of the string, assumed in the form $u(x, t) = y(x)w(t)$, where t is time. (This model will be discussed in great detail in Secs. 12.2–12.4.)

Solution. From (1) and (2) we see that $p = 1$, $q = 0$, $r = 1$ in (1), and $a = 0$, $b = \pi$, $k_1 = l_1 = 1$, $k_2 = l_2 = 0$ in (2). For negative $\lambda = -\nu^2$ a general solution of the ODE in (3) is $y(x) = c_1e^{\nu x} + c_2e^{-\nu x}$. From the boundary conditions we obtain $c_1 = c_2 = 0$, so that $y \equiv 0$, which is not an eigenfunction. For $\lambda = 0$ the situation is similar. For positive $\lambda = \nu^2$ a general solution is

$$y(x) = A \cos \nu x + B \sin \nu x.$$

From the first boundary condition we obtain $y(0) = A = 0$. The second boundary condition then yields

$$y(\pi) = B \sin \nu\pi = 0, \qquad \text{thus} \qquad \nu = 0, \pm 1, \pm 2, \cdots.$$

For $\nu = 0$ we have $y \equiv 0$. For $\lambda = \nu^2 = 1, 4, 9, 16, \cdots$, taking $B = 1$, we obtain

$$y(x) = \sin \nu x \qquad (\nu = 1, 2, \cdots).$$

Hence the eigenvalues of the problem are $\lambda = \nu^2$, where $\nu = 1, 2, \cdots$, and corresponding eigenfunctions are $y(x) = \sin \nu x$, where $\nu = 1, 2, \cdots$. ▩

Existence of Eigenvalues

Eigenvalues of a Sturm–Liouville problem (1), (2), even infinitely many, exist under rather general conditions on p, q, r in (1). (Sufficient are the conditions in Theorem 1, below, together with $p(x) > 0$ and $r(x) > 0$ on $a < x < b$. Proofs are complicated; see Ref. [A3] or [A11] listed in App. 1.)

Reality of Eigenvalues

Furthermore, if p, q, r, and p' in (1) are real-valued and continuous on the interval $a \leqq x \leqq b$ and r is positive throughout that interval (or negative throughout that interval), then all the eigenvalues of the Sturm–Liouville problem (1), (2) are real. (Proof in App. 4.) This is what the engineer would expect since eigenvalues are often related to frequencies, energies, or other physical quantities that must be real.

Orthogonality

The most remarkable and important property of eigenfunctions of Sturm–Liouville problems is their orthogonality, which will be crucial in series developments in terms of eigenfunctions.

DEFINITION

Orthogonality

Functions $y_1(x), y_2(x), \cdots$ defined on some interval $a \leqq x \leqq b$ are called **orthogonal** on this interval with respect to the **weight function** $r(x) > 0$ if for all m and all n different from m,

$$\int_a^b r(x)y_m(x)y_n(x)\,dx = 0 \qquad (m \neq n). \tag{4}$$

The **norm** $\|y_m\|$ of y_m is defined by

$$\|y_m\| = \sqrt{\int_a^b r(x)y_m^2(x)\,dx}\,. \tag{5}$$

Note that this is the square root of the integral in (4) with $n = m$.

The functions $y_1, y_2, \cdots$ are called **orthonormal** on $a \leqq x \leqq b$ if they are orthogonal on this interval and all have norm 1.

If $r(x) = 1$, we more briefly call the functions *orthogonal* instead of orthogonal with respect to $r(x) = 1$; similarly for orthonormality. Then

$$\int_a^b y_m(x)\,y_n(x)\,dx = 0 \quad (m \neq n), \qquad \|y_m\| = \sqrt{\int_a^b y_m^2(x)\,dx}\,.$$

EXAMPLE 3 **Orthogonal Functions. Orthonormal Functions**

The functions $y_m(x) = \sin mx$, $m = 1, 2, \cdots$ form an orthogonal set on the interval $-\pi \leqq x \leqq \pi$, because for $m \neq n$ we obtain by integration [see (11) in App. A3.1]

$$\int_{-\pi}^{\pi} y_m(x)y_n(x)\,dx = \int_{-\pi}^{\pi} \sin mx \sin nx\,dx = \frac{1}{2}\int_{-\pi}^{\pi} \cos(m-n)x\,dx - \frac{1}{2}\int_{-\pi}^{\pi} \cos(m+n)x\,dx = 0.$$

The norm $\|y_m\|$ equals $\sqrt{\pi}$, because

$$\|y_m\|^2 = \int_{-\pi}^{\pi} \sin^2 mx\,dx = \pi \qquad (m = 1, 2, \cdots).$$

Hence the corresponding orthonormal set, obtained by division by the norm, is

$$\frac{\sin x}{\sqrt{\pi}}, \qquad \frac{\sin 2x}{\sqrt{\pi}}, \qquad \frac{\sin 3x}{\sqrt{\pi}}, \qquad \cdots.$$

Orthogonality of Eigenfunctions

THEOREM 1

Orthogonality of Eigenfunctions

Suppose that the functions p, q, r, and p' in the Sturm–Liouville equation (1) *are real-valued and continuous and $r(x) > 0$ on the interval $a \leqq x \leqq b$. Let $y_m(x)$ and $y_n(x)$ be eigenfunctions of the Sturm–Liouville problem* (1), (2) *that correspond to different eigenvalues λ_m and λ_n, respectively. Then y_m, y_n are orthogonal on that interval with respect to the weight function r, that is,*

$$\int_a^b r(x)y_m(x)y_n(x)\,dx = 0 \qquad (m \neq n). \tag{6}$$

If $p(a) = 0$, then (2a) *can be dropped from the problem. If $p(b) = 0$, then* (2b) *can be dropped.* [It is then required that y and y' remain bounded at such a point, and the problem is called **singular,** as opposed to a **regular problem** in which (2) is used.]

If $p(a) = p(b)$, then (2) *can be replaced by the* **"periodic boundary conditions"**

$$y(a) = y(b), \qquad y'(a) = y'(b). \tag{7}$$

The boundary value problem consisting of the Sturm–Liouville equation (1) and the periodic boundary conditions (7) is called a **periodic Sturm–Liouville problem.**

PROOF By assumption, y_m and y_n satisfy the Sturm–Liouville equations

$$(py_m')' + (q + \lambda_m r)y_m = 0$$
$$(py_n')' + (q + \lambda_n r)y_n = 0$$

respectively. We multiply the first equation by y_n, the second by $-y_m$, and add,

$$(\lambda_m - \lambda_n)ry_my_n = y_m(py_n')' - y_n(py_m')' = [(py_n')y_m - (py_m')y_n]'$$

where the last equality can be readily verified by performing the indicated differentiation of the last expression in brackets. This expression is continuous on $a \leqq x \leqq b$ since p and p' are continuous by assumption and y_m, y_n are solutions of (1). Integrating over x from a to b, we thus obtain

$$(\lambda_m - \lambda_n)\int_a^b ry_my_n\,dx = \Big[p(y_n'y_m - y_m'y_n)\Big]_a^b \qquad (a < b). \tag{8}$$

The expression on the right equals the sum of the subsequent Lines 1 and 2,

$$\begin{aligned} &p(b)[y_n'(b)y_m(b) - y_m'(b)y_n(b)] && \text{(Line 1)} \\ &-p(a)[y_n'(a)y_m(a) - y_m'(a)y_n(a)] && \text{(Line 2).} \end{aligned} \tag{9}$$

Hence if (9) is zero, (8) with $\lambda_m - \lambda_n \neq 0$ implies the orthogonality (6). Accordingly, we have to show that (9) is zero, using the boundary conditions (2) as needed.

Case 1. $p(a) = p(b) = 0$. Clearly, (9) is zero, and (2) is not needed.

Case 2. $p(a) \neq 0, p(b) = 0$. Line 1 of (9) is zero. Consider Line 2. From (2a) we have

$$k_1 y_n(a) + k_2 y_n'(a) = 0,$$
$$k_1 y_m(a) + k_2 y_m'(a) = 0.$$

Let $k_2 \neq 0$. We multiply the first equation by $y_m(a)$, the last by $-y_n(a)$ and add,

$$k_2[y_n'(a)y_m(a) - y_m'(a)y_n(a)] = 0.$$

This is k_2 times Line 2 of (9), which thus is zero since $k_2 \neq 0$. If $k_2 = 0$, then $k_1 \neq 0$ by assumption, and the argument of proof is similar.

Case 3. $p(a) = 0, p(b) \neq 0$. Line 2 of (9) is zero. From (2b) it follows that Line 1 of (9) is zero; this is similar to Case 2.

Case 4. $p(a) \neq 0, p(b) \neq 0$. We use both (2a) and (2b) and proceed as in Cases 2 and 3.

Case 5. $p(a) = p(b)$. Then (9) becomes

$$p(b)\big[y_n'(b)y_m(b) - y_m'(b)y_n(b) - y_n'(a)y_m(a) + y_m'(a)y_n(a)\big].$$

The expression in brackets $[\cdots]$ is zero, either by (2) used as before, or more directly by (7). Hence in this case, (7) can be used instead of (2), as claimed. This completes the proof of Theorem 1. ▩

EXAMPLE 4 **Application of Theorem 1. Vibrating Elastic String**

The ODE in Example 2 is a Sturm–Liouville equation with $p = 1$, $q = 0$, and $r = 1$. From Theorem 1 it follows that the eigenfunctions $y_m = \sin mx$ $(m = 1, 2, \cdots)$ are orthogonal on the interval $0 \leqq x \leqq \pi$. ▩

EXAMPLE 5 **Application of Theorem 1. Orthogonality of the Legendre Polynomials**

Legendre's equation is a Sturm–Liouville equation (see Example 1)

$$[(1 - x^2)y']' + \lambda y = 0, \qquad \lambda = n(n + 1)$$

with $p = 1 - x^2$, $q = 0$, and $r = 1$. Since $p(-1) = p(1) = 0$, we need no boundary conditions, but have a ***singular Sturm—Liouville problem*** on the interval $-1 \leqq x \leqq 1$. We know that for $n = 0, 1, \cdots$, hence $\lambda = 0, 1 \cdot 2, 2 \cdot 3, \cdots$, the Legendre polynomials $P_n(x)$ are solutions of the problem. Hence these are the eigenfunctions. From Theorem 1 it follows that they are orthogonal on that interval, that is,

$$\int_{-1}^{1} P_m(x)\,P_n(x)\,dx = 0 \qquad (m \neq n). \tag{10}$$

▩

EXAMPLE 6 **Application of Theorem 1. Orthogonality of the Bessel Functions $J_n(x)$**

The Bessel function $J_n(\tilde{x})$ with fixed integer $n \geqq 0$ satisfies Bessel's equation (Sec. 5.5)

$$\tilde{x}^2 \ddot{J}_n(\tilde{x}) + \tilde{x}\dot{J}_n(\tilde{x}) + (\tilde{x}^2 - n^2)J_n(\tilde{x}) = 0,$$

where $\dot{J}_n = dJ_n/d\tilde{x}$, $\ddot{J}_n = d^2J_n/d\tilde{x}^2$. In Example 1 we transformed this equation, by setting $\tilde{x} = kx$, into a Sturm–Liouville equation

$$[xJ_n'(kx)]' + \left(-\frac{n^2}{x} + k^2 x\right) J_n(kx) = 0$$

with $p(x) = x$, $q(x) = -n^2/x$, $r(x) = x$, and parameter $\lambda = k^2$. Since $p(0) = 0$, Theorem 1 implies orthogonality on an interval $0 \leqq x \leqq R$ (R given, fixed) of those solutions $J_n(kx)$ that are zero at $x = R$, that is,

$$J_n(kR) = 0 \qquad (n \text{ fixed}). \tag{11}$$

[Note that $q(x) = -n^2/x$ is discontinuous at 0, but this does not affect the proof of Theorem 1.] It can be shown (see Ref. [A13]) that $J_n(\tilde{x})$ has infinitely many zeros, say, $\tilde{x} = \alpha_{n,1} < \alpha_{n,2} < \cdots$ (see Fig. 107 in Sec. 5.5 for $n = 0$ and 1). Hence we must have

$$(12) \qquad kR = \alpha_{n,m} \qquad \text{thus} \qquad k_{n,m} = \alpha_{n,m}/R \qquad (m = 1, 2, \cdots).$$

This proves the following orthogonality property.

THEOREM 2

Orthogonality of Bessel Functions

For each fixed nonnegative integer n the sequence of Bessel functions of the first kind $J_n(k_{n,1}x)$, $J_n(k_{n,2}x)$, $\cdots$ with $k_{n,m}$ as in (12) forms an orthogonal set on the interval $0 \leqq x \leqq R$ with respect to the weight function $r(x) = x$, that is,

$$(13) \qquad \int_0^R xJ_n(k_{n,m}x)J_n(k_{n,j}x)\,dx = 0 \qquad (j \neq m, \ n \text{ fixed}).$$

Hence we have obtained ***infinitely many orthogonal sets***, each corresponding to one of the ***fixed*** values n. This also illustrates the importance of the zeros of the Bessel functions. ■

EXAMPLE 7 **Eigenvalues from Graphs**

Solve the Sturm–Liouville problem $y'' + \lambda y = 0$, $y(0) + y'(0) = 0$, $y(\pi) - y'(\pi) = 0$.

Solution. A general solution and its derivative are

$$y = A\cos kx + B\sin kx \qquad \text{and} \qquad y' = -Ak\sin kx + Bk\cos kx, \qquad k = \sqrt{\lambda}.$$

The first boundary condition gives $y(0) + y'(0) = A + Bk = 0$, hence $A = -Bk$. The second boundary condition and substitution of $A = -Bk$ give

$$\begin{aligned} y(\pi) - y'(\pi) &= A\cos \pi k + B\sin \pi k + Ak\sin \pi k - Bk\cos \pi k \\ &= -Bk\cos \pi k + B\sin \pi k - Bk^2\sin \pi k - Bk\cos \pi k = 0. \end{aligned}$$

We must have $B \neq 0$ since otherwise $B = A = 0$, hence $y = 0$, which is not an eigenfunction. Division by $B\cos \pi k$ gives

$$-k + \tan \pi k - k^2 \tan \pi k - k = 0, \qquad \text{thus} \qquad \tan \pi k = \frac{-2k}{k^2 - 1}.$$

The graph in Fig. 110 now shows us where to look for eigenvalues. These correspond to the k-values of the points of intersection of $\tan \pi k$ and the right side $-2k/(k^2 - 1)$ of the last equation. The eigenvalues are $\lambda_m = k_m^2$, where $\lambda_0 = 0$ with eigenfunction $y_0 = 1$ and the other λ_m are located near $2^2, 3^2, 4^2, \cdots$, with eigenfunctions $\cos k_m x$ and $\sin k_m x$, $m = 1, 2, \cdots$. The precise numeric determination of the eigenvalues would require a root-finding method (such as those given in Sec. 19.2). ■

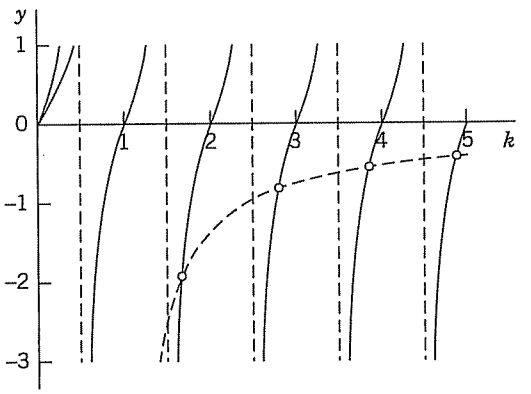


Fig. 110. Example 7. Circles mark the intersections of $\tan \pi k$ and $-2k/(k^2 - 1)$

PROBLEM SET 5.7

1. **(Proof of Theorem 1)** Carry out the details in Cases 3 and 4.
2. **Normalization of eigenfunctions** y_m of (P), (2) means that we multiply y_m by a nonzero constant c_m such that $c_m y_m$ has norm 1. Show that $z_m = cy_m$ with *any* $c \neq 0$ is an eigenfunction for the eigenvalue corresponding to y_m.
3. **(Change of x)** Show that if the functions $y_0(x)$, $y_1(x)$, $\cdots$ form an orthogonal set on an interval $a \leqq x \leqq b$ (with $r(x) = 1$), then the functions $y_0(ct + k)$, $y_1(ct + k)$, $\cdots$, $c > 0$, form an orthogonal set on the interval $(a - k)/c \leqq t \leqq (b - k)/c$.
4. **(Change of x)** Using Prob. 3, derive the orthogonality of $1, \cos \pi x, \sin \pi x, \cos 2\pi x, \sin 2\pi x, \cdots$ on $-1 \leqq x \leqq 1$ ($r(x) = 1$) from that of $1, \cos x, \sin x, \cos 2x, \sin 2x, \cdots$ on $-\pi \leqq x \leqq \pi$.
5. **(Legendre polynomials)** Show that the functions $P_n(\cos \theta)$, $n = 0, 1, \cdots$, form an orthogonal set on the interval $0 \leqq \theta \leqq \pi$ with respect to the weight function $\sin \theta$.
6. **(Tranformation to Sturm–Liouville form)** Show that $y'' + fy' + (g + \lambda h)y = 0$ takes the form (1) if you set $p = \exp(\int f\,dx)$, $q = pg$, $r = hp$. Why would you do such a transformation?

7–19 STURM–LIOUVILLE PROBLEMS

Write the given ODE in the form (1) if it is in a different form. (Use Prob. 6.) Find the eigenvalues and eigenfunctions. Verify orthogonality. (Show the details of your work.)

7. $y'' + \lambda y = 0, \quad y(0) = 0, \quad y(5) = 0$
8. $y'' + \lambda y = 0, \quad y'(0) = 0, \quad y'(\pi) = 0$
9. $y'' + \lambda y = 0, \quad y(0) = 0, \quad y'(L) = 0$
10. $y'' + \lambda y = 0, \quad y(0) = y(1), \quad y'(0) = y'(1)$
11. $y'' + \lambda y = 0, \quad y(0) = y(2\pi), \quad y'(0) = y'(2\pi)$
12. $y'' + \lambda y = 0, \quad y(0) + y'(0) = 0, \quad y(1) + y'(1) = 0$
13. $y'' + \lambda y = 0, \quad y(0) = 0, \quad y(1) + y'(1) = 0$
14. $(xy')' + \lambda x^{-1}y = 0, \quad y(1) = 0, \quad y'(e) = 0.$ (Set $x = e^t$.)
15. $(x^{-1}y')' + (\lambda + 1)x^{-3}y = 0, \quad y(1) = 0, \quad y(e^\pi) = 0.$ (Set $x = e^t$.)
16. $y'' - 2y' + (\lambda + 1)y = 0, \quad y(0) = 0, \quad y(1) = 0$
17. $y'' + 8y' + (\lambda + 16)y = 0, \quad y(0) = 0, \quad y(\pi) = 0$
18. $xy'' + 2y' + \lambda xy = 0, \quad y(\pi) = 0, \quad y(2\pi) = 0.$ (Use a CAS or set $y = x^{-1}u$.)
19. $y'' - 2x^{-1}y' + (k^2 + 2x^{-2})y = 0, \; y(1) = 0, \; y(2) = 0.$ (Use a CAS or set $y = xu$.)
20. **TEAM PROJECT. Special Functions. Orthogonal polynomials** play a great role in applications. For this reason, Legendre polynomials and various other orthogonal polynomials have been studied extensively; see Refs. [GR1], [GR10] in App. 1. Consider some of the most important ones as follows.

(a) Chebyshev polynomials[10] of the first and second kind are defined by

$$T_n(x) = \cos(n \arccos x)$$

$$U_n(x) = \frac{\sin[(n+1)\arccos x]}{\sqrt{1 - x^2}}$$

respectively, where $n = 0, 1, \cdots$. Show that

$$T_0 = 1, \quad T_1(x) = x, \quad T_2(x) = 2x^2 - 1,$$

$$T_3(x) = 4x^3 - 3x,$$

$$U_0 = 1, \quad U_1(x) = 2x, \quad U_2(x) = 4x^2 - 1,$$

$$U_3(x) = 8x^3 - 4x.$$

Show that the Chebyshev polynomials $T_n(x)$ are orthogonal on the interval $-1 \leqq x \leqq 1$ with respect to the weight function $r(x) = 1/\sqrt{1 - x^2}$. (*Hint.* To evaluate the integral, set $\arccos x = \theta$.) Verify that $T_n(x)$, $n = 0, 1, 2, 3$, satisfy the **Chebyshev equation**

$$(1 - x^2)y'' - xy' + n^2 y = 0.$$

(b) Orthogonality on an infinite interval: Laguerre polynomials[11] are defined by $L_0 = 1$, and

$$L_n(x) = \frac{e^x}{n!}\frac{d^n(x^n e^{-x})}{dx^n}, \qquad n = 1, 2, \cdots.$$

Show that

$$L_1(x) = 1 - x, \quad L_2(x) = 1 - 2x + x^2/2,$$

$$L_3(x) = 1 - 3x + 3x^2/2 - x^3/6.$$

Prove that the Laguerre polynomials are orthogonal on the positive axis $0 \leqq x < \infty$ with respect to the weight function $r(x) = e^{-x}$. *Hint.* Since the highest power in L_m is x^m, it suffices to show that $\int e^{-x}x^k L_n\,dx = 0$ for $k < n$. Do this by k integrations by parts.

---

[10] PAFNUTI CHEBYSHEV (1821–1894), Russian mathematician, is known for his work in approximation theory and the theory of numbers. Another transliteration of the name is TCHEBICHEF.

[11] EDMOND LAGUERRE (1834–1886), French mathematician, who did research work in geometry and in the theory of infinite series.

5.8 Orthogonal Eigenfunction Expansions

Orthogonal functions (obtained from Sturm–Liouville problems or otherwise) yield important series developments of given functions, as we shall see. This includes the famous *Fourier series* (to which we devote Chaps. 11 and 12), the daily bread of the physicist and engineer for solving problems in heat conduction, mechanical and electrical vibrations, etc. Indeed, orthogonality is one of the most useful ideas ever introduced in applied mathematics.

Standard Notation for Orthogonality and Orthonormality

The integral (4) in Sec. 5.7 defining orthogonality is denoted by (y_m, y_n). This is standard. Also, **Kronecker's delta**[12] δ_{mn} is defined by $\delta_{mn} = 0$ if $m \neq n$ and $\delta_{mn} = 1$ if $m = n$ (thus $\delta_{nn} = 1$). Hence for orthonormal functions $y_0, y_1, y_2, \cdots$ with respect to weight $r(x)$ (> 0) on $a \leqq x \leqq b$ we can now simply write $(y_m, y_n) = \delta_{mn}$, written out

$$(1) \qquad (y_m, y_n) = \int_a^b r(x) y_m(x) y_n(x)\, dx = \delta_{mn} = \begin{cases} 0 & \text{if } m \neq n \\ 1 & \text{if } m = n. \end{cases}$$

Also, for the norm we can now write

$$(2) \qquad \|y\| = \sqrt{(y_m, y_m)} = \sqrt{\int_a^b r(x) y_m^2(x)\, dx}\,.$$

Write down a few examples of your own, to get used to this practical short notation.

Orthogonal Series

Now comes the instant that shows why orthogonality is a fundamental concept. Let $y_0, y_1, y_2, \cdots$ be an orthogonal set with respect to weight $r(x)$ on an interval $a \leqq x \leqq b$. Let $f(x)$ be a function that can be represented by a convergent series

$$(3) \qquad f(x) = \sum_{m=0}^{\infty} a_m y_m(x) = a_0 y_0(x) + a_1 y_1(x) + \cdots.$$

This is called an **orthogonal expansion** or **generalized Fourier series.** If the y_m are eigenfunctions of a Sturm–Liouville problem, we call (3) an **eigenfunction expansion.** In (3) we use again m for summation since n will be used as a fixed order of Bessel functions.

Given $f(x)$, we have to determine the coefficients in (3), called the **Fourier constants** *of* $f(x)$ *with respect to* $y_0, y_1, \cdots$. Because of the orthogonality this is simple. All we have to do is to multiply both sides of (3) by $r(x)y_n(x)$ (***n fixed***) and then integrate on both sides from a to b. We assume that term-by-term integration is permissible. (This is justified, for instance, in the case of "uniform convergence," as is shown in Sec. 15.5.) Then we obtain

$$(f, y_n) = \int_a^b r f y_n\, dx = \int_a^b r \left(\sum_{m=0}^{\infty} a_m y_m \right) y_n\, dx = \sum_{m=0}^{\infty} a_m (y_m, y_n).$$

[12] LEOPOLD KRONECKER (1823–1891). German mathematician at Berlin University, who made important contributions to algebra, group theory, and number theory.

Because of the orthogonality all the integrals on the right are zero, except when $m = n$. Hence the whole infinite series reduces to the single term

$$a_n(y_n, y_n) = a_n \|y_n\|^2.$$

Assuming that all the functions y_n have nonzero norm, we can divide by $\|y_n\|^2$; writing again m for n, to be in agreement with (3), we get the desired formula for the Fourier constants

$$(4) \qquad a_m = \frac{(f, y_m)}{\|y_m\|^2} = \frac{1}{\|y_m\|^2} \int_a^b r(x) f(x) y_m(x)\, dx \qquad (m = 0, 1, \cdots).$$

EXAMPLE 1 Fourier Series

A most important class of eigenfunction expansions is obtained from the periodic Sturm–Liouville problem

$$y'' + \lambda y = 0, \qquad y(\pi) = y(-\pi), \qquad y'(\pi) = y'(-\pi).$$

A general solution of the ODE is $y = A \cos kx + B \sin kx$, where $k = \sqrt{\lambda}$. Substituting y and its derivative into the boundary conditions, we obtain

$$A \cos k\pi + B \sin k\pi = A \cos(-k\pi) + B \sin(-k\pi)$$

$$-kA \sin k\pi + kB \cos k\pi = -kA \sin(-k\pi) + kB \cos(-k\pi).$$

Since $\cos(-\alpha) = \cos \alpha$, the cosine terms cancel, so that these equations give no condition for these terms. Since $\sin(-\alpha) = -\sin \alpha$, the equations gives the condition $\sin k\pi = 0$, hence $k\pi = m\pi$, $k = m = 0, 1, 2, \cdots$, so that the eigenfunctions are

$$\cos 0 = 1, \qquad \cos x, \qquad \sin x, \qquad \cos 2x, \qquad \sin 2x, \cdots, \qquad \cos mx, \qquad \sin mx, \cdots$$

corresponding pairwise to the eigenvalues $\lambda = k^2 = 0, 1, 4, \cdots, m^2, \cdots$. ($\sin 0 = 0$ is not an eigenfunction.)

By Theorem 1 in Sec. 5.7, any two of these belonging to different eigenvalues are orthogonal on the interval $-\pi \leqq x \leqq \pi$ (note that $r(x) = 1$ for the present ODE). The orthogonality of $\cos mx$ and $\sin mx$ for the same m follows by integration,

$$\int_{-\pi}^{\pi} \cos mx \sin mx\, dx = \frac{1}{2} \int_{-\pi}^{\pi} \sin 2mx\, dx = 0.$$

For the norms we get $\|1\| = \sqrt{2\pi}$, and $\sqrt{\pi}$ for all the others, as you may verify by integrating 1, $\cos^2 x$, $\sin^2 x$, etc., from $-\pi$ to π. This gives the series (with a slight extension of notation since we have two functions for each eigenvalue 1, 4, 9, $\cdots$)

$$(5) \qquad f(x) = a_0 + \sum_{m=1}^{\infty} (a_m \cos mx + b_m \sin mx).$$

According to (4) the coefficients (with $m = 1, 2, \cdots$) are

$$(6) \qquad a_0 = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x)\, dx, \quad a_m = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos mx\, dx, \quad b_m = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin mx\, dx.$$

The series (5) is called the **Fourier series** of $f(x)$. Its coefficients are called the **Fourier coefficients** of $f(x)$, as given by the so-called **Euler formulas** (6) (not to be confused with the Euler formula (11) in Sec. 2.2).

For instance, for the **"periodic rectangular wave"** in Fig. 111, given by

$$f(x) = \begin{cases} -1 & \text{if } -\pi < x < 0 \\ 1 & \text{if } 0 < x < \pi \end{cases} \qquad \text{and} \qquad f(x + 2\pi) = f(x),$$

we get from (6) the values $a_0 = 0$ and

$$a_m = \frac{1}{\pi}\left[\int_{-\pi}^{0}(-1)\cos mx\,dx + \int_{0}^{\pi}1\cdot\cos mx\,dx\right] = 0,$$

$$b_m = \frac{1}{\pi}\left[\int_{-\pi}^{0}(-1)\sin mx\,dx + \int_{0}^{\pi}1\cdot\sin mx\,dx\right]$$

$$= \frac{1}{\pi}\left[\frac{\cos mx}{m}\Big|_{-\pi}^{0} - \frac{\cos mx}{m}\Big|_{0}^{\pi}\right]$$

$$= \frac{1}{\pi m}[1 - 2\cos m\pi + 1] = \begin{cases} 4/(\pi m) & \text{if } m = 1, 3, \cdots, \\ 0 & \text{if } m = 2, 4\cdots. \end{cases}$$

Hence the Fourier series of the periodic rectangular wave is

$$f(x) = \frac{4}{\pi}\left(\sin x + \frac{1}{3}\sin 3x + \frac{1}{5}\sin 5x + \cdots\right).$$

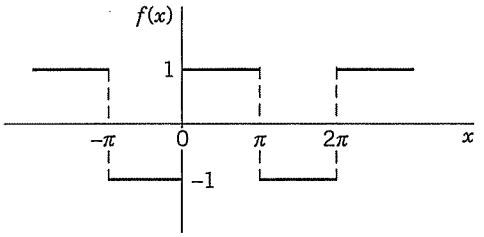


Fig. 111. Periodic rectangular wave in Example 1

Fourier series are by far the most important eigenfunction expansions, so important to the engineer that we shall devote two chapters (11 and 12) to them and their applications, and discuss numerous examples.

Did it surprise you that a series of continuous functions (sine functions) can represent a discontinuous function? More on this in Chap. 11.

EXAMPLE 2 Fourier–Legendre Series

A **Fourier–Legendre series** is an eigenfunction expansion

$$f(x) = \sum_{m=0}^{\infty} a_m P_m(x) = a_0 P_0 + a_1 P_1(x) + a_2 P_2(x) + \cdots = a_0 + a_1 x + a_2\left(\tfrac{3}{2}x^2 - \tfrac{1}{2}\right) + \cdots$$

in terms of Legendre polynomials (Sec. 5.3). The latter are the eigenfunctions of the Sturm–Liouville problem in Example 5 of Sec. 5.7 on the interval $-1 \leqq x \leqq 1$. We have $r(x) = 1$ for Legendre's equation, and (4) gives

$$(7) \qquad a_m = \frac{2m+1}{2}\int_{-1}^{1} f(x)P_m(x)\,dx, \qquad m = 0, 1, \cdots$$

because the norm is

$$(8) \qquad \|P_m\| = \sqrt{\int_{-1}^{1} P_m(x)^2\,dx} = \sqrt{\frac{2}{2m+1}} \qquad (m = 0, 1, \cdots)$$

as we state without proof. (The proof is tricky; it uses Rodrigues's formula in Problem Set 5.3 and a reduction of the resulting integral to a quotient of gamma functions.)

For instance, let $f(x) = \sin \pi x$. Then we obtain the coefficients

$$a_m = \frac{2m+1}{2} \int_{-1}^{1} (\sin \pi x) P_m(x)\, dx, \qquad \text{thus} \qquad a_1 = \frac{3}{2} \int_{-1}^{1} x \sin \pi x\, dx = \frac{3}{\pi} = 0.95493, \qquad \text{etc.}$$

Hence the Fourier–Legendre series of $\sin \pi x$ is

$$\sin \pi x = 0.95493P_1(x) - 1.15824P_3(x) + 0.21429P_5(x) - 0.01664P_7(x) + 0.00068P_9(x) - 0.00002P_{11}(x) + \cdots.$$

The coefficient of P_{13} is about $3 \cdot 10^{-7}$. The sum of the first three nonzero terms gives a curve that practically coincides with the sine curve. Can you see why the even-numbered coefficients are zero? Why a_3 is the absolutely biggest coefficient? ■

EXAMPLE 3 Fourier–Bessel Series

In Example 6 of Sec. 5.7 we obtained infinitely many orthogonal sets of Bessel functions, one for each of J_0, J_1, J_2, $\cdots$. Each set is orthogonal on an interval $0 \leqq x \leqq R$ with a fixed positive R of our choice and with respect to the weight x. The orthogonal set for J_n is $J_n(k_{n,1}x)$, $J_n(k_{n,2}x)$, $J_n(k_{n,3}x)$, $\cdots$, where n is ***fixed*** and $k_{n,m}$ is given in (12), Sec. 5.7. The corresponding Fourier–Bessel series is

$$(9) \qquad f(x) = \sum_{m=1}^{\infty} a_m J_n(k_{n,m}x) = a_1 J_n(k_{n,1}x) + a_2 J_n(k_{n,2}x) + a_3 J_n(k_{n,3}x) + \cdots \qquad (n \text{ fixed}).$$

The coefficients are (with $\alpha_{n,m} = k_{n,m}R$)

$$(10) \qquad a_m = \frac{2}{R^2 J_{n+1}^2(\alpha_{n,m})} \int_0^R x f(x) J_n(k_{n,m}x)\, dx, \qquad m = 1, 2, \cdots$$

because the square of the norm is

$$(11) \qquad \|J_n(k_{n,m}x)\|^2 = \int_0^R x J_n^2(k_{n,m}x)\, dx = \frac{R^2}{2} J_{n+1}^2(k_{n,m}R)$$

as we state without proof (which is tricky; see the discussion beginning on p. 576 of [A13]).

For instance, let us consider $f(x) = 1 - x^2$ and take $R = 1$ and $n = 0$ in the series (9), simply writing λ for $\alpha_{0,m}$. Then $k_{n,m} = \alpha_{0,m} = \lambda = 2.405, 5.520, 8.654, 11.792$, etc. (use a CAS or Table A1 in App. 5). Next we calculate the coefficients a_m by (10),

$$a_m = \frac{2}{J_1^2(\lambda)} \int_0^1 x(1 - x^2) J_0(\lambda x)\, dx.$$

This can be integrated by a CAS or by formulas as follows. First use $[xJ_1(\lambda x)]' = \lambda x J_0(\lambda x)$ from Theorem 3 in Sec. 5.5 and then integration by parts,

$$a_m = \frac{2}{J_1^2(\lambda)} \int_0^1 x(1 - x^2) J_0(\lambda x)\, dx = \frac{2}{J_1^2(\lambda)} \left[\frac{1}{\lambda} (1 - x^2) x J_1(\lambda x) \Big|_0^1 - \frac{1}{\lambda} \int_0^1 x J_1(\lambda x)(-2x)\, dx \right].$$

The integral-free part is zero. The remaining integral can be evaluated by $\left[x^2 J_2(\lambda x)\right]' = \lambda x^2 J_1(\lambda x)$ from Theorem 3 in Sec. 5.5. This gives

$$a_m = \frac{4J_2(\lambda)}{\lambda^2 J_1^2(\lambda)} \qquad (\lambda = \alpha_{0,m}).$$

Numeric values can be obtained from a CAS (or from the table on p. 409 of Ref. [GR1] in App. 1, together with the formula $J_2 = 2x^{-1}J_1 - J_0$ in Theorem 3 of Sec. 5.5). This gives the eigenfunction expansion of $1 - x^2$ in terms of Bessel functions J_0, that is,

$$1 - x^2 = 1.1081J_0(2.405x) - 0.1398J_0(5.520x) + 0.0455J_0(8.654x) - 0.0210J_0(11.792x) + \cdots.$$

A graph would show that the curve of $1 - x^2$ and that of the sum of the first three terms practically coincide. ■

Mean Square Convergence. Completeness of Orthonormal Sets

The remaining part of this section will give an introduction to a convergence suitable in connection with orthogonal series and quite different from the convergence used in calculus for Taylor series.

In practice, one uses only orthonormal sets that consist of "sufficiently many" functions, so that one can represent large classes of functions by a generalized Fourier series (3)—certainly all continuous functions on an interval $a \leqq x \leqq b$, but also functions that do "not have too many" discontinuities (see Example 1). Such orthonormal sets are called "complete" (in the set of functions considered; definition below). For instance, the orthonormal sets corresponding to Examples 1–3 are complete in the set of functions continuous on the intervals considered (or even in more general sets of functions; see Ref. [GR7], Secs. 3.4–3.7, listed in App. 1, where "complete sets" bear the more modern name "total sets").

In this connection, convergence is **convergence in the norm,** also called **mean-square convergence;** that is, a sequence of functions f_k is called **convergent** *with the limit* f if

$$\lim_{k\to\infty} \|f_k - f\| = 0; \tag{12*}$$

written out by (2) (where we can drop the square root, as this does not affect the limit)

$$\lim_{k\to\infty} \int_a^b r(x)[f_k(x) - f(x)]^2\, dx = 0. \tag{12}$$

Accordingly, the series (3) converges and represents f if

$$\lim_{k\to\infty} \int_a^b r(x)[s_k(x) - f(x)]^2\, dx = 0 \tag{13}$$

where s_k is the kth partial sum of (3),

$$s_k(x) = \sum_{m=0}^{k} a_m y_m(x). \tag{14}$$

By definition, an orthonormal set $y_0, y_1, \cdots$ on an interval $a \leqq x \leqq b$ is **complete** *in a set of functions* S defined on $a \leqq x \leqq b$ if we can approximate every f belonging to S arbitrarily closely by a linear combination $a_0y_0 + a_1y_1 + \cdots + a_ky_k$, that is, technically, if for every $\epsilon > 0$ we can find constants $a_0, \cdots, a_k$ (with k large enough) such that

$$\|f - (a_0y_0 + \cdots + a_ky_k)\| < \epsilon. \tag{15}$$

An interesting and basic consequence of the integral in (13) is obtained as follows. Performing the square and using (14), we first have

$$\int_a^b r(x)[s_k(x) - f(x)]^2\, dx = \int_a^b r{s_k}^2\, dx - 2\int_a^b rfs_k\, dx + \int_a^b rf^2\, dx$$

$$= \int_a^b r\left[\sum_{m=0}^{k} a_m y_m\right]^2 dx - 2\sum_{m=0}^{k} a_m \int_a^b rfy_m\, dx + \int_a^b rf^2\, dx.$$

The first integral on the right equals $\Sigma\, {a_m}^2$ because $\int ry_my_l\, dx = 0$ for $m \neq l$, and $\int r{y_m}^2\, dx = 1$. In the second sum on the right, the integral equals a_m, by (4) with

$\|y_m\|^2 = 1$. Hence the first term on the right cancels half of the second term, so that the right side reduces to

$$-\sum_{m=0}^{k} a_m^2 + \int_a^b rf^2\,dx.$$

This is nonnegative because in the previous formula the integrand on the left is nonnegative (recall that the weight $r(x)$ is positive!) and so is the integral on the left. This proves the important **Bessel's inequality**

$$\sum_{m=0}^{k} a_m^2 \leqq \|f\|^2 = \int_a^b r(x)f(x)^2\,dx \qquad (k = 1, 2, \cdots). \tag{16}$$

Here we can let $k \to \infty$, because the left sides form a monotone increasing sequence that is bounded by the right side, so that we have convergence by the familiar Theorem 1 in App. A3.3. Hence

$$\sum_{m=0}^{\infty} a_m^2 \leqq \|f\|^2. \tag{17}$$

Furthermore, if $y_0, y_1, \cdots$ is complete in a set of functions S, then (13) holds for every f belonging to S. By (15) this implies equality in (16) with $k \to \infty$. Hence in the case of completeness every f in S satisfies the so-called **Parseval's equality**

$$\sum_{m=0}^{\infty} a_m^2 = \|f\|^2 = \int_a^b r(x)f(x)^2\,dx. \tag{18}$$

As a consequence of (18) we prove that in the case of *completeness* there is no function orthogonal to *every* function of the orthonormal set, with the trivial exception of a function of zero norm:

THEOREM 1

Completeness

Let $y_0, y_1, \cdots$ be a complete orthonormal set on $a \leqq x \leqq b$ in a set of functions S. Then if a function f belongs to S and is orthogonal to every y_m, it must have norm zero. In particular, if f is continuous, then f must be identically zero.

PROOF Since f is orthogonal to every y_m, the left side of (18) must be zero. If f is continuous, then $\|f\| = 0$ implies $f(x) \equiv 0$, as can be seen directly from (2) with f instead of y_m because $r(x) > 0$. ∎

EXAMPLE 4 **Fourier Series**

The orthonormal set in Example 1 is complete in the set of continuous functions on $-\pi \leqq x \leqq \pi$. Verify directly that $f(x) \equiv 0$ is the only continuous function orthogonal to all the functions of that set.

Solution. Lef f be any continuous function. By the orthogonality (we can omit $\sqrt{2\pi}$ and $\sqrt{\pi}$),

$$\int_{-\pi}^{\pi} 1 \cdot f(x)\,dx = 0, \qquad \int_{-\pi}^{\pi} f(x)\cos mx\,dx = 0, \qquad \int_{-\pi}^{\pi} f(x)\sin mx\,dx = 0.$$

Hence $a_m = 0$ and $b_m = 0$ in (6) for all m, so that (3) reduces to $f(x) \equiv 0$. ∎

This is the end of Chap. 5 on the power series method and the Frobenius method, which are indispensable in solving linear ODEs with variable coefficients, some of the most important of which we have discussed and solved. We have also seen that the latter are important sources of special functions having orthogonality properties that make them suitable for orthogonal series representations of given functions.

PROBLEM SET 5.8

1–4 FOURIER–LEGENDRE SERIES

Showing the details of your calculations, develop:

1. $7x^4 - 6x^2$
2. $(x + 1)^2$
3. $x^3 - x^2 + x - 1$
4. $1, x, x^2, x^3$

5. Prove that if $f(x)$ in Example 2 is even [that is, $f(x) = f(-x)$], its series contains only $P_m(x)$ with even m.

6–16 CAS EXPERIMENTS. FOURIER–LEGENDRE SERIES

Find and graph (on common axes) the partial sums up to that S_{m_0} whose graph practically coincides with that of $f(x)$ within graphical accuracy. State what m_0 is. On what does the size of m_0 seem to depend?

6. $f(x) = \sin \pi x$
7. $f(x) = \sin 2\pi x$
8. $f(x) = \cos \pi x$
9. $f(x) = \cos 2\pi x$
10. $f(x) = \cos 3\pi x$
11. $f(x) = e^x$
12. $f(x) = e^{-x^2}$
13. $f(x) = 1/(1 + x^2)$
14. $f(x) = J_0(\alpha_{0,1}x)$, where $\alpha_{0,1}$ is the first positive zero of J_0
15. $f(x) = J_0(\alpha_{0,2}x)$, where $\alpha_{0,2}$ is the second positive zero of J_0
16. $f(x) = J_1(\alpha_{1,1}x)$, where $\alpha_{1,1}$ is the first positive zero of J_1

17. **CAS EXPERIMENT. Fourier–Bessel Series.** Use Example 3 and again take $n = 10$ and $R = 1$, so that you get the series

$$(19) \quad f(x) = a_1 J_0(\alpha_{0,1}x) + a_2 J_0(\alpha_{0,2}x) + a_3 J_0(\alpha_{0,3}x) + \cdots$$

with the zeros $\alpha_{0,1}$ $\alpha_{0,2}$, $\cdots$ from your CAS (see also Table A1 in App. 5).

(a) Graph the terms $J_0(\alpha_{0,1}x)$, $\cdots$, $J_0(\alpha_{0,10}x)$ for $0 \leqq x \leqq 1$ on common axes.

(b) Write a program for calculating partial sums of (19). Find out for what $f(x)$ your CAS can evaluate the integrals. Take two such $f(x)$ and comment empirically on the speed of convergence by observing the decrease of the coefficients.

(c) Take $f(x) = 1$ in (19) and evaluate the integrals for the coefficients analytically by (24a), Sec. 5.5, with $\nu = 1$. Graph the first few partial sums on common axes.

18. **TEAM PROJECT. Orthogonality on the Entire Real Axis. Hermite Polynomials.**[13] These orthogonal polynomials are defined by $He_0(1) = 1$ and

$$He_n(x) = (-1)^n e^{x^2/2} \frac{d^n}{dx^n}(e^{-x^2/2}), \quad n = 1, 2, \cdots.$$

REMARK. As is true for many special functions, the literature contains more than one notation, and one sometimes defines as Hermite polynomials the functions

$$H_0^* = 1, \qquad H_n^*(x) = (-1)^n e^{x^2} \frac{d^n e^{-x^2}}{dx^n}.$$

This differs from our definition, which is preferred in applications.

(a) Small Values of n. Show that

$$He_1(x) = x, \qquad He_2(x) = x^2 - 1,$$
$$He_3(x) = x^3 - 3x, \qquad He_4(x) = x^4 - 6x^2 + 3.$$

(b) Generating Function. A generating function of the Hermite polynomials is

$$(20) \qquad e^{tx - t^2/2} = \sum_{n=0}^{\infty} a_n(x)t^n$$

because $He_n(x) = n!a_n(x)$. Prove this. *Hint:* Use the formula for the coefficients of a Maclaurin series and note that $tx - \frac{1}{2}t^2 = \frac{1}{2}x^2 - \frac{1}{2}(x - t)^2$.

(c) Derivative. Differentiating the generating function with respect to x, show that

$$(21) \qquad He_n'(x) = nHe_{n-1}(x).$$

[13] CHARLES HERMITE (1822–1901), French mathematician, is known for his work in algebra and number theory. The great HENRI POINCARÉ (1854–1912) was one of his students.

(d) Orthogonality on the x-Axis needs a weight function that goes to zero sufficiently fast as $x \to \pm\infty$. (Why?) Show that the Hermite polynomials are orthogonal on $-\infty < x < \infty$ with respect to the weight function $r(x) = e^{-x^2/2}$. *Hint.* Use integration by parts and (21).

(e) ODEs. Show that

$$(22) \qquad He_n'(x) = xHe_n(x) - He_{n+1}(x).$$

Using this with $n-1$ instead of n and (21), show that $y = He_n(x)$ satisfies the ODE

$$(23) \qquad y'' - xy' + ny = 0.$$

Show that $w = e^{-x^2/4}y$ is a solution of **Weber's equation**[14]

$$(24) \quad w'' + (n + \tfrac{1}{2} - \tfrac{1}{4}x^2)w = 0 \qquad (n = 0, 1, \cdorn).$$

19. WRITING PROJECT. Orthogonality. Write a short report (2–3 pages) about the most important ideas and facts related to orthogonality and orthogonal series and their applications.

CHAPTER 5 REVIEW QUESTIONS AND PROBLEMS

1. What is a power series? Can it contain negative or fractional powers? How would you test for convergence?
2. Why could we use the power series method for Legendre's equation but needed the Frobenius method for Bessel's equation?
3. Why did we introduce two kinds of Bessel functions, J and Y?
4. What is the hypergeometric equation and why did Gauss introduce it?
5. List the three cases of the Frobenius method, giving examples of your own.
6. What is the difference between an initial value problem and a boundary value problem?
7. What does orthogonality of functions mean and how is it used in series expansions? Give examples.
8. What is the Sturm–Liouville theory and its practical importance?
9. What do you remember about the orthogonality of the Legendre polynomials? Of Bessel functions?
10. What is completeness of orthogonal sets? Why is it important?

11–20 SERIES SOLUTIONS

Find a basis of solutions. Try to identify the series as expansions of known functions. (Show the details of your work.)

11. $y'' - 9y = 0$
12. $(1 - x)^2y'' + (1 - x)y' - 3y = 0$
13. $xy'' - (x + 1)y' + y = 0$
14. $x^2y'' - 3xy' + 4y = 0$
15. $y'' + 4xy' + (4x^2 + 2)y = 0$
16. $x^2y'' - 4xy' + (x^2 + 6)y = 0$
17. $xy'' + (2x + 1)y' + (x + 1)y = 0$
18. $(x^2 - 1)y'' - 2xy' + 2y = 0$
19. $(x^2 - 1)y'' + 4xy' + 2y = 0$
20. $x^2y'' + xy' + (4x^4 - 1)y = 0$

21–25 BESSEL'S EQUATION

Find a general solution in terms of Bessel functions. (Use the indicated transformations and show the details.)

21. $x^2y'' + xy' + (36x^2 - 2)y = 0 \qquad (6x = z)$
22. $x^2y'' + 5xy' + (x^2 - 12)y = 0 \qquad (y = u/x^2)$
23. $x^2y'' + xy' + 4(x^4 - 1)y = 0 \qquad (x^2 = z)$
24. $4x^2y'' - 20xy' + (4x^2 + 35)y = 0 \qquad (y = x^3u)$
25. $y'' + k^2x^2y = 0 \qquad (y = u\sqrt{x},\ \tfrac{1}{2}kx^2 = z)$

26–30 BOUNDARY VALUE PROBLEMS

Find the eigenvalues and eigenfunctions.

26. $y'' + \lambda y = 0, \qquad y(0) = 0, \qquad y'(\pi) = 0$
27. $y'' + \lambda y = 0, \qquad y(0) = y(1), \quad y'(0) = y'(1)$
28. $(xy')' + \lambda x^{-1}y = 0, \qquad y(1) = 0, \qquad y(e) = 0.$ (Set $x = e^t$.)
29. $x^2y'' + xy' + (\lambda x^2 - 1)y = 0, \quad y(0) = 0, \quad y(1) = 0$
30. $y'' + \lambda y = 0, \qquad y(0) + y'(0) = 0, \qquad y(2\pi) = 0$

31–35 CAS PROBLEMS

Write a program, develop in a Fourier–Legendre series, and graph the first five partial sums on common axes, together with the given function. Comment on accuracy.

31. $e^{2x} \quad (-1 \leqq x \leqq 1)$
32. $\sin(\pi x^2) \quad (-1 \leqq x \leqq 1)$
33. $1/(1 + |x|) \quad (-1 \leqq x \leqq 1)$
34. $|\cos \pi x| \quad (-1 \leqq x \leqq 1)$
35. x if $0 \leqq x \leqq 1$, 0 if $-1 \leqq x < 0$

[14]HEINRICH WEBER (1842–1913), German mathematician.

SUMMARY OF CHAPTER 5
Series Solution of ODEs. Special Functions

The **power series method** gives solutions of linear ODEs

$$y'' + p(x)y' + q(x)y = 0 \tag{1}$$

with ***variable coefficients*** p and q in the form of a power series (with any center x_0, e.g., $x_0 = 0$)

$$y(x) = \sum_{m=0}^{\infty} a_m(x - x_0)^m = a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + \cdots. \tag{2}$$

Such a solution is obtained by substituting (2) and its derivatives into (1). This gives a **recurrence formula** for the coefficients. You may program this formula (or even obtain and graph the whole solution) on your CAS.

If p and q are **analytic** at x_0 (that is, representable by a power series in powers of $x - x_0$ with positive radius of convergence; Sec. 5.2), then (1) has solutions of this form (2). The same holds if $\tilde{h}$, $\tilde{p}$, $\tilde{q}$ in

$$\tilde{h}(x)y'' + \tilde{p}(x)y' + \tilde{q}(x)y = 0$$

are analytic at x_0 and $\tilde{h}(x_0) \neq 0$, so that we can divide by $\tilde{h}$ and obtain the standard form (1). **Legendre's equation** is solved by the power series method in Sec. 5.3.

The **Frobenius method** (Sec. 5.4) extends the power series method to ODEs

$$y'' + \frac{a(x)}{x - x_0}y' + \frac{b(x)}{(x - x_0)^2}y = 0 \tag{3}$$

whose coefficients are **singular** (i.e., not analytic) at x_0, but are "not too bad," namely, such that a and b are analytic at x_0. Then (3) has at least one solution of the form

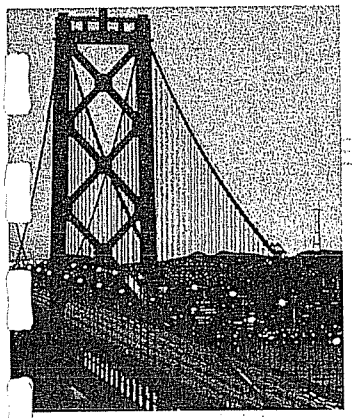
$$y(x) = (x - x_0)^r \sum_{m=0}^{\infty} a_m(x - x_0)^m = a_0(x - x_0)^r + a_1(x - x_0)^{r+1} + \cdots \tag{4}$$

where r can be any real (or even complex) number and is determined by substituting (4) into (3) from the **indicial equation** (Sec. 5.4), along with the coefficients of (4). A second linearly independent solution of (3) may be of a similar form (with different r and a_m's) or may involve a logarithmic term. **Bessel's equation** is solved by the Frobenius method in Secs. 5.5 and 5.6.

"Special functions" is a common name for higher functions, as opposed to the usual functions of calculus. Most of them arise either as nonelementary integrals [see (24)–(44) in App. 3.1] or as solutions of (1) or (3). They get a name and notation and are included in the usual CASs if they are important in application or in theory.

Of this kind, and particularly useful to the engineer and physicist, are **Legendre's equation and polynomials** P_0, P_1, $\cdots$ (Sec. 5.3), **Gauss's hypergeometric equation and functions** $F(a, b, c; x)$ (Sec. 5.4), and **Bessel's equation and functions** J_ν **and** Y_ν (Secs. 5.5, 5.6).

Modeling involving ODEs usually leads to initial value problems (Chaps. 1–3) or **boundary value problems.** Many of the latter can be written in the form of **Sturm–Liouville problems** (Sec. 5.7). These are **eigenvalue problems** involving a parameter λ that is often related to frequencies, energies, or other physical quantities. Solutions of Sturm–Liouville problems, called **eigenfunctions,** have many general properties in common, notably the highly important **orthogonality** (Sec. 5.7), which is useful in **eigenfunction expansions** (Sec. 5.8) in terms of cosine and sine (*"Fourier series"*, the topic of Chap. 11), Legendre polynomials, Bessel functions (Sec. 5.8), and other eigenfunctions.



CHAPTER 6

Laplace Transforms

The Laplace transform method is a powerful method for solving linear ODEs and corresponding initial value problems, as well as systems of ODEs arising in engineering. The process of solution consists of three steps (see Fig. 112).

Step 1. The given ODE is transformed into an algebraic equation **("subsidiary equation").**

Step 2. The subsidiary equation is solved by purely algebraic manipulations.

Step 3. The solution in Step 2 is transformed back, resulting in the solution of the given problem.

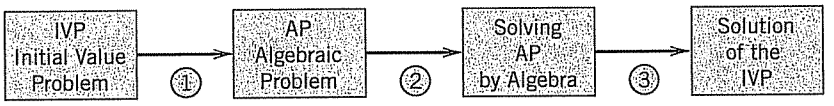


Fig. 112. Solving an IVP by Laplace transforms

Thus solving an ODE is reduced to an *algebraic* problem (plus those transformations). This switching from calculus to algebra is called **operational calculus.** The Laplace transform method is the most important operational method to the engineer. This method has two main advantages over the usual methods of Chaps. 1–4:

A. Problems are solved more directly, initial value problems without first determining a general solution, and nonhomogeneous ODEs without first solving the corresponding homogeneous ODE.

B. More importantly, the use of the **unit step function (Heaviside function** in Sec. 6.3) and **Dirac's delta** (in Sec. 6.4) make the method particularly powerful for problems with inputs (driving forces) that have discontinuities or represent short impulses or complicated periodic functions.

In this chapter we consider the Laplace transform and its application to engineering problems involving ODEs. PDEs will be solved by the Laplace transform in Sec. 12.11.

General formulas are listed in Sec. 6.8, **transforms and inverses** in Sec. 6.9. The usual **CASs** can handle most Laplace transforms.

Prerequisite: Chap. 2
Sections that may be omitted in a shorter course: 6.5, 6.7
References and Answers to Problems: App. 1 Part A, App. 2.

6.1 Laplace Transform. Inverse Transform. Linearity. s-Shifting

If $f(t)$ is a function defined for all $t \geqq 0$, its **Laplace transform**[1] is the integral of $f(t)$ times e^{-st} from $t = 0$ to ∞. It is a function of s, say, $F(s)$, and is denoted by $\mathscr{L}(f)$; thus

$$F(s) = \mathscr{L}(f) = \int_0^\infty e^{-st} f(t)\, dt. \tag{1}$$

Here we must assume that $f(t)$ is such that the integral exists (that is, has some finite value). This assumption is usually satisfied in applications—we shall discuss this near the end of the section.

Not only is the result $F(s)$ called the Laplace transform, but the operation just described, which yields $F(s)$ from a given $f(t)$, is also called the **Laplace transform.** It is an **"integral transform"**

$$F(s) = \int_0^\infty k(s, t) f(t)\, dt$$

with **"kernel"** $k(s, t) = e^{-st}$.

Furthermore, the given function $f(t)$ in (1) is called the **inverse transform** of $F(s)$ and is denoted by $\mathscr{L}^{-1}(F)$; that is, we shall write

$$f(t) = \mathscr{L}^{-1}(F). \tag{1*}$$

Note that (1) and (1*) together imply $\mathscr{L}^{-1}(\mathscr{L}(f)) = f$ and $\mathscr{L}(\mathscr{L}^{-1}(F)) = F$.

Notation

Original functions depend on t and their transforms on s—keep this in mind! Original functions are denoted by ***lowercase letters*** and their transforms by the same ***letters in capital,*** so that $F(s)$ denotes the transform of $f(t)$, and $Y(s)$ denotes the transform of $y(t)$, and so on.

EXAMPLE 1 **Laplace Transform**

Let $f(t) = 1$ when $t \geqq 0$. Find $F(s)$.

Solution. From (1) we obtain by integration

$$\mathscr{L}(f) = \mathscr{L}(1) = \int_0^\infty e^{-st}\, dt = -\frac{1}{s} e^{-st} \Big|_0^\infty = \frac{1}{s} \qquad (s > 0).$$

---

[1]PIERRE SIMON MARQUIS DE LAPLACE (1749–1827), great French mathematician, was a professor in Paris. He developed the foundation of potential theory and made important contributions to celestial mechanics, astronomy in general, special functions, and probability theory. Napoléon Bonaparte was his student for a year. For Laplace's interesting political involvements, see Ref. [GR2], listed in App. 1.

The powerful practical Laplace transform techniques were developed over a century later by the English electrical engineer OLIVER HEAVISIDE (1850–1925) and were often called "Heaviside calculus."

We shall drop variables when this simplifies formulas without causing confusion. For instance, in (1) we wrote $\mathscr{L}(f)$ instead of $\mathscr{L}(f)(s)$ and in (1*) $\mathscr{L}^{-1}(F)$ instead of $\mathscr{L}^{-1}(F)(t)$.

Our notation is convenient, but we should say a word about it. The interval of integration in (1) is infinite. Such an integral is called an **improper integral** and, by definition, is evaluated according to the rule

$$\int_0^\infty e^{-st} f(t)\,dt = \lim_{T\to\infty} \int_0^T e^{-st} f(t)\,dt.$$

Hence our convenient notation means

$$\int_0^\infty e^{-st}\,dt = \lim_{T\to\infty} \left[-\frac{1}{s}\, e^{-st} \right]_0^T = \lim_{T\to\infty} \left[-\frac{1}{s}\, e^{-sT} + \frac{1}{s}\, e^0 \right] = \frac{1}{s} \qquad (s > 0).$$

We shall use this notation throughout this chapter. ■

EXAMPLE 2 **Laplace Transform $\mathcal{L}(e^{at})$ of the Exponential Function e^{at}**

Let $f(t) = e^{at}$ when $t \geqq 0$, where a is a constant. Find $\mathcal{L}(f)$.

Solution. Again by (1),

$$\mathcal{L}(e^{at}) = \int_0^\infty e^{-st} e^{at}\,dt = \frac{1}{a-s}\, e^{-(s-a)t} \Big|_0^\infty;$$

hence, when $s - a > 0$,

$$\mathcal{L}(e^{at}) = \frac{1}{s-a}.$$ ■

Must we go on in this fashion and obtain the transform of one function after another directly from the definition? The answer is no. And the reason is that new transforms can be found from known ones by the use of the many general properties of the Laplace transform. Above all, the Laplace transform is a "linear operation," just as differentiation and integration. By this we mean the following.

THEOREM 1

Linearity of the Laplace Transform

The Laplace transform is a linear operation; that is, for any functions $f(t)$ and $g(t)$ whose transforms exist and any constants a and b the transform of $af(t) + bg(t)$ exists, and

$$\mathcal{L}\{af(t) + bg(t)\} = a\mathcal{L}\{f(t)\} + b\mathcal{L}\{g(t)\}.$$

PROOF By the definition in (1),

$$\begin{aligned}\mathcal{L}\{af(t) + bg(t)\} &= \int_0^\infty e^{-st}[af(t) + bg(t)]\,dt \\ &= a\int_0^\infty e^{-st} f(t)\,dt + b\int_0^\infty e^{-st} g(t)\,dt = a\mathcal{L}\{f(t)\} + b\mathcal{L}\{g(t)\}.\end{aligned}$$ ■

EXAMPLE 3 **Application of Theorem 1: Hyperbolic Functions**

Find the transforms of $\cosh at$ and $\sinh at$.

Solution. Since $\cosh at = \frac{1}{2}(e^{at} + e^{-at})$ and $\sinh at = \frac{1}{2}(e^{at} - e^{-at})$, we obtain from Example 2 and Theorem 1

$$\mathcal{L}(\cosh at) = \frac{1}{2}(\mathcal{L}(e^{at}) + \mathcal{L}(e^{-at})) = \frac{1}{2}\left(\frac{1}{s-a} + \frac{1}{s+a}\right) = \frac{s}{s^2 - a^2}$$

$$\mathcal{L}(\sinh at) = \frac{1}{2}(\mathcal{L}(e^{at}) - \mathcal{L}(e^{-at})) = \frac{1}{2}\left(\frac{1}{s-a} - \frac{1}{s+a}\right) = \frac{a}{s^2 - a^2}.$$ ■

EXAMPLE 4 Cosine and Sine

Derive the formulas

$$\mathcal{L}(\cos \omega t) = \frac{s}{s^2 + \omega^2}, \qquad \mathcal{L}(\sin \omega t) = \frac{\omega}{s^2 + \omega^2}.$$

Solution by Calculus. We write $L_c = \mathcal{L}(\cos \omega t)$ and $L_s = \mathcal{L}(\sin \omega t)$. Integrating by parts and noting that the integral-free parts give no contribution from the upper limit ∞, we obtain

$$L_c = \int_0^\infty e^{-st} \cos \omega t \, dt = \frac{e^{-st}}{-s} \cos \omega t \Big|_0^\infty - \frac{\omega}{s} \int_0^\infty e^{-st} \sin \omega t \, dt = \frac{1}{s} - \frac{\omega}{s} L_s,$$

$$L_s = \int_0^\infty e^{-st} \sin \omega t \, dt = \frac{e^{-st}}{-s} \sin \omega t \Big|_0^\infty + \frac{\omega}{s} \int_0^\infty e^{-st} \cos \omega t \, dt = \frac{\omega}{s} L_c.$$

By substituting L_s into the formula for L_c on the right and then by substituting L_c into the formula for L_s on the right, we obtain

$$L_c = \frac{1}{s} - \frac{\omega}{s}\left(\frac{\omega}{s} L_c\right), \qquad L_c\left(1 + \frac{\omega^2}{s^2}\right) = \frac{1}{s}, \qquad L_c = \frac{s}{s^2 + \omega^2},$$

$$L_s = \frac{\omega}{s}\left(\frac{1}{s} - \frac{\omega}{s} L_s\right), \qquad L_s\left(1 + \frac{\omega^2}{s^2}\right) = \frac{\omega}{s^2}, \qquad L_s = \frac{\omega}{s^2 + \omega^2}.$$

Solution by Transforms Using Derivatives. See next section.

Solution by Complex Methods. In Example 2, if we set $a = i\omega$ with $i = \sqrt{-1}$, we obtain

$$\mathcal{L}(e^{i\omega t}) = \frac{1}{s - i\omega} = \frac{s + i\omega}{(s - i\omega)(s + i\omega)} = \frac{s + i\omega}{s^2 + \omega^2} = \frac{s}{s^2 + \omega^2} + i \frac{\omega}{s^2 + \omega^2}.$$

Now by Theorem 1 and $e^{i\omega t} = \cos \omega t + i \sin \omega t$ [see (11) in Sec. 2.2 with ωt instead of t] we have

$$\mathcal{L}(e^{i\omega t}) = \mathcal{L}(\cos \omega t + i \sin \omega t) = \mathcal{L}(\cos \omega t) + i\mathcal{L}(\sin \omega t).$$

If we equate the real and imaginary parts of this and the previous equation, the result follows. (This formal calculation can be justified in the theory of complex integration.) ▩

Basic transforms are listed in Table 6.1. We shall see that from these almost all the others can be obtained by the use of the general properties of the Laplace transform. Formulas 1–3 are special cases of formula 4, which is proved by induction. Indeed, it is true for $n = 0$ because of Example 1 and $0! = 1$. We make the induction hypothesis that it holds for any integer $n \geqq 0$ and then get it for $n + 1$ directly from (1). Indeed, integration by parts first gives

$$\mathcal{L}(t^{n+1}) = \int_0^\infty e^{-st} t^{n+1} \, dt = -\frac{1}{s} e^{-st} t^{n+1} \Big|_0^\infty + \frac{n + 1}{s} \int_0^\infty e^{-st} t^n \, dt.$$

Now the integral-free part is zero and the last part is $(n + 1)/s$ times $\mathcal{L}(t^n)$. From this and the induction hypothesis,

$$\mathcal{L}(t^{n+1}) = \frac{n + 1}{s} \mathcal{L}(t^n) = \frac{n + 1}{s} \cdot \frac{n!}{s^{n+1}} = \frac{(n + 1)!}{s^{n+2}}.$$

This proves formula 4.

Table 6.1 Some Functions $f(t)$ and Their Laplace Transforms $\mathcal{L}(f)$

	$f(t)$	$\mathcal{L}(f)$		$f(t)$	$\mathcal{L}(f)$
1	1	$1/s$	7	$\cos \omega t$	$\dfrac{s}{s^2 + \omega^2}$
2	t	$1/s^2$	8	$\sin \omega t$	$\dfrac{\omega}{s^2 + \omega^2}$
3	t^2	$2!/s^3$	9	$\cosh at$	$\dfrac{s}{s^2 - a^2}$
4	t^n $(n = 0, 1, \cdorn)$	$\dfrac{n!}{s^{n+1}}$	10	$\sinh at$	$\dfrac{a}{s^2 - a^2}$
5	t^a (a positive)	$\dfrac{\Gamma(a + 1)}{s^{a+1}}$	11	$e^{at} \cos \omega t$	$\dfrac{s - a}{(s - a)^2 + \omega^2}$
6	e^{at}	$\dfrac{1}{s - a}$	12	$e^{at} \sin \omega t$	$\dfrac{\omega}{(s - a)^2 + \omega^2}$

$\Gamma(a + 1)$ in formula 5 is the so-called *gamma function* [(15) in Sec. 5.5 or (24) in App. A3.1]. We get formula 5 from (1), setting $st = x$:

$$\mathcal{L}(t^a) = \int_0^\infty e^{-st} t^a \, dt = \int_0^\infty e^{-x} \left(\frac{x}{s}\right)^a \frac{dx}{s} = \frac{1}{s^{a+1}} \int_0^\infty e^{-x} x^a \, dx$$

where $s > 0$. The last integral is precisely that defining $\Gamma(a + 1)$, so we have $\Gamma(a + 1)/s^{a+1}$, as claimed. (**CAUTION!** $\Gamma(a + 1)$ has x^a in the integral, not x^{a+1}.)

Note the formula 4 also follows from 5 because $\Gamma(n + 1) = n!$ for integer $n \geqq 0$.

Formulas 6–10 were proved in Examples 2–4. Formulas 11 and 12 will follow from 7 and 8 by "shifting," to which we turn next.

s-Shifting: Replacing s by $s - a$ in the Transform

The Laplace transform has the very useful property that if we know the transform of $f(t)$, we can immediately get that of $e^{at} f(t)$, as follows.

THEOREM 2

First Shifting Theorem, s-Shifting

If $f(t)$ has the transform $F(s)$ (where $s > k$ for some k), then $e^{at} f(t)$ has the transform $F(s - a)$ (where $s - a > k$). In formulas,

$$\mathcal{L}\{e^{at} f(t)\} = F(s - a)$$

or, if we take the inverse on both sides,

$$e^{at} f(t) = \mathcal{L}^{-1}\{F(s - a)\}.$$

PROOF We obtain $F(s-a)$ by replacing s with $s-a$ in the integral in (1), so that

$$F(s-a)=\int_0^\infty e^{-(s-a)t}f(t)\,dt=\int_0^\infty e^{-st}[e^{at}f(t)]\,dt=\mathcal{L}\{e^{at}f(t)\}.$$

If $F(s)$ exists (i.e., is finite) for s greater than some k, then our first integral exists for $s-a>k$. Now take the inverse on both sides of this formula to obtain the second formula in the theorem. (**CAUTION!** $-a$ in $F(s-a)$ but $+a$ in $e^{at}f(t)$.) ▩

EXAMPLE 5 **s-Shifting: Damped Vibrations. Completing the Square**

From Example 4 and the first shifting theorem we immediately obtain formulas 11 and 12 in Table 6.1,

$$\mathcal{L}\{e^{at}\cos\omega t\}=\frac{s-a}{(s-a)^2+\omega^2},\qquad \mathcal{L}\{e^{at}\sin\omega t\}=\frac{\omega}{(s-a)^2+\omega^2}.$$

For instance, use these formulas to find the inverse of the transform

$$\mathcal{L}(f)=\frac{3s-137}{s^2+2s+401}.$$

Solution. Applying the inverse transform, using its linearity (Prob. 28), and completing the square, we obtain

$$f=\mathcal{L}^{-1}\left\{\frac{3(s+1)-140}{(s+1)^2+400}\right\}=3\mathcal{L}^{-1}\left\{\frac{s+1}{(s+1)^2+20^2}\right\}-7\mathcal{L}^{-1}\left\{\frac{20}{(s+1)^2+20^2}\right\}.$$

We now see that the inverse of the right side is the damped vibration (Fig. 113)

$$f(t)=e^{-t}(3\cos 20t-7\sin 20t).$$ ▩

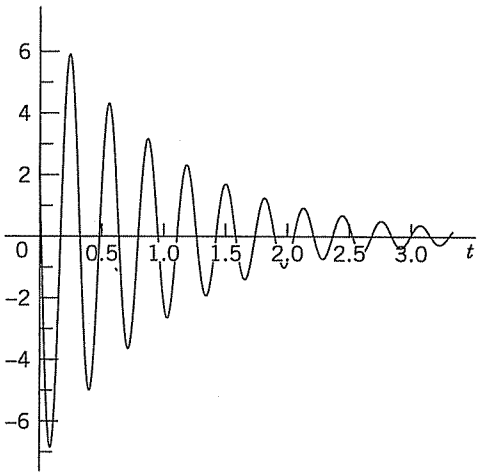


Fig. 113. Vibrations in Example 5

Existence and Uniqueness of Laplace Transforms

This is not a big *practical* problem because in most cases we can check the solution of an ODE without too much trouble. Nevertheless we should be aware of some basic facts.

A function $f(t)$ has a Laplace transform if it does not grow too fast, say, if for all $t\geqq 0$ and some constants M and k it satisfies the **"growth restriction"**

$$|f(t)|\leqq Me^{kt}. \tag{2}$$

(The growth restriction (2) is sometimes called "growth of exponential order," which may be misleading since it hides that the exponent must be kt, not kt^2 or similar.)

$f(t)$ need not be continuous, but it should not be too bad. The technical term (generally used in mathematics) is *piecewise continuity*. $f(t)$ is **piecewise continuous** on a finite interval $a \leqq t \leqq b$ where f is defined, if this interval can be divided into *finitely many* subintervals in each of which f is continuous and has finite limits as t approaches either endpoint of such a subinterval from the interior. This then gives **finite jumps** as in Fig. 114 as the only possible discontinuities, but this suffices in most applications, and so does the following theorem.

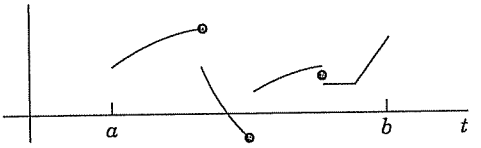


Fig. 114. Example of a piecewise continuous function $f(t)$. (The dots mark the function values at the jumps.)

THEOREM 3

Existence Theorem for Laplace Transforms

If $f(t)$ is defined and piecewise continuous on every finite interval on the semi-axis $t \geqq 0$ and satisfies (2) for all $t \geqq 0$ and some constants M and k, then the Laplace transform $\mathcal{L}(f)$ exists for all $s > k$.

PROOF Since $f(t)$ is piecewise continuous, $e^{-st}f(t)$ is integrable over any finite interval on the t-axis. From (2), assuming that $s > k$ (to be needed for the existence of the last of the following integrals), we obtain the proof of the existence of $\mathcal{L}(f)$ from

$$|\mathcal{L}(f)| = \left|\int_0^\infty e^{-st}f(t)\,dt\right| \leqq \int_0^\infty |f(t)|e^{-st}\,dt \leqq \int_0^\infty Me^{kt}e^{-st}\,dt = \frac{M}{s-k}. \qquad \blacksquare$$

Note that (2) can be readily checked. For instance, $\cosh t < e^t$, $t^n < n!e^t$ (because $t^n/n!$ is a single term of the Maclaurin series), and so on. A function that does not satisfy (2) for any M and k is e^{t^2} (take logarithms to see it). We mention that the conditions in Theorem 3 are sufficient rather than necessary (see Prob. 22).

Uniqueness. If the Laplace transform of a given function exists, it is uniquely determined. Conversely, it can be shown that if two functions (both defined on the positive real axis) have the same transform, these functions cannot differ over an interval of positive length, although they may differ at isolated points (see Ref. [A14] in App. 1). Hence we may say that the inverse of a given transform is essentially unique. In particular, if two *continuous* functions have the same transform, they are completely identical.

PROBLEM SET 6.1

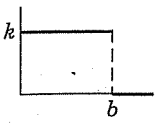
1–20 LAPLACE TRANSFORMS

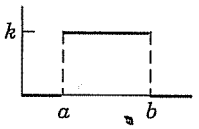
Find the Laplace transforms of the following functions. Show the details of your work. (a, b, k, ω, θ are constants.)

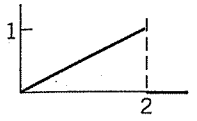
1. $t^2 - 2t$
2. $(t^2 - 3)^2$
3. $\cos 2\pi t$
4. $\sin^2 4t$
5. $e^{2t}\cosh t$
6. $e^{-t}\sinh 5t$
7. $\cos(\omega t + \theta)$
8. $\sin(3t - \frac{1}{2})$
9. e^{3a-2bt}
10. $-8\sin 0.2t$

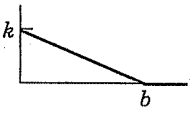
11. $\sin t \cos t$

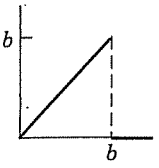
12. $(t + 1)^3$

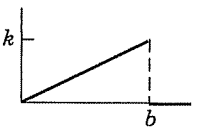
13.


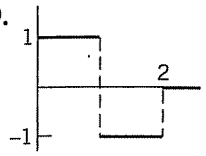
14.


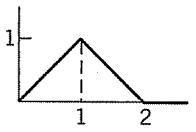
15.


16.


17.


18.


19.


20.


21. Using $\mathcal{L}(f)$ in Prob. 13, find $\mathcal{L}(f_1)$, where $f_1(t) = 0$ if $t \leqq 2$ and $f_1(t) = 1$ if $t > 2$.

22. (Existence) Show that $\mathcal{L}(1/\sqrt{t}) = \sqrt{\pi/s}$. [Use (30) $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ in App. 3.1.] Conclude from this that the conditions in Theorem 3 are sufficient but not necessary for the existence of a Laplace transform.

23. (Change of scale) If $\mathcal{L}(f(t)) = F(s)$ and c is any positive constant, show that $\mathcal{L}(f(ct)) = F(s/c)/c$. (*Hint:* Use (1).) Use this to obtain $\mathcal{L}(\cos \omega t)$ from $\mathcal{L}(\cos t)$.

24. (Nonexistence) Show that e^{t^2} does not satisfy a condition of the form (2).

25. (Nonexistence) Give simple examples of functions (defined for all $x \geqq 0$) that have no Laplace transform.

26. (Table 6.1) Derive formula 6 from formulas 9 and 10.

27. (Table 6.1) Convert Table 6.1 from a table for finding transforms to a table for finding inverse transforms (with obvious changes, e.g., $\mathcal{L}^{-1}(1/s^n) = t^{n-1}/(n-1)!$, etc.).

28. (Inverse transform) Prove that $\mathcal{L}^{-1}$ is linear. *Hint.* Use the fact that $\mathcal{L}$ is linear.

29–40 INVERSE LAPLACE TRANSFORMS

Given $F(s) = \mathcal{L}(f)$, find $f(t)$. Show the details. (L, n, k, a, b are constants.)

29. $\dfrac{4s - 3\pi}{s^2 + \pi^2}$

30. $\dfrac{2s + 16}{s^2 - 16}$

31. $\dfrac{s^4 - 3s^2 + 12}{s^5}$

32. $\dfrac{10}{2s + \sqrt{2}}$

33. $\dfrac{n\pi L}{L^2 s^2 + n^2\pi^2}$

34. $\dfrac{20}{(s-1)(s+4)}$

35. $\dfrac{8}{s^2 + 4s}$

36. $\displaystyle\sum_{k=1}^{4} \frac{(k+1)^2}{s + k^2}$

37. $\dfrac{1}{(s - \sqrt{3})(s + \sqrt{5})}$

38. $\dfrac{18s - 12}{9s^2 - 1}$

39. $\dfrac{1}{s^2 + 5} - \dfrac{1}{s + 5}$

40. $\dfrac{1}{(s+a)(s+b)}$

41–54 APPLICATIONS OF THE FIRST SHIFTING THEOREM (s-SHIFTING)

In Probs. 41–46 find the transform. In Probs. 47–54 find the inverse transform. Show the details.

41. $3.8te^{2.4t}$

42. $-3t^4e^{-0.5t}$

43. $5e^{-at} \sin \omega t$

44. $e^{-3t} \cos \pi t$

45. $e^{-kt}(a \cos t + b \sin t)$

46. $e^{-t}(a_0 + a_1 t + \cdots + a_n t^n)$

47. $\dfrac{7}{(s-1)^3}$

48. $\dfrac{\pi}{(s + \pi)^2}$

49. $\dfrac{\sqrt{8}}{(s + \sqrt{2})^3}$

50. $\dfrac{s - 6}{(s-1)^2 + 4}$

51. $\dfrac{15}{s^2 + 4s + 29}$

52. $\dfrac{4s - 2}{s^2 - 6s + 18}$

53. $\dfrac{\pi}{s^2 + 10\pi s + 24\pi^2}$

54. $\dfrac{2s - 56}{s^2 - 4s - 12}$

6.2 Transforms of Derivatives and Integrals. ODEs

The Laplace transform is a method of solving ODEs and initial value problems. The crucial idea is that ***operations of calculus on functions are replaced by operations of algebra on transforms.*** Roughly, *differentiation* of $f(t)$ will correspond to *multiplication* of $\mathcal{L}(f)$ by s (see Theorems 1 and 2) and *integration* of $f(t)$ to *division* of $\mathcal{L}(f)$ by s. To solve ODEs, we must first consider the Laplace transform of derivatives.

THEOREM 1

Laplace Transform of Derivatives

The transforms of the first and second derivatives of $f(t)$ satisfy

$$\mathcal{L}(f') = s\mathcal{L}(f) - f(0) \tag{1}$$

$$\mathcal{L}(f'') = s^2\mathcal{L}(f) - sf(0) - f'(0). \tag{2}$$

Formula (1) *holds if $f(t)$ is continuous for all $t \geqq 0$ and satisfies the growth restriction* (2) *in Sec.* 6.1 *and $f'(t)$ is piecewise continuous on every finite interval on the semi-axis $t \geqq 0$. Similarly,* (2) *holds if f and f' are continuous for all $t \geqq 0$ and satisfy the growth restriction and f'' is piecewise continuous on every finite interval on the semi-axis $t \geqq 0$.*

PROOF We prove (1) first under the *additional assumption* that f' is continuous. Then by the definition and integration by parts,

$$\mathcal{L}(f') = \int_0^\infty e^{-st}f'(t)\,dt = [e^{-st}f(t)]\Big|_0^\infty + s\int_0^\infty e^{-st}f(t)\,dt.$$

Since f satisfies (2) in Sec. 6.1, the integrated part on the right is zero at the upper limit when $s > k$, and at the lower limit it contributes $-f(0)$. The last integral is $\mathcal{L}(f)$. It exists for $s > k$ because of Theorem 3 in Sec. 6.1. Hence $\mathcal{L}(f')$ exists when $s > k$ and (1) holds.

If f' is merely piecewise continuous, the proof is similar. In this case the interval of integration of f' must be broken up into parts such that f' is continuous in each such part.

The proof of (2) now follows by applying (1) to f'' and then substituting (1), that is

$$\mathcal{L}(f'') = s\mathcal{L}(f') - f'(0) = s[s\mathcal{L}(f) - f(0)] = s^2\mathcal{L}(f) - sf(0) - f'(0). \qquad \blacksquare$$

Continuing by substitution as in the proof of (2) and using induction, we obtain the following extension of Theorem 1.

THEOREM 2

Laplace Transform of the Derivative $f^{(n)}$ of Any Order

Let $f, f', \cdots, f^{(n-1)}$ be continuous for all $t \geqq 0$ and satisfy the growth restriction (2) *in Sec.* 6.1. *Furthermore, let $f^{(n)}$ be piecewise continuous on every finite interval on the semi-axis $t \geqq 0$. Then the transform of $f^{(n)}$ satisfies*

$$\mathcal{L}(f^{(n)}) = s^n\mathcal{L}(f) - s^{n-1}f(0) - s^{n-2}f'(0) - \cdots - f^{(n-1)}(0). \tag{3}$$

EXAMPLE 1 **Transform of a Resonance Term (Sec. 2.8)**

Let $f(t) = t \sin \omega t$. Then $f(0) = 0$, $f'(t) = \sin \omega t + \omega t \cos \omega t$, $f'(0) = 0$, $f'' = 2\omega \cos \omega t - \omega^2 t \sin \omega t$. Hence by (2),

$$\mathcal{L}(f'') = 2\omega\frac{s}{s^2 + \omega^2} - \omega^2\mathcal{L}(f) = s^2\mathcal{L}(f), \qquad \text{thus} \qquad \mathcal{L}(f) = \mathcal{L}(t \sin \omega t) = \frac{2\omega s}{(s^2 + \omega^2)^2}. \qquad \blacksquare$$

EXAMPLE 2 **Formulas 7 and 8 in Table 6.1, Sec. 6.1**

This is a third derivation of $\mathcal{L}(\cos \omega t)$ and $\mathcal{L}(\sin \omega t)$; cf. Example 4 in Sec. 6.1. Let $f(t) = \cos \omega t$. Then $f(0) = 1$, $f'(0) = 0$, $f''(t) = -\omega^2 \cos \omega t$. From this and (2) we obtain

$$\mathcal{L}(f'') = s^2\mathcal{L}(f) - s = -\omega^2\mathcal{L}(f). \qquad \text{By algebra,} \qquad \mathcal{L}(\cos \omega t) = \frac{s}{s^2 + \omega^2}.$$

Similarly, let $g = \sin \omega t$. Then $g(0) = 0$, $g' = \omega \cos \omega t$. From this and (1) we obtain

$$\mathcal{L}(g') = s\mathcal{L}(g) = \omega\mathcal{L}(\cos \omega t). \qquad \text{Hence} \qquad \mathcal{L}(\sin \omega t) = \frac{\omega}{s}\mathcal{L}(\cos \omega t) = \frac{\omega}{s^2 + \omega^2}. \qquad \blacksquare$$

Laplace Transform of the Integral of a Function

Differentiation and integration are inverse operations, and so are multiplication and division. Since differentiation of a function $f(t)$ (roughly) corresponds to multiplication of its transform $\mathcal{L}(f)$ by s, we expect integration of $f(t)$ to correspond to division of $\mathcal{L}(f)$ by s:

THEOREM 3 **Laplace Transform of Integral**

Let $F(s)$ denote the transform of a function $f(t)$ which is piecewise continuous for $t \geqq 0$ and satisfies a growth restriction (2), *Sec.* 6.1. *Then, for $s > 0$, $s > k$, and $t > 0$,*

(4) $$\mathcal{L}\left\{\int_0^t f(\tau)\, d\tau\right\} = \frac{1}{s}F(s), \qquad \text{thus} \qquad \int_0^t f(\tau)\, d\tau = \mathcal{L}^{-1}\left\{\frac{1}{s}F(s)\right\}.$$

PROOF Denote the integral in (4) by $g(t)$. Since $f(t)$ is piecewise continuous, $g(t)$ is continuous, and (2), Sec. 6.1, gives

$$|g(t)| = \left|\int_0^t f(\tau)\, d\tau\right| \leqq \int_0^t |f(\tau)|\, d\tau \leqq M\int_0^t e^{k\tau}\, d\tau = \frac{M}{k}(e^{kt} - 1) \leqq \frac{M}{k}e^{kt} \qquad (k > 0).$$

This shows that $g(t)$ also satisfies a growth restriction. Also, $g'(t) = f(t)$, except at points at which $f(t)$ is discontinuous. Hence $g'(t)$ is piecewise continuous on each finite interval and, by Theorem 1, since $g(0) = 0$ (the integral from 0 to 0 is zero)

$$\mathcal{L}\{f(t)\} = \mathcal{L}\{g'(t)\} = s\mathcal{L}\{g(t)\} - g(0) = s\mathcal{L}\{g(t)\}.$$

Division by s and interchange of the left and right sides gives the first formula in (4), from which the second follows by taking the inverse transform on both sides. $\blacksquare$

EXAMPLE 3 **Application of Theorem 3: Formulas 19 and 20 in the Table of Sec. 6.9**

Using Theorem 3, find the inverse of $\dfrac{1}{s(s^2 + \omega^2)}$ and $\dfrac{1}{s^2(s^2 + \omega^2)}$.

Solution. From Table 6.1 in Sec. 6.1 and the integration in (4) (second formula with the sides interchanged) we obtain

$$\mathcal{L}^{-1}\left\{\frac{1}{s^2 + \omega^2}\right\} = \frac{\sin \omega t}{\omega}, \qquad \mathcal{L}^{-1}\left\{\frac{1}{s(s^2 + \omega^2)}\right\} = \int_0^t \frac{\sin \omega\tau}{\omega}\, d\tau = \frac{1}{\omega^2}(1 - \cos \omega t).$$

This is formula 19 in Sec. 6.9. Integrating this result again and using (4) as before, we obtain formula 20 in Sec. 6.9:

$$\mathscr{L}^{-1}\left\{\frac{1}{s^2(s^2+\omega^2)}\right\} = \frac{1}{\omega^2}\int_0^t (1-\cos\omega\tau)\,d\tau = \left[\frac{\tau}{\omega^2} - \frac{\sin\omega\tau}{\omega^3}\right]_0^t = \frac{t}{\omega^2} - \frac{\sin\omega t}{\omega^3}.$$

It is typical that results such as these can be found in several ways. In this example, try partial fraction reduction.

Differential Equations, Initial Value Problems

We shall now discuss how the Laplace transform method solves ODEs and initial value problems. We consider an initial value problem

$$y'' + ay' + by = r(t), \qquad y(0) = K_0, \qquad y'(0) = K_1 \tag{5}$$

where a and b are constant. Here $r(t)$ is the given **input** (*driving force*) applied to the mechanical or electrical system and $y(t)$ is the **output** (*response to the input*) to be obtained. In Laplace's method we do three steps:

Step 1. Setting up the subsidiary equation. This is an algebraic equation for the transform $Y = \mathscr{L}(y)$ obtained by transforming (5) by means of (1) and (2), namely,

$$[s^2Y - sy(0) - y'(0)] + a[sY - y(0)] + bY = R(s)$$

where $R(s) = \mathscr{L}(r)$. Collecting the Y-terms, we have the subsidiary equation

$$(s^2 + as + b)Y = (s + a)y(0) + y'(0) + R(s).$$

Step 2. Solution of the subsidiary equation by algebra. We divide by $s^2 + as + b$ and use the so-called **transfer function**

$$Q(s) = \frac{1}{s^2 + as + b} = \frac{1}{(s + \frac{1}{2}a)^2 + b - \frac{1}{4}a^2}. \tag{6}$$

(Q is often denoted by H, but we need H much more frequently for other purposes.) This gives the solution

$$Y(s) = [(s + a)y(0) + y'(0)]Q(s) + R(s)Q(s). \tag{7}$$

If $y(0) = y'(0) = 0$, this is simply $Y = RQ$; hence

$$Q = \frac{Y}{R} = \frac{\mathscr{L}(\text{output})}{\mathscr{L}(\text{input})}$$

and this explains the name of Q. Note that ***Q depends neither on r(t) nor on the initial conditions*** (but only on a and b).

Step 3. Inversion of Y to obtain $y = \mathscr{L}^{-1}(Y)$. We reduce (7) (usually by ***partial fractions*** as in calculus) to a sum of terms whose inverses can be found from the tables (e.g., in Sec. 6.1 or Sec. 6.9) or by a CAS, so that we obtain the solution $y(t) = \mathscr{L}^{-1}(Y)$ *of* (5).

EXAMPLE 4 Initial Value Problem: The Basic Laplace Steps

Solve

$$y'' - y = t, \qquad y(0) = 1, \qquad y'(0) = 1.$$

Solution. *Step 1.* From (2) and Table 6.1 we get the subsidiary equation [with $Y = \mathcal{L}(y)$]

$$s^2Y - sy(0) - y'(0) - Y = 1/s^2, \qquad \text{thus} \qquad (s^2 - 1)Y = s + 1 + 1/s^2.$$

Step 2. The transfer function is $Q = 1/(s^2 - 1)$, and (7) becomes

$$Y = (s + 1)Q + \frac{1}{s^2}\,Q = \frac{s + 1}{s^2 - 1} + \frac{1}{s^2(s^2 - 1)}.$$

Simplification and **partial fraction expansion** gives

$$Y = \frac{1}{s - 1} + \left(\frac{1}{s^2 - 1} - \frac{1}{s^2}\right).$$

Step 3. From this expression for Y and Table 6.1 we obtain the solution

$$y(t) = \mathcal{L}^{-1}(Y) = \mathcal{L}^{-1}\left\{\frac{1}{s - 1}\right\} + \mathcal{L}^{-1}\left\{\frac{1}{s^2 - 1}\right\} - \mathcal{L}^{-1}\left\{\frac{1}{s^2}\right\} = e^t + \sinh t - t.$$

The diagram in Fig. 115 summarizes our approach. ■

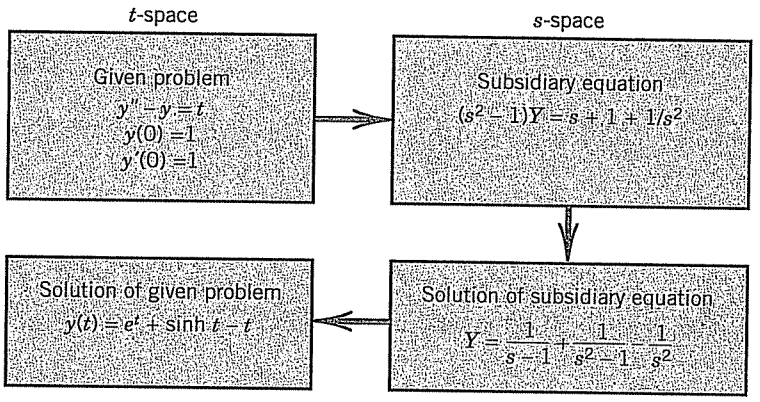


Fig. 115. Laplace transform method

EXAMPLE 5 Comparison with the Usual Method

Solve the initial value problem

$$y'' + y' + 9y = 0, \qquad y(0) = 0.16, \qquad y'(0) = 0.$$

Solution. From (1) and (2) we see that the subsidiary equation is

$$s^2Y - 0.16s + sY - 0.16 + 9Y = 0, \qquad \text{thus} \qquad (s^2 + s + 9)Y = 0.16(s + 1).$$

The solution is

$$Y = \frac{0.16(s + 1)}{s^2 + s + 9} = \frac{0.16(s + \frac{1}{2}) + 0.08}{(s + \frac{1}{2})^2 + \frac{35}{4}}.$$

Hence by the first shifting theorem and the formulas for cos and sin in Table 6.1 we obtain

$$y(t) = \mathcal{L}^{-1}(Y) = e^{-t/2}\left(0.16\cos\sqrt{\frac{35}{4}}\,t + \frac{0.08}{\frac{1}{2}\sqrt{35}}\sin\sqrt{\frac{35}{4}}\,t\right)$$

$$= e^{-0.5t}(0.16\cos 2.96t + 0.027\sin 2.96t).$$

This agrees with Example 2, Case (III) in Sec. 2.4. The work was less. ■

Advantages of the Laplace Method

1. *Solving a nonhomogeneous ODE does not require first solving the homogeneous ODE.* See Example 4.
2. *Initial values are automatically taken care of.* See Examples 4 and 5.
3. *Complicated inputs* $r(t)$ (right sides of linear ODEs) *can be handled very efficiently,* as we show in the next sections.

EXAMPLE 6 Shifted Data Problems

This means initial value problems with initial conditions given at some $t = t_0 > 0$ instead of $t = 0$. For such a problem set $t = \tilde{t} + t_0$, so that $t = t_0$ gives $\tilde{t} = 0$ and the Laplace transform can be applied. For instance, solve

$$y'' + y = 2t, \qquad y(\tfrac{1}{4}\pi) = \tfrac{1}{2}\pi, \qquad y'(\tfrac{1}{4}\pi) = 2 - \sqrt{2}.$$

Solution. We have $t_0 = \frac{1}{4}\pi$ and we set $t = \tilde{t} + \frac{1}{4}\pi$. Then the problem is

$$\tilde{y}'' + \tilde{y} = 2(\tilde{t} + \tfrac{1}{4}\pi), \qquad \tilde{y}(0) = \tfrac{1}{2}\pi, \qquad \tilde{y}'(0) = 2 - \sqrt{2}$$

where $\tilde{y}(\tilde{t}) = y(t)$. Using (2) and Table 6.1 and denoting the transform of $\tilde{y}$ by $\tilde{Y}$, we see that the subsidiary equation of the "shifted" initial value problem is

$$s^2\tilde{Y} - s \cdot \tfrac{1}{2}\pi - (2 - \sqrt{2}) + \tilde{Y} = \frac{2}{s^2} + \frac{\frac{1}{2}\pi}{s}, \qquad \text{thus} \qquad (s^2 + 1)\tilde{Y} = \frac{2}{s^2} + \frac{\frac{1}{2}\pi}{s} + \frac{1}{2}\pi s + 2 - \sqrt{2}.$$

Solving this algebraically for $\tilde{Y}$, we obtain

$$\tilde{Y} = \frac{2}{(s^2 + 1)s^2} + \frac{\frac{1}{2}\pi}{(s^2 + 1)s} + \frac{\frac{1}{2}\pi s}{s^2 + 1} + \frac{2 - \sqrt{2}}{s^2 + 1}.$$

The inverse of the first two terms can be seen from Example 3 (with $\omega = 1$), and the last two terms give cos and sin,

$$\tilde{y} = \mathcal{L}^{-1}(\tilde{Y}) = 2(\tilde{t} - \sin\tilde{t}) + \tfrac{1}{2}\pi(1 - \cos\tilde{t}) + \tfrac{1}{2}\pi\cos\tilde{t} + (2 - \sqrt{2})\sin\tilde{t}$$

$$= 2\tilde{t} + \tfrac{1}{2}\pi - \sqrt{2}\sin\tilde{t}.$$

Now $\tilde{t} = t - \frac{1}{4}\pi$, $\sin\tilde{t} = \dfrac{1}{\sqrt{2}}(\sin t - \cos t)$, so that the answer (the solution) is

$$y = 2t - \sin t + \cos t.$$

PROBLEM SET 6.2

1–8 OBTAINING TRANSFORMS BY DIFFERENTIATION

Using (1) or (2), find $\mathcal{L}(f)$ if $f(t)$ equals:

1. te^{kt}
2. $t\cos 5t$
3. $\sin^2 \omega t$
4. $\cos^2 \pi t$
5. $\sinh^2 at$
6. $\cosh^2 \frac{1}{2}t$
7. $t\sin \frac{1}{2}\pi t$
8. $\sin^4 t$ (Use Prob. 3.)

9. **(Derivation by different methods)** It is typical that various transforms can be obtained by several methods. Show this for Prob. 1. Show it for $\mathcal{L}(\cos^2 \frac{1}{2}t)$ (a) by expressing $\cos^2 \frac{1}{2}t$ in terms of $\cos t$, (b) by using Prob. 3.

10–24 INITIAL VALUE PROBLEMS

Solve the following initial value problems by the Laplace transform. (If necessary, use partial fraction expansion as in Example 4. Show all details.)

10. $y' + 4y = 0, \qquad y(0) = 2.8$
11. $y' + \frac{1}{2}y = 17\sin 2t, \qquad y(0) = -1$
12. $y'' - y' - 6y = 0, \qquad y(0) = 6, \quad y'(0) = 13$

13. $y'' - \frac{1}{4}y = 0, \quad y(0) = 4, \quad y'(0) = 0$
14. $y'' - 4y' + 4y = 0, \quad y(0) = 2.1, \quad y'(0) = 3.9$
15. $y'' + 2y' + 2y = 0, \quad y(0) = 1, \quad y'(0) = -3$
16. $y'' + ky' - 2k^2y = 0, \quad y(0) = 2, \quad y'(0) = 2k$
17. $y'' + 7y' + 12y = 21e^{3t}, \quad y(0) = 3.5, \quad y'(0) = -10$
18. $y'' + 9y = 10e^{-t}, \quad y(0) = 0, \quad y'(0) = 0$
19. $y'' + 3y' + 2.25y = 9t^3 + 64, \quad y(0) = 1, \quad y'(0) = 31.5$
20. $y'' - 6y' + 5y = 29 \cos 2t, \quad y(0) = 3.2, \quad y'(0) = 6.2$
21. **(Shifted data)** $y' - 6y = 0, \quad y(2) = 4$
22. $y'' - 2y' - 3y = 0, \quad y(1) = -3, \quad y'(1) = -17$
23. $y'' + 3y' - 4y = 6e^{2t-2}, \quad y(1) = 4, \quad y'(1) = 5$
24. $y'' + 2y' + 5y = 50t - 150, \quad y(3) = -4, \quad y'(3) = 14$

25. **PROJECT. Comments on Sec. 6.2.** **(a)** Give reasons why Theorems 1 and 2 are more important than Theorem 3.

(b) Extend Theorem 1 by showing that if $f(t)$ is continuous, except for an ordinary discontinuity (finite jump) at some $t = a$ (> 0), the other conditions remaining as in Theorem 1, then (see Fig. 116)

(1*) $\mathcal{L}(f') = s\mathcal{L}(f) - f(0) - [f(a + 0) - f(a - 0)]e^{-as}$.

(c) Verify (1*) for $f(t) = e^{-t}$ if $0 < t < 1$ and 0 if $t > 1$.

(d) Verify (1*) for two more complicated functions of your choice.

(e) Compare the Laplace transform of solving ODEs with the method in Chap. 2. Give examples of your own to illustrate the advantages of the present method (to the extent we have seen them so far).

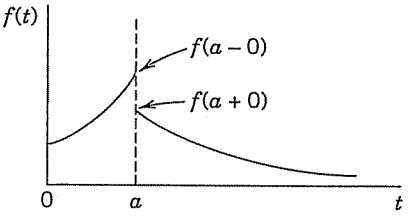


Fig. 116. Formula (1*)

26. **PROJECT. Further Results by Differentiation.** Proceeding as in Example 1, obtain

$$\text{(a)} \quad \mathcal{L}(t \cos \omega t) = \frac{s^2 - \omega^2}{(s^2 + \omega^2)^2}$$

and from this and Example 1: (b) formula 21, (c) 22, (d) 23 in Sec. 6.9,

$$\text{(e)} \quad \mathcal{L}(t \cosh at) = \frac{s^2 + a^2}{(s^2 - a^2)^2},$$

$$\text{(f)} \quad \mathcal{L}(t \sinh at) = \frac{2as}{(s^2 - a^2)^2}.$$

27–34 OBTAINING TRANSFORMS BY INTEGRATION

Using Theorem 3, find $f(t)$ if $\mathcal{L}(f)$ equals:

27. $\dfrac{1}{s^2 + s/2}$
28. $\dfrac{10}{s^3 - \pi s^2}$
29. $\dfrac{1}{s^3 - ks^2}$
30. $\dfrac{1}{s^4 + s^2}$
31. $\dfrac{5}{s^3 - 5s}$
32. $\dfrac{2}{s^3 + 9s}$
33. $\dfrac{1}{s^4 - 4s^2}$
34. $\dfrac{1}{s^4 + \pi^2 s^2}$

35. **(Partial fractions)** Solve Probs. 27, 29, and 31 by using partial fractions.

6.3 Unit Step Function. t-Shifting

This section and the next one are extremely important because we shall now reach the point where the Laplace transform method shows its real power in applications and its superiority over the classical approach of Chap. 2. The reason is that we shall introduce two auxiliary functions, the *unit step function* or *Heaviside function* $u(t - a)$ (below) and *Dirac's delta* $\delta(t - a)$ (in Sec. 6.4). These functions are suitable for solving ODEs with complicated right sides of considerable engineering interest, such as single waves, inputs (driving forces) that are discontinuous or act for some time only, periodic inputs more general than just cosine and sine, or impulsive forces acting for an instant (hammerblows, for example).

Unit Step Function (Heaviside Function) $u(t-a)$

The **unit step function** or **Heaviside function** $u(t-a)$ is 0 for $t<a$, has a jump of size 1 at $t=a$ (where we can leave it undefined), and is 1 for $t>a$, in a formula:

(1)
$$u(t-a)=\begin{cases}0 & \text{if } t<a\\ 1 & \text{if } t>a\end{cases} \qquad (a \geqq 0).$$

Figure 117 shows the special case $u(t)$, which has its jump at zero, and Fig. 118 the general case $u(t-a)$ for an arbitrary positive a. (For Heaviside see Sec. 6.1.)

The transform of $u(t-a)$ follows directly from the defining integral in Sec. 6.1,

$$\mathcal{L}\{u(t-a)\}=\int_0^\infty e^{-st}u(t-a)\,dt=\int_a^\infty e^{-st}\cdot 1\,dt=-\frac{e^{-st}}{s}\Bigg|_{t=a}^{\infty};$$

here the integration begins at $t=a$ ($\geqq 0$) because $u(t-a)$ is 0 for $t<a$. Hence

(2)
$$\mathcal{L}\{u(t-a)\}=\frac{e^{-as}}{s} \qquad (s>0).$$

The unit step function is a typical "engineering function" made to measure for engineering applications, which often involve functions (mechanical or electrical driving forces) that are either "off" or "on." Multiplying functions $f(t)$ with $u(t-a)$, we can produce all sorts of effects. The simple basic idea is illustrated in Figs. 119 and 120. In Fig. 119 the given function is shown in (A). In (B) it is switched off between $t=0$ and $t=2$ (because $u(t-2)=0$ when $t<2$) and is switched on beginning at $t=2$. In (C) it is shifted to the right by 2 units, say, for instance, by 2 secs, so that it begins 2 secs later in the same fashion as before. More generally we have the following.

Let $f(t)=0$ for all negative t. Then $f(t-a)u(t-a)$ with $a>0$ is $f(t)$ ***shifted*** *(translated) to the right by the amount a.*

Figure 120 shows the effect of many unit step functions, three of them in (A) and infinitely many in (B) when continued periodically to the right; this is the effect of a rectifier that clips off the negative half-waves of a sinuosidal voltage. CAUTION! Make sure that you fully understand these figures, in particular the difference between parts (B) and (C) of Figure 119. Figure 119(C) will be applied next.

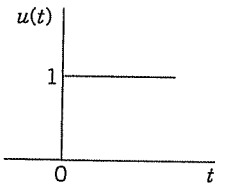


Fig. 117. Unit step function $u(t)$

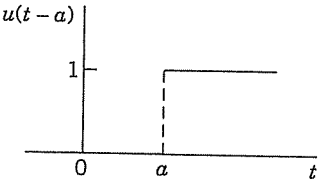
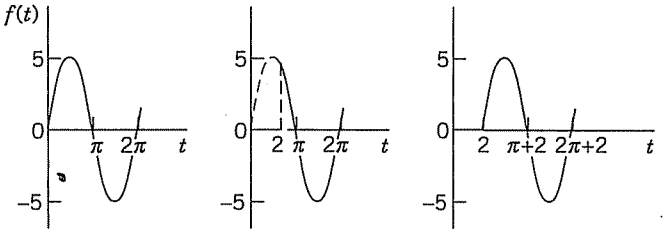
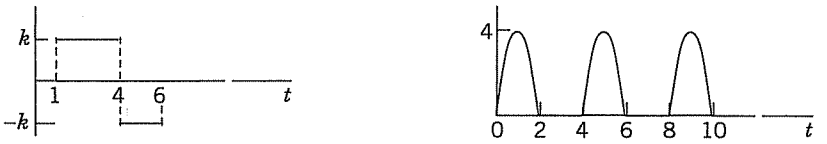


Fig. 118. Unit step function $u(t-a)$



(A) $f(t) = 5 \sin t$ (B) $f(t)u(t-2)$ (C) $f(t-2)u(t-2)$

Fig. 119. Effects of the unit step function: (A) Given function. (B) Switching off and on. (C) Shift.



(A) $k[u(t-1) - 2u(t-4) + u(t-6)]$ (B) $4 \sin(\frac{1}{2}\pi t)[u(t) - u(t-2) + u(t-4) - + \cdots]$

Fig. 120. Use of many unit step functions.

Time Shifting (t-Shifting): Replacing t by $t - a$ in $f(t)$

The first shifting theorem ("s-shifting") in Sec. 6.1 concerned transforms $F(s) = \mathcal{L}\{f(t)\}$ and $F(s - a) = \mathcal{L}\{e^{at}f(t)\}$. The second shifting theorem will concern functions $f(t)$ and $f(t - a)$. Unit step functions are just tools, and the theorem will be needed to apply them in connection with any other functions.

THEOREM 1

Second Shifting Theorem; Time Shifting

If $f(t)$ has the transform $F(s)$, then the **"shifted function"**

$$\tilde{f}(t) = f(t-a)u(t-a) = \begin{cases} 0 & \text{if } t < a \\ f(t-a) & \text{if } t > a \end{cases} \tag{3}$$

has the transform $e^{-as}F(s)$. That is, if $\mathcal{L}\{f(t)\} = F(s)$, then

$$\mathcal{L}\{f(t-a)u(t-a)\} = e^{-as}F(s). \tag{4}$$

Or, if we take the inverse on both sides, we can write

$$f(t-a)u(t-a) = \mathcal{L}^{-1}\{e^{-as}F(s)\}. \tag{4*}$$

Practically speaking, if we know $F(s)$, we can obtain the transform of (3) by multiplying $F(s)$ by e^{-as}. In Fig. 119, the transform of $5 \sin t$ is $F(s) = 5/(s^2 + 1)$, hence the shifted function $5 \sin(t - 2)\, u(t - 2)$ shown in Fig. 119(C) has the transform

$$e^{-2s}F(s) = 5e^{-2s}/(s^2 + 1).$$

PROOF We prove Theorem 1. In (4) on the right we use the definition of the Laplace transform, writing τ for t (to have t available later). Then, taking e^{-as} inside the integral, we have

$$e^{-as}F(s) = e^{-as}\int_0^\infty e^{-s\tau}f(\tau)\,d\tau = \int_0^\infty e^{-s(\tau+a)}f(\tau)\,d\tau.$$

Substituting $\tau + a = t$, thus $\tau = t - a$, $d\tau = dt$, in the integral (CAUTION, the lower limit changes!), we obtain

$$e^{-as}F(s) = \int_a^\infty e^{-st}f(t-a)\,dt.$$

To make the right side into a Laplace transform, we must have an integral from 0 to ∞, not from a to ∞. But this is easy. We multiply the integrand by $u(t-a)$. Then for t from 0 to a the integrand is 0, and we can write, with $\tilde{f}$ as in (3),

$$e^{-as}F(s) = \int_0^\infty e^{-st}f(t-a)u(t-a)\,dt = \int_0^\infty e^{-st}\tilde{f}(t)\,dt.$$

(Do you now see why $u(t-a)$ appears?) This integral is the left side of (4), the Laplace transform of $\tilde{f}(t)$ in (3). This completes the proof. ∎

EXAMPLE 1 **Application of Theorem 1. Use of Unit Step Functions**

Write the following function using unit step functions and find its transform.

$$f(t) = \begin{cases} 2 & \text{if } 0 < t < 1 \\ \frac{1}{2}t^2 & \text{if } 1 < t < \frac{1}{2}\pi \\ \cos t & \text{if } \quad t > \frac{1}{2}\pi. \end{cases} \qquad \text{(Fig. 121)}$$

Solution. ***Step 1.*** In terms of unit step functions,

$$f(t) = 2(1 - u(t-1)) + \tfrac{1}{2}t^2(u(t-1) - u(t-\tfrac{1}{2}\pi)) + (\cos t)u(t-\tfrac{1}{2}\pi).$$

Indeed, $2(1-u(t-1))$ gives $f(t)$ for $0 < t < 1$, and so on.

Step 2. To apply Theorem 1, we must write each term in $f(t)$ in the form $f(t-a)u(t-a)$. Thus, $2(1-u(t-1))$ remains as it is and gives the transform $2(1-e^{-s})/s$. Then

$$\mathscr{L}\left\{\frac{1}{2}t^2u(t-1)\right\} = \mathscr{L}\left(\frac{1}{2}(t-1)^2 + (t-1) + \frac{1}{2}\right)u(t-1)\Big\} = \left(\frac{1}{s^3} + \frac{1}{s^2} + \frac{1}{2s}\right)e^{-s}$$

$$\mathscr{L}\left\{\frac{1}{2}t^2u\left(t-\frac{1}{2}\pi\right)\right\} = \mathscr{L}\left\{\frac{1}{2}\left(t-\frac{1}{2}\pi\right)^2 + \frac{\pi}{2}\left(t-\frac{1}{2}\pi\right) + \frac{\pi^2}{8}\right)u\left(t-\frac{1}{2}\pi\right)\Big\}$$

$$= \left(\frac{1}{s^3} + \frac{\pi}{2s^2} + \frac{\pi^2}{8s}\right)e^{-\pi s/2}$$

$$\mathscr{L}\left\{(\cos t)u\left(t-\frac{1}{2}\pi\right)\right\} = \mathscr{L}\left\{-\left(\sin\left(t-\frac{1}{2}\pi\right)\right)u\left(t-\frac{1}{2}\pi\right)\right\} = -\frac{1}{s^2+1}e^{-\pi s/2}.$$

Together,

$$\mathscr{L}(f) = \frac{2}{s} - \frac{2}{s}e^{-s} + \left(\frac{1}{s^3} + \frac{1}{s^2} + \frac{1}{2s}\right)e^{-s} - \left(\frac{1}{s^3} + \frac{\pi}{2s^2} + \frac{\pi^2}{8s}\right)e^{-\pi s/2} - \frac{1}{s^2+1}e^{-\pi s/2}.$$

If the conversion of $f(t)$ to $f(t-a)$ is inconvenient, replace it by

(4)** $$\mathcal{L}\{f(t)u(t-a)\} = e^{-as}\mathcal{L}\{f(t+a)\}.$$

(4**) follows from (4) by writing $f(t-a) = g(t)$, hence $f(t) = g(t+a)$ and then again writing f for g. Thus,

$$\mathcal{L}\left\{\frac{1}{2}t^2u(t-1)\right\} = e^{-s}\mathcal{L}\left\{\frac{1}{2}(t+1)^2\right\} = e^{-s}\mathcal{L}\left\{\frac{1}{2}t^2 + t + \frac{1}{2}\right\} = e^{-s}\left(\frac{1}{s^3} + \frac{1}{s^2} + \frac{1}{2s}\right)$$

as before. Similarly for $\mathcal{L}\{\frac{1}{2}t^2u(t-\frac{1}{2}\pi)\}$. Finally, by (4**),

$$\mathcal{L}\left\{\cos t\, u\left(t - \frac{1}{2}\pi\right)\right\} = e^{-\pi s/2}\mathcal{L}\left\{\cos\left(t + \frac{1}{2}\pi\right)\right\} = e^{-\pi s/2}\mathcal{L}\{-\sin t\} = -e^{-\pi s/2}\frac{1}{s^2+1}. \quad \blacksquare$$

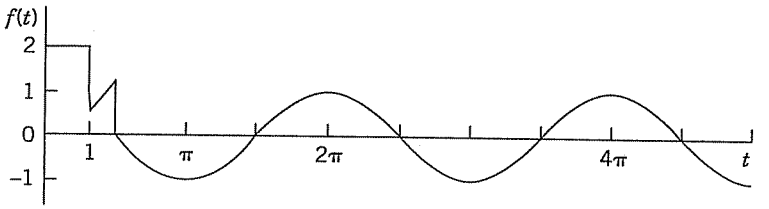


Fig. 121. $f(t)$ in Example 1

EXAMPLE 2 **Application of Both Shifting Theorems. Inverse Transform**

Find the inverse transform $f(t)$ of

$$F(s) = \frac{e^{-s}}{s^2+\pi^2} + \frac{e^{-2s}}{s^2+\pi^2} + \frac{e^{-3s}}{(s+2)^2}.$$

Solution. Without the exponential functions in the numerator the three terms of $F(s)$ would have the inverses $(\sin \pi t)/\pi$, $(\sin \pi t)/\pi$, and te^{-2t} because $1/s^2$ has the inverse t, so that $1/(s+2)^2$ has the inverse te^{-2t} by the first shifting theorem in Sec. 6.1. Hence by the second shifting theorem (t-shifting),

$$f(t) = \frac{1}{\pi}\sin(\pi(t-1))\,u(t-1) + \frac{1}{\pi}\sin(\pi(t-2))\,u(t-2) + (t-3)e^{-2(t-3)}\,u(t-3).$$

Now $\sin(\pi t - \pi) = -\sin \pi t$ and $\sin(\pi t - 2\pi) = \sin \pi t$, so that the second and third terms cancel each other when $t > 2$. Hence we obtain $f(t) = 0$ if $0 < t < 1$, $-(\sin \pi t)/\pi$ if $1 < t < 2$, 0 if $2 < t < 3$, and $(t-3)e^{-2(t-3)}$ if $t > 3$. See Fig. 122. $\blacksquare$

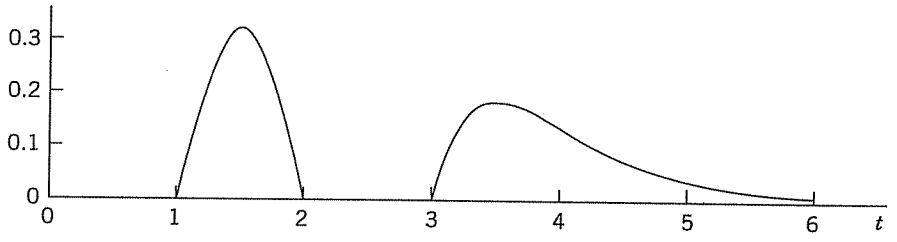


Fig. 122. $f(t)$ in Example 2

EXAMPLE 3 **Response of an *RC*-Circuit to a Single Rectangular Wave**

Find the current $i(t)$ in the *RC*-circuit in Fig. 123 if a single rectangular wave with voltage V_0 is applied. The circuit is assumed to be quiescent before the wave is applied.

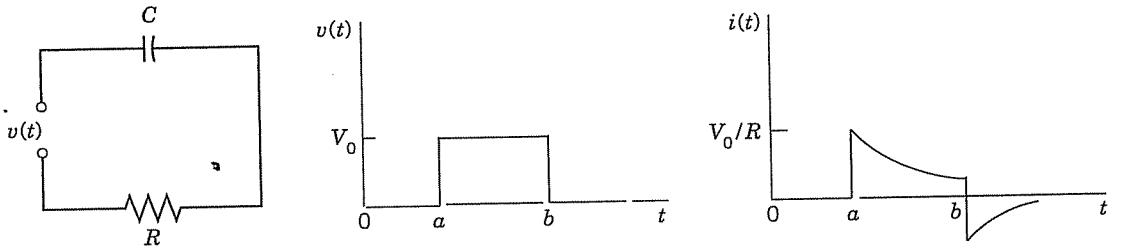


Fig. 123. *RC*-circuit, electromotive force $v(t)$, and current in Example 3

Solution. The input is $V_0[u(t-a) - u(t-b)]$. Hence the circuit is modeled by the integro-differential equation (see Sec. 2.9 and Fig. 123)

$$Ri(t) + \frac{q(t)}{C} = Ri(t) + \frac{1}{C}\int_0^t i(\tau)\,d\tau = v(t) = V_0[u(t-a) - u(t-b)].$$

Using Theorem 3 in Sec. 6.2 and formula (1) in this section, we obtain the subsidiary equation

$$RI(s) + \frac{I(s)}{sC} = \frac{V_0}{s}\,[e^{-as} - e^{-bs}].$$

Solving this equation algebraically for $I(s)$, we get

$$I(s) = F(s)(e^{-as} - e^{-bs}) \qquad \text{where} \qquad F(s) = \frac{V_0/R}{s + 1/(RC)} \qquad \text{and} \qquad \mathcal{L}^{-1}(F) = \frac{V_0}{R}\,e^{-t/(RC)},$$

the last expression being obtained from Table 6.1 in Sec. 6.1. Hence Theorem 1 yields the solution (Fig. 123)

$$i(t) = \mathcal{L}^{-1}(I) = \mathcal{L}^{-1}\{e^{-as}F(s)\} - \mathcal{L}^{-1}\{e^{-bs}F(s)\} = \frac{V_0}{R}\,[e^{-(t-a)/(RC)}u(t-a) - e^{-(t-b)/(RC)}u(t-b)];$$

that is, $i(t) = 0$ if $t < a$, and

$$i(t) = \begin{cases} K_1 e^{-t/(RC)} & \text{if } a < t < b \\ (K_1 - K_2)e^{-t/(RC)} & \text{if } a > b \end{cases}$$

where $K_1 = V_0 e^{a/(RC)}/R$ and $K_2 = V_0 e^{b/(RC)}/R$.

EXAMPLE 4 **Response of an *RLC*-Circuit to a Sinusoidal Input Acting Over a Time Interval**

Find the response (the current) of the *RLC*-circuit in Fig. 124, where $E(t)$ is sinusoidal, acting for a short time interval only, say,

$$E(t) = 100 \sin 400t \quad \text{if } 0 < t < 2\pi \qquad \text{and} \qquad E(t) = 0 \text{ if } t > 2\pi$$

and current and charge are initially zero.

Solution. The electromotive force $E(t)$ can be represented by $(100 \sin 400t)(1 - u(t - 2\pi))$. Hence the model for the current $i(t)$ in the circuit is the integro-differential equation (see Sec. 2.9)

$$0.1i' + 11i + 100\int_0^t i(\tau)\,d\tau = (100 \sin 400t)(1 - u(t - 2\pi)), \qquad i(0) = 0, \quad i'(0) = 0.$$

From Theorems 2 and 3 in Sec. 6.2 we obtain the subsidiary equation for $I(s) = \mathcal{L}(i)$

$$0.1sI + 11I + 100\,\frac{I}{s} = \frac{100 \cdot 400s}{s^2 + 400^2}\left(\frac{1}{s} - \frac{e^{-2\pi s}}{s}\right).$$

Solving it algebraically and noting that $s^2 + 110s + 1000 = (s + 10)(s + 100)$, we obtain

$$I(s) = \frac{1000 \cdot 400}{(s + 10)(s + 100)} \left(\frac{s}{s^2 + 400^2} - \frac{se^{-2\pi s}}{s^2 + 400^2} \right).$$

For the first term in the parentheses $(\cdots)$ times the factor in front of them we use the partial fraction expansion

$$\frac{400\,000s}{(s + 10)(s + 100)(s^2 + 400^2)} = \frac{A}{s + 10} + \frac{B}{s + 100} + \frac{Ds + K}{s^2 + 400^2}.$$

Now determine A, B, D, K by your favorite method or by a CAS or as follows. Multiplication by the common denominator gives

$$400\,000s = A(s + 100)(s^2 + 400^2) + B(s + 10)(s^2 + 400^2) + (Ds + K)(s + 10)(s + 100).$$

We set $s = -10$ and -100 and then equate the sums of the s^3 and s^2 terms to zero, obtaining (all values rounded)

$(s = -10)$	$-4\,000\,000 = 90(10^2 + 400^2)A,$	$A = -0.27760$
$(s = -100)$	$-40\,000\,000 = -90(100^2 + 400^2)B,$	$B = 2.6144$
(s^3-terms)	$0 = A + B + D,$	$D = -2.3368$
(s^2-terms)	$0 = 100A + 10B + 110D + K,$	$K = 258.66.$

Since $K = 258.66 = 0.6467 \cdot 400$, we thus obtain for the first term I_1 in $I = I_1 - I_2$

$$I_1 = -\frac{0.2776}{s + 10} + \frac{2.6144}{s + 100} - \frac{2.3368s}{s^2 + 400^2} + \frac{0.6467 \cdot 400}{s^2 + 400^2}.$$

From Table 6.1 in Sec. 6.1 we see that its inverse is

$$i_1(t) = -0.2776e^{-10t} + 2.6144e^{-100t} - 2.3368 \cos 400t + 0.6467 \sin 400t.$$

This is the current $i(t)$ when $0 < t < 2\pi$. It agrees for $0 < t < 2\pi$ with that in Example 1 of Sec. 2.9 (except for notation), which concerned the same *RLC*-circuit. Its graph in Fig. 62 in Sec. 2.9 shows that the exponential terms decrease very rapidly. Note that the present amount of work was substantially less.

The second term I_1 of I differs from the first term by the factor $e^{-2\pi s}$. Since $\cos 400(t - 2\pi) = \cos 400t$ and $\sin 400(t - 2\pi) = \sin 400t$, the second shifting theorem (Theorem 1) gives the inverse $i_2(t) = 0$ if $0 < t < 2\pi$, and for $> 2\pi$ it gives

$$i_2(t) = -0.2776e^{-10(t-2\pi)} + 2.6144e^{-100(t-2\pi)} - 2.3368 \cos 400t + 0.6467 \sin 400t.$$

Hence in $i(t)$ the cosine and sine terms cancel, and the current for $t > 2\pi$ is

$$i(t) = -0.2776(e^{-10t} - e^{-10(t-2\pi)}) + 2.6144(e^{-100t} - e^{-100(t-2\pi)}).$$

It goes to zero very rapidly, practically within 0.5 sec. ▩

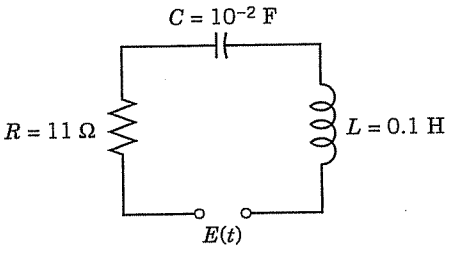


Fig. 124. *RLC*-circuit in Example 4

PROBLEM SET 6.3

1. **WRITING PROJECT. Shifting Theorem.** Explain and compare the different roles of the two shifting theorems, using your own formulations and examples.

2–13 UNIT STEP FUNCTION AND SECOND SHIFTING THEOREM

Sketch or graph the given function (which is assumed to be zero outside the given interval). Represent it using unit step functions. Find its transform. Show the details of your work.

2. $t \ (0 < t < 1)$
3. $e^t \ (0 < t < 2)$
4. $\sin 3t \ (0 < t < \pi)$
5. $t^2 \ (1 < t < 2)$
6. $t^2 \ (t > 3)$
7. $\cos \pi t \ (1 < t < 4)$
8. $1 - e^{-t} \ (0 < t < \pi)$
9. $t \ (5 < t < 10)$
10. $\sin \omega t \ (t > 6\pi/\omega)$
11. $20 \cos \pi t \ (3 < t < 6)$
12. $\sinh t \ (0 < t < 2)$
13. $e^{\pi t} \ (2 < t < 4)$

14–22 INVERSE TRANSFORMS BY THE SECOND SHIFTING THEOREM

Find and sketch or graph $f(t)$ if $\mathcal{L}(f)$ equals:

14. $se^{-s}/(s^2 + \omega^2)$
15. e^{-4s}/s^2
16. $s^{-2} - (s^{-2} + s^{-1})e^{-s}$
17. $(e^{-2\pi s} - e^{-8\pi s})/(s^2 + 1)$
18. $e^{-\pi s}/(s^2 + 2s + 2)$
19. e^{-2s}/s^5
20. $(1 - e^{-s+k})/(s - k)$
21. $se^{-3s}/(s^2 - 4)$
22. $2.5(e^{-3.8s} - e^{-2.6s})/s$

23–34 INITIAL VALUE PROBLEMS, SOME WITH DISCONTINUOUS INPUTS

Using the Laplace transform and showing the details, solve:

23. $y'' + 2y' + 2y = 0, \quad y(0) = 0, \quad y'(0) = 1$
24. $9y'' - 6y' + y = 0, \quad y(0) = 3, \quad y'(0) = 1$
25. $y'' + 4y' + 13y = 145 \cos 2t, \quad y(0) = 10, \quad y'(0) = 14$
26. $y'' + 10y' + 24y = 144t^2, \quad y(0) = \frac{19}{12}, \quad y'(0) = -5$
27. $y'' + 9y = r(t)$, $r(t) = 8 \sin t$ if $0 < t < \pi$ and 0 if $t > \pi$; $y(0) = 0, y'(0) = 4$
28. $y'' + 3y' + 2y = r(t)$, $r(t) = 1$ if $0 < t < 1$ and 0 if $t > 1$; $y(0) = 0, y'(0) = 0$
29. $y'' + y = r(t), \quad r(t) = t$ if $0 < t < 1$ and 0 if $t > 1$; $y(0) = y'(0) = 0$
30. $y'' - 16y = r(t), \quad r(t) = 48e^{2t}$ if $0 < t < 4$ and 0 if $t > 4$; $y(0) = 3, \quad y'(0) = -4$
31. $y'' + y' - 2y = r(t)$, $r(t) = 3 \sin t - \cos t$ if $0 < t < 2\pi$ and $3 \sin 2t - \cos 2t$ if $t > 2\pi$; $y(0) = 1, \quad y'(0) = 0$
32. $y'' + 8y' + 15y = r(t)$, $r(t) = 35e^{2t}$ if $0 < t < 2$ and 0 if $t > 2$; $y(0) = 3$, $y'(0) = -8$
33. **(Shifted data)** $y'' + 4y = 8t^2$ if $0 < t < 5$ and 0 if $t > 5$; $y(1) = 1 + \cos 2$, $y'(1) = 4 - 2 \sin 2$
34. $y'' + 2y' + 5y = 10 \sin t$ if $0 < t < 2\pi$ and 0 if $t > 2\pi$; $y(\pi) = 1$, $y'(\pi) = 2e^{-\pi} - 2$

MODELS OF ELECTRIC CIRCUITS

35. **(Discharge)** Using the Laplace transform, find the charge $q(t)$ on the capacitor of capacitance C in Fig. 125 if the capacitor is charged so that its potential is V_0 and the switch is closed at $t = 0$.

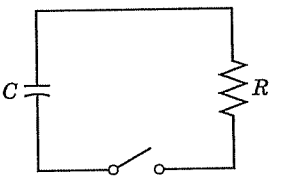


Fig. 125. Problem 35

36–38 RC-CIRCUIT

Using the Laplace transform and showing the details, find the current $i(t)$ in the circuit in Fig. 126 with $R = 10\ \Omega$ and $C = 10^{-2}$ F, where the current at $t = 0$ is assumed to be zero, and:

36. $v(t) = 100$ V if $0.5 < t < 0.6$ and 0 otherwise. Why does $i(t)$ have jumps?
37. $v = 0$ if $t < 2$ and $100(t - 2)$ V if $t > 2$
38. $v = 0$ if $t < 4$ and $14 \cdot 10^6 e^{-3t}$ V if $t > 4$

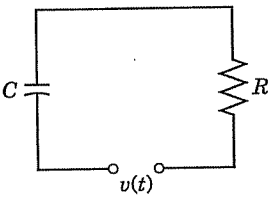


Fig. 126. Problems 36–38

39–41 RL-CIRCUIT

Using the Laplace transform and showing the details, find the current $i(t)$ in the circuit in Fig. 127, assuming $i(0) = 0$ and:

39. $R = 10\ \Omega$, $L = 0.5$ H, $v = 200t$ V if $0 < t < 2$ and 0 if $t > 2$

40. $R = 1\ \text{k}\Omega$ $(= 1000\ \Omega)$, $L = 1$ H, $v = 0$ if $0 < t < \pi$, and $40 \sin t$ V if $t > \pi$

41. $R = 25\ \Omega$, $L = 0.1$ H, $v = 490e^{-5t}$ V if $0 < t < 1$ and 0 if $t > 1$

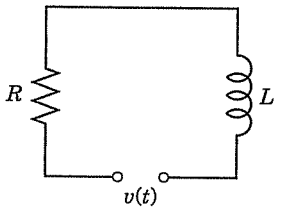


Fig. 127. Problems 39–41

42–44 LC-CIRCUIT

Using the Laplace transform and showing the details, find the current $i(t)$ in the circuit in Fig. 128, assuming zero initial current and charge on the capacitor and:

42. $L = 1$ H, $C = 0.25$ F, $v = 200(t - \frac{1}{3}t^3)$ V if $0 < t < 1$ and 0 if $t > 1$

43. $L = 1$ H, $C = 10^{-2}$ F, $v = -9900 \cos t$ V if $\pi < t < 3\pi$ and 0 otherwise

44. $L = 0.5$ H, $C = 0.05$ F, $v = 78 \sin t$ V if $0 < t < \pi$ and 0 if $t > \pi$

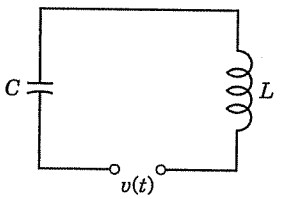


Fig. 128. Problems 42–44

45–47 RLC-CIRCUIT

Using the Laplace transform and showing the details, find the current $i(t)$ in the circuit in Fig. 129, assuming zero initial current and charge and:

45. $R = 2\ \Omega$, $L = 1$ H, $C = 0.5$ F, $v(t) = 1$ kV if $0 < t < 2$ and 0 if $t > 2$

46. $R = 4\ \Omega$, $L = 1$ H, $C = 0.05$ F, $v = 34e^{-t}$ V if $0 < t < 4$ and 0 if $t > 4$

47. $R = 2\ \Omega$, $L = 1$ H, $C = 0.1$ F, $v = 255 \sin t$ V if $0 < t < 2\pi$ and 0 if $t > 2\pi$

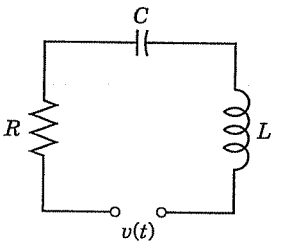


Fig. 129. Problems 45–47

6.4 Short Impulses. Dirac's Delta Function. Partial Fractions

Phenomena of an impulsive nature, such as the action of forces or voltages over short intervals of time, arise in various applications, for instance, if a mechanical system is hit by a hammerblow, an airplane makes a "hard" landing, a ship is hit by a single high wave, or we hit a tennisball by a racket, and so on. Our goal is to show how such problems are modeled by "Dirac's delta function" and can be solved very efficiently by the Laplace transform.

To model situations of that type, we consider the function

$$(1) \qquad f_k(t-a) = \begin{cases} 1/k & \text{if } a \leqq t \leqq a+k \\ 0 & \text{otherwise} \end{cases} \qquad \text{(Fig. 130)}$$

(and later its limit as $k \to 0$). This function represents, for instance, a force of magnitude $1/k$ acting from $t = a$ to $t = a + k$, where k is positive and small. In mechanics, the integral of a force acting over a time interval $a \leqq t \leqq a + k$ is called the **impulse** of the

force; similarly for electromotive forces $E(t)$ acting on circuits. Since the blue rectangle in Fig. 130 has area 1, the impulse of f_k in (1) is

$$(2) \qquad I_k = \int_0^\infty f_k(t-a)\,dt = \int_a^{a+k} \frac{1}{k}\,dt = 1.$$

To find out what will happen if k becomes smaller and smaller, we take the limit of f_k as $k \to 0$ $(k > 0)$. This limit is denoted by $\delta(t-a)$, that is,

$$\delta(t-a) = \lim_{k\to 0} f_k(t-a).$$

$\delta(t-a)$ is called the **Dirac delta function**[2] or the **unit impulse function.**

$\delta(t-a)$ is not a function in the ordinary sense as used in calculus, but a so-called *generalized function.*[2] To see this, we note that the impulse I_k of f_k is 1, so that from (1) and (2) by taking the limit as $k \to 0$ we obtain

$$(3) \qquad \delta(t-a) = \begin{cases} \infty & \text{if } t = a \\ 0 & \text{otherwise} \end{cases} \qquad \text{and} \qquad \int_0^\infty \delta(t-a)\,dt = 1,$$

but from calculus we know that a function which is everywhere 0 except at a single point must have the integral equal to 0. Nevertheless, in impulse problems it is convenient to operate on $\delta(t-a)$ as though it were an ordinary function. In particular, for a *continuous* function $g(t)$ one uses the property [often called the **sifting property** of $\delta(t-a)$, not to be confused with *shifting*]

$$(4) \qquad \int_0^\infty g(t)\,\delta(t-a)\,dt = g(a)$$

which is plausible by (2).

To obtain the Laplace transform of $\delta(t-a)$, we write

$$f_k(t-a) = \frac{1}{k}\,[u(t-a) - u(t-(a+k))]$$

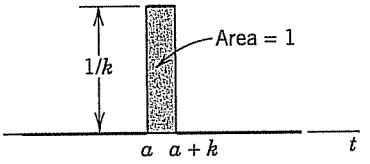


Fig. 130. The function $f_k(t-a)$ in (1)

[2]PAUL DIRAC (1902–1984), English physicist, was awarded the Nobel Prize [jointly with the Austrian ERWIN SCHRÖDINGER (1887–1961)] in 1933 for his work in quantum mechanics.

Generalized functions are also called **distributions.** Their theory was created in 1936 by the Russian mathematician SERGEI L'VOVICH SOBOLEV (1908–1989), and in 1945, under wider aspects, by the French mathematician LAURENT SCHWARTZ (1915–2002).

and take the transform [see (2)]

$$\mathcal{L}\{f_k(t-a)\} = \frac{1}{ks}\left[e^{-as} - e^{-(a+k)s}\right] = e^{-as}\,\frac{1-e^{-ks}}{ks}.$$

We now take the limit as $k \to 0$. By l'Hôpital's rule the quotient on the right has the limit 1 (differentiate the numerator and the denominator separately with respect to k, obtaining se^{-ks} and s, respectively, and use $se^{-ks}/s \to 1$ as $k \to 0$). Hence the right side has the limit e^{-as}. This suggests defining the transform of $\delta(t-a)$ by this limit, that is,

(5) $$\mathcal{L}\{\delta(t-a)\} = e^{-as}.$$

The unit step and unit impulse functions can now be used on the right side of ODEs modeling mechanical or electrical systems, as we illustrate next.

EXAMPLE 1 **Mass–Spring System Under a Square Wave**

Determine the response of the damped mass–spring system (see Sec. 2.8) under a square wave, modeled by (see Fig. 131)

$$y'' + 3y' + 2y = r(t) = u(t-1) - u(t-2), \qquad y(0) = 0, \qquad y'(0) = 0.$$

Solution. From (1) and (2) in Sec. 6.2 and (2) and (4) in this section we obtain the subsidiary equation

$$s^2Y + 3sY + 2Y = \frac{1}{s}\,(e^{-s} - e^{-2s}). \qquad \text{Solution} \qquad Y(s) = \frac{1}{s(s^2+3s+2)}\,(e^{-s} - e^{-2s}).$$

Using the notation $F(s)$ and partial fractions, we obtain

$$F(s) = \frac{1}{s(s^2+3s+2)} = \frac{1}{s(s+1)(s+2)} = \frac{1/2}{s} - \frac{1}{s+1} + \frac{1/2}{s+2}.$$

From Table 6.1 in Sec. 6.1, we see that the inverse is

$$f(t) = \mathcal{L}^{-1}(F) = \tfrac{1}{2} - e^{-t} + \tfrac{1}{2}e^{-2t}.$$

Therefore, by Theorem 1 in Sec. 6.3 (t-shifting) we obtain the square-wave response shown in Fig. 131,

$$\begin{aligned} y &= \mathcal{L}^{-1}(F(s)e^{-s} - F(s)e^{-2s}) \\ &= f(t-1)u(t-1) - f(t-2)u(t-2) \\ &= \begin{cases} 0 & (0 < t < 1) \\ \frac{1}{2} - e^{-(t-1)} + \frac{1}{2}e^{-2(t-1)} & (1 < t < 2) \\ -e^{-(t-1)} + e^{-(t-2)} + \frac{1}{2}e^{-2(t-1)} - \frac{1}{2}e^{-2(t-2)} & (t > 2). \end{cases} \end{aligned}$$

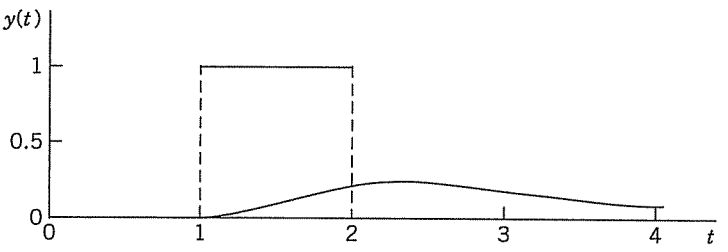


Fig. 131. Square wave and response in Example 1

EXAMPLE 2 **Hammerblow Response of a Mass–Spring System**

Find the response of the system in Example 1 with the square wave replaced by a unit impulse at time $t = 1$.

Solution. We now have the ODE and the subsidiary equation

$$y'' + 3y' + 2y = \delta(t - 1), \qquad \text{and} \qquad (s^2 + 3s + 2)Y = e^{-s}.$$

Solving algebraically gives

$$Y(s) = \frac{e^{-s}}{(s + 1)(s + 2)} = \left(\frac{1}{s + 1} - \frac{1}{s + 2}\right) e^{-s}.$$

By Theorem 1 the inverse is

$$y(t) = \mathcal{L}^{-1}(Y) = \begin{cases} 0 & \text{if } 0 < t < 1 \\ e^{-(t-1)} - e^{-2(t-1)} & \text{if } \quad t > 1. \end{cases}$$

$y(t)$ is shown in Fig. 132. Can you imagine how Fig. 131 approaches Fig. 132 as the wave becomes shorter and shorter, the area of the rectangle remaining 1? ■

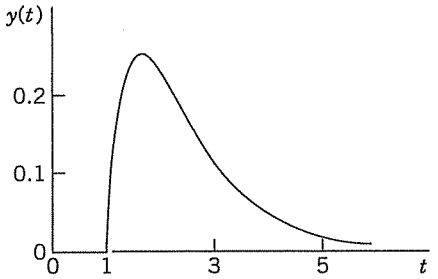


Fig. 132. Response to a hammerblow in Example 2

EXAMPLE 3 **Four-Terminal *RLC*-Network**

Find the output voltage response in Fig. 133 if $R = 20\ \Omega$, $L = 1$ H, $C = 10^{-4}$ F, the input is $\delta(t)$ (a unit impulse at time $t = 0$), and current and charge are zero at time $t = 0$.

Solution. To understand what is going on, note that the network is an *RLC*-circuit to which two wires at A and B are attached for recording the voltage $v(t)$ on the capacitor. Recalling from Sec. 2.9 that current $i(t)$ and charge $q(t)$ are related by $i = q' = dq/dt$, we obtain the model

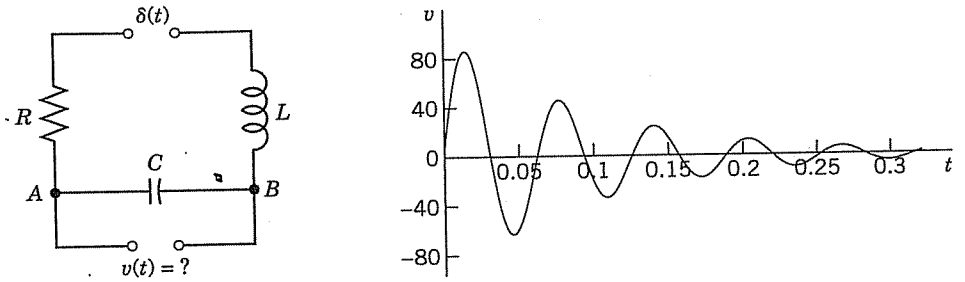
$$Li' + Ri + \frac{q}{C} = Lq'' + Rq' + \frac{q}{C} = q'' + 20q' + 10\,000q = \delta(t).$$

From (1) and (2) in Sec. 6.2 and (5) in this section we obtain the subsidiary equation for $Q(s) = \mathcal{L}(q)$

$$(s^2 + 20s + 10\,000)Q = 1. \qquad \text{Solution} \qquad Q = \frac{1}{(s + 10)^2 + 9900}.$$

By the first shifting theorem in Sec. 6.1 we obtain from Q damped oscillations for q and v; rounding $9900 \approx 99.50^2$, we get (Fig. 133)

$$q = \mathcal{L}^{-1}(Q) = \frac{1}{99.50} e^{-10t} \sin 99.50t \qquad \text{and} \qquad v = \frac{q}{C} = 100.5e^{-10t} \sin 99.50t. \quad ■$$



Network

Voltage on the capacitor

Fig. 133. Network and output voltage in Example 3

More on Partial Fractions

We have seen that the solution Y of a subsidiary equation usually appears as a quotient of polynomials $Y(s) = F(s)/G(s)$, so that a partial fraction representation leads to a sum of expressions whose inverses we can obtain from a table, aided by the first shifting theorem (Sec. 6.1). These representations are sometimes called **Heaviside expansions.**

An *unrepeated factor* $s - a$ in $G(s)$ requires a single partial fraction $A/(s - a)$. See Examples 1 and 2 on pp. 243, 244. *Repeated real factors* $(s - a)^2$, $(s - a)^3$, etc., require partial fractions

$$\frac{A_2}{(s-a)^2} + \frac{A_1}{s-a}, \qquad \frac{A_3}{(s-a)^3} + \frac{A_2}{(s-a)^2} + \frac{A_1}{s-a}, \quad \text{etc.,}$$

The inverses are $(A_2t + A_1)e^{at}$, $(\frac{1}{2}A_3t^2 + A_2t + A_1)e^{at}$, etc.

Unrepeated complex factors $(s - a)(s - \bar{a})$, $a = \alpha + i\beta$, $\bar{a} = \alpha - i\beta$, require a partial fraction $(As + B)/[(s - \alpha)^2 + \beta^2]$. For an application, see Example 4 in Sec. 6.3. A further one is the following.

EXAMPLE 4 **Unrepeated Complex Factors. Damped Forced Vibrations**

Solve the initial value problem for a damped mass–spring system acted upon by a sinusoidal force for some time interval (Fig. 134),

$$y'' + 2y' + 2y = r(t), \quad r(t) = 10 \sin 2t \text{ if } 0 < t < \pi \text{ and } 0 \text{ if } t > \pi; \qquad y(0) = 1, \quad y'(0) = -5.$$

Solution. From Table 6.1, (1), (2) in Sec. 6.2, and the second shifting theorem in Sec. 6.3, we obtain the subsidiary equation

$$(s^2Y - s + 5) + 2(sY - 1) + 2Y = 10\,\frac{2}{s^2 + 4}\,(1 - e^{-\pi s}).$$

We collect the Y-terms, $(s^2 + 2s + 2)Y$, take $-s + 5 - 2 = -s + 3$ to the right, and solve,

$$Y = \frac{20}{(s^2+4)(s^2+2s+2)} - \frac{20e^{-\pi s}}{(s^2+4)(s^2+2s+2)} + \frac{s-3}{s^2+2s+2}. \tag{6}$$

For the last fraction we get from Table 6.1 and the first shifting theorem

$$\mathcal{L}^{-1}\left\{\frac{s+1-4}{(s+1)^2+1}\right\} = e^{-t}(\cos t - 4 \sin t). \tag{7}$$

In the first fraction in (6) we have unrepeated complex roots, hence a partial fraction representation

$$\frac{20}{(s^2+4)(s^2+2s+2)} = \frac{As+B}{s^2+4} + \frac{Ms+N}{s^2+2s+2}.$$

Multiplication by the common denominator gives

$$20 = (As+B)(s^2+2s+2) + (Ms+N)(s^2+4).$$

We determine A, B, M, N. Equating the coefficients of each power of s on both sides gives the four equations

(a) $[s^3]$: $0 = A + M$ (b) $[s^2]$: $0 = 2A + B + N$

(c) $[s]$: $0 = 2A + 2B + 4M$ (d) $[s^0]$: $20 = 2B + 4N$.

We can solve this, for instance, obtaining $M = -A$ from (a), then $A = B$ from (c), then $N = -3A$ from (b), and finally $A = -2$ from (d). Hence $A = -2$, $B = -2$, $M = 2$, $N = 6$, and the first fraction in (6) has the representation

(8) $$\frac{-2s-2}{s^2+4} + \frac{2(s+1)+6-2}{(s+1)^2+1}.$$ Inverse transform: $-2\cos 2t - \sin 2t + e^{-t}(2\cos t + 4\sin t)$.

The sum of this and (7) is the solution of the problem for $0 < t < \pi$, namely (the sines cancel),

(9) $$y(t) = 3e^{-t}\cos t - 2\cos 2t - \sin 2t \qquad \text{if } 0 < t < \pi.$$

In the second fraction in (6) taken with the minus sign we have the factor $e^{-\pi s}$, so that from (8) and the second shifting theorem (Sec. 6.3) we get the inverse transform

$$+2\cos(2t-2\pi) + \sin(2t-2\pi) - e^{-(t-\pi)}[2\cos(t-\pi) + 4\sin(t-\pi)]$$
$$= 2\cos 2t + \sin 2t + e^{-(t-\pi)}(2\cos t + 4\sin t).$$

The sum of this and (9) is the solution for $t > \pi$,

(10) $$y(t) = e^{-t}[(3+2e^{\pi})\cos t + 4e^{\pi}\sin t] \qquad \text{if } t > \pi.$$

Figure 134 shows (9) (for $0 < t < \pi$) and (10) (for $t > \pi$), a beginning vibration, which goes to zero rapidly because of the damping and the absence of a driving force after $t = \pi$.

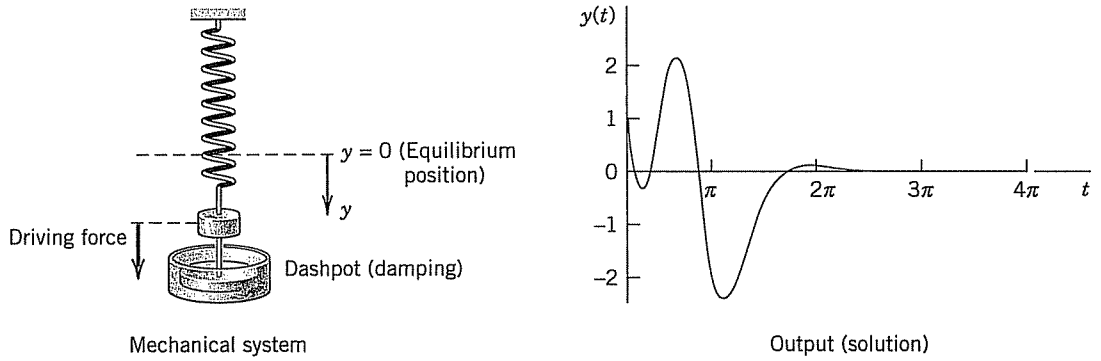


Fig. 134. Example 4

The case of repeated complex factors $[(s-a)(s-\bar{a})]^2$, which is important in connection with resonance, will be handled by "convolution" in the next section.

PROBLEM SET 6.4

1–12 EFFECT OF DELTA FUNCTION ON VIBRATING SYSTEMS

Showing the details, find, graph, and discuss the solution.

1. $y'' + y = \delta(t - 2\pi), \quad y(0) = 10, \quad y'(0) = 0$
2. $y'' + 2y' + 2y = e^{-t} + 5\delta(t - 2), \quad y(0) = 0, \quad y'(0) = 1$
3. $y'' - y = 10\delta(t - \frac{1}{2}) - 100\delta(t - 1), \quad y(0) = 10, \quad y'(0) = 1$
4. $y'' + 3y' + 2y = 10(\sin t + \delta(t - 1)), \quad y(0) = 1, \quad y'(0) = -1$
5. $y'' + 4y' + 5y = [1 - u(t - 10)]e^t - e^{10}\delta(t - 10), \quad y(0) = 0, \quad y'(0) = 1$
6. $y'' + 2y' - 3y = 100\delta(t - 2) + 100\delta(t - 3), \quad y(0) = 1, \quad y'(0) = 0$
7. $y'' + 2y' + 10y = 10[1 - u(t - 4)] - 10\delta(t - 5), \quad y(0) = 1, \quad y'(0) = 1$
8. $y'' + 5y' + 6y = \delta(t - \frac{1}{2}\pi) + u(t - \pi)\cos t, \quad y(0) = 0, \quad y'(0) = 0$
9. $y'' + 2y' + 5y = 25t - 100\delta(t - \pi), \quad y(0) = -2, \quad y'(0) = 5$
10. $y'' + 5y = 25t - 100\delta(t - \pi), \quad y(0) = -2, \quad y'(0) = 5$. (Compare with Prob. 9.)
11. $y'' + 3y' - 4y = 2e^t - 8e^2\delta(t - 2), \quad y(0) = 2, \quad y'(0) = 0$
12. $y'' + y = -2\sin t + 10\delta(t - \pi), \quad y(0) = 0, \quad y'(0) = 1$

13. **CAS PROJECT. Effect of Damping.** Consider a vibrating system of your choice modeled by

$$y'' + cy' + ky = r(t)$$

with $r(t)$ involving a δ-function. **(a)** Using graphs of the solution, describe the effect of continuously decreasing the damping to 0, keeping k constant.

(b) What happens if c is kept constant and k is continuously increased, starting from 0?

(c) Extend your results to a system with two δ-functions on the right, acting at different times.

14. **CAS PROJECT. Limit of a Rectangular Wave. Effects of Impulse.**

(a) In Example 1, take a rectangular wave of area 1 from 1 to $1 + k$. Graph the responses for a sequence of values of k approaching zero, illustrating that for smaller and smaller k those curves approach the curve shown in Fig. 132. *Hint:* If your CAS gives no solution for the differential equation. involving k, take specific k's from the beginning.

(b) Experiment on the response of the ODE in Example 1 (or of another ODE of your choice) to an impulse $\delta(t - a)$ for various systematically chosen a (> 0); choose initial conditions $y(0) \neq 0$, $y'(0) = 0$. Also consider the solution if no impulse is applied. Is there a dependence of the response on a? On b if you choose $b\delta(t - a)$? Would $-\delta(t - \tilde{a})$ with $\tilde{a} > a$ annihilate the effect of $\delta(t - a)$? Can you think of other questions that one could consider experimentally by inspecting graphs?

15. **PROJECT. Heaviside Formulas. (a)** Show that for a simple root a and fraction $A/(s - a)$ in $F(s)/G(s)$ we have the *Heaviside formula*

$$A = \lim_{s \to a} \frac{(s - a)F(s)}{G(s)}.$$

(b) Similarly, show that for a root a of order m and fractions in

$$\frac{F(s)}{G(s)} = \frac{A_m}{(s - a)^m} + \frac{A_{m-1}}{(s - a)^{m-1}} + \cdots + \frac{A_1}{s - a} + \text{further fractions}$$

we have the *Heaviside formulas* for the first coefficient

$$A_m = \lim_{s \to a} \frac{(s - a)^m F(s)}{G(s)}$$

and for the other coefficients

$$A_k = \frac{1}{(m - k)!} \lim_{s \to a} \frac{d^{m-k}}{ds^{m-k}} \left[\frac{(s - a)^m F(s)}{G(s)} \right], \quad k = 1, \cdots, m - 1.$$

16. **TEAM PROJECT. Laplace Transform of Periodic Functions**

(a) Theorem. *The Laplace transform of a piecewise continuous function $f(t)$ with period p is*

$$\text{(11)} \quad \mathcal{L}(f) = \frac{1}{1 - e^{-ps}} \int_0^p e^{-st} f(t)\, dt \qquad (s > 0).$$

Prove this theorem. *Hint:* Write $\int_0^\infty = \int_0^p + \int_p^{2p} + \cdots$. Set $t = (n - 1)p$ in the nth integral. Take out $e^{-(n-1)ps}$ from under the integral sign. Use the sum formula for the geometric series.

(b) Half-wave rectifier. Using (11), show that the half-wave rectification of $\sin \omega t$ in Fig. 135 has the Laplace transform

$$\mathscr{L}(f) = \frac{\omega(1 + e^{-\pi s/\omega})}{(s^2 + \omega^2)(1 - e^{-2\pi s/\omega})} = \frac{\omega}{(s^2 + \omega^2)(1 - e^{-\pi s/\omega})}.$$

(A *half-wave rectifier* clips the negative portions of the curve. A *full-wave rectifier* converts them to positive; see Fig. 136.)

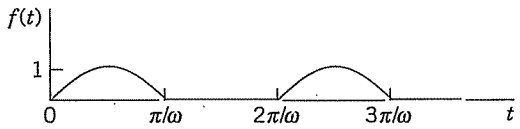


Fig. 135. Half-wave rectification

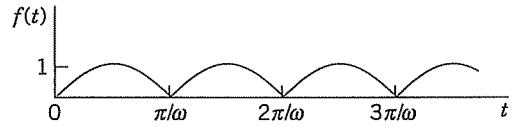


Fig. 136. Full-wave rectification

(c) Full-wave rectifier. Show that the Laplace transform of the full-wave rectification of $\sin \omega t$ is

$$\frac{\omega}{s^2 + \omega^2} \coth \frac{\pi s}{2\omega}.$$

(d) Saw-tooth wave. Find the Laplace transform of the saw-tooth wave in Fig. 137.

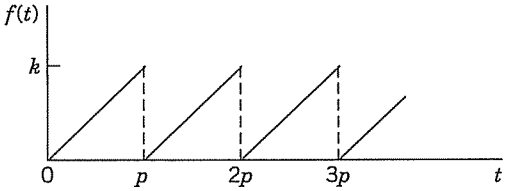


Fig. 137. Saw-tooth wave

(e) Staircase function. Find the Laplace transform of the staircase function in Fig. 138 by noting that it is the difference of kt/p and the function in (d).

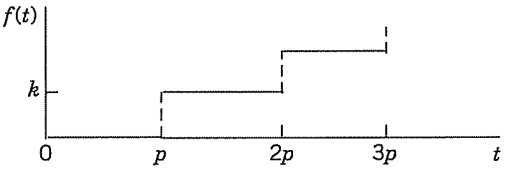


Fig. 138. Staircase function

6.5 Convolution. Integral Equations

Convolution has to do with the multiplication of transforms. The situation is as follows. *Addition* of transforms provides no problem; we know that $\mathscr{L}(f + g) = \mathscr{L}(f) + \mathscr{L}(g)$. Now **multiplication of transforms** occurs frequently in connection with ODEs, integral equations, and elsewhere. Then we usually know $\mathscr{L}(f)$ and $\mathscr{L}(g)$ and would like to know the function whose transform is the product $\mathscr{L}(f)\mathscr{L}(g)$. We might perhaps guess that it is fg, but this is false. *The transform of a product is generally different from the product of the transforms of the factors,*

$$\mathscr{L}(fg) \neq \mathscr{L}(f)\mathscr{L}(g) \qquad \text{in general.}$$

To see this take $f = e^t$ and $g = 1$. Then $fg = e^t$, $\mathscr{L}(fg) = 1/(s - 1)$, but $\mathscr{L}(f) = 1/(s - 1)$ and $\mathscr{L}(1) = 1/s$ give $\mathscr{L}(f)\mathscr{L}(g) = 1/(s^2 - s)$.

According to the next theorem, the correct answer is that $\mathscr{L}(f)\mathscr{L}(g)$ is the transform of the **convolution** of f and g, denoted by the standard notation $f * g$ and defined by the integral

(1) $$h(t) = (f * g)(t) = \int_0^t f(\tau)g(t - \tau)\, d\tau.$$

THEOREM 1

Convolution Theorem

If two functions f and g satisfy the assumption in the existence theorem in Sec. 6.1, so that their transforms F and G exist, the product $H = FG$ is the transform of h given by (1). (Proof after Example 2.)

EXAMPLE 1 **Convolution**

Let $H(s) = 1/[(s - a)s]$. Find $h(t)$.

Solution. $1/(s - a)$ has the inverse $f(t) = e^{at}$, and $1/s$ has the inverse $g(t) = 1$. With $f(\tau) = e^{a\tau}$ and $g(t - \tau) \equiv 1$ we thus obtain from (1) the answer

$$h(t) = e^{at} * 1 = \int_0^t e^{a\tau} \cdot 1 \, d\tau = \frac{1}{a}\,(e^{at} - 1).$$

To check, calculate

$$H(s) = \mathcal{L}(h)(s) = \frac{1}{a}\left(\frac{1}{s-a} - \frac{1}{s}\right) = \frac{1}{a} \cdot \frac{a}{s^2 - as} = \frac{1}{s-a} \cdot \frac{1}{s} = \mathcal{L}(e^{at})\,\mathcal{L}(1). \qquad \blacksquare$$

EXAMPLE 2 **Convolution**

Let $H(s) = 1/(s^2 + \omega^2)^2$. Find $h(t)$.

Solution. The inverse of $1/(s^2 + \omega^2)$ is $(\sin \omega t)/\omega$. Hence from (1) and the trigonometric formula (11) in App. 3.1 with $x = \frac{1}{2}(\omega t + \omega\tau)$ and $y = \frac{1}{2}(\omega t - \omega\tau)$ we obtain

$$\begin{aligned} h(t) = \frac{\sin \omega t}{\omega} * \frac{\sin \omega t}{\omega} &= \frac{1}{\omega^2} \int_0^t \sin \omega\tau \sin \omega(t - \tau)\, d\tau \\ &= \frac{1}{2\omega^2} \int_0^t [-\cos \omega t + \cos \omega\tau]\, d\tau \\ &= \frac{1}{2\omega^2} \left[-\tau \cos \omega t + \frac{\sin \omega\tau}{\omega}\right]_{\tau=0}^{t} \\ &= \frac{1}{2\omega^2} \left[-t \cos \omega t + \frac{\sin \omega t}{\omega}\right] \end{aligned}$$

in agreement with formula 21 in the table in Sec. 6.9. $\blacksquare$

PROOF We prove the Convolution Theorem 1. CAUTION! Note which ones are the variables of integration! We can denote them as we want, for instance, by τ and p, and write

$$F(s) = \int_0^\infty e^{-s\tau} f(\tau)\, d\tau \qquad \text{and} \qquad G(s) = \int_0^\infty e^{-sp} g(p)\, dp.$$

We now set $t = p + \tau$, where τ is at first constant. Then $p = t - \tau$, and t varies from τ to ∞. Thus

$$G(s) = \int_\tau^\infty e^{-s(t-\tau)} g(t - \tau)\, dt = e^{s\tau} \int_\tau^\infty e^{-st} g(t - \tau)\, dt.$$

τ in F and t in G vary independently. Hence we can insert the G-integral into the F-integral. Cancellation of $e^{-s\tau}$ and $e^{s\tau}$ then gives

$$F(s)G(s) = \int_0^\infty e^{-s\tau} f(\tau) e^{s\tau} \int_\tau^\infty e^{-st} g(t-\tau)\, dt\, d\tau = \int_0^\infty f(\tau) \int_\tau^\infty e^{-st} g(t-\tau)\, dt\, d\tau.$$

Here we integrate for fixed τ over t from τ to ∞ and then over τ from 0 to ∞. This is the blue region in Fig. 139. Under the assumption on f and g the order of integration can be reversed (see Ref. [A5] for a proof using uniform convergence). We then integrate first over τ from 0 to t and then over t from 0 to ∞, that is,

$$F(s)G(s) = \int_0^\infty e^{-st} \int_0^t f(\tau) g(t-\tau)\, d\tau\, dt = \int_0^\infty e^{-st} h(t)\, dt = \mathcal{L}(h) = H(s).$$

This completes the proof. ∎

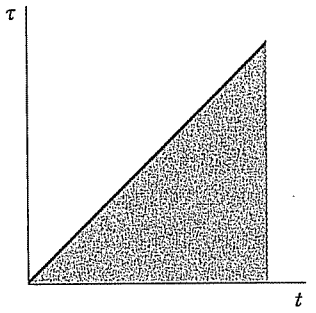


Fig. 139. Region of integration in the $t\tau$-plane in the proof of Theorem 1

From the definition it follows almost immediately that convolution has the properties

$$f * g = g * f \qquad \text{(commutative law)}$$

$$f * (g_1 + g_2) = f * g_1 + f * g_2 \qquad \text{(distributive law)}$$

$$(f * g) * v = f * (g * v) \qquad \text{(associative law)}$$

$$f * 0 = 0 * f = 0$$

similar to those of the multiplication of numbers. Unusual are the following two properties.

EXAMPLE 3 Unusual Properties of Convolution

$f * 1 \neq f$ in general. For instance,

$$t * 1 = \int_0^t \tau \cdot 1\, d\tau = \tfrac{1}{2} t^2 \neq t.$$

$(f * f)(t) \geqq 0$ may not hold. For instance, Example 2 with $\omega = 1$ gives

$$\sin t * \sin t = -\tfrac{1}{2} t \cos t + \tfrac{1}{2} \sin t \qquad \text{(Fig. 140).}$$ ∎

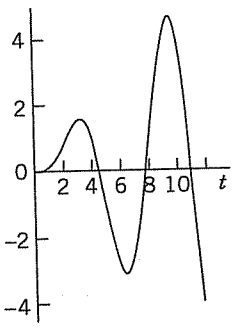


Fig. 140. Example 3

We shall now take up the case of a complex double root (left aside in the last section in connection with partial fractions) and find the solution (the inverse transform) directly by convolution.

EXAMPLE 4 **Repeated Complex Factors. Resonance**

In an undamped mass–spring system, resonance occurs if the frequency of the driving force equals the natural frequency of the system. Then the model is (see Sec. 2.8)

$$y'' + \omega_0^2 y = K \sin \omega_0 t$$

where $\omega_0^2 = k/m$, k is the spring constant, and m is the mass of the body attached to the spring. We assume $y(0) = 0$ and $y'(0) = 0$, for simplicity. Then the subsidiary equation is

$$s^2 Y + \omega_0^2 Y = \frac{K\omega_0}{s^2 + \omega_0^2}. \qquad \text{Its solution is} \qquad Y = \frac{K\omega_0}{(s^2 + \omega_0^2)^2}.$$

This is a transform as in Example 2 with $\omega = \omega_0$ and multiplied by $K\omega_0$. Hence from Example 2 we can see directly that the solution of our problem is

$$y(t) = \frac{K\omega_0}{2\omega_0^2}\left(-t\cos\omega_0 t + \frac{\sin\omega_0 t}{\omega_0}\right) = \frac{K}{2\omega_0^2}(-\omega_0 t\cos\omega_0 t + \sin\omega_0 t).$$

We see that the first term grows without bound. Clearly, in the case of resonance such a term must occur. (See also a similar kind of solution in Fig. 54 in Sec. 2.8.) ■

Application to Nonhomogeneous Linear ODEs

Nonhomogeneous linear ODEs can now be solved by a general method based on convolution by which the solution is obtained in the form of an integral. To see this, recall from Sec. 6.2 that the subsidiary equation of the ODE

$$y'' + ay' + by = r(t) \qquad (a, b \text{ constant}) \tag{2}$$

has the solution [(7) in Sec. 6.2]

$$Y(s) = [(s + a)y(0) + y'(0)]Q(s) + R(s)Q(s)$$

with $R(s) = \mathcal{L}(r)$ and $Q(s) = 1/(s^2 + as + b)$ the transfer function. Inversion of the first term $[\cdots]$ provides no difficulty; depending on whether $\frac{1}{4}a^2 - b$ is positive, zero, or negative, its inverse will be a linear combination of two exponential functions, or of the

form $(c_1 + c_2 t)e^{-at/2}$, or a damped oscillation, respectively. The interesting term is $R(s)Q(s)$ because $r(t)$ can have various forms of practical importance, as we shall see. If $y(0) = 0$ and $y'(0) = 0$, then $Y = RQ$, and the convolution theorem gives the solution

$$y(t) = \int_0^t q(t - \tau)r(\tau)\, d\tau. \tag{3}$$

EXAMPLE 5 **Response of a Damped Vibrating System to a Single Square Wave**

Using convolution, determine the response of the damped mass–spring system modeled by

$$y'' + 3y' + 2y = r(t), \qquad r(t) = 1 \text{ if } 1 < t < 2 \text{ and 0 otherwise}, \qquad y(0) = y'(0) = 0.$$

This system with an **input** (a driving force) ***that acts for some time only*** (Fig. 141) has been solved by partial fraction reduction in Sec. 6.4 (Example 1).

Solution by Convolution. The transfer function and its inverse are

$$Q(s) = \frac{1}{s^2 + 3s + 2} = \frac{1}{(s + 1)(s + 2)} = \frac{1}{s + 1} - \frac{1}{s + 2}, \qquad \text{hence} \qquad q(t) = e^{-t} - e^{-2t}.$$

Hence the convolution integral (3) is (except for the limits of integration)

$$y(t) = \int q(t - \tau) \cdot 1\, d\tau = \int \left[e^{-(t-\tau)} - e^{-2(t-\tau)}\right] d\tau = e^{-(t-\tau)} - \tfrac{1}{2}e^{-2(t-\tau)}.$$

Now comes an important point in handling convolution. $r(\tau) = 1$ if $1 < \tau < 2$ only. Hence if $t < 1$, the integral is zero. If $1 < t < 2$, we have to integrate from $\tau = 1$ (not 0) to t. This gives (with the first two terms from the upper limit)

$$y(t) = e^{-0} - \tfrac{1}{2}e^{-0} - (e^{-(t-1)} - \tfrac{1}{2}e^{-2(t-1)}) = \tfrac{1}{2} - e^{-(t-1)} + \tfrac{1}{2}e^{-2(t-1)}.$$

If $t > 2$, we have to integrate from $\tau = 1$ to 2 (not to t). This gives

$$y(t) = e^{-(t-2)} - \tfrac{1}{2}e^{-2(t-2)} - (e^{-(t-1)} - \tfrac{1}{2}e^{-2(t-1)}).$$

Figure 141 shows the input (the square wave) and the interesting output, which is zero from 0 to 1, then increases, reaches a maximum (near 2.6) after the input has become zero (why?), and finally decreases to zero in a monotone fashion.

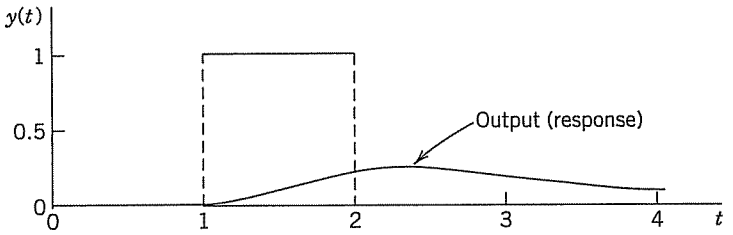


Fig. 141. Square wave and response in Example 5

Integral Equations

Convolution also helps in solving certain **integral equations,** that is, equations in which the unknown function $y(t)$ appears in an integral (and perhaps also outside of it). This concerns equations with an integral of the form of a convolution. Hence these are special and it suffices to explain the idea in terms of two examples and add a few problems in the problem set.

EXAMPLE 6 A Volterra Integral Equation of the Second Kind

Solve the Volterra integral equation of the second kind[3]

$$y(t) - \int_0^t y(\tau) \sin(t - \tau)\, d\tau = t.$$

Solution. From (1) we see that the given equation can be written as a convolution, $y - y * \sin t = t$. Writing $Y = \mathcal{L}(y)$ and applying the convolution theorem, we obtain

$$Y(s) - Y(s)\frac{1}{s^2 + 1} = Y(s)\frac{s^2}{s^2 + 1} = \frac{1}{s^2}.$$

The solution is

$$Y(s) = \frac{s^2 + 1}{s^4} = \frac{1}{s^2} + \frac{1}{s^4} \qquad \text{and gives the answer} \qquad y(t) = t + \frac{t^3}{6}.$$

Check the result by a CAS or by substitution and repeated integration by parts (which will need patience). ■

EXAMPLE 7 Another Volterra Integral Equation of the Second Kind

Solve the Volterra integral equation

$$y(t) - \int_0^t (1 + \tau)\, y(t - \tau)\, d\tau = 1 - \sinh t.$$

Solution. By (1) we can write $y - (1 + t) * y = 1 - \sinh t$. Writing $Y = \mathcal{L}(y)$, we obtain by using the convolution theorem and then taking common denominators

$$Y(s)\left[1 - \left(\frac{1}{s} + \frac{1}{s^2}\right)\right] = \frac{1}{s} - \frac{1}{s^2 - 1}, \qquad \text{hence} \qquad Y(s) \cdot \frac{s^2 - s - 1}{s^2} = \frac{s^2 - 1 - s}{s(s^2 - 1)}.$$

$(s^2 - s - 1)/s$ cancels on both sides, so that solving for Y simply gives

$$Y(s) = \frac{s}{s^2 - 1} \qquad \text{and the solution is} \qquad y(t) = \cosh t.$$

■

PROBLEM SET 6.5

1–8 CONVOLUTIONS BY INTEGRATION

Find by integration:

1. $1 * 1$
2. $t * t$
3. $t * e^t$
4. $e^{at} * e^{bt}$ $(a \neq b)$
5. $1 * \cos \omega t$
6. $1 * f(t)$
7. $e^{kt} * e^{-kt}$
8. $\sin t * \cos t$

9–16 INVERSE TRANSFORMS BY CONVOLUTION

Find $f(t)$ if $\mathcal{L}(f)$ equals:

9. $\dfrac{1}{(s - 3)(s + 5)}$
10. $\dfrac{1}{s(s - 1)}$
11. $\dfrac{1}{s(s^2 + 4)}$
12. $\dfrac{1}{s^2(s - 2)}$
13. $\dfrac{1}{s^2(s^2 + 1)}$
14. $\dfrac{s}{(s^2 + 16)^2}$
15. $\dfrac{1}{s(s^2 - 9)}$
16. $\dfrac{5}{(s^2 + 1)(s^2 + 25)}$
17. **(Partial fractions)** Solve Probs. 9, 11, and 13 by using partial fractions. Comment on the amount of work.

18–25 SOLVING INITIAL VALUE PROBLEMS

Using the convolution theorem, solve:

18. $y'' + y = \sin t, \quad y(0) = 0, \quad y'(0) = 0$
19. $y'' + 4y = \sin 3t, \quad y(0) = 0, \quad y'(0) = 0$
20. $y'' + 5y' + 4y = 2e^{-2t}, \quad y(0) = 0, \quad y'(0) = 0$

[3] If the upper limit of integration is *variable,* the equation is named after the Italian mathematician VITO VOLTERRA (1860–1940), and if that limit is *constant,* the equation is named after the Swedish mathematician IVAR FREDHOLM (1866–1927). "Of the second kind (first kind)" indicates that y occurs (does not occur) outside of the integral.

21. $y'' + 9y = 8 \sin t$ if $0 < t < \pi$ and 0 if $t > \pi$; $y(0) = 0, \quad y'(0) = 4$

22. $y'' + 3y' + 2y = 1$ if $0 < t < a$ and 0 if $t > a$; $y(0) = 0, \quad y'(0) = 0$

23. $y'' + 4y = 5u(t - 1); \qquad y(0) = 0, \quad y'(0) = 0$

24. $y'' + 5y' + 6y = \delta(t - 3); \quad y(0) = 1, \quad y'(0) = 0$

25. $y'' + 6y' + 8y = 2\delta(t - 1) + 2\delta(t - 2); \quad y(0) = 1, \quad y'(0) = 0$

26. TEAM PROJECT. Properties of Convolution. Prove:

(a) Commutativity, $f * g = g * f$

(b) Associativity, $(f * g) * v = f * (g * v)$

(c) Distributivity, $f * (g_1 + g_2) = f * g_1 + f * g_2$

(d) Dirac's delta. Derive the sifting formula (4) in Sec. 6.4 by using f_k with $a = 0$ [(1), Sec. 6.4] and applying the mean value theorem for integrals.

(e) Unspecified driving force. Show that forced vibrations governed by

$$y'' + \omega^2 y = r(t), \qquad y(0) = K_1, \qquad y'(0) = K_2$$

with $\omega \neq 0$ and an unspecified driving force $r(t)$ can be written in convolution form,

$$y = \frac{1}{\omega} \sin \omega t * r(t) + K_1 \cos \omega t + \frac{K_2}{\omega} \sin \omega t.$$

27–34 INTEGRAL EQUATIONS

Using Laplace transforms and showing the details, solve:

27. $y(t) - \int_0^t y(\tau)\, d\tau = 1$

28. $y(t) + \int_0^t y(\tau) \cosh (t - \tau)\, d\tau = t + e^t$

29. $y(t) - \int_0^t y(\tau) \sin (t - \tau)\, d\tau = \cos t$

30. $y(t) + 2\int_0^t y(\tau) \cos (t - \tau)\, d\tau = \cos t$

31. $y(t) + \int_0^t (t - \tau) y(\tau)\, d\tau = 1$

32. $y(t) - \int_0^t y(\tau)(t - \tau)\, d\tau = 2 - \frac{1}{2}t^2$

33. $y(t) + 2e^t \int_0^t e^{-\tau} y(\tau)\, d\tau = te^t$

34. $y(t) + \int_0^t e^{2(t-\tau)} y(\tau)\, d\tau = t^2 - t - \frac{1}{2} + \frac{1}{2}e^{2t}$

35. CAS EXPERIMENT. Variation of a Parameter.
(a) Replace 2 in Prob. 33 by a parameter k and investigate graphically how the solution curve changes if you vary k, in particular near $k = -2$.

(b) Make similar experiments with an integral equation of your choice whose solution is oscillating.

6.6 Differentiation and Integration of Transforms. ODEs with Variable Coefficients

The variety of methods for obtaining transforms and inverse transforms and their application in solving ODEs is surprisingly large. We have seen that they include direct integration, the use of linearity (Sec. 6.1), shifting (Secs. 6.1, 6.3), convolution (Sec. 6.5), and differentiation and integration of functions $f(t)$ (Sec. 6.2). But this is not all. In this section we shall consider operations of somewhat lesser importance, namely, differentiation and integration of *transforms* $F(s)$ and corresponding operations for functions $f(t)$, with applications to ODEs with variable coefficients.

Differentiation of Transforms

It can be shown that if a function $f(t)$ satisfies the conditions of the existence theorem in Sec. 6.1, then the derivative $F'(s) = dF/ds$ of the transform $F(s) = \mathcal{L}(f)$ can be obtained by differentiating $F(s)$ under the integral sign with respect to s (proof in Ref. [GR4] listed in App. 1). Thus, if

$$F(s) = \int_0^\infty e^{-st} f(t)\, dt, \qquad \text{then} \qquad F'(s) = -\int_0^\infty e^{-st} t f(t)\, dt.$$

Consequently, if $\mathcal{L}(f) = F(s)$, then

$$\text{(1)} \qquad \mathcal{L}\{tf(t)\} = -F'(s), \qquad \text{hence} \qquad \mathcal{L}^{-1}\{F'(s)\} = -tf(t)$$

where the second formula is obtained by applying $\mathcal{L}^{-1}$ on both sides of the first formula. In this way, *differentiation of the transform of a function corresponds to the multiplication of the function by* $-t$.

EXAMPLE 1 **Differentiation of Transforms. Formulas 21–23 in Sec. 6.9**

We shall derive the following three formulas.

	$\mathcal{L}(f)$	$f(t)$
(2)	$\dfrac{1}{(s^2 + \beta^2)^2}$	$\dfrac{1}{2\beta^3}(\sin \beta t - \beta t \cos \beta t)$
(3)	$\dfrac{s}{(s^2 + \beta^2)^2}$	$\dfrac{t}{2\beta} \sin \beta t$
(4)	$\dfrac{s^2}{(s^2 + \beta^2)^2}$	$\dfrac{1}{2\beta}(\sin \beta t + \beta t \cos \beta t)$

Solution. From (1) and formula 8 (with $\omega = \beta$) in Table 6.1 of Sec. 6.1 we obtain by differentiation (CAUTION! Chain rule!)

$$\mathcal{L}(t \sin \beta t) = \frac{2\beta s}{(s^2 + \beta^2)^2}.$$

Dividing by 2β and using the linearity of $\mathcal{L}$, we obtain (3).

Formulas (2) and (4) are obtained as follows. From (1) and formula 7 (with $\omega = \beta$) in Table 6.1 we find

$$\text{(5)} \qquad \mathcal{L}(t \cos \beta t) = -\frac{(s^2 + \beta^2) - 2s^2}{(s^2 + \beta^2)^2} = \frac{s^2 - \beta^2}{(s^2 + \beta^2)^2}.$$

From this and formula 8 (with $\omega = \beta$) in Table 6.1 we have

$$\mathcal{L}\left(t \cos \beta t \pm \frac{1}{\beta} \sin \beta t\right) = \frac{s^2 - \beta^2}{(s^2 + \beta^2)^2} \pm \frac{1}{s^2 + \beta^2}.$$

On the right we now take the common denominator. Then we see that for the plus sign the numerator becomes $s^2 - \beta^2 + s^2 + \beta^2 = 2s^2$, so that (4) follows by division by 2. Similarly, for the minus sign the numerator takes the form $s^2 - \beta^2 - s^2 - \beta^2 = -2\beta^2$, and we obtain (2). This agrees with Example 2 in Sec. 6.5. ■

Integration of Transforms

Similarly, if $f(t)$ satisfies the conditions of the existence theorem in Sec. 6.1 and the limit of $f(t)/t$, as t approaches 0 from the right, exists, then for $s > k$,

$$\text{(6)} \qquad \mathcal{L}\left\{\frac{f(t)}{t}\right\} = \int_s^\infty F(\tilde{s})\, d\tilde{s} \qquad \text{hence} \qquad \mathcal{L}^{-1}\left\{\int_s^\infty F(\tilde{s})\, d\tilde{s}\right\} = \frac{f(t)}{t}.$$

In this way, *integration of the transform of a function* $f(t)$ *corresponds to the division of* $f(t)$ *by* t.

We indicate how (6) is obtained. From the definition it follows that

$$\int_s^\infty F(\tilde{s})\,d\tilde{s} = \int_s^\infty \left[\int_0^\infty e^{-\tilde{s}t} f(t)\,dt\right] d\tilde{s},$$

and it can be shown (see Ref. [GR4] in App. 1) that under the above assumptions we may reverse the order of integration, that is,

$$\int_s^\infty F(\tilde{s})\,d\tilde{s} = \int_0^\infty \left[\int_s^\infty e^{-\tilde{s}t} f(t)\,d\tilde{s}\right] dt = \int_0^\infty f(t)\left[\int_s^\infty e^{-\tilde{s}t}\,d\tilde{s}\right] dt.$$

Integration of $e^{-\tilde{s}t}$ with respect to $\tilde{s}$ gives $e^{-\tilde{s}t}/(-t)$. Here the integral over $\tilde{s}$ on the right equals e^{-st}/t. Therefore,

$$\int_s^\infty F(\tilde{s})\,d\tilde{s} = \int_0^\infty e^{-st}\,\frac{f(t)}{t}\,dt = \mathcal{L}\left\{\frac{f(t)}{t}\right\} \qquad (s > k). \quad \blacksquare$$

EXAMPLE 2 Differentiation and Integration of Transforms

Find the inverse transform of $\ln\left(1 + \dfrac{\omega^2}{s^2}\right) = \ln\dfrac{s^2 + \omega^2}{s^2}$.

Solution. Denote the given transform by $F(s)$. Its derivative is

$$F'(s) = \frac{d}{ds}\left(\ln(s^2 + \omega^2) - \ln s^2\right) = \frac{2s}{s^2 + \omega^2} - \frac{2s}{s^2}.$$

Taking the inverse transform and using (1), we obtain

$$\mathcal{L}^{-1}\{F'(s)\} = \mathcal{L}^{-1}\left\{\frac{2s}{s^2 + \omega^2} - \frac{2}{s}\right\} = 2\cos\omega t - 2 = -tf(t).$$

Hence the inverse $f(t)$ of $F(s)$ is $f(t) = 2(1 - \cos\omega t)/t$. This agrees with formula 42 in Sec. 6.9.

Alternatively, if we let

$$G(s) = \frac{2s}{s^2 + \omega^2} - \frac{2}{s}, \qquad \text{then} \qquad g(t) = \mathcal{L}^{-1}(G) = 2(\cos\omega t - 1).$$

From this and (6) we get, in agreement with the answer just obtained,

$$\ln\frac{s^2 + \omega^2}{s^2} = \int_s^\infty G(s)\,ds = -\frac{g(t)}{t} = \frac{2}{t}(1 - \cos\omega t),$$

the minus occurring since s is the lower limit of integration.

In a similar way we obtain formula 43 in Sec. 6.9,

$$\mathcal{L}^{-1}\left\{\ln\left(1 - \frac{a^2}{s^2}\right)\right\} = \frac{2}{t}(1 - \cosh at). \quad \blacksquare$$

Special Linear ODEs with Variable Coefficients

Formula (1) can be used to solve certain ODEs with variable coefficients. The idea is this. Let $\mathcal{L}(y) = Y$. Then $\mathcal{L}(y') = sY - y(0)$ (see Sec. 6.2). Hence by (1),

$$\mathcal{L}(ty') = -\frac{d}{ds}[sY - y(0)] = -Y - s\frac{dY}{ds}. \tag{7}$$

Similarly, $\mathcal{L}(y'') = s^2Y - sy(0) - y'(0)$ and by (1)

$$(8) \qquad \mathcal{L}(ty'') = -\frac{d}{ds}\left[s^2Y - sy(0) - y'(0)\right] = -2sY - s^2\frac{dY}{ds} + y(0).$$

Hence if an ODE has coefficients such as $at + b$, the subsidiary equation is a first-order ODE for Y, which is sometimes simpler than the given second-order ODE. But if the latter has coefficients $at^2 + bt + c$, then two applications of (1) would give a second-order ODE for Y, and this shows that the present method works well only for rather special ODEs with variable coefficients. An important ODE for which the method is advantageous is the following.

EXAMPLE 3 Laguerre's Equation. Laguerre Polynomials

Laguerre's ODE is

$$(9) \qquad ty'' + (1 - t)y' + ny = 0.$$

We determine a solution of (9) with $n = 0, 1, 2, \cdots$. From (7)–(9) we get the subsidiary equation

$$\left[-2sY - s^2\frac{dY}{ds} + y(0)\right] + sY - y(0) - \left(-Y - s\frac{dY}{ds}\right) + nY = 0.$$

Simplification gives

$$(s - s^2)\frac{dY}{ds} + (n + 1 - s)Y = 0.$$

Separating variables, using partial fractions, integrating (with the constant of integration taken zero), and taking exponentials, we get

$$(10^*) \qquad \frac{dY}{Y} = -\frac{n + 1 - s}{s - s^2}\,ds = \left(\frac{n}{s - 1} - \frac{n + 1}{s}\right)ds \qquad \text{and} \qquad Y = \frac{(s - 1)^n}{s^{n+1}}.$$

We write $l_n = \mathcal{L}^{-1}(Y)$ and prove **Rodrigues's formula**

$$(10) \qquad l_0 = 1, \qquad l_n(t) = \frac{e^t}{n!}\frac{d^n}{dt^n}(t^ne^{-t}), \qquad n = 1, 2, \cdots.$$

These are polynomials because the exponential terms cancel if we perform the indicated differentiations. They are called **Laguerre polynomials** and are usually denoted by L_n (see Problem Set 5.7, but we continue to reserve capital letters for transforms). We prove (10). By Table 6.1 and the first shifting theorem (s-shifting),

$$\mathcal{L}(t^ne^{-t}) = \frac{n!}{(s + 1)^{n+1}}, \qquad \text{hence by (3) in Sec. 6.2} \qquad \mathcal{L}\left\{\frac{d^n}{dt^n}(t^ne^{-t})\right\} = \frac{n!s^n}{(s + 1)^{n+1}}$$

because the derivatives up to the order $n - 1$ are zero at 0. Now make another shift and divide by $n!$ to get [see (10) and then (10*)]

$$\mathcal{L}(l_n) = \frac{(s - 1)^n}{s^{n+1}} = Y.$$

PROBLEM SET 6.6

1–12 TRANSFORMS BY DIFFERENTIATION

Showing the details of your work, find $\mathcal{L}(f)$ if $f(t)$ equals:

1. $4te^t$
2. $-t\cosh 2t$
3. $t\sin \omega t$
4. $t\cos(t + k)$
5. $te^{-2t}\sin t$
6. $t^2\sin 3t$
7. $t^2\sinh 4t$
8. t^ne^{kt}
9. $t^2\sin \omega t$
10. $t\cos \omega t$
11. $t\sin(t + k)$
12. $te^{-kt}\sin t$

13–20 INVERSE TRANSFORMS

Using differentiation, integration, s-shifting, or convolution (and showing the details), find $f(t)$ if $\mathcal{L}(f)$ equals:

13. $\dfrac{6}{(s+1)^2}$

14. $\dfrac{s}{(s^2+16)^2}$

15. $\dfrac{2(s+2)}{[(s+2)^2+1]^2}$

16. $\dfrac{s}{(s^2-1)^2}$

17. $\dfrac{2}{(s-k)^3}$

18. $\ln\dfrac{s+a}{s+b}$

19. $\ln\dfrac{s}{s-1}$

20. $\operatorname{arccot}\dfrac{s}{\omega}$

21. WRITING PROJECT. Differentiation and Integration of Functions and Transforms. Make a short draft of these four operations from memory. Then compare your notes with the text and write a report of 2–3 pages on these operations and their significance in applications.

22. CAS PROJECT. Laguerre Polynomials. (a) Write a CAS program for finding $l_n(t)$ in explicit form from (10). Apply it to calculate $l_0, \cdots, l_{10}$. Verify that $l_0, \cdots, l_{10}$ satisfy Laguerre's differential equation (9).

(b) Show that

$$l_n(t) = \sum_{m=0}^{n} \frac{(-1)^m}{m!} \binom{n}{m} t^m$$

and calculate $l_0, \cdots, l_{10}$ from this formula.

(c) Calculate $l_0, \cdots, l_{10}$ recursively from $l_0 = 1$, $l_1 = 1 - t$ by

$$(n+1)l_{n+1} = (2n + 1 - t)l_n - nl_{n-1}.$$

(d) Experiment with the graphs of $l_0, \cdots, l_{10}$, finding out empirically how the first maximum, first minimum, $\cdots$ is moving with respect to its location as a function of n. Write a short report on this.

(e) A **generating function** (definition in Problem Set 5.3) for the Laguerre polynomials is

$$\sum_{n=0}^{\infty} l_n(t)x^n = (1-x)^{-1}e^{tx/(x-1)}.$$

Obtain $l_0, \cdots, l_{10}$ from the corresponding partial sum of this power series in x and compare the l_n with those in (a), (b), or (c).

6.7 Systems of ODEs

The Laplace transform method may also be used for solving systems of ODEs, as we shall explain in terms of typical applications. We consider a first-order linear system with constant coefficients (as discussed in Sec. 4.1)

$$\begin{aligned} y_1' &= a_{11}y_1 + a_{12}y_2 + g_1(t) \\ y_2' &= a_{21}y_1 + a_{22}y_2 + g_2(t). \end{aligned} \tag{1}$$

Writing $Y_1 = \mathcal{L}(y_1)$, $Y_2 = \mathcal{L}(y_2)$, $G_1 = \mathcal{L}(g_1)$, $G_2 = \mathcal{L}(g_2)$, we obtain from (1) in Sec. 6.2 the subsidiary system

$$\begin{aligned} sY_1 - y_1(0) &= a_{11}Y_1 + a_{12}Y_2 + G_1(s) \\ sY_2 - y_2(0) &= a_{21}Y_1 + a_{22}Y_2 + G_2(s). \end{aligned}$$

By collecting the Y_1- and Y_2-terms we have

$$\begin{aligned} (a_{11} - s)Y_1 + \quad a_{12}Y_2 \quad &= -y_1(0) - G_1(s) \\ a_{21}Y_1 \quad + (a_{22} - s)Y_2 &= -y_2(0) - G_2(s). \end{aligned} \tag{2}$$

By solving this system algebraically for $Y_1(s)$, $Y_2(s)$ and taking the inverse transform we obtain the solution $y_1 = \mathcal{L}^{-1}(Y_1)$, $y_2 = \mathcal{L}^{-1}(Y_2)$ of the given system (1).

Note that (1) and (2) may be written in vector form (and similarly for the systems in the examples); thus, setting $\mathbf{y} = [y_1 \quad y_2]^T$, $\mathbf{A} = [a_{jk}]$, $\mathbf{g} = [g_1 \quad g_2]^T$, $\mathbf{Y} = [Y_1 \quad Y_2]^T$, $\mathbf{G} = [G_1 \quad G_2]^T$ we have

$$\mathbf{y}' = \mathbf{A}\mathbf{y} + \mathbf{g} \qquad \text{and} \qquad (\mathbf{A} - s\mathbf{I})\mathbf{Y} = -\mathbf{y}(0) - \mathbf{G}.$$

EXAMPLE 1 **Mixing Problem Involving Two Tanks**

Tank T_1 in Fig. 142 contains initially 100 gal of pure water. Tank T_2 contains initially 100 gal of water in which 150 lb of salt are dissolved. The inflow into T_1 is 2 gal/min from T_2 and 6 gal/min containing 6 lb of salt from the outside. The inflow into T_2 is 8 gal/min from T_1. The outflow from T_2 is 2 + 6 = 8 gal/min, as shown in the figure. The mixtures are kept uniform by stirring. Find and plot the salt contents $y_1(t)$ and $y_2(t)$ in T_1 and T_2, respectively.

Solution. The model is obtained in the form of two equations

$$\text{Time rate of change} = \text{Inflow/min} - \text{Outflow/min}$$

for the two tanks (see Sec. 4.1). Thus,

$$y_1' = -\frac{8}{100}\,y_1 + \frac{2}{100}\,y_2 + 6, \qquad y_2' = \frac{8}{100}\,y_1 - \frac{8}{100}\,y_2.$$

The initial conditions are $y_1(0) = 0$, $y_2(0) = 150$. From this we see that the subsidiary system (2) is

$$\begin{aligned}(-0.08 - s)Y_1 + \quad 0.02Y_2 \quad &= -\frac{6}{s} \\ 0.08Y_1 \quad + (-0.08 - s)Y_2 &= -150.\end{aligned}$$

We solve this algebraically for Y_1 and Y_2 by elimination (or by Cramer's rule in Sec. 7.7), and we write the solutions in terms of partial fractions,

$$Y_1 = \frac{9s + 0.48}{s(s + 0.12)(s + 0.04)} = \frac{100}{s} - \frac{62.5}{s + 0.12} - \frac{37.5}{s + 0.04}$$

$$Y_2 = \frac{150s^2 + 12s + 0.48}{s(s + 0.12)(s + 0.04)} = \frac{100}{s} + \frac{125}{s + 0.12} - \frac{75}{s + 0.04}.$$

By taking the inverse transform we arrive at the solution

$$y_1 = 100 - 62.5e^{-0.12t} - 37.5e^{-0.04t}$$

$$y_2 = 100 + 125e^{-0.12t} - 75e^{-0.04t}.$$

Figure 142 shows the interesting plot of these functions. Can you give physical explanations for their main features? Why do they have the limit 100? Why is y_2 not monotone, whereas y_1 is? Why is y_1 from some time on suddenly larger than y_2? Etc. ■

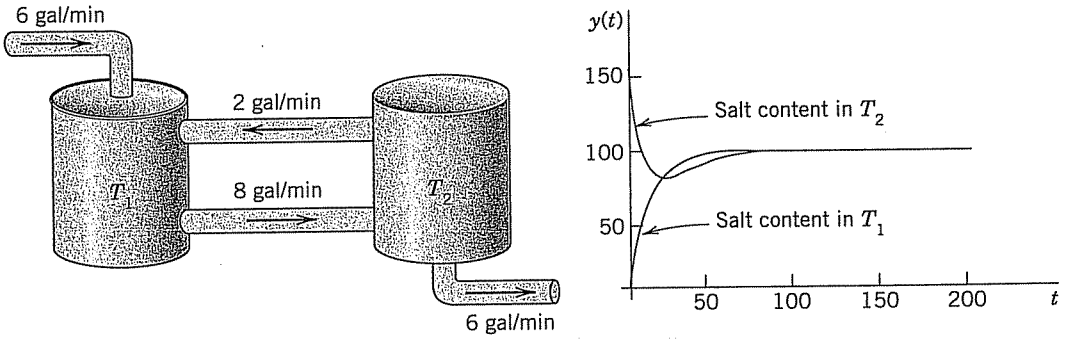


Fig. 142. Mixing problem in Example 1

Other systems of ODEs of practical importance can be solved by the Laplace transform method in a similar way, and eigenvalues and eigenvectors as we had to determine them in Chap. 4 will come out automatically, as we have seen in Example 1.

EXAMPLE 2 Electrical Network

Find the currents $i_1(t)$ and $i_2(t)$ in the network in Fig. 143 with L and R measured in terms of the usual units (see Sec. 2.9), $v(t) = 100$ volts if $0 \leqq t \leqq 0.5$ sec and 0 thereafter, and $i(0) = 0$, $i'(0) = 0$.

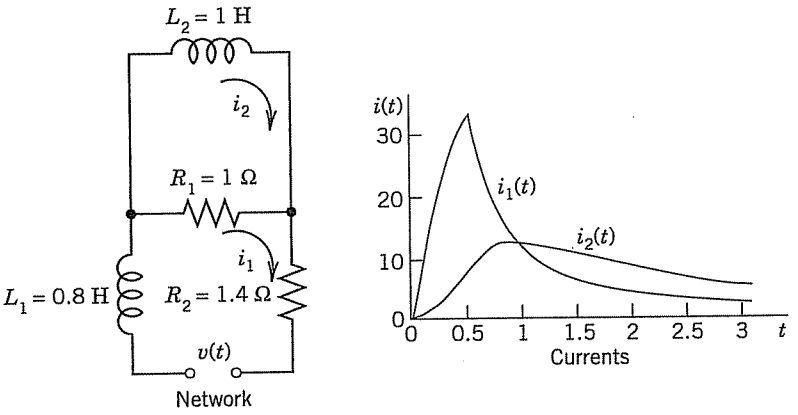


Fig. 143. Electrical network in Example 2

Solution. The model of the network is obtained from Kirchhoff's voltage law as in Sec. 2.9. For the lower circuit we obtain

$$0.8\,i_1' + 1(i_1 - i_2) + 1.4\,i_1 = 100\left[1 - u\left(t - \tfrac{1}{2}\right)\right]$$

and for the upper

$$1 \cdot i_2' + 1(i_2 - i_1) \qquad = 0.$$

Division by 0.8 and ordering gives for the lower circuit

$$i_1' + 3i_1 - 1.25\,i_2 = 125\left[1 - u\left(t - \tfrac{1}{2}\right)\right]$$

and for the upper

$$i_2' - \ i_1 + \qquad i_2 = 0.$$

With $i_1(0) = 0$, $i_2(0) = 0$ we obtain from (1) in Sec. 6.2 and the second shifting theorem the subsidiary system

$$(s + 3)I_1 - \ 1.25\,I_2 \ = 125\left(\frac{1}{s} - \frac{e^{-s/2}}{s}\right)$$
$$-I_1 + (s + 1)I_2 = 0.$$

Solving algebraically for I_1 and I_2 gives

$$I_1 = \frac{125(s + 1)}{s(s + \frac{1}{2})(s + \frac{7}{2})}\,(1 - e^{-s/2}),$$

$$I_2 = \frac{125}{s(s + \frac{1}{2})(s + \frac{7}{2})}\,(1 - e^{-s/2}).$$

The right sides without the factor $1 - e^{-s/2}$ have the partial fraction expansions

$$\frac{500}{7s} - \frac{125}{3(s + \frac{1}{2})} - \frac{625}{21(s + \frac{7}{2})}$$

and

$$\frac{500}{7s} - \frac{250}{3(s+\frac{1}{2})} + \frac{250}{21(s+\frac{7}{2})},$$

respectively. The inverse transform of this gives the solution for $0 \leqq t \leqq \frac{1}{2}$,

$$i_1(t) = -\frac{125}{3}e^{-t/2} - \frac{625}{21}e^{-7t/2} + \frac{500}{7}$$
$$i_2(t) = -\frac{250}{3}e^{-t/2} + \frac{250}{21}e^{-7t/2} + \frac{500}{7} \qquad (0 \leqq t \leqq \tfrac{1}{2}).$$

According to the second shifting theorem the solution for $t > \frac{1}{2}$ is $i_1(t) - i_1(t-\frac{1}{2})$ and $i_2(t) - i_2(t-\frac{1}{2})$, that is,

$$i_1(t) = -\frac{125}{3}(1-e^{1/4})e^{-t/2} - \frac{625}{21}(1-e^{7/4})e^{-7t/2}$$
$$i_2(t) = -\frac{250}{3}(1-e^{1/4})e^{-t/2} + \frac{250}{21}(1-e^{7/4})e^{-7t/2} \qquad (t > \tfrac{1}{2})$$

Can you explain physically why both currents eventually go to zero, and why $i_1(t)$ has a sharp cusp whereas $i_2(t)$ has a continuous tangent direction at $t = \frac{1}{2}$? ■

Systems of ODEs of higher order can be solved by the Laplace transform method in a similar fashion. As an important application, typical of many similar mechanical systems, we consider coupled vibrating masses on springs.

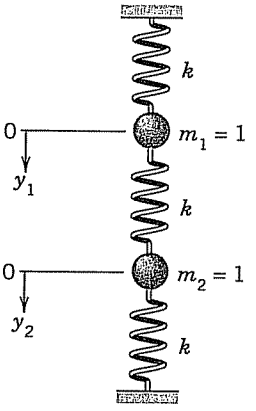


Fig. 144. Example 3

EXAMPLE 3 Model of Two Masses on Springs (Fig. 144)

The mechanical system in Fig. 144 consists of two bodies of mass 1 on three springs of the same spring constant k and of negligibly small masses of the springs. Also damping is assumed to be practically zero. Then the model of the physical system is the system of ODEs

$$y_1'' = -ky_1 + k(y_2 - y_1)$$
$$y_2'' = -k(y_2 - y_1) - ky_2. \qquad (3)$$

Here y_1 and y_2 are the displacements of the bodies from their positions of static equilibrium. These ODEs follow from **Newton's second law,** *Mass* $\times$ *Acceleration* = *Force,* as in Sec. 2.4 for a single body. We again regard downward forces as positive and upward as negative. On the upper body, $-ky_1$ is the force of the upper spring and $k(y_2 - y_1)$ that of the middle spring, $y_2 - y_1$ being the net change in spring length—think this over before going on. On the lower body, $-k(y_2 - y_1)$ is the force of the middle spring and $-ky_2$ that of the lower spring.

We shall determine the solution corresponding to the initial conditions $y_1(0) = 1$, $y_2(0) = 1$, $y_1'(0) = \sqrt{3k}$, $y_2'(0) = -\sqrt{3k}$. Let $Y_1 = \mathcal{L}(y_1)$ and $Y_2 = \mathcal{L}(y_2)$. Then from (2) in Sec. 6.2 and the initial conditions we obtain the subsidiary system

$$s^2Y_1 - s - \sqrt{3k} = -kY_1 + k(Y_2 - Y_1)$$
$$s^2Y_2 - s + \sqrt{3k} = -k(Y_2 - Y_1) - kY_2.$$

This system of linear algebraic equations in the unknowns Y_1 and Y_2 may be written

$$(s^2 + 2k)Y_1 - kY_2 = s + \sqrt{3k}$$
$$-kY_1 + (s^2 + 2k)Y_2 = s - \sqrt{3k}.$$

Elimination (or Cramer's rule in Sec. 7.7) yields the solution, which we can expand in terms of partial fractions,

$$Y_1 = \frac{(s + \sqrt{3k})(s^2 + 2k) + k(s - \sqrt{3k})}{(s^2 + 2k)^2 - k^2} = \frac{s}{s^2 + k} + \frac{\sqrt{3k}}{s^2 + 3k}$$

$$Y_2 = \frac{(s^2 + 2k)(s - \sqrt{3k}) + k(s + \sqrt{3k})}{(s^2 + 2k)^2 - k^2} = \frac{s}{s^2 + k} - \frac{\sqrt{3k}}{s^2 + 3k}.$$

Hence the solution of our initial value problem is (Fig. 145)

$$y_1(t) = \mathcal{L}^{-1}(Y_1) = \cos\sqrt{k}t + \sin\sqrt{3k}t$$
$$y_2(t) = \mathcal{L}^{-1}(Y_2) = \cos\sqrt{k}t - \sin\sqrt{3k}t.$$

We see that the motion of each mass is harmonic (the system is undamped!), being the superposition of a "slow" oscillation and a "rapid" oscillation. ▩

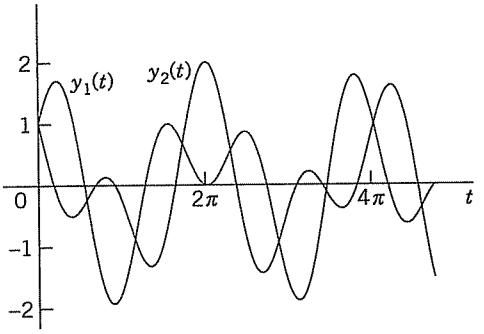


Fig. 145. Solutions in Example 3

PROBLEM SET 6.7

1–20 SYSTEMS OF ODES

Using the Laplace transform and showing the details of your work, solve the initial value problem:

1. $y_1' = -y_1 - y_2$, $y_2' = y_1 - y_2$, $y_1(0) = 0$, $y_2(0) = 1$
2. $y_1' = 5y_1 + y_2$, $y_2' = y_1 + 5y_2$, $y_1(0) = 1$, $y_2(0) = -3$
3. $y_1' = -6y_1 + 4y_2$, $y_2' = -4y_1 + 4y_2$, $y_1(0) = -2$, $y_2(0) = -7$
4. $y_1' + y_2 = 0$, $y_1 + y_2' = 2\cos t$, $y_1(0) = 1$, $y_2(0) = 0$
5. $y_1' = -4y_1 - 2y_2 + t$, $y_2' = 3y_1 + y_2 - t$, $y_1(0) = 5.75$, $y_2(0) = -6.75$
6. $y_1' = 4y_2 - 8\cos 4t$, $y_2' = -3y_1 - 9\sin 4t$, $y_1(0) = 0$, $y_2(0) = 3$

7. $y_1' = 5y_1 - 4y_2 - 9t^2 + 2t$,
$y_2' = 10y_1 - 7y_2 - 17t^2 - 2t$,
$y_1(0) = 2, \quad y_2(0) = 0$

8. $y_1' = 6y_1 + y_2, \quad y_2' = 9y_1 + 6y_2$,
$y_1(0) = -3, \quad y_2(0) = -3$

9. $y_1' = 5y_1 + 5y_2 - 15\cos t + 27\sin t$,
$y_2' = -10y_1 - 5y_2 - 150\sin t$,
$y_1(0) = 2, \quad y_2(0) = 2$

10. $y_1' = -2y_1 + 3y_2, \quad y_2' = 4y_1 - y_2$,
$y_1(0) = 4, \quad y_2(0) = 3$

11. $y_1' = y_2 + 1 - u(t-1)$,
$y_2' = -y_1 + 1 - u(t-1), \quad y_1(0) = 0$,
$y_2(0) = 0$

12. $y_1' = 2y_1 + y_2, \quad y_2' = 4y_1 + 2y_2 + 64tu(t-1)$,
$y_1(0) = 2, \quad y_2(0) = 0$

13. $y_1' = y_1 + 6u(t-2)e^{4t}, \quad y_2' = y_1 + 2y_2$,
$y_1(0) = 0, \quad y_2(0) = 1$

14. $y_1' = -y_2, \quad y_2' = -y_1 + 2[1 - u(t-2\pi)]\cos t$,
$y_1(0) = 1, \quad y_2(0) = 0$

15. $y_1' = -3y_1 + y_2 + u(t-1)e^t$,
$y_2' = -4y_1 + 2y_2 + u(t-1)e^t$,
$y_1(0) = 0, \quad y_2(0) = 3$

16. $y_1'' = -2y_1 + 2y_2, \quad y_2'' = 2y_1 - 5y_2$,
$y_1(0) = 1, \quad y_1'(0) = 0, \quad y_2(0) = 3, \quad y_2'(0) = 0$

17. $y_1'' = 4y_1 + 8y_2, \quad y_2'' = 5y_1 + y_2$,
$y_1(0) = 8, \quad y_1'(0) = -18, \quad y_2(0) = 5$,
$y_2'(0) = -21$

18. $y_1'' + y_2 = -101\sin 10t, \quad y_2'' + y_1 = 101\sin 10t$,
$y_1(0) = 0, \quad y_1'(0) = 6, \quad y_2(0) = 8$,
$y_2'(0) = -6$

19. $y_1' + y_2' = 2e^t + e^{-t}, \quad y_2' + y_3' = 2\sinh t$,
$y_3' + y_1' = e^t$
$y_1(0) = 0, \quad y_2(0) = 1, \quad y_3(0) = 1$

20. $4y_1' + y_2' - 2y_3' = 0, \quad -2y_1' + y_3' = 1$,
$2y_2' - 4y_3' = -16t$
$y_1(0) = 2, \quad y_2(0) = 0, \quad y_3(0) = 0$

21. **TEAM PROJECT. Comparison of Methods for Linear Systems of ODEs.**

(a) Models. Solve the models in Examples 1 and 2 of Sec. 4.1 by Laplace transforms and compare the amount of work with that in Sec. 4.1. (Show the details of your work.)

(b) Homogeneous Systems. Solve the systems (8), (11)–(13) in Sec. 4.3 by Laplace transforms. (Show the details.)

(c) Nonhomogeneous System. Solve the system (3) in Sec. 4.6 by Laplace transforms. (Show the details.)

FURTHER APPLICATIONS

22. **(Forced vibrations of two masses)** Solve the model in Example 3 with $k = 4$ and initial conditions $y_1(0) = 1$, $y_1'(0) = 1$, $y_2(0) = 1$, $y_2'(0) = -1$ under the assumption that the force $11\sin t$ is acting on the first body and the force $-11\sin t$ on the second. Graph the two curves on common axes and explain the motion physically.

23. **CAS Experiment. Effect of Initial Conditions.** In Prob. 22, vary the initial conditions systematically, describe and explain the graphs physically. The great variety of curves will surprise you. Are they always periodic? Can you find empirical laws for the changes in terms of continuous changes of those conditions?

24. **(Mixing problem)** What will happen in Example 1 if you double all flows (in particular, an increase to 12 gal/min containing 12 lb of salt from the outside), leaving the size of the tanks and the initial conditions as before? First guess, then calculate. Can you relate the new solution to the old one?

25. **(Electrical network)** Using Laplace transforms, find the currents $i_1(t)$ and $i_2(t)$ in Fig. 146, where $v(t) = 390\cos t$ and $i_1(0) = 0$, $i_2(0) = 0$. How soon will the currents practically reach their steady state?

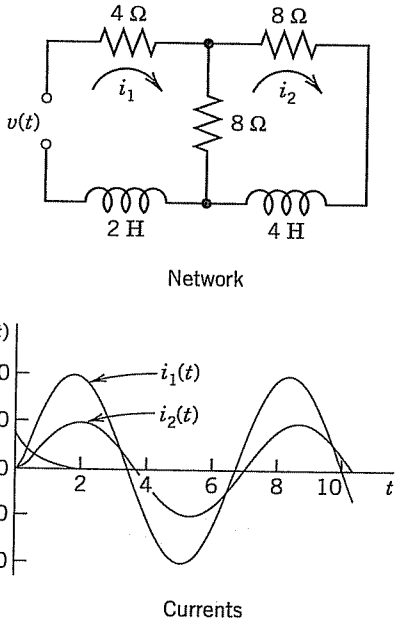


Fig. 146. Electrical network and currents in Problem 25

26. **(Single cosine wave)** Solve Prob. 25 when the EMF (electromotive force) is acting from 0 to 2π only. Can you do this just by looking at Prob. 25, practically without calculation?

6.8 Laplace Transform: General Formulas

Formula	Name, Comments	Sec.
$F(s) = \mathcal{L}\{f(t)\} = \int_0^\infty e^{-st} f(t)\, dt$ $f(t) = \mathcal{L}^{-1}\{F(s)\}$	Definition of Transform Inverse Transform	6.1
$\mathcal{L}\{af(t) + bg(t)\} = a\mathcal{L}\{f(t)\} + b\mathcal{L}\{g(t)\}$	Linearity	6.1
$\mathcal{L}\{e^{at} f(t)\} = F(s - a)$ $\mathcal{L}^{-1}\{F(s - a)\} = e^{at} f(t)$	s-Shifting (First Shifting Theorem)	6.1
$\mathcal{L}(f') = s\mathcal{L}(f) - f(0)$ $\mathcal{L}(f'') = s^2\mathcal{L}(f) - sf(0) - f'(0)$ $\mathcal{L}(f^{(n)}) = s^n\mathcal{L}(f) - s^{(n-1)}f(0) - \cdots$ $\cdots - f^{(n-1)}(0)$ $\mathcal{L}\left\{\int_0^t f(\tau)\, d\tau\right\} = \frac{1}{s}\mathcal{L}(f)$	Differentiation of Function Integration of Function	6.2
$(f * g)(t) = \int_0^t f(\tau) g(t - \tau)\, d\tau$ $= \int_0^t f(t - \tau) g(\tau)\, d\tau$ $\mathcal{L}(f * g) = \mathcal{L}(f)\mathcal{L}(g)$	Convolution	6.5
$\mathcal{L}\{f(t - a)\, u(t - a)\} = e^{-as} F(s)$ $\mathcal{L}^{-1}\{e^{-as} F(s)\} = f(t - a)\, u(t - a)$	t-Shifting (Second Shifting Theorem)	6.3
$\mathcal{L}\{tf(t)\} = -F'(s)$ $\mathcal{L}\left\{\frac{f(t)}{t}\right\} = \int_s^\infty F(\tilde{s})\, d\tilde{s}$	Differentiation of Transform Integration of Transform	6.6
$\mathcal{L}(f) = \frac{1}{1 - e^{-ps}} \int_0^p e^{-st} f(t)\, dt$	f Periodic with Period p	6.4 Project 16

6.9 Table of Laplace Transforms

For more extensive tables, see Ref.* [A9] *in Appendix 1.

	$F(s) = \mathcal{L}\{f(t)\}$	$f(t)$	Sec.
1	$1/s$	1	6.1
2	$1/s^2$	t	
3	$1/s^n \quad (n = 1, 2, \cdots)$	$t^{n-1}/(n-1)!$	
4	$1/\sqrt{s}$	$1/\sqrt{\pi t}$	
5	$1/s^{3/2}$	$2\sqrt{t/\pi}$	
6	$1/s^a \quad (a > 0)$	$t^{a-1}/\Gamma(a)$	
7	$\dfrac{1}{s-a}$	e^{at}	6.1
8	$\dfrac{1}{(s-a)^2}$	te^{at}	
9	$\dfrac{1}{(s-a)^n} \quad (n = 1, 2, \cdots)$	$\dfrac{1}{(n-1)!}\,t^{n-1}e^{at}$	
10	$\dfrac{1}{(s-a)^k} \quad (k > 0)$	$\dfrac{1}{\Gamma(k)}\,t^{k-1}e^{at}$	
11	$\dfrac{1}{(s-a)(s-b)} \quad (a \neq b)$	$\dfrac{1}{(a-b)}\,(e^{at} - e^{bt})$	
12	$\dfrac{s}{(s-a)(s-b)} \quad (a \neq b)$	$\dfrac{1}{(a-b)}\,(ae^{at} - be^{bt})$	
13	$\dfrac{1}{s^2+\omega^2}$	$\dfrac{1}{\omega}\sin \omega t$	6.1
14	$\dfrac{s}{s^2+\omega^2}$	$\cos \omega t$	
15	$\dfrac{1}{s^2-a^2}$	$\dfrac{1}{a}\sinh at$	
16	$\dfrac{s}{s^2-a^2}$	$\cosh at$	
17	$\dfrac{1}{(s-a)^2+\omega^2}$	$\dfrac{1}{\omega}\,e^{at}\sin \omega t$	
18	$\dfrac{s-a}{(s-a)^2+\omega^2}$	$e^{at}\cos \omega t$	
19	$\dfrac{1}{s(s^2+\omega^2)}$	$\dfrac{1}{\omega^2}\,(1 - \cos \omega t)$	6.2
20	$\dfrac{1}{s^2(s^2+\omega^2)}$	$\dfrac{1}{\omega^3}\,(\omega t - \sin \omega t)$	
21	$\dfrac{1}{(s^2+\omega^2)^2}$	$\dfrac{1}{2\omega^3}\,(\sin \omega t - \omega t \cos \omega t)$	6.6

(continued)

Table of Laplace Transforms (*continued*)

	$F(s) = \mathcal{L}\{f(t)\}$	$f(t)$	Sec.
22	$\dfrac{s}{(s^2 + \omega^2)^2}$	$\dfrac{t}{2\omega} \sin \omega t$	6.6
23	$\dfrac{s^2}{(s^2 + \omega^2)^2}$	$\dfrac{1}{2\omega} (\sin \omega t + \omega t \cos \omega t)$	6.6
24	$\dfrac{s}{(s^2 + a^2)(s^2 + b^2)} \quad (a^2 \neq b^2)$	$\dfrac{1}{b^2 - a^2} (\cos at - \cos bt)$	6.6
25	$\dfrac{1}{s^4 + 4k^4}$	$\dfrac{1}{4k^3} (\sin kt \cos kt - \cos kt \sinh kt)$	
26	$\dfrac{s}{s^4 + 4k^4}$	$\dfrac{1}{2k^2} \sin kt \sinh kt$	
27	$\dfrac{1}{s^4 - k^4}$	$\dfrac{1}{2k^3} (\sinh kt - \sin kt)$	
28	$\dfrac{s}{s^4 - k^4}$	$\dfrac{1}{2k^2} (\cosh kt - \cos kt)$	
29	$\sqrt{s - a} - \sqrt{s - b}$	$\dfrac{1}{2\sqrt{\pi t^3}} (e^{bt} - e^{at})$	
30	$\dfrac{1}{\sqrt{s + a}\sqrt{s + b}}$	$e^{-(a+b)t/2} I_0\left(\dfrac{a - b}{2} t\right)$	5.6
31	$\dfrac{1}{\sqrt{s^2 + a^2}}$	$J_0(at)$	5.5
32	$\dfrac{s}{(s - a)^{3/2}}$	$\dfrac{1}{\sqrt{\pi t}} e^{at}(1 + 2at)$	
33	$\dfrac{1}{(s^2 - a^2)^k} \quad (k > 0)$	$\dfrac{\sqrt{\pi}}{\Gamma(k)} \left(\dfrac{t}{2a}\right)^{k-1/2} I_{k-1/2}(at)$	5.6
34	e^{-as}/s	$u(t - a)$	6.3
35	e^{-as}	$\delta(t - a)$	6.4
36	$\dfrac{1}{s} e^{-k/s}$	$J_0(2\sqrt{kt})$	5.5
37	$\dfrac{1}{\sqrt{s}} e^{-k/s}$	$\dfrac{1}{\sqrt{\pi t}} \cos 2\sqrt{kt}$	
38	$\dfrac{1}{s^{3/2}} e^{k/s}$	$\dfrac{1}{\sqrt{\pi k}} \sinh 2\sqrt{kt}$	
39	$e^{-k\sqrt{s}} \quad (k > 0)$	$\dfrac{k}{2\sqrt{\pi t^3}} e^{-k^2/4t}$	
40	$\dfrac{1}{s} \ln s$	$-\ln t - \gamma \quad (\gamma \approx 0.5772)$	5.6

(*continued*)

Table of Laplace Transforms (*continued*)

	$F(s) = \mathcal{L}\{f(t)\}$	$f(t)$	Sec.
41	$\ln \dfrac{s-a}{s-b}$	$\dfrac{1}{t}(e^{bt} - e^{at})$	
42	$\ln \dfrac{s^2+\omega^2}{s^2}$	$\dfrac{2}{t}(1 - \cos \omega t)$	6.6
43	$\ln \dfrac{s^2-a^2}{s^2}$	$\dfrac{2}{t}(1 - \cosh at)$	
44	$\arctan \dfrac{\omega}{s}$	$\dfrac{1}{t}\sin \omega t$	
45	$\dfrac{1}{s}\operatorname{arccot} s$	$\mathrm{Si}(t)$	App. A3.1

CHAPTER 6 REVIEW QUESTIONS AND PROBLEMS

1. What do we mean by operational calculus?
2. What are the steps needed in solving an ODE by Laplace transform? What is the subsidiary equation?
3. The Laplace transform is a linear operation. What does this mean? Why is it important?
4. For what problems is the Laplace transform preferable over the usual method? Explain.
5. What are the unit step and Dirac's delta functions? Give examples.
6. What is the difference between the two shifting theorems? When do they apply?
7. Is $\mathcal{L}\{f(t)g(t)\} = \mathcal{L}\{f(t)\}\mathcal{L}\{g(t)\}$? Explain.
8. Can a discontinuous function have a Laplace transform? Does every continuous function have a Laplace transform? Give reasons.
9. State the transforms of a few simple functions from memory.
10. If two different continuous functions have transforms, the latter are different. Why is this practically important?

11–22 LAPLACE TRANSFORMS

Find the transform (showing the details of your work and indicating the method or formula you are using):

11. te^{3t}
12. $e^{-t}\sin 2t$
13. $\sin^2 t$
14. $\cos^2 4t$
15. $tu(t-\pi)$
16. $u(t-2\pi)\sin t$
17. $e^t * \cos 2t$
18. $(\sin \omega t) * (\cos \omega t)$
19. $\sin t + \sinh t$
20. $\cosh t - \cos t$
21. $e^{at} - e^{bt} \quad (a \neq b)$
22. $\cosh 2t - \cosh t$

23–34 INVERSE LAPLACE TRANSFORMS

Find the inverse transform (showing the details of your work and indicating the method or formula used):

23. $\dfrac{10s}{s^2+2}$
24. $\dfrac{15}{s^2-4}$
25. $\dfrac{12}{s^2+4s+20}$
26. $\dfrac{3s}{s^2-2s+2}$
27. $\dfrac{5s+4}{s^2}e^{-2s}$
28. $\dfrac{2s-10}{s^3}e^{-5s}$
29. $\dfrac{2s+4}{(s^2+4s+5)^2}$
30. $\dfrac{s^2-16}{(s^2+16)^2}$
31. $\left(\dfrac{2}{s^2}+\dfrac{2}{s^3}\right)e^{-s}$
32. $\dfrac{180+18s^2+3s^4}{s^7}$
33. $\dfrac{\pi}{s^2(s^2+\omega^2)}$
34. $\dfrac{2}{2s^2+2s+1}$

35–50 SINGLE ODEs AND SYSTEMS OF ODEs

Solve by Laplace transforms, showing the details and graphing the solution:

35. $y'' + y = u(t - 1)$, $y(0) = 0$, $y'(0) = 20$

36. $y'' + 16y = 4\delta(t - \pi)$, $y(0) = -1$, $y'(0) = 0$

37. $y'' + 4y = 8\delta(t - 5)$, $y(0) = 10$, $y'(0) = -1$

38. $y'' + y = u(t - 2)$, $y(0) = 0$, $y'(0) = 0$

39. $y'' + 2y' + 10y = 0$, $y(0) = 7$, $y'(0) = -1$

40. $y'' + 4y' + 5y = 50t$, $y(0) = 5$, $y'(0) = -5$

41. $y'' - y' - 2y = 12u(t - \pi)\sin t$, $y(0) = 1$, $y'(0) = -1$

42. $y'' - 2y' + y = t\delta(t - 1)$, $y(0) = 0$, $y'(0) = 0$

43. $y'' - 4y' + 4y = \delta(t - 1) - \delta(t - 2)$, $y(0) = 0$, $y'(0) = 0$

44. $y'' + 4y = \delta(t - \pi) - \delta(t - 2\pi)$, $y(0) = 1$, $y'(0) = 0$

45. $y_1' + y_2 = \sin t$, $y_2' + y_1 = -\sin t$, $y_1(0) = 1$, $y_2(0) = 0$

46. $y_1' = -3y_1 + y_2 - 12t$, $y_2' = -4y_1 + 2y_2 + 12t$, $y_1(0) = 0$, $y_2(0) = 0$

47. $y_1' = y_2$, $y_2' = -5y_1 - 2y_2$, $y_1(0) = 0$, $y_2(0) = 1$

48. $y_1' = y_2$, $y_2' = -4y_1 + \delta(t - \pi)$, $y_1(0) = 0$, $y_2(0) = 0$

49. $y_1'' = 4y_2 - 4e^t$, $y_2'' = 3y_1 + y_2$, $y_1(0) = 1$, $y_1'(0) = 2$, $y_2(0) = 2$, $y_2'(0) = 3$

50. $y_1'' = 16y_2$, $y_2'' = 16y_1$, $y_1(0) = 2$, $y_1'(0) = 12$, $y_2(0) = 6$, $y_2'(0) = 4$

MODELS OF CIRCUITS AND NETWORKS

51. (RC-circuit) Find and graph the current $i(t)$ in the RC-circuit in Fig. 147, where $R = 100\ \Omega$, $C = 10^{-3}$ F, $v(t) = 100t$ V if $0 < t < 2$, $v(t) = 200$ V if $t > 2$ and the initial charge on the capacitor is 0.

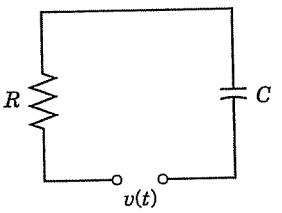


Fig. 147. RC-circuit

52. (LC-circuit) Find and graph the charge $q(t)$ and the current $i(t)$ in the LC-circuit in Fig. 148, where $L = 0.5$ H, $C = 0.02$ F, $v(t) = 1425 \sin 5t$ V if $0 < t < \pi$, $v(t) = 0$ if $t > \pi$, and current and charge at $t = 0$ are 0.

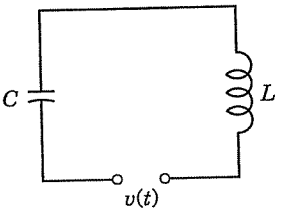


Fig. 148. LC-circuit

53. (RLC-circuit) Find and graph the current $i(t)$ in the RLC-circuit in Fig. 149, where $R = 1\ \Omega$, $L = 0.25$ H, $C = 0.2$ F, $v(t) = 377 \sin 20t$ V, and current and charge at $t = 0$ are 0.

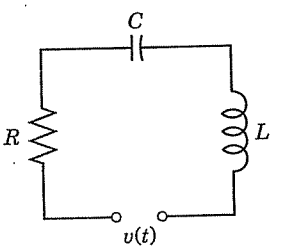


Fig. 149. RLC-circuit

54. (Network) Show that by Kirchhoff's voltage law (Sec. 2.9), the currents in the network in Fig. 150 are obtained from the system

$$Li_1' + R(i_1 - i_2) = v(t)$$

$$R(i_2' - i_1') + \frac{1}{C}\, i_2 = 0.$$

Solve this system, where $R = 1\ \Omega$, $L = 2$ H, $C = 0.5$ F, $v(t) = 90e^{-t/4}$ V, $i_1(0) = 0$, $i_2(0) = 2$ A.

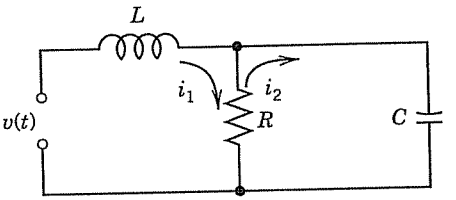


Fig. 150. Network in Problem 54

55. (Network) Set up the model of the network in Fig. 151 and find and graph the currents, assuming that the currents and the charge on the capacitor are 0 when the switch is closed at $t = 0$.

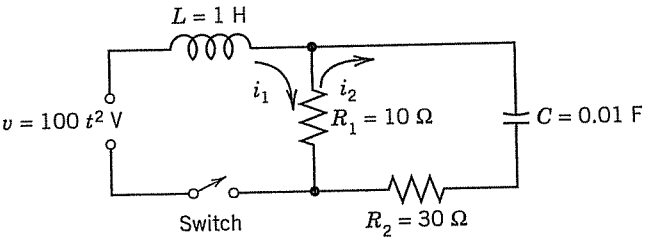


Fig. 151. Network in Problem 55

SUMMARY OF CHAPTER 6
Laplace Transforms

The main purpose of Laplace transforms is the solution of differential equations and systems of such equations, as well as corresponding initial value problems. The **Laplace transform** $F(s) = \mathcal{L}(f)$ of a function $f(t)$ is defined by

$$(1) \qquad F(s) = \mathcal{L}(f) = \int_0^\infty e^{-st} f(t)\, dt \qquad \text{(Sec. 6.1).}$$

This definition is motivated by the property that the differentiation of f with respect to t corresponds to the multiplication of the transform F by s; more precisely,

$$(2) \qquad \begin{aligned} \mathcal{L}(f') &= s\mathcal{L}(f) - f(0) \\ \mathcal{L}(f'') &= s^2\mathcal{L}(f) - sf(0) - f'(0) \end{aligned} \qquad \text{(Sec. 6.2)}$$

etc. Hence by taking the transform of a given differential equation

$$(3) \qquad y'' + ay' + by = r(t) \qquad (a, b \text{ constant})$$

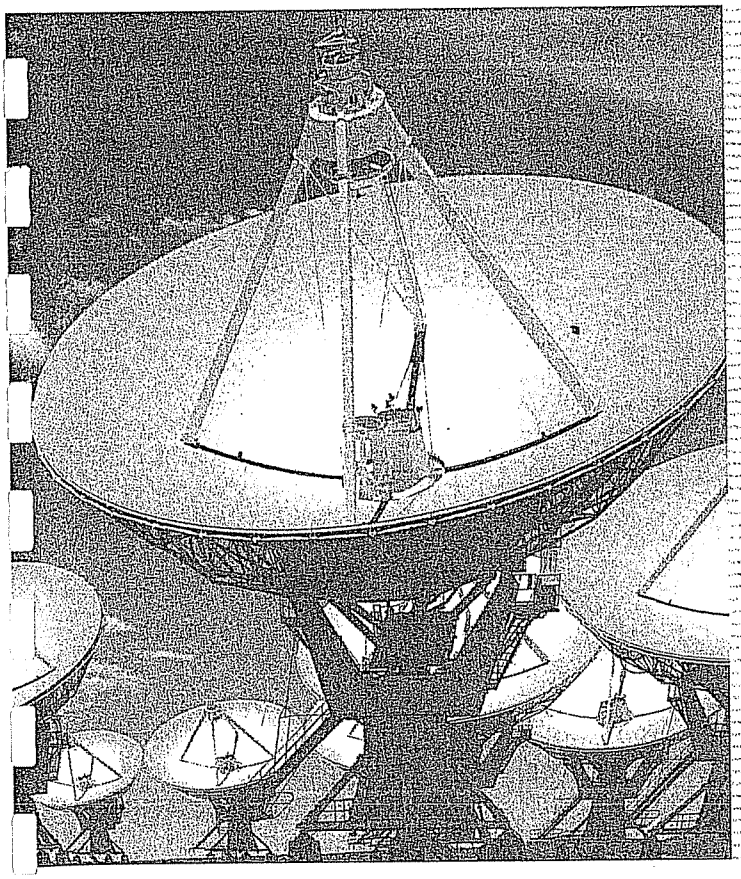
and writing $\mathcal{L}(y) = Y(s)$, we obtain the **subsidiary equation**

$$(4) \qquad (s^2 + as + b)Y = \mathcal{L}(r) + sf(0) + f'(0) + af(0).$$

Here, in obtaining the transform $\mathcal{L}(r)$ we can get help from the small table in Sec. 6.1 or the larger table in Sec. 6.9. This is the first step. In the second step we solve the subsidiary equation *algebraically* for $Y(s)$. In the third step we determine the **inverse transform** $y(t) = \mathcal{L}^{-1}(Y)$, that is, the solution of the problem. This is generally the hardest step, and in it we may again use one of those two tables. $Y(s)$ will often be a rational function, so that we can obtain the inverse $\mathcal{L}^{-1}(Y)$ by partial fraction reduction (Sec. 6.4) if we see no simpler way.

The Laplace method avoids the determination of a general solution of the homogeneous ODE, and we also need not determine values of arbitrary constants in a general solution from initial conditions; instead, we can insert the latter directly into (4). Two further facts account for the practical importance of the Laplace transform. First, it has some basic properties and resulting techniques that simplify the determination of transforms and inverses. The most important of these properties are listed in Sec. 6.8, together with references to the corresponding sections. More on the use of unit step functions and Dirac's delta can be found in Secs. 6.3 and 6.4, and more on convolution in Sec. 6.5. Second, due to these properties, the present method is particularly suitable for handling right sides $r(t)$ given by different expressions over different intervals of time, for instance, when $r(t)$ is a square wave or an impulse or of a form such as $r(t) = \cos t$ if $0 \leqq t \leqq 4\pi$ and 0 elsewhere.

The application of the Laplace transform to systems of ODEs is shown in Sec. 6.7. (The application to PDEs follows in Sec. 12.11.)



PART B

Linear Algebra. Vector Calculus

CHAPTER 7 **Linear Algebra: Matrices, Vectors, Determinants. Linear Systems**

CHAPTER 8 **Linear Algebra: Matrix Eigenvalue Problems**

CHAPTER 9 **Vector Differential Calculus. Grad, Div, Curl**

CHAPTER 10 **Vector Integral Calculus. Integral Theorems**

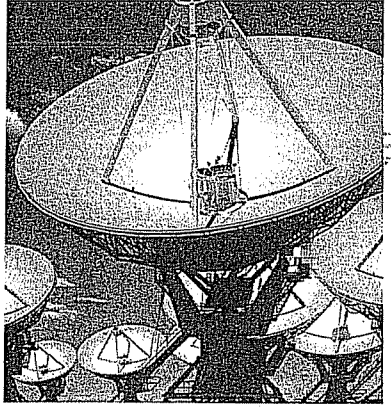
Linear algebra in Chaps. 7 and 8 consists of the theory and application of vectors and matrices, mainly related to linear systems of equations, eigenvalue problems, and linear transformations.

Linear algebra is of growing importance in engineering research and teaching because it forms a foundation of numeric methods (see Chaps. 20–22), and its main instruments, matrices, can hold enormous amounts of data—think of a net of millions of telephone connections—in a form readily accessible by the computer.

Linear analysis in Chaps. 9 and 10, usually called **vector calculus,** extends *differentiation* of functions of one variable to functions of several variables—this includes the vector differential operations grad, div, and curl. And it generalizes *integration* to integrals over curves, surfaces, and solids, with transformations of these integrals into one another, by the basic theorems of Gauss, Green, and Stokes (Chap. 10).

Software suitable for linear algebra (Lapack, Maple, Mathematica, Matlab) can be found in the list at the opening of Part E of the book if needed.

Numeric linear algebra (Chap. 20) ***can be studied directly after Chap. 7 or 8*** because Chap. 20 is independent of the other chapters in Part E on numerics.



CHAPTER 7

Linear Algebra: Matrices, Vectors, Determinants. Linear Systems

This is the first of two chapters on **linear algebra,** which concerns mainly systems of linear equations and linear transformations (to be discussed in this chapter) and eigenvalue problems (to follow in Chap. 8).

Systems of linear equations, briefly called **linear systems,** arise in electrical networks, mechanical frameworks, economic models, optimization problems, numerics for differential equations, as we shall see in Chaps. 21–23, and so on.

As main tools, linear algebra uses **matrices** (rectangular arrays of numbers or functions) and **vectors.** Calculations with matrices handle matrices as single objects, denote them by single letters, and calculate with them in a very compact form, almost as with numbers, so that matrix calculations constitute a powerful "mathematical shorthand".

Calculations with matrices and vectors are defined and explained in Secs. 7.1–7.2. Sections 7.3–7.8 center around linear systems, with a thorough discussion of Gauss elimination, the role of rank, the existence and uniqueness problem for solutions (Sec. 7.5), and matrix inversion. This also includes determinants (Cramer's rule) in Sec. 7.6 (for quick reference) and Sec. 7.7. Applications are considered throughout this chapter. The last section (Sec. 7.9) on vector spaces, inner product spaces, and linear transformations is more abstract. Eigenvalue problems follow in Chap. 8.

COMMENT. ***Numeric linear algebra (Secs. 20.1–20.5) can be studied immediately after this chapter.***

Prerequisite: None.
Sections that may be omitted in a short course: 7.5, 7.9.
References and Answers to Problems: App. 1 Part B, and App. 2.

7.1 Matrices, Vectors: Addition and Scalar Multiplication

In this section and the next one we introduce the basic concepts and rules of matrix and vector algebra. The main application to **linear systems** (systems of linear equations) begins in Sec. 7.3.

A **matrix** is a rectangular array of numbers (or functions) enclosed in brackets. These numbers (or functions) are called the **entries** (or sometimes the *elements*) of the matrix. For example,

$$(1)\qquad \begin{bmatrix} 0.3 & 1 & -5 \\ 0 & -0.2 & 16 \end{bmatrix}, \quad \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix},$$

$$\begin{bmatrix} e^{-x} & 2x^2 \\ e^{6x} & 4x \end{bmatrix}, \quad [a_1 \quad a_2 \quad a_3], \quad \begin{bmatrix} 4 \\ \frac{1}{2} \end{bmatrix}$$

are matrices. The first matrix has two **rows** (horizontal lines of entries) and three **columns** (vertical lines). The second and third matrices are **square matrices,** that is, each has as many rows as columns (3 and 2, respectively). The entries of the second matrix have two indices giving the location of the entry. The first index is the number of the row and the second is the number of the column in which the entry stands. Thus, a_{23} (read *a two three*) is in Row 2 and Column 3, etc. This notation is standard, regardless of whether a matrix is square or not.

Matrices having just a single row or column are called **vectors.** Thus the fourth matrix in (1) has just one row and is called a **row vector.** The last matrix in (1) has just one column and is called a **column vector.**

We shall see that matrices are practical in various applications for storing and processing data. As a first illustration let us consider two simple but typical examples.

EXAMPLE 1 **Linear Systems, a Major Application of Matrices**

In a system of linear equations, briefly called a **linear system,** such as

$$\begin{aligned} 4x_1 + 6x_2 + 9x_3 &= 6 \\ 6x_1 \qquad - 2x_3 &= 20 \\ 5x_1 - 8x_2 + x_3 &= 10 \end{aligned}$$

the coefficients of the **unknowns** x_1, x_2, x_3 are the entries of the **coefficient matrix,** call it **A,**

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

The matrix

$$\tilde{\mathbf{A}} = \begin{bmatrix} 4 & 6 & 9 & 6 \\ 6 & 0 & -2 & 20 \\ 5 & -8 & 1 & 10 \end{bmatrix}$$

is obtained by augmenting **A** by the right sides of the linear system and is called the **augmented matrix** of the system. In **A** the coefficients of the system are displayed in the pattern of the equations. That is, their position in **A** corresponds to that in the system when written as shown. The same is true for $\tilde{\mathbf{A}}$.

We shall see that the augmented matrix $\tilde{\mathbf{A}}$ contains all the information about the solutions of a system, so that we can solve a system just by calculations on its augmented matrix. We shall discuss this in great detail, beginning in Sec. 7.3. Meanwhile you may verify by substitution that the solution is $x_1 = 3$, $x_2 = \frac{1}{2}$, $x_3 = -1$.

The notation x_1, x_2, x_3 for the unknowns is practical but not essential; we could choose x, y, z or some other letters. ■

EXAMPLE 2 Sales Figures in Matrix Form

Sales figures for three products I, II, III in a store on Monday (M), Tuesday (T), · · · may for each week be arranged in a matrix

$$\mathbf{A} = \begin{array}{c} \begin{matrix} \text{M} & \text{T} & \text{W} & \text{Th} & \text{F} & \text{S} \end{matrix} \\ \begin{bmatrix} 400 & 330 & 810 & 0 & 210 & 470 \\ 0 & 120 & 780 & 500 & 500 & 960 \\ 100 & 0 & 0 & 270 & 430 & 780 \end{bmatrix} \end{array} \begin{matrix} \\ \text{I} \\ \text{II} \\ \text{III} \end{matrix}$$

If the company has ten stores, we can set up ten such matrices, one for each store. Then by adding corresponding entries of these matrices we can get a matrix showing the total sales of each product on each day. Can you think of other data for which matrices are feasible? For instance, in transportation or storage problems? Or in recording phone calls, or in listing distances in a network of roads? ■

General Concepts and Notations

We shall denote matrices by capital boldface letters **A, B, C,** · · · , or by writing the general entry in brackets; thus $\mathbf{A} = [a_{jk}]$, and so on. By an $\boldsymbol{m \times n}$ **matrix** (read *m by n matrix*) we mean a matrix with m rows and n columns—rows come always first! $m \times n$ is called the **size** of the matrix. Thus an $m \times n$ matrix is of the form

(2)
$$\mathbf{A} = [a_{jk}] = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}.$$

The matrices in (1) are of sizes 2×3, 3×3, 2×2, 1×3, and 2×1, respectively.

Each entry in (2) has two subscripts. The first is the *row number* and the second is the *column number.* Thus a_{21} is the entry in Row 2 and Column 1.

If $m = n$, we call **A** an $n \times n$ **square matrix.** Then its diagonal containing the entries $a_{11}, a_{22}, \cdots, a_{nn}$ is called the **main diagonal** of **A.** Thus the main diagonals of the two square matrices in (1) are a_{11}, a_{22}, a_{33} and e^{-x}, $4x$, respectively.

Square matrices are particularly important, as we shall see. A matrix that is not square is called a **rectangular matrix.**

Vectors

A **vector** is a matrix with only one row or column. Its entries are called the **components** of the vector. We shall denote vectors by ***lowercase*** boldface letters **a, b,** · · · or by its general component in brackets, $\mathbf{a} = [a_j]$, and so on. Our special vectors in (1) suggest that a (general) **row vector** is of the form

$$\mathbf{a} = [a_1 \quad a_2 \quad \cdots, \quad a_n]. \qquad \text{For instance,} \qquad \mathbf{a} = [-2 \quad 5 \quad 0.8 \quad 0 \quad 1].$$

A **column vector** is of the form

$$\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}. \qquad \text{For instance,} \qquad \mathbf{b} = \begin{bmatrix} 4 \\ 0 \\ -7 \end{bmatrix}.$$

Matrix Addition and Scalar Multiplication

What makes matrices and vectors really useful and particularly suitable for computers is the fact that we can calculate with them almost as easily as with numbers. Indeed, we now introduce rules for addition and for scalar multiplication (multiplication by numbers) that were suggested by practical applications. (Multiplication of matrices by matrices follows in the next section.) We first need the concept of equality.

DEFINITION

Equality of Matrices

Two matrices $\mathbf{A} = [a_{jk}]$ and $\mathbf{B} = [b_{jk}]$ are **equal,** written $\mathbf{A} = \mathbf{B}$, if and only if they have the same size and the corresponding entries are equal, that is, $a_{11} = b_{11}$, $a_{12} = b_{12}$, and so on. Matrices that are not equal are called **different.** Thus, matrices of different sizes are always different.

EXAMPLE 3 **Equality of Matrices**

Let

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} 4 & 0 \\ 3 & -1 \end{bmatrix}.$$

Then

$$\mathbf{A} = \mathbf{B} \quad \text{if and only if} \quad \begin{aligned} a_{11} &= 4, & a_{12} &= 0, \\ a_{21} &= 3, & a_{22} &= -1. \end{aligned}$$

The following matrices are all different. Explain!

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

DEFINITION

Addition of Matrices

The **sum** of two matrices $\mathbf{A} = [a_{jk}]$ and $\mathbf{B} = [b_{jk}]$ ***of the same size*** is written $\mathbf{A} + \mathbf{B}$ and has the entries $a_{jk} + b_{jk}$ obtained by adding the corresponding entries of $\mathbf{A}$ and $\mathbf{B}$. Matrices of different sizes cannot be added.

As a special case, the **sum a + b** of two row vectors or two column vectors, which must have the same number of components, is obtained by adding the corresponding components.

EXAMPLE 4 **Addition of Matrices and Vectors**

$$\text{If} \quad \mathbf{A} = \begin{bmatrix} -4 & 6 & 3 \\ 0 & 1 & 2 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} 5 & -1 & 0 \\ 3 & 1 & 0 \end{bmatrix}, \quad \text{then} \quad \mathbf{A} + \mathbf{B} = \begin{bmatrix} 1 & 5 & 3 \\ 3 & 2 & 2 \end{bmatrix}.$$

$\mathbf{A}$ in Example 3 and our present $\mathbf{A}$ cannot be added. If $\mathbf{a} = [5 \quad 7 \quad 2]$ and $\mathbf{b} = [-6 \quad 2 \quad 0]$, then $\mathbf{a} + \mathbf{b} = [-1 \quad 9 \quad 2]$.

An application of matrix addition was suggested in Example 2. Many others will follow. ■

DEFINITION

Scalar Multiplication (Multiplication by a Number)

The **product** of any $m \times n$ matrix $\mathbf{A} = [a_{jk}]$ and any **scalar** c (number c) is written $c\mathbf{A}$ and is the $m \times n$ matrix $c\mathbf{A} = [ca_{jk}]$ obtained by multiplying each entry of $\mathbf{A}$ by c.

Here $(-1)\mathbf{A}$ is simply written $-\mathbf{A}$ and is called the **negative** of $\mathbf{A}$. Similarly, $(-k)\mathbf{A}$ is written $-k\mathbf{A}$. Also, $\mathbf{A} + (-\mathbf{B})$ is written $\mathbf{A} - \mathbf{B}$ and is called the **difference** of $\mathbf{A}$ and $\mathbf{B}$ (which must have the same size!).

EXAMPLE 5 **Scalar Multiplication**

$$\text{If} \quad \mathbf{A} = \begin{bmatrix} 2.7 & -1.8 \\ 0 & 0.9 \\ 9.0 & -4.5 \end{bmatrix}, \quad \text{then} \quad -\mathbf{A} = \begin{bmatrix} -2.7 & 1.8 \\ 0 & -0.9 \\ -9.0 & 4.5 \end{bmatrix}, \quad \frac{10}{9}\mathbf{A} = \begin{bmatrix} 3 & -2 \\ 0 & 1 \\ 10 & -5 \end{bmatrix}, \quad 0\mathbf{A} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

If a matrix $\mathbf{B}$ shows the distances between some cities in miles, $1.609\mathbf{B}$ gives these distances in kilometers. ■

Rules for Matrix Addition and Scalar Multiplication. From the familiar laws for the addition of numbers we obtain similar laws for the addition of matrices of the same size $m \times n$, namely,

(3)
$$\begin{aligned} &\text{(a)} \quad \mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A} \\ &\text{(b)} \quad (\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C}) \quad \text{(written } \mathbf{A} + \mathbf{B} + \mathbf{C}\text{)} \\ &\text{(c)} \quad \mathbf{A} + \mathbf{0} = \mathbf{A} \\ &\text{(d)} \quad \mathbf{A} + (-\mathbf{A}) = \mathbf{0}. \end{aligned}$$

Here $\mathbf{0}$ denotes the **zero matrix** (of size $m \times n$), that is, the $m \times n$ matrix with all entries zero. (The last matrix in Example 5 is a zero matrix.)

Hence matrix addition is *commutative* and *associative* [by (3a) and (3b)].

Similarly, for scalar multiplication we obtain the rules

(4)
$$\begin{aligned} &\text{(a)} \quad c(\mathbf{A} + \mathbf{B}) = c\mathbf{A} + c\mathbf{B} \\ &\text{(b)} \quad (c + k)\mathbf{A} = c\mathbf{A} + k\mathbf{A} \\ &\text{(c)} \quad c(k\mathbf{A}) = (ck)\mathbf{A} \quad \text{(written } ck\mathbf{A}\text{)} \\ &\text{(d)} \quad 1\mathbf{A} = \mathbf{A}. \end{aligned}$$

PROBLEM SET 7.1

1–8 ADDITION AND SCALAR MULTIPLICATION OF MATRICES AND VECTORS

Let

$$\mathbf{A} = \begin{bmatrix} 3 & 0 & 4 \\ -1 & 2 & 2 \\ 6 & 5 & -4 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 & -5 & -3 \\ -5 & 2 & 4 \\ -3 & 4 & 0 \end{bmatrix},$$

$$\mathbf{C} = \begin{bmatrix} 0 & 2 \\ 2 & 4 \\ 1 & 3 \end{bmatrix}, \quad \mathbf{D} = \begin{bmatrix} 6 & 1 \\ -4 & 7 \\ -8 & 3 \end{bmatrix},$$

$$\mathbf{u} = \begin{bmatrix} 2 \\ 0 \\ -1 \end{bmatrix}, \quad \mathbf{v} = \begin{bmatrix} -4.5 \\ 0.8 \\ 1.2 \end{bmatrix}.$$

Find the following expressions or give reasons why they are undefined.

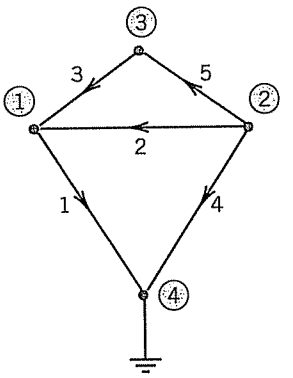
1. $\mathbf{C} + \mathbf{D}$, $\mathbf{D} + \mathbf{C}$, $6(\mathbf{D} - \mathbf{C})$, $6\mathbf{C} - 6\mathbf{D}$
2. $4\mathbf{C}$, $2\mathbf{D}$, $4\mathbf{C} + 2\mathbf{D}$, $8\mathbf{C} - 0\mathbf{D}$
3. $\mathbf{A} + \mathbf{C} - \mathbf{D}$, $\mathbf{C} - \mathbf{D}$, $\mathbf{D} - \mathbf{C}$, $\mathbf{B} + 2\mathbf{C} + 4\mathbf{D}$
4. $2(\mathbf{A} + \mathbf{B})$, $2\mathbf{A} + 2\mathbf{B}$, $5\mathbf{A} - \frac{1}{2}\mathbf{B}$, $\mathbf{A} + \mathbf{B} + \mathbf{C}$
5. $3\mathbf{C} - 8\mathbf{D}$, $4(3\mathbf{A})$, $(4 \cdot 3)\mathbf{A}$, $\mathbf{B} - \frac{1}{10}\mathbf{A}$
6. $5\mathbf{A} - 3\mathbf{C}$, $\mathbf{A} - \mathbf{B} + \mathbf{D}$, $4(\mathbf{B} - 6\mathbf{A})$, $4\mathbf{B} - 24\mathbf{A}$
7. $33\mathbf{u}$, $4\mathbf{v} + 9\mathbf{u}$, $4(\mathbf{v} + 2.25\mathbf{u})$, $\mathbf{u} - \mathbf{v}$
8. $\mathbf{A} + \mathbf{u}$, $12\mathbf{u} + 10\mathbf{v}$, $0(\mathbf{B} - \mathbf{v})$, $0\mathbf{B} + \mathbf{u}$

9. **(Linear system)** Write down a linear system (as in Example 1) whose augmented matrix is the matrix **B** in this problem set.
10. **(Scalar multiplication)** The matrix **A** in Example 2 shows the numbers of items sold. Find the matrix showing the number of units sold if a unit consists of (a) 5 items, (b) 10 items?
11. **(Double subscript notation)** Write the entries of **A** in Example 2 in the general notation shown in (2).
12. **(Sizes, diagonal)** What sizes do **A**, **B**, **C**, **D**, **u**, **v** in this problem set have? What are the main diagonals of **A** and **B**, and what about **C**?
13. **(Equality)** Give reasons why the five matrices in Example 3 are different.
14. **(Addition of vectors)** Can you add (a) row vectors whose numbers of components are different, (b) a row and a column vector with the same number of components, (c) a vector and a scalar?
15. **(General rules)** Prove (3) and (4) for general 3×2 matrices and scalars c and k.
16. **TEAM PROJECT. Matrices in Modeling Networks.** Matrices have various applications, as we shall see, in a form that these problems can be efficiently handled on the computer. For instance, they can be used to characterize connections in electrical networks, in nets of roads, in production processes, etc., as follows.

(a) Nodal incidence matrix. The network in Fig. 152 consists of 5 *branches* or *edges* (connections, numbered $1, 2, \cdots, 5$) and 4 *nodes* (points where two or more branches come together), with one node being grounded. We number the nodes and branches and give each branch a direction shown by an arrow. This we do arbitrarily. The network can now be described by a *"nodal incidence matrix"* $\mathbf{A} = [a_{jk}]$, where

$$a_{jk} = \begin{cases} +1 \text{ if branch } k \text{ leaves node } (j) \\ -1 \text{ if branch } k \text{ enters node } (j) \\ 0 \text{ if branch } k \text{ does not touch } (j). \end{cases}$$

Show that for the network in Fig. 152 the matrix **A** has the given form



Branch	1	2	3	4	5
Node ①	1	−1	−1	0	0
Node ②	0	1	0	1	1
Node ③	0	0	1	0	−1
Node ④	−1	0	0	−1	0

Fig. 152. Network and nodal incidence matrix in Team Project 16(a)

(b) Find the nodal incidence matrices of the networks in Fig. 153.

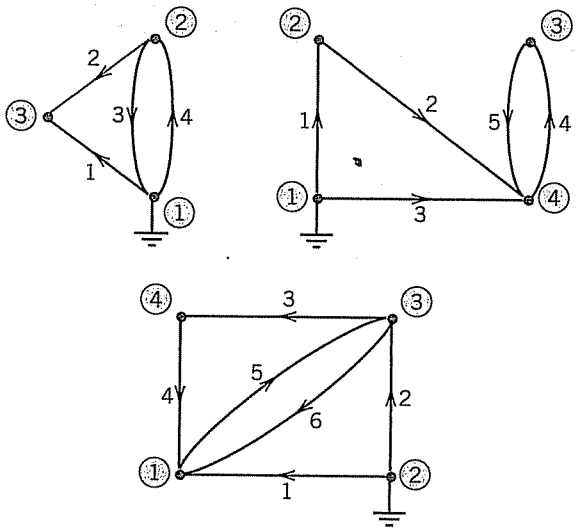


Fig. 153. Networks in Team Project 16(b)

(c) Graph the three networks corresponding to the nodal incidence matrices

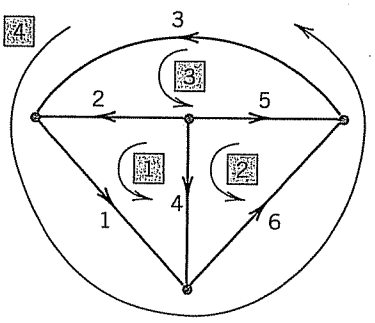
$$\begin{bmatrix} 1 & -1 & 1 & -1 \\ 0 & 0 & -1 & 1 \\ -1 & 1 & 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 1 \\ -1 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix},$$

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

(d) Mesh incidence matrix. A network can also be characterized by the *mesh incidence matrix* $\mathbf{M} = [m_{jk}]$, where

$$m_{jk} = \begin{cases} +1 \text{ if branch } k \text{ is in mesh } \boxed{j} \\ \text{and has the same orientation} \\ -1 \text{ if branch } k \text{ is in mesh } \boxed{j} \\ \text{and has the opposite orientation} \\ 0 \text{ if branch } k \text{ is not in mesh } \boxed{j} \end{cases}$$

and a mesh is a loop with no branch in its interior (or in its exterior). Here, the meshes are numbered and directed (oriented) in an arbitrary fashion. Show that in Fig. 154 the matrix $\mathbf{M}$ corresponds to the given figure, where Row 1 corresponds to mesh 1, etc.



$$\mathbf{M} = \begin{bmatrix} 1 & 1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 1 \\ 0 & -1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}$$

Fig. 154. Network and matrix $\mathbf{M}$ in Team Project 16(d)

(e) Number the nodes in Fig. 154 from left to right 1, 2, 3 and the low node by 4. Find the corresponding nodal incidence matrix.

7.2 Matrix Multiplication

Matrix multiplication means multiplication of matrices by matrices. This is the last algebraic operation to be defined (except for *transposition,* which is of lesser importance). Now matrices are *added* by *adding* corresponding entries. In *multiplication,* do we *multiply* corresponding entries? The answer is no. Why not? Such an operation would not be of much use in applications. The standard definition of multiplication looks artificial, but will be fully motivated later in this section by the use of matrices in "linear transformations," by which this multiplication is suggested.

DEFINITION

Multiplication of a Matrix by a Matrix

The **product** $\mathbf{C} = \mathbf{AB}$ (in this order) of an $m \times n$ matrix $\mathbf{A} = [a_{jk}]$ times an $r \times p$ matrix $\mathbf{B} = [b_{jk}]$ is defined if and only if $r = n$ and is then the $m \times p$ matrix $\mathbf{C} = [c_{jk}]$ with entries

$$(1) \qquad c_{jk} = \sum_{l=1}^{n} a_{jl}b_{lk} = a_{j1}b_{1k} + a_{j2}b_{2k} + \cdots + a_{jn}b_{nk} \qquad \begin{matrix} j = 1, \cdots, m \\ k = 1, \cdots, p. \end{matrix}$$

The condition $r = n$ means that the second factor, **B,** must have as many rows as the first factor has columns, namely n. As a diagram of sizes (denoted as shown):

$$\begin{matrix} \mathbf{A} & \mathbf{B} & = & \mathbf{C} \\ [m \times n] & [n \times r] & = & [m \times r]. \end{matrix}$$

c_{jk} in (1) is obtained by multiplying each entry in the jth row of **A** by the corresponding entry in the kth column of **B** and then adding these n products. For instance, $c_{21} = a_{21}b_{11} + a_{22}b_{21} + \cdots + a_{2n}b_{n1}$, and so on. One calls this briefly a ***"multiplication of rows into columns."*** See the illustration in Fig. 155, where $n = 3$.

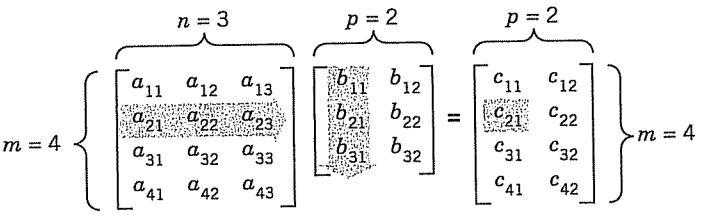


Fig. 155. Notations in a product **AB** = **C**

EXAMPLE 1 **Matrix Multiplication**

$$\mathbf{AB} = \begin{bmatrix} 3 & 5 & -1 \\ 4 & 0 & 2 \\ -6 & -3 & 2 \end{bmatrix} \begin{bmatrix} 2 & -2 & 3 & 1 \\ 5 & 0 & 7 & 8 \\ 9 & -4 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 22 & -2 & 43 & 42 \\ 26 & -16 & 14 & 6 \\ -9 & 4 & -37 & -28 \end{bmatrix}$$

Here $c_{11} = 3 \cdot 2 + 5 \cdot 5 + (-1) \cdot 9 = 22$, and so on. The entry in the box is $c_{23} = 4 \cdot 3 + 0 \cdot 7 + 2 \cdot 1 = 14$. The product **BA** is not defined. ■

EXAMPLE 2 **Multiplication of a Matrix and a Vector**

$$\begin{bmatrix} 4 & 2 \\ 1 & 8 \end{bmatrix} \begin{bmatrix} 3 \\ 5 \end{bmatrix} = \begin{bmatrix} 4 \cdot 3 + 2 \cdot 5 \\ 1 \cdot 3 + 8 \cdot 5 \end{bmatrix} = \begin{bmatrix} 22 \\ 43 \end{bmatrix} \quad \text{whereas} \quad \begin{bmatrix} 3 \\ 5 \end{bmatrix} \begin{bmatrix} 4 & 2 \\ 1 & 8 \end{bmatrix} \quad \text{is undefined.} \blacksquare$$

EXAMPLE 3 **Products of Row and Column Vectors**

$$[3 \quad 6 \quad 1] \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix} = [19], \qquad \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix} [3 \quad 6 \quad 1] = \begin{bmatrix} 3 & 6 & 1 \\ 6 & 12 & 2 \\ 12 & 24 & 4 \end{bmatrix}. \blacksquare$$

EXAMPLE 4 **CAUTION! Matrix Multiplication Is Not Commutative, $AB \neq BA$ in General**

This is illustrated by Examples 1 and 2, where one of the two products is not even defined, and by Example 3, where the two products have different sizes. But it also holds for square matrices. For instance,

$$\begin{bmatrix} 1 & 1 \\ 100 & 100 \end{bmatrix}\begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \quad \text{but} \quad \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 100 & 100 \end{bmatrix} = \begin{bmatrix} 99 & 99 \\ -99 & -99 \end{bmatrix}.$$

It is interesting that this also shows that $\mathbf{AB} = \mathbf{0}$ does *not* necessarily imply $\mathbf{BA} = \mathbf{0}$ or $\mathbf{A} = \mathbf{0}$ or $\mathbf{B} = \mathbf{0}$. We shall discuss this further in Sec. 7.8, along with reasons when this happens. ▩

Our examples show that the ***order of factors*** in matrix products ***must always be observed very carefully.*** Otherwise matrix multiplication satisfies rules similar to those for numbers, namely.

(2)
$$\begin{aligned} &\text{(a)} \quad (k\mathbf{A})\mathbf{B} = k(\mathbf{AB}) = \mathbf{A}(k\mathbf{B}) && \textit{written } k\mathbf{AB} \textit{ or } \mathbf{A}k\mathbf{B} \\ &\text{(b)} \quad \mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C} && \textit{written } \mathbf{ABC} \\ &\text{(c)} \quad (\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC} \\ &\text{(d)} \quad \mathbf{C}(\mathbf{A} + \mathbf{B}) = \mathbf{CA} + \mathbf{CB} \end{aligned}$$

provided **A, B,** and **C** are such that the expressions on the left are defined; here, k is any scalar. (2b) is called the **associative law.** (2c) and (2d) are called the **distributive laws.**

Since matrix multiplication is a multiplication of rows into columns, we can write the defining formula (1) more compactly as

(3)
$$c_{jk} = \mathbf{a}_j\mathbf{b}_k, \qquad j = 1, \cdots, m; \quad k = 1, \cdots, p,$$

where $\mathbf{a}_j$ is the jth row vector of $\mathbf{A}$ and $\mathbf{b}_k$ is the kth column vector of $\mathbf{B}$, so that in agreement with (1),

$$\mathbf{a}_j\mathbf{b}_k = [a_{j1} \quad a_{j2} \quad \cdots \quad a_{jn}]\begin{bmatrix} b_{1k} \\ \vdots \\ b_{nk} \end{bmatrix} = a_{j1}b_{1k} + a_{j2}b_{2k} + \cdots + a_{jn}b_{nk}.$$

EXAMPLE 5 **Product in Terms of Row and Column Vectors**

If $\mathbf{A} = [a_{jk}]$ is of size 3×3 and $\mathbf{B} = [b_{jk}]$ is of size 3×4, then

(4)
$$\mathbf{AB} = \begin{bmatrix} \mathbf{a}_1\mathbf{b}_1 & \mathbf{a}_1\mathbf{b}_2 & \mathbf{a}_1\mathbf{b}_3 & \mathbf{a}_1\mathbf{b}_4 \\ \mathbf{a}_2\mathbf{b}_1 & \mathbf{a}_2\mathbf{b}_2 & \mathbf{a}_2\mathbf{b}_3 & \mathbf{a}_2\mathbf{b}_4 \\ \mathbf{a}_3\mathbf{b}_1 & \mathbf{a}_3\mathbf{b}_2 & \mathbf{a}_3\mathbf{b}_3 & \mathbf{a}_3\mathbf{b}_4 \end{bmatrix}.$$

Taking $\mathbf{a}_1 = [3 \quad 5 \quad -1]$, $\mathbf{a}_2 = [4 \quad 0 \quad 2]$, etc., verify (4) for the product in Example 1. ▩

Parallel processing of products on the computer is facilitated by a variant of (3) for computing $\mathbf{C} = \mathbf{AB}$, which is used by standard algorithms (such as in Lapack). In this method, **A** is used as given, **B** is taken in terms of its column vectors, and the product is computed columnwise; thus,

(5)
$$\mathbf{AB} = \mathbf{A}[\mathbf{b}_1 \quad \mathbf{b}_2 \quad \cdots \quad \mathbf{b}_p] = [\mathbf{Ab}_1 \quad \mathbf{Ab}_2 \quad \cdots \quad \mathbf{Ab}_p].$$

Columns of **B** are then assigned to different processors (individually or several to each processor), which simultaneously compute the columns of the product matrix $\mathbf{Ab}_1$, $\mathbf{Ab}_2$, etc.

EXAMPLE 6 Computing Products Columnwise by (5)

To obtain

$$\mathbf{AB} = \begin{bmatrix} 4 & 1 \\ -5 & 2 \end{bmatrix} \begin{bmatrix} 3 & 0 & 7 \\ -1 & 4 & 6 \end{bmatrix} = \begin{bmatrix} 11 & 4 & 34 \\ -17 & 8 & -23 \end{bmatrix}$$

from (5), calculate the columns

$$\begin{bmatrix} 4 & 1 \\ -5 & 2 \end{bmatrix} \begin{bmatrix} 3 \\ -1 \end{bmatrix} = \begin{bmatrix} 11 \\ -17 \end{bmatrix}, \quad \begin{bmatrix} 4 & 1 \\ -5 & 2 \end{bmatrix} \begin{bmatrix} 0 \\ 4 \end{bmatrix} = \begin{bmatrix} 4 \\ 8 \end{bmatrix}, \quad \begin{bmatrix} 4 & 1 \\ -5 & 2 \end{bmatrix} \begin{bmatrix} 7 \\ 6 \end{bmatrix} = \begin{bmatrix} 34 \\ -23 \end{bmatrix}$$

of **AB** and then write them as a single matrix, as shown in the first formula on the right. ■

Motivation of Multiplication by Linear Transformations

Let us now motivate the "unnatural" matrix multiplication by its use in **linear transformations.** For $n = 2$ variables these transformations are of the form

$$\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 \\ y_2 &= a_{21}x_1 + a_{22}x_2 \end{aligned} \tag{6*}$$

and suffice to explain the idea. (For general n they will be discussed in Sec. 7.9.) For instance, (6*) may relate an x_1x_2-coordinate system to a y_1y_2-coordinate system in the plane. In vectorial form we can write (6*) as

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \mathbf{Ax} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 \\ a_{21}x_1 + a_{22}x_2 \end{bmatrix}. \tag{6}$$

Now suppose further that the x_1x_2-system is related to a w_1w_2-system by another linear transformation, say,

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \mathbf{Bw} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} b_{11}w_1 + b_{12}w_2 \\ b_{21}w_1 + b_{22}w_2 \end{bmatrix}. \tag{7}$$

Then the y_1y_2-system is related to the w_1w_2-system indirectly via the x_1x_2-system, and we wish to express this relation directly. Substitution will show that this direct relation is a linear transformation, too, say,

$$\mathbf{y} = \mathbf{Cw} = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} c_{11}w_1 + c_{12}w_2 \\ c_{21}w_1 + c_{22}w_2 \end{bmatrix}. \tag{8}$$

Indeed, substituting (7) into (6), we obtain

$$\begin{aligned} y_1 &= a_{11}(b_{11}w_1 + b_{12}w_2) + a_{12}(b_{21}w_1 + b_{22}w_2) \\ &= (a_{11}b_{11} + a_{12}b_{21})w_1 + (a_{11}b_{12} + a_{12}b_{22})w_2 \\ y_2 &= a_{21}(b_{11}w_1 + b_{12}w_2) + a_{22}(b_{21}w_1 + b_{22}w_2) \\ &= (a_{21}b_{11} + a_{22}b_{21})w_1 + (a_{21}b_{12} + a_{22}b_{22})w_2. \end{aligned}$$

Comparing this with (8), we see that

$$c_{11} = a_{11}b_{11} + a_{12}b_{21} \qquad c_{12} = a_{11}b_{12} + a_{12}b_{22}$$
$$c_{21} = a_{21}b_{11} + a_{22}b_{21} \qquad c_{22} = a_{21}b_{12} + a_{22}b_{22}.$$

This proves that $\mathbf{C} = \mathbf{AB}$ with the product defined as in (1). For larger matrix sizes the idea and result are exactly the same. Only the number of variables changes. We then have m variables y and n variables x and p variables w. The matrices **A, B,** and $\mathbf{C} = \mathbf{AB}$ then have sizes $m \times n$, $n \times p$, and $m \times p$, respectively. And the requirement that **C** be the product **AB** leads to formula (1) in its general form. *This motivates matrix multiplication completely.*

Transposition

Transposition provides a transition from row vectors to column vectors and conversely. More generally, it gives us a choice to work either with a matrix or with its transpose, whatever will be more practical in a specific situation.

DEFINITION

Transposition of Matrices and Vectors

The transpose of an $m \times n$ matrix $\mathbf{A} = [a_{jk}]$ is the $n \times m$ matrix $\mathbf{A}^T$ (read *A transpose*) that has the first *row* of **A** as its first *column,* the second *row* of **A** as its second *column,* and so on. Thus the transpose of **A** in (2) is $\mathbf{A}^T = [a_{kj}]$, written out

$$\mathbf{A}^T = [a_{kj}] = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \cdot & \cdot & \cdots & \cdot \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{bmatrix}. \tag{9}$$

As a special case, transposition converts row vectors to column vectors and conversely.

EXAMPLE 7 **Transposition of Matrices and Vectors**

If $\mathbf{A} = \begin{bmatrix} 5 & -8 & 1 \\ 4 & 0 & 0 \end{bmatrix}$, then $\mathbf{A}^T = \begin{bmatrix} 5 & 4 \\ -8 & 0 \\ 1 & 0 \end{bmatrix}$.

A little more compactly, we can write

$$\begin{bmatrix} 5 & -8 & 1 \\ 4 & 0 & 0 \end{bmatrix}^T = \begin{bmatrix} 5 & 4 \\ -8 & 0 \\ 1 & 0 \end{bmatrix}, \qquad \begin{bmatrix} 3 & 0 \\ 8 & -1 \end{bmatrix}^T = \begin{bmatrix} 3 & 8 \\ 0 & -1 \end{bmatrix}$$

$$[6 \quad 2 \quad 3]^T = \begin{bmatrix} 6 \\ 2 \\ 3 \end{bmatrix}, \qquad \begin{bmatrix} 6 \\ 2 \\ 3 \end{bmatrix}^T = [6 \quad 2 \quad 3].$$

Note that for a square matrix, the transpose is obtained by interchanging entries that are symmetrically positioned with respect to the main diagonal, e.g., a_{12} and a_{21}, and so on. ▩

Rules for transposition are

$$\textbf{(10)}\qquad \begin{aligned} &\text{(a)} && (\mathbf{A}^T)^T = \mathbf{A} \\ &\text{(b)} && (\mathbf{A}+\mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T \\ &\text{(c)} && (c\mathbf{A})^T = c\mathbf{A}^T \\ &\text{(d)} && (\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T. \end{aligned}$$

CAUTION! Note that in (10d) the transposed matrices are ***in reversed order.*** We leave the proofs to the student. (See Prob. 22.)

Special Matrices

Certain kinds of matrices will occur quite frequently in our work, and we now list the most important ones of them.

Symmetric and Skew-Symmetric Matrices. Transposition gives rise to two useful classes of matrices, as follows. *Symmetric matrices* and *skew-symmetric matrices* are ***square*** matrices whose transpose equals the matrix itself or minus the matrix, respectively:

$$\textbf{(11)}\qquad \underbrace{\mathbf{A}^T = \mathbf{A} \quad (\text{thus } a_{kj} = a_{jk}),}_{\text{Symmetric Matrix}} \qquad \underbrace{\mathbf{A}^T = -\mathbf{A} \quad (\text{thus } a_{kj} = -a_{jk}, \text{ hence } a_{jj} = 0).}_{\text{Skew-Symmetric Matrix}}$$

EXAMPLE 8 **Symmetric and Skew-Symmetric Matrices**

$$\mathbf{A} = \begin{bmatrix} 20 & 120 & 200 \\ 120 & 10 & 150 \\ 200 & 150 & 30 \end{bmatrix} \text{ is symmetric, and } \quad \mathbf{B} = \begin{bmatrix} 0 & 1 & -3 \\ -1 & 0 & -2 \\ 3 & 2 & 0 \end{bmatrix} \text{ is skew-symmetric.}$$

For instance, if a company has three building supply centers C_1, C_2, C_3, then $\mathbf{A}$ could show costs, say, a_{jj} for handling 1000 bags of cement on center C_j, and a_{jk} $(j \neq k)$ the cost of shipping 1000 bags from C_j to C_k. Clearly, $a_{jk} = a_{kj}$ because shipping in the opposite direction will usually cost the same.

Symmetric matrices have several general properties which make them important. This will be seen as we proceed. ■

Triangular Matrices. **Upper triangular matrices** are square matrices that can have nonzero entries only on and ***above*** the main diagonal, whereas any entry below the diagonal must be zero. Similarly, **lower triangular matrices** can have nonzero entries only on and ***below*** the main diagonal. Any entry on the main diagonal of a triangular matrix may be zero or not.

EXAMPLE 9 **Upper and Lower Triangular Matrices**

$$\underbrace{\begin{bmatrix} 1 & 3 \\ 0 & 2 \end{bmatrix}, \quad \begin{bmatrix} 1 & 4 & 2 \\ 0 & 3 & 2 \\ 0 & 0 & 6 \end{bmatrix}}_{\text{Upper triangular}}, \qquad \underbrace{\begin{bmatrix} 2 & 0 & 0 \\ 8 & -1 & 0 \\ 7 & 6 & 8 \end{bmatrix}, \quad \begin{bmatrix} 3 & 0 & 0 & 0 \\ 9 & -3 & 0 & 0 \\ 1 & 0 & 2 & 0 \\ 1 & 9 & 3 & 6 \end{bmatrix}}_{\text{Lower triangular}}.$$

■

Diagonal Matrices. These are square matrices that can have nonzero entries only on the main diagonal. Any entry above or below the main diagonal must be zero.

If all the diagonal entries of a diagonal matrix **S** are equal, say, c, we call **S** a **scalar matrix** because multiplication of any square matrix **A** of the same size by **S** has the same effect as the multiplication by a scalar, that is,

$$\mathbf{AS} = \mathbf{SA} = c\mathbf{A}. \tag{12}$$

In particular, a scalar matrix whose entries on the main diagonal are all 1 is called a **unit matrix** (or **identity matrix**) and is denoted by $\mathbf{I}_n$ or simply by **I.** For **I,** formula (12) becomes

$$\mathbf{AI} = \mathbf{IA} = \mathbf{A}. \tag{13}$$

EXAMPLE 10 **Diagonal Matrix D. Scalar Matrix S. Unit Matrix I**

$$\mathbf{D} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & -3 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} c & 0 & 0 \\ 0 & c & 0 \\ 0 & 0 & c \end{bmatrix}, \quad \mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Applications of Matrix Multiplication

Matrix multiplication will play a crucial role in connection with linear systems of equations, beginning in the next section. For the time being we mention some other simple applications that need no lengthy explanations.

EXAMPLE 11 **Computer Production. Matrix Times Matrix**

Supercomp Ltd produces two computer models PC1086 and PC1186. The matrix **A** shows the cost per computer (in thousands of dollars) and **B** the production figures for the year 2005 (in multiples of 10000 units.) Find a matrix **C** that shows the shareholders the cost per quarter (in millions of dollars) for raw material, labor, and miscellaneous.

$$\mathbf{A} = \begin{array}{c} \\ \text{Raw Components} \\ \text{Labor} \\ \text{Miscellaneous} \end{array} \begin{array}{c} \begin{array}{cc} \text{PC1086} & \text{PC1186} \end{array} \\ \begin{bmatrix} 1.2 & 1.6 \\ 0.3 & 0.4 \\ 0.5 & 0.6 \end{bmatrix} \end{array} \qquad \mathbf{B} = \begin{array}{c} \\ \text{PC1086} \\ \text{PC1186} \end{array} \begin{array}{c} \begin{array}{cccc} \text{Quarter} & & & \\ 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 3 & 8 & 6 & 9 \\ 6 & 2 & 4 & 3 \end{bmatrix} \end{array}$$

Solution.

$$\mathbf{C} = \mathbf{AB} = \begin{array}{c} \\ \text{Raw Components} \\ \text{Labor} \\ \text{Miscellaneous} \end{array} \begin{array}{c} \begin{array}{cccc} \text{Quarter} & & & \\ 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 13.2 & 12.8 & 13.6 & 15.6 \\ 3.3 & 3.2 & 3.4 & 3.9 \\ 5.1 & 5.2 & 5.4 & 6.3 \end{bmatrix} \end{array}$$

Since cost is given in multiples of \$1000 and production in multiples of 10 000 units, the entries of **C** are multiples of \$10 millions; thus $c_{11} = 13.2$ means \$132 million, etc.

EXAMPLE 12 **Weight Watching. Matrix Times Vector**

Suppose that in a weight-watching program, a person of 185 lb burns 350 cal/hr in walking (3 mph), 500 in bycycling (13 mph) and 950 in jogging (5.5 mph). Bill, weighing 185 lb, plans to exercise according to the matrix shown. Verify the calculations (W = Walking, B = Bicycling, J = Jogging).

$$\begin{array}{c} \\ \text{MON} \\ \text{WED} \\ \text{FRI} \\ \text{SAT} \end{array} \begin{array}{c} \begin{array}{ccc} \text{W} & \text{B} & \text{J} \end{array} \\ \begin{bmatrix} 1.0 & 0 & 0.5 \\ 1.0 & 1.0 & 0.5 \\ 1.5 & 0 & 0.5 \\ 2.0 & 1.5 & 1.0 \end{bmatrix} \end{array} \begin{bmatrix} 350 \\ 500 \\ 950 \end{bmatrix} = \begin{bmatrix} 825 \\ 1325 \\ 1000 \\ 2400 \end{bmatrix} \begin{array}{l} \\ \text{MON} \\ \text{WED} \\ \text{FRI} \\ \text{SAT} \end{array}$$

EXAMPLE 13 **Markov Process. Powers of a Matrix. Stochastic Matrix**

Suppose that the 2004 state of land use in a city of 60 mi^2 of built-up area is

C: Commercially Used 25% I: Industrially Used 20% R: Residentially Used 55%.

Find the states in 2009, 2014, and 2019, assuming that the transition probabilities for 5-year intervals are given by the matrix **A** and remain practically the same over the time considered.

$$\mathbf{A} = \begin{array}{c} \begin{array}{ccc} \text{From C} & \text{From I} & \text{From R} \end{array} \\ \begin{bmatrix} 0.7 & 0.1 & 0 \\ 0.2 & 0.9 & 0.2 \\ 0.1 & 0 & 0.8 \end{bmatrix} \end{array} \begin{array}{l} \text{To C} \\ \text{To I} \\ \text{To R} \end{array}$$

A is a **stochastic matrix,** that is, a square matrix with all entries nonnegative and all column sums equal to 1. Our example concerns a **Markov process**[1], that is, a process for which the probability of entering a certain state depends only on the last state occupied (and the matrix **A**), not on any earlier state.

Solution. From the matrix **A** and the 2004 state we can compute the 2009 state,

$$\begin{array}{c} \text{C} \\ \text{I} \\ \text{R} \end{array} \begin{bmatrix} 0.7 \cdot 25 + 0.1 \cdot 20 + 0 \cdot 55 \\ 0.2 \cdot 25 + 0.9 \cdot 20 + 0.2 \cdot 55 \\ 0.1 \cdot 25 + 0 \cdot 20 + 0.8 \cdot 55 \end{bmatrix} = \begin{bmatrix} 0.7 & 0.1 & 0 \\ 0.2 & 0.9 & 0.2 \\ 0.1 & 0 & 0.8 \end{bmatrix} \begin{bmatrix} 25 \\ 20 \\ 55 \end{bmatrix} = \begin{bmatrix} 19.5 \\ 34.0 \\ 46.5 \end{bmatrix}.$$

To explain: The 2009 figure for C equals 25% times the probability 0.7 that C goes into C, plus 20% times the probability 0.1 that I goes into C, plus 55% times the probability 0 that R goes into C. Together,

$$25 \cdot 0.7 + 20 \cdot 0.1 + 55 \cdot 0 = 19.5 \text{ [\%]}. \qquad \text{Also} \qquad 25 \cdot 0.2 + 20 \cdot 0.9 + 55 \cdot 0.2 = 34 \text{ [\%]}.$$

Similarly, the new R is 46.5%. We see that the 2009 state vector is the column vector

$$\mathbf{y} = [19.5 \quad 34.0 \quad 46.5]^T = \mathbf{Ax} = \mathbf{A}\,[25 \quad 20 \quad 55]^T$$

where the column vector $\mathbf{x} = [25 \quad 20 \quad 55]^T$ is the given 2004 state vector. Note that the sum of the entries of **y** is 100 [%]. Similarly, you may verify that for 2014 and 2019 we get the state vectors

$$\mathbf{z} = \mathbf{Ay} = \mathbf{A}(\mathbf{Ax}) = \mathbf{A}^2\mathbf{x} = [17.05 \quad 43.80 \quad 39.15]^T$$

$$\mathbf{u} = \mathbf{Az} = \mathbf{A}^2\mathbf{y} = \mathbf{A}^3\mathbf{x} = [16.315 \quad 50.660 \quad 33.025]^T.$$

[1]ANDREI ANDREJEVITCH MARKOV (1856–1922), Russian mathematician, known for his work in probability theory.

Answer. In 2009 the commercial area will be 19.5% (11.7 mi^2), the industrial 34% (20.4 mi^2) and the residential 46.5% (27.9 mi^2). For 2014 the corresponding figures are 17.05%, 43.80%, 39.15%. For 2019 they are 16.315%, 50.660%, 33.025%. (In Sec. 8.2 we shall see what happens in the limit, assuming that those probabilities remain the same. In the meantime, can you experiment or guess?) ■

PROBLEM SET 7.2

1–14 MULTIPLICATION, ADDITION, AND TRANSPOSITION OF MATRICES AND VECTORS

Let

$$\mathbf{A} = \begin{bmatrix} 6 & -2 & -2 \\ 10 & -3 & 1 \\ -10 & 5 & 1 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 9 & 4 & -4 \\ 4 & 7 & 0 \\ -4 & 0 & 11 \end{bmatrix},$$

$$\mathbf{C} = \begin{bmatrix} 3 & 1 \\ 0 & -2 \\ 4 & 0 \end{bmatrix}, \quad \mathbf{a} = \begin{bmatrix} 5 \\ 1 \\ 2 \end{bmatrix}, \quad \mathbf{b} = [3 \quad 0 \quad 8].$$

Calculate the following products and sums or give reasons why they are not defined. (Show all intermediate results.)

1. $\mathbf{Aa}$, $\mathbf{Ab}$, $\mathbf{Ab}^T$, $\mathbf{AB}$
2. $\mathbf{Ab}^T + \mathbf{Bb}^T$, $(\mathbf{A} + \mathbf{B})\mathbf{b}^T$, $\mathbf{bA}$, $\mathbf{B} - \mathbf{B}^T$
3. $\mathbf{AB}$, $\mathbf{BA}$, $\mathbf{AA}^T$, $\mathbf{A}^T\mathbf{A}$
4. $\mathbf{A}^2$, $\mathbf{B}^2$, $(\mathbf{A}^T)^2$, $(\mathbf{A}^2)^T$
5. $\mathbf{a}^T\mathbf{A}$, $\mathbf{bA}$, $5\mathbf{B}(3\mathbf{a} + 2\mathbf{b}^T)$, $15\mathbf{Ba} + 10\mathbf{Bb}^T$
6. $\mathbf{A}^T\mathbf{b}$, $\mathbf{b}^T\mathbf{B}$, $(3\mathbf{A} - 2\mathbf{B})^T\mathbf{a}$, $\mathbf{a}^T(3\mathbf{A} - 2\mathbf{B})$
7. $\mathbf{ab}$, $\mathbf{ba}$, $(\mathbf{ab})\mathbf{A}$, $\mathbf{a}(\mathbf{bA})$
8. $\mathbf{ab} - \mathbf{ba}$, $-(4\mathbf{b})(7\mathbf{a})$, $-28\mathbf{ba}$, $5\mathbf{abB}$
9. $(\mathbf{A} + \mathbf{B})^2$, $\mathbf{A}^2 + \mathbf{AB} + \mathbf{BA} + \mathbf{B}^2$, $\mathbf{A}^2 + 2\mathbf{AB} + \mathbf{B}^2$
10. $(\mathbf{A} + \mathbf{B})(\mathbf{A} - \mathbf{B})$, $\mathbf{A}^2 - \mathbf{AB} + \mathbf{BA} - \mathbf{B}^2$, $\mathbf{A}^2 - \mathbf{B}^2$
11. $\mathbf{A}^2\mathbf{B}$, $\mathbf{A}^3$, $(\mathbf{AB})^2$, $\mathbf{A}^2\mathbf{B}^2$
12. $\mathbf{B}^3$, $\mathbf{BC}$, $(\mathbf{BC})^2$, $(\mathbf{BC})(\mathbf{BC})^T$
13. $\mathbf{a}^T\mathbf{Aa}$, $\mathbf{a}^T(\mathbf{A} + \mathbf{A}^T)\mathbf{a}$, $\mathbf{bBb}^T$, $\mathbf{b}(\mathbf{B} - \mathbf{B}^T)\mathbf{b}^T$
14. $\mathbf{a}^T\mathbf{CC}^T\mathbf{a}$, $\mathbf{a}^T\mathbf{C}^2\mathbf{a}$, $\mathbf{bC}^T\mathbf{Cb}^T$, $\mathbf{bCC}^T\mathbf{b}^T$

15. **(General rules)** Prove (2) for 2×2 matrices $\mathbf{A} = [a_{jk}]$, $\mathbf{B} = [b_{jk}]$, $\mathbf{C} = [c_{jk}]$ and a general scalar.
16. **(Commutativity)** Find all 2×2 matrices $\mathbf{A} = [a_{jk}]$ that commute with $\mathbf{B} = [b_{jk}]$, where $b_{jk} = j + k$.
17. **(Product)** Write $\mathbf{AB}$ in Probs. 1–14 in terms of row and column vectors.
18. **(Product)** Calculate $\mathbf{AB}$ in Prob. 1 columnwise. (See Example 6.)
19. **TEAM PROJECT. Symmetric and Skew-Symmetric Matrices.** These matrices occur quite frequently in applications, so it is worthwhile to study some of their most important properties.

    (a) Verify the claims in (11) that $a_{kj} = a_{jk}$ for a symmetric matrix, and $a_{kj} = -a_{jk}$ for a skew-symmetric matrix. Give examples.

    (b) Show that for every square matrix $\mathbf{C}$ the matrix $\mathbf{C} + \mathbf{C}^T$ is symmetric and $\mathbf{C} - \mathbf{C}^T$ is skew-symmetric. Write $\mathbf{C}$ in the form $\mathbf{C} = \mathbf{S} + \mathbf{T}$, where $\mathbf{S}$ is symmetric and $\mathbf{T}$ is skew-symmetric and find $\mathbf{S}$ and $\mathbf{T}$ in terms of $\mathbf{C}$. Represent $\mathbf{A}$ and $\mathbf{B}$ in Probs. 1–14 in this form.

    (c) A **linear combination** of matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\cdots$, $\mathbf{M}$ of the same size is an expression of the form

    $$(14) \qquad a\mathbf{A} + b\mathbf{B} + c\mathbf{C} + \cdots + m\mathbf{M},$$

    where $a, \cdots, m$ are any scalars. Show that if these matrices are square and symmetric, so is (14); similarly, if they are skew-symmetric, so is (14).

    (d) Show that $\mathbf{AB}$ with symmetric $\mathbf{A}$ and $\mathbf{B}$ is symmetric if and only if $\mathbf{A}$ and $\mathbf{B}$ **commute,** that is, $\mathbf{AB} = \mathbf{BA}$.

    (e) Under what condition is the product of skew-symmetric matrices skew-symmetric?
20. **(Idempotent and nilpotent matrices)** By definition, $\mathbf{A}$ is **idempotent** if $\mathbf{A}^2 = \mathbf{A}$, and $\mathbf{B}$ is **nilpotent** if $\mathbf{B}^m = \mathbf{0}$ for some positive integer m. Give examples (different from $\mathbf{0}$ or $\mathbf{I}$). Also give examples such that $\mathbf{A}^2 = \mathbf{I}$ (the unit matrix).
21. **(Triangular matrices)** Let $\mathbf{U}_1$, $\mathbf{U}_2$ be upper triangular and $\mathbf{L}_1$, $\mathbf{L}_2$ lower triangular. Which of the following are triangular? Give examples. How can you save half of your work by transposition?

    $$\mathbf{U}_1 + \mathbf{U}_2, \mathbf{U}_1\mathbf{U}_2, \mathbf{U}_1^2, \mathbf{U}_1 + \mathbf{L}_1, \mathbf{U}_1\mathbf{L}_1, \mathbf{L}_1 + \mathbf{L}_2,$$

    $$\mathbf{L}_1\mathbf{L}_2, \mathbf{L}_1^2$$
22. **(Transposition of products)** Prove (10a)–(10c). Illustrate the basic formula (10d) by examples of your own. Then prove it.

APPLICATIONS

23. **(Markov process)** If the transition matrix $\mathbf{A}$ has the entries $a_{11} = 0.5$, $a_{12} = 0.3$, $a_{21} = 0.5$, $a_{22} = 0.7$ and the initial state is $[1 \quad 1]^T$, what will the next three states be?
24. **(Concert subscription)** In a community of 300 000 adults, subscribers to a concert series tend to renew their subscription with probability 90% and persons presently not subscribing will subscribe for the next season with probability 0.1%. If the present number of subscribers is 2000, can one predict an increase, decrease, or no change over each of the next three seasons?

CAS Experiment. Markov Process. Write a program for a Markov process. Use it to calculate further steps in Example 13 of the text. Experiment with other stochastic 3×3 matrices, also using different starting values.

(Production) In a production process, let N mean "no trouble" and T "trouble." Let the transition probabilities from one day to the next be 0.9 for $N \to N$, hence 0.1 for $N \to T$, and 0.5 for $T \to N$, hence 0.5 for $T \to T$. If today there is no trouble, what is the probability of N two days after today? Three days after today?

(Profit vector) Two factory outlets F_1 and F_2 in New York and Los Angeles sell sofas (S), chairs (C), and tables (T) with a profit of \$110, \$45, and \$80, respectively. Let the sales in a certain week be given by the matrix

$$\mathbf{A} = \begin{array}{c} \begin{matrix} S & C & T \end{matrix} \\ \begin{bmatrix} 600 & 400 & 100 \\ 300 & 820 & 205 \end{bmatrix} \end{array} \begin{matrix} \\ F_1 \\ F_2 \end{matrix}$$

Introduce a "profit vector" $\mathbf{p}$ such that the components of $\mathbf{v} = \mathbf{Ap}$ give the total profits of F_1 and F_2.

8. TEAM PROJECT. Special Linear Transformations. Rotations have various applications. We show in this project how they can be handled by matrices.

(a) Rotation in the plane. Show that the linear transformation $\mathbf{y} = \mathbf{Ax}$ with matrix

$$\mathbf{A} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \quad \text{and} \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}.$$

is a counterclockwise rotation of the Cartesian x_1x_2-coordinate system in the plane about the origin, where θ is the angle of rotation.

(b) Rotation through $n\theta$. Show that in (a)

$$\mathbf{A}^n = \begin{bmatrix} \cos n\theta & -\sin n\theta \\ \sin n\theta & \cos n\theta \end{bmatrix}.$$

Is this plausible? Explain this in words.

(c) Addition formulas for cosine and sine. By geometry we should have

$$\begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix} \begin{bmatrix} \cos\beta & -\sin\beta \\ \sin\beta & \cos\beta \end{bmatrix} = \begin{bmatrix} \cos(\alpha+\beta) & -\sin(\alpha+\beta) \\ \sin(\alpha+\beta) & \cos(\alpha+\beta) \end{bmatrix}.$$

Derive from this the addition formulas (6) in App. A3.1.

(d) Computer graphics. To visualize a three-dimensional object with plane faces (e.g., a cube), we may store the position vectors of the vertices with respect to a suitable $x_1x_2x_3$-coordinate system (and a list of the connecting edges) and then obtain a two-dimensional image on a video screen by projecting the object onto a coordinate plane, for instance, onto the x_1x_2-plane by setting $x_3 = 0$. To change the appearance of the image, we can impose a linear transformation on the position vectors stored. Show that a diagonal matrix $\mathbf{D}$ with main diagonal entries 3, 1, $\frac{1}{2}$ gives from an $\mathbf{x} = [x_j]$ the new position vector $\mathbf{y} = \mathbf{Dx}$, where $y_1 = 3x_1$ (stretch in the x_1-direction by a factor 3), $y_2 = x_2$ (unchanged), $y_3 = \frac{1}{2}x_3$ (contraction in the x_3-direction). What effect would a scalar matrix have?

(e) Rotations in space. Explain $\mathbf{y} = \mathbf{Ax}$ geometrically when $\mathbf{A}$ is one of the three matrices

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix},$$

$$\begin{bmatrix} \cos\varphi & 0 & -\sin\varphi \\ 0 & 1 & 0 \\ \sin\varphi & 0 & \cos\varphi \end{bmatrix}, \quad \begin{bmatrix} \cos\psi & -\sin\psi & 0 \\ \sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

What effect would these transformations have in situations such as that described in (d)?

7.3 Linear Systems of Equations. Gauss Elimination

The most important use of matrices occurs in the solution of *systems of linear equations*, briefly called **linear systems.** Such systems model various problems, for instance, in frameworks, electrical networks, traffic flow, economics, statistics, and many others. In this section we show an important solution method, the *Gauss elimination.* General properties of solutions will be discussed in the next sections.

Linear System, Coefficient Matrix, Augmented Matrix

A **linear system of m equations in n unknowns** $x_1, \cdots, x_n$ is a set of equations of the form

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + \cdots + a_{2n}x_n &= b_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\ a_{m1}x_1 + \cdots + a_{mn}x_n &= b_m \end{aligned} \tag{1}$$

The system is called *linear* because each variable x_j appears in the first power only, just as in the equation of a straight line. $a_{11}, \cdots, a_{mn}$ are given numbers, called the **coefficients** of the system. $b_1, \cdots, b_m$ on the right are also given numbers. If all the b_j are zero, then (1) is called a **homogeneous system.** If at least one b_j is not zero, then (1) is called a **nonhomogeneous system.**

A **solution** of (1) is a set of numbers $x_1, \cdots, x_n$ that satisfies all the m equations. A **solution vector** of (1) is a vector $\mathbf{x}$ whose components form a solution of (1). If the system (1) is homogeneous, it has at least the **trivial solution** $x_1 = 0, \cdots, x_n = 0$.

Matrix Form of the Linear System (1). From the definition of matrix multiplication we see that the m equations of (1) may be written as a single vector equation

$$\mathbf{Ax} = \mathbf{b} \tag{2}$$

where the **coefficient matrix** $\mathbf{A} = [a_{jk}]$ is the $m \times n$ matrix

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \quad \text{and} \quad \mathbf{x} = \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} b_1 \\ \cdot \\ \cdot \\ \cdot \\ b_m \end{bmatrix}$$

are column vectors. We assume that the coefficients a_{jk} are not all zero, so that $\mathbf{A}$ is not a zero matrix. Note that $\mathbf{x}$ has n components, whereas $\mathbf{b}$ has m components. The matrix

$$\tilde{\mathbf{A}} = \left[\begin{array}{ccc|c} a_{11} & \cdots & a_{1n} & b_1 \\ \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdots & \cdot & \cdot \\ a_{m1} & \cdots & a_{mn} & b_m \end{array}\right]$$

is called the **augmented matrix** of the system (1). The dashed vertical line could be omitted (as we shall do later); it is merely a reminder that the last column of $\tilde{\mathbf{A}}$ does not belong to $\mathbf{A}$.

The augmented matrix $\tilde{\mathbf{A}}$ determines the system* (1) *completely because it contains all the given numbers appearing in (1).

EXAMPLE 1 **Geometric Interpretation. Existence and Uniqueness of Solutions**

If $m = n = 2$, we have two equations in two unknowns x_1, x_2

$$a_{11}x_1 + a_{12}x_2 = b_1$$
$$a_{21}x_1 + a_{22}x_2 = b_2.$$

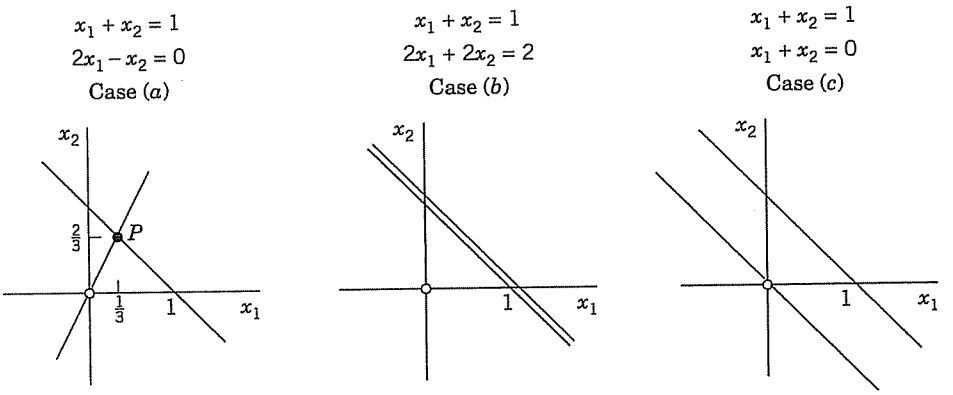
If we interpret x_1, x_2 as coordinates in the x_1x_2-plane, then each of the two equations represents a straight line, and (x_1, x_2) is a solution if and only if the point P with coordinates x_1, x_2 lies on both lines. Hence there are three possible cases:

(a) Precisely one solution if the lines intersect.

(b) Infinitely many solutions if the lines coincide.

(c) No solution if the lines are parallel

For instance,



If the system is homogenous, Case (c) cannot happen, because then those two straight lines pass through the origin, whose coordinates 0, 0 constitute the trivial solution. If you wish, consider three equations in three unknowns as representations of three planes in space and discuss the various possible cases in a similar fashion. See Fig. 156. ▩

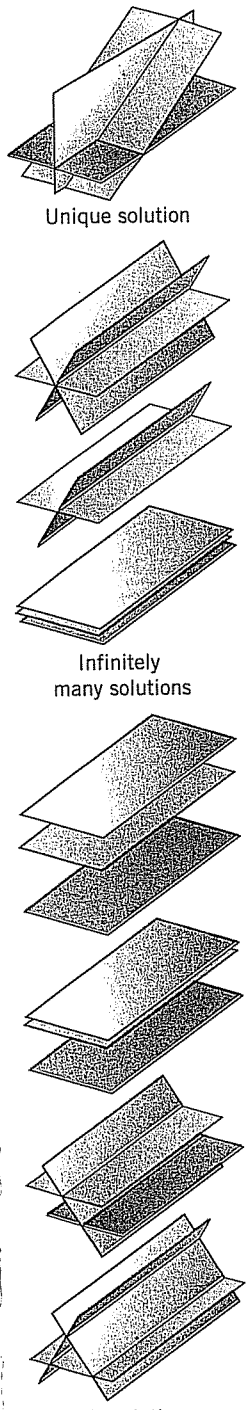


Fig. 156. Three equations in three unknowns interpreted as planes in space

Our simple example illustrates that a system (1) may perhaps have no solution. This poses the following problem. Does a given system (1) have a solution? Under what conditions does it have precisely one solution? If it has more than one solution, how can we characterize the set of all solutions? How can we actually obtain the solutions? Perhaps the last question is the most immediate one from a practical viewpoint. We shall answer it first and discuss the other questions in Sec. 7.5.

Gauss Elimination and Back Substitution

This is a standard elimination method for solving linear systems that proceeds systematically irrespective of particular features of the coefficients. It is a method of great practical importance and is reasonable with respect to computing time and storage demand (two aspects we shall consider in Sec. 20.1 in the chapter on numeric linear algebra). We begin by motivating the method. If a system is in **"triangular form,"** say,

$$2x_1 + 5x_2 = 2$$
$$13x_2 = -26$$

we can solve it by **"back substitution,"** that is, solve the last equation for the variable, $x_2 = -26/13 = -2$, and then work backward, substituting $x_2 = -2$ into the first equation

and solve it for x_1, obtaining $x_1 = \frac{1}{2}(2 - 5x_2) = \frac{1}{2}(2 - 5 \cdot (-2)) = 6$. This gives us the idea of first reducing a general system to triangular form. For instance, let the given system be

$$\begin{aligned} 2x_1 + 5x_2 &= 2 \\ -4x_1 + 3x_2 &= -30. \end{aligned} \qquad \text{Its augmented matrix is} \qquad \begin{bmatrix} 2 & 5 & 2 \\ -4 & 3 & -30 \end{bmatrix}.$$

We leave the first equation as it is. We eliminate x_1 from the second equation, to get a triangular system. For this we add twice the first equation to the second, and we do the same operation on the ***rows*** of the augmented matrix. This gives $-4x_1 + 4x_1 + 3x_2 + 10x_2 = -30 + 2 \cdot 2$, that is,

$$\begin{aligned} 2x_1 + 5x_2 &= 2 \\ 13x_2 &= -26 \end{aligned} \qquad \text{Row 2 + 2 Row 1} \qquad \begin{bmatrix} 2 & 5 & 2 \\ 0 & 13 & -26 \end{bmatrix}$$

where Row 2 + 2 Row 1 means "Add twice Row 1 to Row 2" in the original matrix. This is the **Gauss elimination** (for 2 equations in 2 unknowns) giving the triangular form, from which back substitution now yields $x_2 = -2$ and $x_1 = 6$, as before.

Since a linear system is completely determined by its augmented matrix, ***Gauss elimination can be done by merely considering the matrices,*** as we have just indicated. We do this again in the next example, emphasizing the matrices by writing them first and the equations behind them, just as a help in order not to lose track.

EXAMPLE 2 **Gauss Elimination. Electrical Network**

Solve the linear system

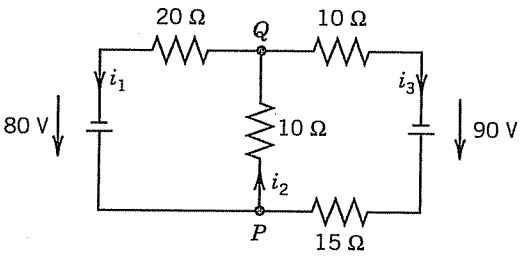
$$\begin{aligned} x_1 - x_2 + x_3 &= 0 \\ -x_1 + x_2 - x_3 &= 0 \\ 10x_2 + 25x_3 &= 90 \\ 20x_1 + 10x_2 &= 80. \end{aligned}$$

Derivation from the circuit in Fig. 157 (Optional). This is the system for the unknown currents $x_1 = i_1$, $x_2 = i_2$, $x_3 = i_3$ in the electrical network in Fig. 157. To obtain it, we label the currents as shown, choosing directions arbitrarily; if a current will come out negative, this will simply mean that the current flows against the direction of our arrow. The current entering each battery will be the same as the current leaving it. The equations for the currents result from Kirchhoff's laws:

Kirchhoff's current law (KCL). *At any point of a circuit, the sum of the inflowing currents equals the sum of the outflowing currents.*

Kirchhoff's voltage law (KVL). *In any closed loop, the sum of all voltage drops equals the impressed electromotive force.*

Node P gives the first equation, node Q the second, the right loop the third, and the left loop the fourth, as indicated in the figure.



Node P: $i_1 - i_2 + i_3 = 0$

Node Q: $-i_1 + i_2 - i_3 = 0$

Right loop: $10i_2 + 25i_3 = 90$

Left loop: $20i_1 + 10i_2 = 80$

Fig. 157. Network in Example 2 and equations relating the currents

Solution by Gauss Elimination. This system could be solved rather quickly by noticing its particular form. But this is not the point. The point is that the Gauss elimination is systematic and will work in general, also for large systems. We apply it to our system and then do back substitution. As indicated let us write the augmented matrix of the system first and then the system itself:

Augmented Matrix $\tilde{\mathbf{A}}$

Pivot 1 ⟶
Eliminate ⟶

$$\left[\begin{array}{rrr|r} \textcircled{1} & -1 & 1 & 0 \\ \boxed{-1} & 1 & -1 & 0 \\ 0 & 10 & 25 & 90 \\ \boxed{20} & 10 & 0 & 80 \end{array}\right]$$

Equations

Pivot 1 ⟶
Eliminate ⟶

$$\begin{aligned} \textcircled{x_1} - x_2 + x_3 &= 0 \\ \boxed{-x_1} + x_2 - x_3 &= 0 \\ 10x_2 + 25x_3 &= 90 \\ \boxed{20x_1} + 10x_2 &= 80. \end{aligned}$$

Step 1. Elimination of x_1

Call the first row of **A** the **pivot row** and the first equation the **pivot equation.** Call the coefficient 1 of its x_1-term the **pivot** in this step. Use this equation to eliminate x_1 (get rid of x_1) in the other equations. For this, do:

Add 1 times the pivot equation to the second equation.

Add −20 times the pivot equation to the fourth equation.

This corresponds to **row operations** on the augmented matrix as indicated in BLUE behind the ***new matrix*** in (3). So the operations are performed on the ***preceding matrix.*** The result is

(3)

$$\left[\begin{array}{rrr|r} 1 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 10 & 25 & 90 \\ 0 & 30 & -20 & 80 \end{array}\right] \qquad \begin{array}{l} \\ \text{Row 2 + Row 1} \\ \\ \text{Row 4 − 20 Row 1} \end{array} \qquad \begin{aligned} x_1 - x_2 + x_3 &= 0 \\ 0 &= 0 \\ 10x_2 + 25x_3 &= 90 \\ 30x_2 - 20x_3 &= 80. \end{aligned}$$

Step 2. Elimination of x_2

The first equation remains as it is. We want the new second equation to serve as the next pivot equation. But since it has no x_2-term (in fact, it is 0 = 0), we must first change the order of the equations and the corresponding rows of the new matrix. We put 0 = 0 at the end and move the third equation and the fourth equation one place up. This is called **partial pivoting** (as opposed to the rarely used *total pivoting*, in which also the order of the unknowns is changed). It gives

Pivot 10 ⟶
Eliminate 30 ⟶

$$\left[\begin{array}{rrr|r} 1 & -1 & 1 & 0 \\ 0 & \textcircled{10} & 25 & 90 \\ 0 & \boxed{30} & -20 & 80 \\ 0 & 0 & 0 & 0 \end{array}\right]$$

$$\begin{aligned} x_1 - x_2 + x_3 &= 0 \\ \textcircled{10x_2} + 25x_3 &= 90 \\ \boxed{30x_2} - 20x_3 &= 80 \\ 0 &= 0 \end{aligned}$$

To eliminate x_2, do:

Add −3 times the pivot equation to the third equation.

The result is

(4)

$$\left[\begin{array}{rrr|r} 1 & -1 & 1 & 0 \\ 0 & 10 & 25 & 90 \\ 0 & 0 & -95 & -190 \\ 0 & 0 & 0 & 0 \end{array}\right] \qquad \text{Row 3 − 3 Row 2} \qquad \begin{aligned} x_1 - x_2 + x_3 &= 0 \\ 10x_2 + 25x_3 &= 90 \\ -95x_3 &= -190 \\ 0 &= 0 \end{aligned}$$

Back Substitution. ***Determination of x_3, x_2, x_1 (in this order)***

Working backward from the last to the first equation of this "triangular" system (4), we can now readily find x_3, then x_2, and then x_1:

$$\begin{aligned} -95x_3 &= -190 \\ 10x_2 + 25x_3 &= 90 \\ x_1 - x_2 + x_3 &= 0 \end{aligned} \qquad \begin{aligned} x_3 &= i_3 = 2 \text{ [A]} \\ x_2 &= \tfrac{1}{10}(90 - 25x_3) = i_2 = 4 \text{ [A]} \\ x_1 &= x_2 - x_3 = i_1 = 2 \text{ [A]} \end{aligned}$$

where A stands for "amperes." This is the answer to our problem. The solution is unique. ■

Elementary Row Operations. Row-Equivalent Systems

Example 2 illustrates the operations of the Gauss elimination. These are the first two of three operations, which are called

Elementary Row Operations for Matrices:

Interchange of two rows

Addition of a constant multiple of one row to another row

Multiplication of a row by a ***nonzero*** *constant c.*

CAUTION! These operations are for rows, ***not for columns!*** They correspond to the following

Elementary Operations for Equations:

Interchange of two equations

Addition of a constant multiple of one equation to another equation

Multiplication of an equation by a ***nonzero*** *constant c.*

Clearly, the interchange of two equations does not alter the solution set. Neither does that addition because we can undo it by a corresponding subtraction. Similarly for that multiplication, which we can undo by multiplying the new equation by $1/c$ (since $c \neq 0$), producing the original equation.

We now call a linear system S_1 **row-equivalent** to a linear system S_2 if S_1 can be obtained from S_2 by (finitely many!) row operations. Thus we have proved the following result, which also justifies the Gauss elimination.

THEOREM 1

Row-Equivalent Systems

Row-equivalent linear systems have the same set of solutions.

Because of this theorem, systems having the same solution sets are often called *equivalent systems.* But note well that we are dealing with ***row operations.*** No column operations on the augmented matrix are permitted in this context because they would generally alter the solution set.

A linear system (1) is called **overdetermined** if it has more equations than unknowns, as in Example 2, **determined** if $m = n$, as in Example 1, and **underdetermined** if it has fewer equations than unknowns.

Furthermore, a system (1) is called **consistent** if it has at least one solution (thus, one solution or infinitely many solutions), but **inconsistent** if it has no solutions at all, as $x_1 + x_2 = 1$, $x_1 + x_2 = 0$ in Example 1.

Gauss Elimination: The Three Possible Cases of Systems

The Gauss elimination can take care of linear systems with a unique solution (see Example 2), with infinitely many solutions (Example 3, below), and without solutions (inconsistent systems; see Example 4).

EXAMPLE 3 **Gauss Elimination if Infinitely Many Solutions Exist**

Solve the following linear systems of three equations in four unknowns whose augmented matrix is

$$(5)\qquad \left[\begin{array}{cccc|c} 3.0 & 2.0 & 2.0 & -5.0 & 8.0 \\ 0.6 & 1.5 & 1.5 & -5.4 & 2.7 \\ 1.2 & -0.3 & -0.3 & 2.4 & 2.1 \end{array}\right].$$

Thus,

$$\begin{aligned} (3.0x_1) + 2.0x_2 + 2.0x_3 - 5.0x_4 &= 8.0 \\ \boxed{0.6x_1} + 1.5x_2 + 1.5x_3 - 5.4x_4 &= 2.7 \\ \boxed{1.2x_1} - 0.3x_2 - 0.3x_3 + 2.4x_4 &= 2.1. \end{aligned}$$

Solution. As in the previous example, we circle pivots and box terms of equations and corresponding entries to be eliminated. We indicate the operations in terms of equations and operate on both equations and matrices.

Step 1. Elimination of x_1 from the second and third equations by adding

$-0.6/3.0 = -0.2$ times the first equation to the second equation,

$-1.2/3.0 = -0.4$ times the first equation to the third equation.

This gives the following, in which the pivot of the next step is circled.

$$(6)\qquad \left[\begin{array}{cccc|c} 3.0 & 2.0 & 2.0 & -5.0 & 8.0 \\ 0 & 1.1 & 1.1 & -4.4 & 1.1 \\ 0 & -1.1 & -1.1 & 4.4 & -1.1 \end{array}\right] \begin{array}{l} \\ \text{Row 2} - 0.2\ \text{Row 1} \\ \text{Row 3} - 0.4\ \text{Row 1} \end{array}$$

$$\begin{aligned} 3.0x_1 + 2.0x_2 + 2.0x_3 - 5.0x_4 &= 8.0 \\ (1.1x_2) + 1.1x_3 - 4.4x_4 &= 1.1 \\ \boxed{-1.1x_2} - 1.1x_3 + 4.4x_4 &= -1.1 \end{aligned}$$

Step 2. Elimination of x_2 from the third equation of (6) by adding

$1.1/1.1 = 1$ times the second equation to the third equation.

This gives

$$(7)\qquad \left[\begin{array}{cccc|c} 3.0 & 2.0 & 2.0 & -5.0 & 8.0 \\ 0 & 1.1 & 1.1 & -4.4 & 1.1 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right] \begin{array}{l} \\ \\ \text{Row 3} + \text{Row 2} \end{array}$$

$$\begin{aligned} 3.0x_1 + 2.0x_2 + 2.0x_3 - 5.0x_4 &= 8.0 \\ 1.1x_2 + 1.1x_3 - 4.4x_4 &= 1.1 \\ 0 &= 0\ . \end{aligned}$$

Back Substitution. From the second equation, $x_2 = 1 - x_3 + 4x_4$. From this and the first equation, $x_1 = 2 - x_4$. Since x_3 and x_4 remain arbitrary, we have infinitely many solutions. If we choose a value of x_3 and a value of x_4, then the corresponding values of x_1 and x_2 are uniquely determined.

On Notation. If unknowns remain arbitrary, it is also customary to denote them by other letters $t_1, t_2, \cdots$. In this example we may thus write $x_1 = 2 - x_4 = 2 - t_2$, $x_2 = 1 - x_3 + 4x_4 = 1 - t_1 + 4t_2$, $x_3 = t_1$ (first arbitrary unknown), $x_4 = t_2$ (second arbitrary unknown). ■

EXAMPLE 4 **Gauss Elimination if no Solution Exists**

What will happen if we apply the Gauss elimination to a linear system that has no solution? The answer is that in this case the method will show this fact by producing a contradiction. For instance, consider

$$\left[\begin{array}{ccc|c} 3 & 2 & 1 & 3 \\ 2 & 1 & 1 & 0 \\ 6 & 2 & 4 & 6 \end{array}\right]$$

$$\begin{aligned} (3x_1) + 2x_2 + x_3 &= 3 \\ \boxed{2x_1} + x_2 + x_3 &= 0 \\ \boxed{6x_1} + 2x_2 + 4x_3 &= 6. \end{aligned}$$

Step 1. Elimination of x_1 from the second and third equations by adding

$-\frac{2}{3}$ times the first equation to the second equation,

$-\frac{6}{3} = -2$ times the first equation to the third equation.

This gives

$$\left[\begin{array}{ccc|c} 3 & 2 & 1 & 3 \\ 0 & -\frac{1}{3} & \frac{1}{3} & -2 \\ 0 & -2 & 2 & 0 \end{array}\right] \begin{array}{l} \\ \text{Row 2} - \frac{2}{3}\ \text{Row 1} \\ \text{Row 3} - 2\ \text{Row 1} \end{array}$$

$$\begin{aligned} 3x_1 + 2x_2 + x_3 &= 3 \\ \boxed{-\tfrac{1}{3}x_2} + \tfrac{1}{3}x_3 &= -2 \\ \boxed{-2x_2} + 2x_3 &= 0. \end{aligned}$$

Step 2. Elimination of x_2 from the third equation gives

$$\left[\begin{array}{ccc|c} 3 & 2 & 1 & 3 \\ 0 & -\frac{1}{3} & \frac{1}{3} & -2 \\ 0 & 0 & 0 & 12 \end{array}\right] \quad \text{Row 3} - 6 \text{ Row 2} \qquad \begin{aligned} 3x_1 + 2x_2 + x_3 &= 3 \\ -\tfrac{1}{3}x_2 + \tfrac{1}{3}x_3 &= -2 \\ 0 &= 12. \end{aligned}$$

The false statement $0 = 12$ shows that the system has no solution. ∎

Row Echelon Form and Information From It

At the end of the Gauss elimination the form of the coefficient matrix, the augmented matrix, and the system itself are called the **row echelon form.** In it, rows of zeros, if present, are the last rows, and in each nonzero row the leftmost nonzero entry is farther to the right than in the previous row. For instance, in Example 4 the coefficient matrix and its augmented in row echelon form are

$$\left[\begin{array}{ccc} 3 & 2 & 1 \\ 0 & -\frac{1}{3} & \frac{1}{3} \\ 0 & 0 & 0 \end{array}\right] \quad \text{and} \quad \left[\begin{array}{ccc|c} 3 & 2 & 1 & 3 \\ 0 & -\frac{1}{3} & \frac{1}{3} & -2 \\ 0 & 0 & 0 & 12 \end{array}\right].$$

Note that we do not require that the leftmost nonzero entries be 1 since this would have no theoretic or numeric advantage. (The so-called *reduced echelon form*, in which those entries *are* 1, will be discussed in Sec. 7.8.)

At the end of the Gauss elimination (before the back substitution) the row echelon form of the augmented matrix will be

(8)

$$\left[\begin{array}{cccc|c} a_{11} & a_{12} & \cdots\cdots & a_{1n} & b_1 \\ & c_{22} & \cdots\cdots & c_{2n} & \widetilde{b}_2 \\ & & \ddots & \vdots & \vdots \\ & & k_{rr} \quad \cdots & k_{rn} & \widetilde{b}_r \\ & & & & \widetilde{b}_{r+1} \\ & & & & \vdots \\ & & & & \widetilde{b}_m \end{array}\right].$$

Here, $r \leqq m$ and $a_{11} \neq 0$, $c_{22} \neq 0, \cdots, k_{rr} \neq 0$, and all the entries in the blue triangle as well as in the blue rectangle are zero. From this we see that with respect to solutions of the system with augmented matrix (8) (and thus with respect to the originally given system) there are three possible cases:

(a) ***Exactly one solution*** if $r = n$ and $\widetilde{b}_{r+1}, \cdots, \widetilde{b}_m$, if present, are zero. To get the solution, solve the nth equation corresponding to (8) (which is $k_{nn}x_n = \widetilde{b}_n$) for x_n, then the $(n-1)$st equation for x_{n-1}, and so on up the line. See Example 2, where $r = n = 3$ and $m = 4$.

(b) ***Infinitely many solutions*** if $r < n$ and $\widetilde{b}_{r+1}, \cdots, \widetilde{b}_m$, if present, are zero. To obtain any of these solutions, choose values of $x_{r+1}, \cdots, x_n$ arbitrarily. Then solve the rth equation for x_r, then the $(r-1)$st equation for x_{r-1}, and so on up the line. See Example 3.

(c) ***No solution*** if $r < m$ and one of the entries $\widetilde{b}_{r+1}, \cdots, \widetilde{b}_m$ is not zero. See Example 4, where $r = 2 < m = 3$ and $\widetilde{b}_{r+1} = \widetilde{b}_3 = 12$.

PROBLEM SET 7.3

1–16 GAUSS ELIMINATION AND BACK SUBSTITUTION

Solve the following systems or indicate the nonexistence of solutions. (Show the details of your work.)

1.
$$5x - 2y = 20.9$$
$$-x + 4y = -19.3$$

2.
$$3.0x + 6.2y = 0.2$$
$$2.1x + 8.5y = 4.3$$

3.
$$0.5x + 3.5y = 5.7$$
$$-x + 5.0y = 7.8$$

4.
$$4y - 2z = 2$$
$$6x - 2y + z = 29$$
$$4x + 8y - 4z = 24$$

5.
$$0.8x + 1.2y - 0.6z = -7.8$$
$$2.6x + 1.7z = 15.3$$
$$4.0x - 7.3y - 1.5z = 1.1$$

6.
$$14x - 2y - 4z = 0$$
$$18x - 2y - 6z = 0$$
$$4x + 8y - 14z = 0$$

7.
$$y + z = -2$$
$$4y + 6z = -12$$
$$x + y + z = 2$$

8.
$$2x + y - 3z = 8$$
$$5x + 2z = 3$$
$$8x - y + 7z = 0$$

9.
$$4y + 4z = 24$$
$$3x - 11y - 2z = -6$$
$$6x - 17y + z = 18$$

10.
$$0.6x + 0.3y - 0.4z = -1.9$$
$$-4.6x + 0.5y + 1.2z = -1.3$$

11.
$$2x - y + 3z = -1$$
$$-4x + 2y - 6z = 2$$

12.
$$-2y - 2z = -8$$
$$3x + 4y - 5z = 13$$

13.
$$x + y - 2z = 0$$
$$-4w - x - y + 2z = -4$$
$$-2w + 3x + 3y - 6z = -2$$

14.
$$w - 2x + 5y - 3z = 0$$
$$-3w + 6x + y + z = 0$$
$$2w - 4x + 3y - z = 3$$

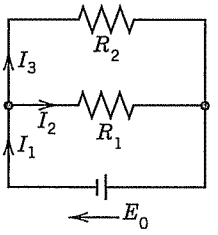
15.
$$3x + 7y - 4z = -46$$
$$5w + 4x + 8y + z = 7$$
$$8w + 4y - 2z = 0$$
$$-w + 6x + 2z = 13$$

16.
$$-2w - 17x + 4y + 3z = 0$$
$$7w + 3y - 2z = 0$$
$$2x + 8y - 6z = -20$$
$$5w - 13x - y + 5z = 16$$

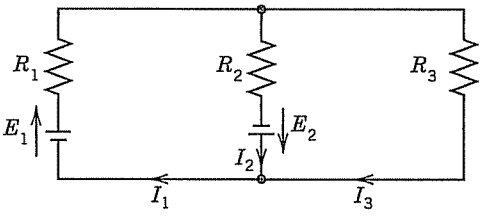
17–19 MODELS OF ELECTRICAL NETWORKS

Using Kirchhoff's laws (see Example 2), find the currents. (Show the details of your work.)

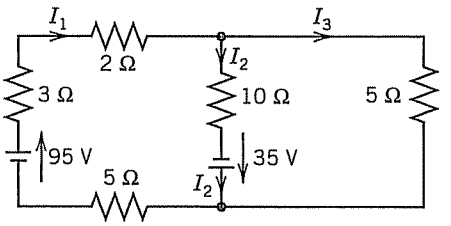
17.

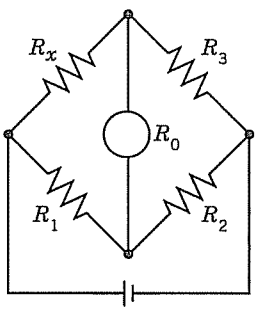


18.



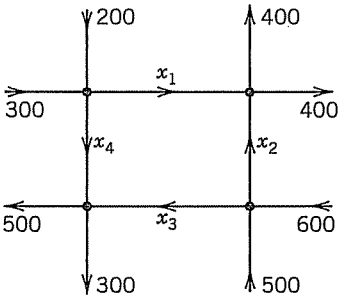
19.





Wheatstone bridge

(Prob. 20, next page)



Net of one-way streets

(Prob. 21, next page)

20. (Wheatstone bridge) Show that if $R_x/R_3 = R_1/R_2$ in the figure, then $I = 0$. (R_0 is the resistance of the instrument by which I is measured.) This bridge is a method for determining R_x. R_1, R_2, R_3 are known. R_3 is variable. To get R_x, make $I = 0$ by varing R_3. Then calculate $R_x = R_3R_1/R_2$.

21. (Traffic flow) Methods of electrical circuit analysis have applications to other fields. For instance, applying the analog of Kirchhoff's current law, find the traffic flow (cars per hour) in the net of one-way streets (in the directions indicated by the arrows) shown in the figure. Is the solution unique?

22. (Models of markets) Determine the equilibrium solution ($D_1 = S_1$, $D_2 = S_2$) of the two-commodity market with linear model (D, S, P = demand, supply, price; index 1 = first commodity, index 2 = second commodity)

$$D_1 = 60 - 2P_1 - P_2, \quad S_1 = 4P_1 - 2P_2 + 14$$
$$D_2 = 4P_1 - P_2 + 10, \quad S_2 = 5P_2 - 2.$$

23. (Equivalence relation) By definition, an *equivalence relation* on a set is a relation satisfying three conditions (named as indicated):

(i) Each element A of the set is equivalent to itself (*"Reflexivity"*).

(ii) If A is equivalent to B, then B is equivalent to A (*"Symmetry"*).

(iii) If A is equivalent to B and B is equivalent to C, then A is equivalent to C (*"Transitivity"*).

Show that row equivalence of matrices satisfies these three conditions. *Hint.* Show that for each of the three elementary row operations these conditions hold.

24. PROJECT. Elementary Matrices. The idea is that elementary operations can be accomplished by matrix multiplication. If $\mathbf{A}$ is an $m \times n$ matrix on which we want to do an elementary operation, then there is a matrix $\mathbf{E}$ such that $\mathbf{EA}$ is the new matrix after the operation. Such an $\mathbf{E}$ is called an **elementary matrix.** This idea can be helpful, for instance, in the design of algorithms. (*Computationally,* it is generally preferable to do row operations *directly,* rather than by multiplication by $\mathbf{E}$.)

(a) Show that the following are elementary matrices, for interchanging Rows 2 and 3, for adding -5 times the first row to the third, and for multiplying the fourth row by 8.

$$\mathbf{E}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{E}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -5 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$
$$\mathbf{E}_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 8 \end{bmatrix}.$$

Apply $\mathbf{E}_1$, $\mathbf{E}_2$, $\mathbf{E}_3$ to a vector and to a 4×3 matrix of your choice. Find $\mathbf{B} = \mathbf{E}_3\mathbf{E}_2\mathbf{E}_1\mathbf{A}$, where $\mathbf{A} = [a_{jk}]$ is the general 4×2 matrix. Is $\mathbf{B}$ equal to $\mathbf{C} = \mathbf{E}_1\mathbf{E}_2\mathbf{E}_3\mathbf{A}$?

(b) Conclude that $\mathbf{E}_1$, $\mathbf{E}_2$, $\mathbf{E}_3$ are obtained by doing the corresponding elementary operations on the 4×4 unit matrix. Prove that *if* $\mathbf{M}$ *is obtained from* $\mathbf{A}$ *by an elementary row operation, then*

$$\mathbf{M} = \mathbf{EA},$$

where $\mathbf{E}$ *is obtained from the* $n \times n$ *unit matrix* $\mathbf{I}_n$ *by the same row operation.*

25. CAS PROJECT. Gauss Elimination and Back Substitution. Write a program for Gauss elimination and back substitution (a) that does not include pivoting, (b) that does include pivoting. Apply the programs to Probs. 13–16 and to some larger systems of your choice.

7.4 Linear Independence. Rank of a Matrix. Vector Space

In the last section we explained the Gauss elimination with back substitution, the most important numeric solution method for linear systems of equations. It appeared that such a system may have a unique solution or infinitely many solutions, or it may be inconsistent, that is, have no solution at all. Hence we are confronted with the questions of **existence and uniqueness of solutions.** We shall answer these questions in the next section. As the

key concept for this (and other questions) we introduce the ***rank of a matrix.*** To define rank, we first need the following concepts, which are of general importance.

Linear Independence and Dependence of Vectors

Given any set of m vectors $\mathbf{a}_{(1)}, \cdots, \mathbf{a}_{(m)}$ (with the same number of components), a **linear combination** of these vectors is an expression of the form

$$c_1\mathbf{a}_{(1)} + c_2\mathbf{a}_{(2)} + \cdots + c_m\mathbf{a}_{(m)}$$

where $c_1, c_2, \cdots, c_m$ are any scalars. Now consider the equation

$$c_1\mathbf{a}_{(1)} + c_2\mathbf{a}_{(2)} + \cdots + c_m\mathbf{a}_{(m)} = \mathbf{0}. \tag{1}$$

Clearly, this vector equation (1) holds if we choose all c_j's zero, because then it becomes $\mathbf{0} = \mathbf{0}$. If this is the only m-tuple of scalars for which (1) holds, then our vectors $\mathbf{a}_{(1)}, \cdots, \mathbf{a}_{(m)}$ are said to form a *linearly independent set* or, more briefly, we call them **linearly independent.** Otherwise, if (1) also holds with scalars not all zero, we call these vectors **linearly dependent,** because then we can express (at least) one of them as a linear combination of the others. For instance, if (1) holds with, say, $c_1 \neq 0$, we can solve (1) for $\mathbf{a}_{(1)}$:

$$\mathbf{a}_{(1)} = k_2\mathbf{a}_{(2)} + \cdots + k_m\mathbf{a}_{(m)} \qquad \text{where } k_j = -c_j/c_1.$$

(Some k_j's may be zero. Or even all of them, namely, if $\mathbf{a}_{(1)} = \mathbf{0}$.)

Why is this important? Well, in the case of linear dependence we can get rid of some of the vectors until we arrive at a linearly independent set that is optimal to work with because it is smallest possible in the sense that it consists only of the "really essential" vectors, which can no longer be expressed linearly in terms of each other. This motivates the idea of a "basis" used in various contexts, notably later in our present section.

EXAMPLE 1 **Linear Independence and Dependence**

The three vectors

$$\mathbf{a}_{(1)} = [\ 3 \quad 0 \quad 2 \quad 2]$$
$$\mathbf{a}_{(2)} = [-6 \quad 42 \quad 24 \quad 54]$$
$$\mathbf{a}_{(3)} = [\ 21 \quad -21 \quad 0 \quad -15]$$

are linearly dependent because

$$6\mathbf{a}_{(1)} - \tfrac{1}{2}\mathbf{a}_{(2)} - \mathbf{a}_{(3)} = \mathbf{0}.$$

Although this is easily checked (do it!), it is not so easy to discover. However, a systematic method for finding out about linear independence and dependence follows below.

The first two of the three vectors are linearly independent because $c_1\mathbf{a}_{(1)} + c_2\mathbf{a}_{(2)} = \mathbf{0}$ implies $c_2 = 0$ (from the second components) and then $c_1 = 0$ (from any other component of $\mathbf{a}_{(1)}$). ■

Rank of a Matrix

DEFINITION

The **rank** of a matrix $\mathbf{A}$ is the maximum number of linearly independent row vectors of $\mathbf{A}$. It is denoted by rank $\mathbf{A}$.

Our further discussion will show that the rank of a matrix is an important key concept for understanding general properties of matrices and linear systems of equations.

EXAMPLE 2 Rank

The matrix

$$\mathbf{A} = \begin{bmatrix} 3 & 0 & 2 & 2 \\ -6 & 42 & 24 & 54 \\ 21 & -21 & 0 & -15 \end{bmatrix} \tag{2}$$

has rank 2, because Example 1 shows that the first two row vectors are linearly independent, whereas all three row vectors are linearly dependent.

Note further that rank $\mathbf{A} = 0$ if and only if $\mathbf{A} = \mathbf{0}$. This follows directly from the definition. ▩

We call a matrix $\mathbf{A}_1$ **row-equivalent** to a matrix $\mathbf{A}_2$ if $\mathbf{A}_1$ can be obtained from $\mathbf{A}_2$ by (finitely many!) elementary row operations.

Now the maximum number of linearly independent row vectors of a matrix does not change if we change the order of rows or multiply a row by an nonzero c or take a linear combination by adding a multiple of a row to another row. This proves that rank is **invariant** under elementary row operations:

THEOREM 1 Row-Equivalent Matrices

Row-equivalent matrices have the same rank.

Hence we can determine the rank of a matrix by reduction to row-echelon form (Sec. 7.3) and then see the rank directly.

EXAMPLE 3 Determination of Rank

For the matrix in Example 2 we obtain successively

$$\mathbf{A} = \begin{bmatrix} 3 & 0 & 2 & 2 \\ -6 & 42 & 24 & 54 \\ 21 & -21 & 0 & -15 \end{bmatrix} \quad \text{(given)}$$

$$\begin{bmatrix} 3 & 0 & 2 & 2 \\ 0 & 42 & 28 & 58 \\ 0 & -21 & -14 & -29 \end{bmatrix} \quad \begin{matrix} \\ \text{Row 2} + 2\ \text{Row 1} \\ \text{Row 3} - 7\ \text{Row 1} \end{matrix}$$

$$\begin{bmatrix} 3 & 0 & 2 & 2 \\ 0 & 42 & 28 & 58 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{Row 3} + \tfrac{1}{2}\ \text{Row 2}$$ ▩

Since rank is defined in terms of two vectors, we immediately have the useful

THEOREM 2 Linear Independence and Dependence of Vectors

p vectors with n components each are linearly independent if the matrix with these vectors as row vectors has rank p, but they are linearly dependent if that rank is less than p.

Further important properties will result from the basic

THEOREM 3

Rank in Terms of Column Vectors

The rank r of a matrix $\mathbf{A}$ equals the maximum number of linearly independent ***column*** *vectors of* $\mathbf{A}$.

Hence $\mathbf{A}$ *and its transpose* $\mathbf{A}^T$ *have the same rank.*

PROOF In this proof we write simply "rows" and "columns" for row and column vectors. Let $\mathbf{A}$ be an $m \times n$ matrix of rank $\mathbf{A} = r$. Then by definition of rank, $\mathbf{A}$ has r linearly independent rows which we denote by $\mathbf{v}_{(1)}, \cdots, \mathbf{v}_{(r)}$ (regardless of their position in $\mathbf{A}$), and all the rows $\mathbf{a}_{(1)}, \cdots, \mathbf{a}_{(m)}$ of $\mathbf{A}$ are linear combinations of those, say,

$$
\begin{aligned}
\mathbf{a}_{(1)} &= c_{11}\mathbf{v}_{(1)} + c_{12}\mathbf{v}_{(2)} + \cdots + c_{1r}\mathbf{v}_{(r)} \\
\mathbf{a}_{(2)} &= c_{21}\mathbf{v}_{(1)} + c_{22}\mathbf{v}_{(2)} + \cdots + c_{2r}\mathbf{v}_{(r)} \\
&\vdots \qquad\quad \vdots \qquad\quad \vdots \qquad\qquad \vdots \\
\mathbf{a}_{(m)} &= c_{m1}\mathbf{v}_{(1)} + c_{m2}\mathbf{v}_{(2)} + \cdots + c_{mr}\mathbf{v}_{(r)}.
\end{aligned} \tag{3}
$$

These are vector equations for rows. To switch to columns, we write (3) in terms of components as n such systems, with $k = 1, \cdots, n$,

$$
\begin{aligned}
a_{1k} &= c_{11}v_{1k} + c_{12}v_{2k} + \cdots + c_{1r}v_{rk} \\
a_{2k} &= c_{21}v_{1k} + c_{22}v_{2k} + \cdots + c_{2r}v_{rk} \\
&\vdots \qquad\quad \vdots \qquad\quad \vdots \qquad\qquad \vdots \\
a_{mk} &= c_{m1}v_{1k} + c_{m2}v_{2k} + \cdots + c_{mr}v_{rk}
\end{aligned} \tag{4}
$$

and collect components in columns. Indeed, we can write (4) as

$$
\begin{bmatrix} a_{1k} \\ a_{2k} \\ \vdots \\ a_{mk} \end{bmatrix} = v_{1k} \begin{bmatrix} c_{11} \\ c_{21} \\ \vdots \\ c_{m1} \end{bmatrix} + v_{2k} \begin{bmatrix} c_{12} \\ c_{22} \\ \vdots \\ c_{m2} \end{bmatrix} + \cdots + v_{rk} \begin{bmatrix} c_{1r} \\ c_{2r} \\ \vdots \\ c_{mr} \end{bmatrix} \tag{5}
$$

where $k = 1, \cdots, n$. Now the vector on the left is the kth column vector of $\mathbf{A}$. We see that each of these n columns is a linear combination of the same r columns on the right. Hence $\mathbf{A}$ cannot have more linearly independent columns than rows, whose number is rank $\mathbf{A} = r$. Now rows of $\mathbf{A}$ are columns of the transpose $\mathbf{A}^T$. For $\mathbf{A}^T$ our conclusion is that $\mathbf{A}^T$ cannot have more linearly independent columns than rows, so that $\mathbf{A}$ cannot have more linearly independent rows than columns. Together, the number of linearly independent columns of $\mathbf{A}$ must be r, the rank of $\mathbf{A}$. This completes the proof. ∎

EXAMPLE 4 **Illustration of Theorem 3**

The matrix in (2) has rank 2. From Example 3 we see that the first two row vectors are linearly independent and by "working backward" we can verify that Row 3 = 6 Row 1 $-\frac{1}{2}$ Row 2. Similarly, the first two columns are linearly independent, and by reducing the last matrix in Example 3 by columns we find that

$$\text{Column 3} = \tfrac{2}{3}\,\text{Column 1} + \tfrac{2}{3}\,\text{Column 2} \qquad \text{and} \qquad \text{Column 4} = \tfrac{2}{3}\,\text{Column 1} + \tfrac{29}{21}\,\text{Column 2}.$$ ∎

Combining Theorems 2 and 3 we obtain

THEOREM 4

Linear Dependence of Vectors

p vectors with $n < p$ components are always linearly dependent.

PROOF The matrix $\mathbf{A}$ with those p vectors as row vectors has p rows and $n < p$ columns; hence by Theorem 3 it has rank $\mathbf{A} \leqq n < p$, which implies linear dependence by Theorem 2. ∎

Vector Space

The following related concepts are of general interest in linear algebra. In the present context they provide a clarification of essential properties of matrices and their role in connection with linear systems.

A **vector space** is a (nonempty) set V of vectors such that with any two vectors $\mathbf{a}$ and $\mathbf{b}$ in V all their linear combinations $\alpha\mathbf{a} + \beta\mathbf{b}$ (α, β any real numbers) are elements of V, and these vectors satisfy the laws (3) and (4) in Sec. 7.1 (written in lowercase letters **a, b, u,** $\cdots$, which is our notation for vectors). (This definition is presently sufficient. General vector spaces will be discussed in Sec. 7.9.)

The maximum number of linearly independent vectors in V is called the **dimension** of V and is denoted by dim V. Here we assume the dimension to be finite; infinite dimension will be defined in Sec. 7.9.

A linearly independent set in V consisting of a maximum possible number of vectors in V is called a **basis** for V. Thus the number of vectors of a basis for V equals dim V.

The set of all linear combinations of given vectors $\mathbf{a}_{(1)}, \cdots, \mathbf{a}_{(p)}$ with the same number of components is called the **span** of these vectors. Obviously, a span is a vector space.

By a **subspace** of a vector space V we mean a nonempty subset of V (including V itself) that forms itself a vector space with respect to the two algebraic operations (addition and scalar multiplication) defined for the vectors of V.

EXAMPLE 5 **Vector Space, Dimension, Basis**

The span of the three vectors in Example 1 is a vector space of dimension 2, and a basis is $\mathbf{a}_{(1)}$, $\mathbf{a}_{(2)}$, for instance, or $\mathbf{a}_{(1)}$, $\mathbf{a}_{(3)}$, etc. ∎

We further note the simple

THEOREM 5

Vector Space R^n

The vector space R^n consisting of all vectors with n components (n real numbers) has dimension n.

PROOF A basis of n vectors is $\mathbf{a}_{(1)} = [1 \quad 0 \quad \cdots \quad 0]$, $\mathbf{a}_{(2)} = [0 \quad 1 \quad 0 \quad \cdots \quad 0], \cdots,$ $\mathbf{a}_{(n)} = [0 \quad \cdots \quad 0 \quad 1]$. ∎

In the case of a matrix $\mathbf{A}$ we call the span of the row vectors the **row space** of $\mathbf{A}$ and the span of the column vectors the **column space** of $\mathbf{A}$.

Now, Theorem 3 shows that a matrix **A** has as many linearly independent rows as columns. By the definition of dimension, their number is the dimension of the row space or the column space of **A.** This proves

THEOREM 6

Row Space and Column Space

The row space and the column space of a matrix **A** *have the same dimension, equal to rank* **A.**

Finally, for a given matrix **A** the solution set of the homogeneous system $\mathbf{Ax} = \mathbf{0}$ is a vector space, called the **null space** of **A,** and its dimension is called the **nullity** of **A.** In the next section we motivate and prove the basic relation

(6) $\quad \text{rank } \mathbf{A} + \text{nullity } \mathbf{A} = \text{Number of columns of } \mathbf{A}.$

PROBLEM SET 7.4

1–12 RANK, ROW SPACE, COLUMN SPACE

Find the rank and a basis for the row space and for the column space. *Hint.* Row-reduce the matrix and its transpose. (You may omit obvious factors from the vectors of these bases.)

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

2. $\begin{bmatrix} 8 & 2 & 5 \\ 16 & 6 & 29 \\ 4 & 0 & -7 \end{bmatrix}$

3. $\begin{bmatrix} 0 & -2 & 1 & 3 \\ 1 & 4 & 0 & 7 \\ 5 & 5 & 5 & 5 \end{bmatrix}$

4. $\begin{bmatrix} a & b & c \\ b & a & c \end{bmatrix}$

5. $\begin{bmatrix} 0 & 3 & 4 \\ -3 & 0 & -5 \\ -4 & 5 & 0 \end{bmatrix}$

6. $\begin{bmatrix} 1 & 1 & a \\ 1 & a & 1 \\ a & 1 & 1 \end{bmatrix}$

7. $\begin{bmatrix} 8 & 0 & 4 \\ 0 & 2 & 0 \\ 4 & 0 & 2 \\ 0 & 4 & 0 \end{bmatrix}$

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

9. $\begin{bmatrix} 1 & 0 & 3 & 0 \\ 0 & 5 & 8 & -37 \\ 3 & 8 & 7 & 0 \\ 0 & -37 & 0 & 37 \end{bmatrix}$

10. $\begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 5 \\ 3 & 4 & 5 & 6 \\ 4 & 5 & 6 & 7 \end{bmatrix}$

11. $\begin{bmatrix} 2 & 4 & 8 & 16 \\ 16 & 8 & 4 & 2 \\ 4 & 8 & 16 & 2 \\ 2 & 16 & 8 & 4 \end{bmatrix}$

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

13–20 LINEAR INDEPENDENCE

Are the following sets of vectors linearly independent? (Show the details.)

13. [3 −2 0 4], [5 0 0 1], [−6 1 0 1], [2 0 0 3]

14. [1 1 0], [1 0 0], [1 1 1]

15. [6 0 3 1 4 2], [0 −1 2 7 0 5], [12 3 0 −19 8 −11]

16. [3 4 7], [2 0 3], [8 2 3], [5 5 6]

17. [0.2 1.2 5.3 2.8 1.6], [4.3 3.4 0.9 2.0 −4.3]

18. [3 2 1], [0 0 0], [4 3 6]

19. $[1\ \frac{1}{2}\ \frac{1}{3}\ \frac{1}{4}]$, $[\frac{1}{2}\ \frac{1}{3}\ \frac{1}{4}\ \frac{1}{5}]$, $[\frac{1}{3}\ \frac{1}{4}\ \frac{1}{5}\ \frac{1}{6}]$, $[\frac{1}{4}\ \frac{1}{5}\ \frac{1}{6}\ \frac{1}{7}]$

20. [1 2 3 4], [2 3 4 5], [3 4 5 6], [4 5 6 7]

21. CAS Experiment. Rank. (a) Show experimentally that the $n \times n$ matrix $\mathbf{A} = [a_{jk}]$ with $a_{jk} = j + k - 1$ has rank 2 for any n. (Problem 20 shows $n = 4$.) Try to prove it.

(b) Do the same when $a_{jk} = j + k + c$, where c is any positive integer.

(c) What is rank $\mathbf{A}$ if $a_{jk} = 2^{j+k-2}$? Try to find other large matrices of low rank independent of n.

22–26 PROPERTIES OF RANK AND CONSEQUENCES

Show the following.

22. rank $\mathbf{B}^T\mathbf{A}^T$ = rank $\mathbf{AB}$. (Note the order!)

23. rank $\mathbf{A}$ = rank $\mathbf{B}$ does *not* imply rank $\mathbf{A}^2$ = rank $\mathbf{B}^2$. (Give a counterexample.)

24. If $\mathbf{A}$ is not square, either the row vectors or the column vectors of $\mathbf{A}$ are linearly dependent.

25. If the row vectors of a square matrix are linearly independent, so are the column vectors, and conversely.

26. Give examples showing that the rank of a product of matrices cannot exceed the rank of either factor.

27–36 VECTOR SPACES

Is the given set of vectors a vector space? (Give reason.) If your answer is yes, determine the dimension and find a basis. ($v_1, v_2, \cdots$ denote components.)

27. All vectors in R^3 such that $v_1 + v_2 = 0$

28. All vectors in R^4 such that $2v_2 - 3v_4 = k$

29. All vectors in R^3 with $v_1 \geqq 0$, $v_2 = -4v_3$

30. All vectors in R^2 with $v_1 \leqq v_2$

31. All vectors in R^3 with $4v_1 + v_3 = 0$, $3v_2 = v_3$

32. All vectors in R^4 with $v_1 - v_2 = 0$, $v_3 = 5v_1$, $v_4 = 0$

33. All vectors in R^n with $|v_j| \leqq 1$ for $j = 1, \cdots, n$

34. All ordered quadruples of positive real numbers

35. All vectors in R^5 with $v_1 = 2v_2 = 3v_3 = 4v_4 = 5v_5$

36. All vectors in R^4 with $3v_1 - v_3 = 0$, $2v_1 + 3v_2 - 4v_4 = 0$

7.5 Solutions of Linear Systems: Existence, Uniqueness

Rank as just defined gives complete information about existence, uniqueness, and general structure of the solution set of linear systems as follows.

A linear system of equations in n unknowns has a unique solution if the coefficient matrix and the augmented matrix have the same rank n, and infinitely many solution if that common rank is less than n. The system has no solution if those two matrices have different rank.

To state this precisely and prove it, we shall use the (generally important) concept of a **submatrix** of $\mathbf{A}$. By this we mean any matrix obtained from $\mathbf{A}$ by omitting some rows or columns (or both). By definition this includes $\mathbf{A}$ itself (as the matrix obtained by omitting no rows or columns); this is practical.

THEOREM 1

Fundamental Theorem for Linear Systems

(a) Existence. *A linear system of m equations in n unknowns $x_1, \cdots, x_n$*

$$
\begin{aligned}
a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\
a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\
&\cdots\cdots\cdots\cdots \\
a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m
\end{aligned}
\tag{1}
$$

is **consistent,** *that is, has solutions, if and only if the coefficient matrix* $\mathbf{A}$ *and the augmented matrix* $\tilde{\mathbf{A}}$ *have the same rank. Here,*

$$\mathbf{A} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \quad \text{and} \quad \tilde{\mathbf{A}} = \left[\begin{array}{ccc|c} a_{11} & \cdots & a_{1n} & b_1 \\ \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdots & \cdot & \cdot \\ a_{m1} & \cdots & a_{mn} & b_m \end{array}\right].$$

(b) Uniqueness. *The system* (1) *has precisely one solution if and only if this common rank* r *of* $\mathbf{A}$ *and* $\tilde{\mathbf{A}}$ *equals* n.

(c) Infinitely many solutions. *If this common rank* r *is less than* n, *the system* (1) *has infinitely many solutions. All of these solutions are obtained by determining* r *suitable unknowns (whose submatrix of coefficients must have rank* r*) in terms of the remaining* $n - r$ *unknowns, to which arbitrary values can be assigned.* (See Example 3 in Sec. 7.3.)

(d) Gauss elimination (Sec. 7.3). *If solutions exist, they can all be obtained by the Gauss elimination.* (This method will automatically reveal whether or not solutions exist; see Sec. 7.3.)

PROOF **(a)** We can write the system (1) in vector form $\mathbf{Ax} = \mathbf{b}$ or in terms of column vectors $\mathbf{c}_{(1)}, \cdots, \mathbf{c}_{(n)}$ of $\mathbf{A}$:

$$\mathbf{c}_{(1)}x_1 + \mathbf{c}_{(2)}x_2 + \cdots + \mathbf{c}_{(n)}x_n = \mathbf{b}. \tag{2}$$

$\tilde{\mathbf{A}}$ is obtained by augmenting $\mathbf{A}$ by a single column $\mathbf{b}$. Hence, by Theorem 3 in Sec. 7.4, rank $\tilde{\mathbf{A}}$ equals rank $\mathbf{A}$ or rank $\mathbf{A} + 1$. Now if (1) has a solution $\mathbf{x}$, then (2) shows that $\mathbf{b}$ must be a linear combination of those column vectors, so that $\tilde{\mathbf{A}}$ and $\mathbf{A}$ have the same maximum number of linearly independent column vectors and thus the same rank.

Conversely, if rank $\tilde{\mathbf{A}}$ = rank $\mathbf{A}$, then $\mathbf{b}$ must be a linear combination of the column vectors of $\mathbf{A}$, say,

$$\mathbf{b} = \alpha_1\mathbf{c}_{(1)} + \cdots + \alpha_n\mathbf{c}_{(n)} \tag{2*}$$

since otherwise rank $\tilde{\mathbf{A}}$ = rank $\mathbf{A} + 1$. But (2*) means that (1) has a solution, namely, $x_1 = \alpha_1, \cdots, x_n = \alpha_n$, as can be seen by comparing (2*) and (2).

(b) If rank $\mathbf{A} = n$, the n column vectors in (2) are linearly independent by Theorem 3 in Sec. 7.4. We claim that then the representation (2) of $\mathbf{b}$ is unique because otherwise

$$\mathbf{c}_{(1)}x_1 + \cdots + \mathbf{c}_{(n)}x_n = \mathbf{c}_{(1)}\tilde{x}_1 + \cdots + \mathbf{c}_{(n)}\tilde{x}_n.$$

This would imply (take all terms to the left, with a minus sign)

$$(x_1 - \tilde{x}_1)\mathbf{c}_{(1)} + \cdots + (x_n - \tilde{x}_n)\mathbf{c}_{(n)} = \mathbf{0}$$

and $x_1 - \tilde{x}_1 = 0, \cdots, x_n - \tilde{x}_n = 0$ by linear independence. But this means that the scalars $x_1, \cdots, x_n$ in (2) are uniquely determined, that is, the solution of (1) is unique.

(c) If rank $\mathbf{A}$ = rank $\widetilde{\mathbf{A}} = r < n$, then by Theorem 3 in Sec. 7.4 there is a linearly independent set K of r column vectors of $\mathbf{A}$ such that the other $n - r$ column vectors of $\mathbf{A}$ are linear combinations of those vectors. We renumber the columns and unknowns, denoting the renumbered quantities by $\hat{}$, so that $\{\hat{\mathbf{c}}_{(1)}, \cdots, \hat{\mathbf{c}}_{(r)}\}$ is that linearly independent set K. Then (2) becomes

$$\hat{\mathbf{c}}_{(1)}\hat{x}_1 + \cdots + \hat{\mathbf{c}}_{(r)}\hat{x}_r + \hat{\mathbf{c}}_{(r+1)}\hat{x}_{r+1} + \cdots + \hat{\mathbf{c}}_{(n)}\hat{x}_n = \mathbf{b},$$

$\hat{\mathbf{c}}_{(r+1)}, \cdots, \hat{\mathbf{c}}_{(n)}$ are linear combinations of the vectors of K, and so are the vectors $\hat{x}_{r+1}\hat{\mathbf{c}}_{(r+1)}, \cdots, \hat{x}_n\hat{\mathbf{c}}_{(n)}$. Expressing these vectors in terms of the vectors of K and collecting terms, we can thus write the system in the form

$$\hat{\mathbf{c}}_{(1)}y_1 + \cdots + \hat{\mathbf{c}}_{(r)}y_r = \mathbf{b} \tag{3}$$

with $y_j = \hat{x}_j + \beta_j$, where β_j results from the $n - r$ terms $\hat{\mathbf{c}}_{(r+1)}\hat{x}_{r+1}, \cdots, \hat{\mathbf{c}}_{(n)}\hat{x}_n$; here, $j = 1, \cdots, r$. Since the system has a solution, there are $y_1, \cdots, y_r$ satisfying (3). These scalars are unique since K is linearly independent. Choosing $\hat{x}_{r+1}, \cdots, \hat{x}_n$ fixes the β_j and corresponding $\hat{x}_j = y_j - \beta_j$, where $j = 1, \cdots, r$.

(d) This was discussed in Sec. 7.3 and is restated here as a reminder. ∎

The theorem is illustrated in Sec. 7.3. In Example 2 there is a unique solution since rank $\widetilde{\mathbf{A}}$ = rank $\mathbf{A} = n = 3$ (as can be seen from the last matrix in the example). In Example 3 we have rank $\widetilde{\mathbf{A}}$ = rank $\mathbf{A} = 2 < n = 4$ and can choose x_3 and x_4 arbitrarily. In Example 4 there is no solution because rank $\mathbf{A} = 2 <$ rank $\widetilde{\mathbf{A}} = 3$.

Homogeneous Linear System

Recall from Sec. 7.3 that a linear system (1) is called **homogeneous** if all the b_j's are zero, and **nonhomogeneous** if one or several b_j's are not zero. For the homogeneous system we obtain from the Fundamental Theorem the following results.

THEOREM 2

Homogeneous Linear System

A homogeneous linear system

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= 0 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= 0 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= 0 \end{aligned} \tag{4}$$

always has the **trivial solution** $x_1 = 0, \cdots, x_n = 0$. *Nontrivial solutions exist if and only if rank* $\mathbf{A} < n$. *If* rank $\mathbf{A} = r < n$, *these solutions, together with* $\mathbf{x} = \mathbf{0}$, *form a vector space* (see Sec. 7.4) *of dimension* $n - r$, called the **solution space** of (4).

In particular, if $\mathbf{x}_{(1)}$ *and* $\mathbf{x}_{(2)}$ *are solution vectors of* (4), *then* $\mathbf{x} = \mathbf{c}_1\mathbf{x}_{(1)} + \mathbf{c}_2\mathbf{x}_{(2)}$ *with any scalars* c_1 *and* c_2 *is a solution vector of* (4). (This ***does not hold*** for nonhomogeneous systems. Also, the term *solution space* is used for *homogeneous* systems only.)

PROOF The first proposition can be seen directly from the system. It agrees with the fact that $\mathbf{b} = \mathbf{0}$ implies that rank $\tilde{\mathbf{A}}$ = rank $\mathbf{A}$, so that a homogeneous system is always ***consistent.*** If rank $\mathbf{A} = n$, the trivial solution is the unique solution according to (b) in Theorem 1. If rank $\mathbf{A} < n$, there are nontrivial solutions according to (c) in Theorem 1. The solutions form a vector space because if $\mathbf{x}_{(1)}$ and $\mathbf{x}_{(2)}$ are any of them, then $\mathbf{A}\mathbf{x}_{(1)} = \mathbf{0}$, $\mathbf{A}\mathbf{x}_{(2)} = \mathbf{0}$, and this implies $\mathbf{A}(\mathbf{x}_{(1)} + \mathbf{x}_{(2)}) = \mathbf{A}\mathbf{x}_{(1)} + \mathbf{A}\mathbf{x}_{(2)} = \mathbf{0}$ as well as $\mathbf{A}(c\mathbf{x}_{(1)}) = c\mathbf{A}\mathbf{x}_{(1)} = \mathbf{0}$, where c is arbitrary. If rank $\mathbf{A} = r < n$, Theorem 1 (c) implies that we can choose $n - r$ suitable unknowns, call them $x_{r+1}, \cdots, x_n$, in an arbitrary fashion, and every solution is obtained in this way. Hence a basis for the solution space, briefly called a **basis of solutions** of (4), is $\mathbf{y}_{(1)}, \cdots, \mathbf{y}_{(n-r)}$, where the basis vector $\mathbf{y}_{(j)}$ is obtained by choosing $x_{r+j} = 1$ and the other $x_{r+1}, \cdots, x_n$ zero; the corresponding first r components of this solution vector are then determined. Thus the solution space of (4) has dimension $n - r$. This proves Theorem 2. ∎

The solution space of (4) is also called the **null space** of $\mathbf{A}$ because $\mathbf{A}\mathbf{x} = \mathbf{0}$ for every $\mathbf{x}$ in the solution space of (4). Its dimension is called the **nullity** of $\mathbf{A}$. Hence Theorem 2 states that

$$\text{rank } \mathbf{A} + \text{nullity } \mathbf{A} = n \tag{5}$$

where n is the number of unknowns (number of columns of $\mathbf{A}$).

Furthermore, by the definition of rank we have rank $\mathbf{A} \leqq m$ in (4). Hence if $m < n$, then rank $\mathbf{A} < n$. By Theorem 2 this gives the practically important

THEOREM 3

Homogeneous Linear System with Fewer Equations Than Unknowns

A homogeneous linear system with fewer equations than unknowns has always nontrivial solutions.

Nonhomogeneous Linear Systems

The characterization of all solutions of the linear system (1) is now quite simple, as follows.

THEOREM 4

Nonhomogeneous Linear System

If a nonhomogeneous linear system (1) *is consistent, then all of its solutions are obtained as*

$$\mathbf{x} = \mathbf{x}_0 + \mathbf{x}_h \tag{6}$$

where $\mathbf{x}_0$ *is any* (fixed) *solution of* (1) *and* $\mathbf{x}_h$ *runs through all the solutions of the corresponding homogeneous system* (4).

PROOF The difference $\mathbf{x}_h = \mathbf{x} - \mathbf{x}_0$ of any two solutions of (1) is a solution of (4) because $\mathbf{A}\mathbf{x}_h = \mathbf{A}(\mathbf{x} - \mathbf{x}_0) = \mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}_0 = \mathbf{b} - \mathbf{b} = \mathbf{0}$. Since $\mathbf{x}$ is any solution of (1), we get all the solutions of (1) if in (6) we take any solution $\mathbf{x}_0$ of (1) and let $\mathbf{x}_h$ vary throughout the solution space of (4). ∎

7.6 For Reference: Second- and Third-Order Determinants

We explain these determinants separately from the general theory in Sec. 7.7 because they will be sufficient for many of our examples and problems. Since this section is for reference, ***go on to the next section, consulting this material only when needed.***

A **determinant of second order** is denoted and defined by

$$D = \det \mathbf{A} = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}. \tag{1}$$

So here we have ***bars*** (whereas a matrix has ***brackets***).

Cramer's rule for solving linear systems of two equations in two unknowns

$$\begin{aligned} &\text{(a)} \quad a_{11}x_1 + a_{12}x_2 = b_1 \\ &\text{(b)} \quad a_{21}x_1 + a_{22}x_2 = b_2 \end{aligned} \tag{2}$$

is

$$x_1 = \frac{\begin{vmatrix} b_1 & a_{12} \\ b_2 & a_{22} \end{vmatrix}}{D} = \frac{b_1a_{22} - a_{12}b_2}{D}, \tag{3}$$

$$x_2 = \frac{\begin{vmatrix} a_{11} & b_1 \\ a_{21} & b_2 \end{vmatrix}}{D} = \frac{a_{11}b_2 - b_1a_{21}}{D}.$$

with D as in (1), provided

$$D \neq 0.$$

The value $D = 0$ appears for inconsistent nonhomogeneous systems and for homogeneous systems with nontrivial solutions.

PROOF We prove (3). To eliminate x_2, multiply (2a) by a_{22} and (2b) by $-a_{12}$ and add,

$$(a_{11}a_{22} - a_{12}a_{21})x_1 = b_1a_{22} - a_{12}b_2.$$

Similarly, to eliminate x_1, multiply (2a) by $-a_{21}$ and (2b) by a_{11} and add,

$$(a_{11}a_{22} - a_{12}a_{21})x_2 = a_{11}b_2 - b_1a_{21}.$$

Assuming that $D = a_{11}a_{22} - a_{12}a_{21} \neq 0$, dividing, and writing the right sides of these two equations as determinants, we obtain (3). ∎

EXAMPLE 1 Cramer's Rule for Two Equations

If $4x_1 + 3x_2 = 12$, $2x_1 + 5x_2 = -8$ then $x_1 = \dfrac{\begin{vmatrix} 12 & 3 \\ -8 & 5 \end{vmatrix}}{\begin{vmatrix} 4 & 3 \\ 2 & 5 \end{vmatrix}} = \dfrac{84}{14} = 6$, $x_2 = \dfrac{\begin{vmatrix} 4 & 12 \\ 2 & -8 \end{vmatrix}}{\begin{vmatrix} 4 & 3 \\ 2 & 5 \end{vmatrix}} = \dfrac{-56}{14} = -4.$ ■

Third-Order Determinants

A **determinant of third order** can be defined by

$$(4) \qquad D = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{21} \begin{vmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{vmatrix} + a_{31} \begin{vmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{vmatrix}.$$

Note the following. The signs on the right are $+ \; - \; +$. Each of the three terms on the right is an entry in the first column of D times its **minor,** that is, the second-order determinant obtained from D by deleting the row and column of that entry; thus, for a_{11} delete the first row and first column, and so on.

If we write out the minors in (4), we obtain

$$(4^*) \qquad D = a_{11}a_{22}a_{33} - a_{11}a_{23}a_{32} + a_{21}a_{13}a_{32} - a_{21}a_{12}a_{33} + a_{31}a_{12}a_{23} - a_{31}a_{13}a_{22}.$$

Cramer's Rule for Linear Systems of Three Equations

$$(5) \qquad \begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3 \end{aligned}$$

is

$$(6) \qquad x_1 = \frac{D_1}{D}, \qquad x_2 = \frac{D_2}{D}, \qquad x_3 = \frac{D_3}{D} \qquad (D \neq 0)$$

with the *determinant D of the system* given by (4) and

$$D_1 = \begin{vmatrix} b_1 & a_{12} & a_{13} \\ b_2 & a_{22} & a_{23} \\ b_3 & a_{32} & a_{33} \end{vmatrix}, \quad D_2 = \begin{vmatrix} a_{11} & b_1 & a_{13} \\ a_{21} & b_2 & a_{23} \\ a_{31} & b_3 & a_{33} \end{vmatrix}, \quad D_3 = \begin{vmatrix} a_{11} & a_{12} & b_1 \\ a_{21} & a_{22} & b_2 \\ a_{31} & a_{32} & b_3 \end{vmatrix}.$$

Note that D_1, D_2, D_3 are obtained by replacing Columns 1, 2, 3, respectively, by the column of the right sides of (5).

Cramer's rule (6) can be derived by eliminations similar to those for (3), but it also follows from the general case (Theorem 4) in the next section.

7.7 Determinants. Cramer's Rule

Determinants were originally introduced for solving linear systems. Although ***impractical in computations,*** they have important engineering applications in eigenvalue problems (Sec. 8.1), differential equations, vector algebra (Sec. 9.3), and so on. They can be introduced in several equivalent ways. Our definition is particularly practical in connection with linear systems.

A **determinant of order** n is a scalar associated with an $n \times n$ (hence ***square!***) matrix $\mathbf{A} = [a_{jk}]$, which is written

(1)
$$D = \det \mathbf{A} = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix}$$

and is defined for $n = 1$ by

(2)
$$D = a_{11}$$

and for $n \geqq 2$ by

(3a)
$$D = a_{j1}C_{j1} + a_{j2}C_{j2} + \cdots + a_{jn}C_{jn} \qquad (j = 1, 2, \cdots, \text{or } n)$$

or

(3b)
$$D = a_{1k}C_{1k} + a_{2k}C_{2k} + \cdots + a_{nk}C_{nk} \qquad (k = 1, 2, \cdots, \text{or } n)$$

Here,

$$C_{jk} = (-1)^{j+k}M_{jk}$$

and M_{jk} is a determinant of order $n - 1$, namely, the determinant of the submatrix of $\mathbf{A}$ obtained from $\mathbf{A}$ by omitting the row and column of the entry a_{jk}, that is, the jth row and the kth column.

In this way, D is defined in terms of n determinants of order $n - 1$, each of which is, in turn, defined in terms of $n - 1$ determinants of order $n - 2$, and so on; we finally arrive at second-order determinants, in which those submatrices consist of single entries whose determinant is defined to be the entry itself.

From the definition it follows that *we may* ***expand*** *D by any row or column,* that is, choose in (3) the entries in any row or column, similarly when expanding the C_{jk}'s in (3), and so on.

This definition is unambiguous, that is, yields the same value for D no matter which columns or rows we choose in expanding. A proof is given in App. 4.

Terms used in connection with determinants are taken from matrices. In D we have n^2 **entries** a_{jk}, also n **rows** and n **columns,** and a **main diagonal** on which $a_{11}, a_{22}, \cdots, a_{nn}$ stand. Two terms are new:

M_{jk} is called the **minor** *of* a_{jk} *in* D, and C_{jk} the **cofactor** *of* a_{jk} *in* D.

For later use we note that (3) may also be written in terms of minors

(4a)
$$D = \sum_{k=1}^{n} (-1)^{j+k} a_{jk} M_{jk} \qquad (j = 1, 2, \cdots, \text{or } n)$$

(4b)
$$D = \sum_{j=1}^{n} (-1)^{j+k} a_{jk} M_{jk} \qquad (k = 1, 2, \cdots, \text{or } n).$$

EXAMPLE 1 **Minors and Cofactors of a Third-Order Determinant**

In (4) of the previous section the minors and cofactors of the entries in the first column can be seen directly. For the entries in the second row the minors are

$$M_{21} = \begin{vmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{vmatrix}, \qquad M_{22} = \begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix}, \qquad M_{23} = \begin{vmatrix} a_{11} & a_{12} \\ a_{31} & a_{32} \end{vmatrix}$$

and the cofactors are $C_{21} = -M_{21}$, $C_{22} = +M_{22}$, and $C_{23} = -M_{23}$. Similarly for the third row—write these down yourself. And verify that the signs in C_{jk} form a **checkerboard pattern**

$$\begin{matrix} + & - & + \\ - & + & - \\ + & - & + \end{matrix}$$

EXAMPLE 2 **Expansions of a Third-Order Determinant**

$$D = \begin{vmatrix} 1 & 3 & 0 \\ 2 & 6 & 4 \\ -1 & 0 & 2 \end{vmatrix} = 1 \begin{vmatrix} 6 & 4 \\ 0 & 2 \end{vmatrix} - 3 \begin{vmatrix} 2 & 4 \\ -1 & 2 \end{vmatrix} + 0 \begin{vmatrix} 2 & 6 \\ -1 & 0 \end{vmatrix}$$

$$= 1(12 - 0) - 3(4 + 4) + 0(0 + 6) = -12.$$

This is the expansion by the first row. The expansion by the third column is

$$D = 0 \begin{vmatrix} 2 & 6 \\ -1 & 0 \end{vmatrix} - 4 \begin{vmatrix} 1 & 3 \\ -1 & 0 \end{vmatrix} + 2 \begin{vmatrix} 1 & 3 \\ 2 & 6 \end{vmatrix} = 0 - 12 + 0 = -12,$$

Verify that the other four expansions also give the value -12.

EXAMPLE 3 **Determinant of a Triangular Matrix**

$$\begin{vmatrix} -3 & 0 & 0 \\ 6 & 4 & 0 \\ -1 & 2 & 5 \end{vmatrix} = -3 \begin{vmatrix} 4 & 0 \\ 2 & 5 \end{vmatrix} = -3 \cdot 4 \cdot 5 = -60.$$

Inspired by this, can you formulate a little theorem on determinants of triangular matrices? Of diagonal matrices?

General Properties of Determinants

To obtain the value of a determinant (1), we can first simplify it systematically by elementary row operations, similar to those for matrices in Sec. 7.3, as follows.

THEOREM 1

Behavior of an *n*th-Order Determinant under Elementary Row Operations

(a) *Interchange of two rows multiplies the value of the determinant by* -1.

(b) *Addition of a multiple of a row to another row does not alter the value of the determinant.*

(c) *Multiplication of a row by a nonzero constant c multiplies the value of the determinant by c.* (This holds also when $c = 0$, but gives no longer an elementary row operation.)

PROOF **(a)** By induction. The statement holds for $n = 2$ because

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc, \qquad \text{but} \qquad \begin{vmatrix} c & d \\ a & b \end{vmatrix} = bc - ad.$$

We now make the induction hypothesis that (a) holds for determinants of order $n - 1 \geqq 2$ and show that it then holds for determinants of order n. Let D be of order n. Let E be obtained from D by the interchange of two rows. Expand D and E by a row that is ***not*** one of those interchanged, call it the jth row. Then by (4a),

$$(5) \qquad D = \sum_{k=1}^{n} (-1)^{j+k} a_{jk} M_{jk}, \qquad E = \sum_{k=1}^{n} (-1)^{j+k} a_{jk} N_{jk}$$

where N_{jk} is obtained from the minor M_{jk} of a_{jk} in D by the interchange of those two rows which have been interchanged in D (and which N_{jk} must both contain because we expand by another row!). Now these minors are of order $n - 1$. Hence the induction hypothesis applies and gives $N_{jk} = -M_{jk}$. Thus $E = -D$ by (5).

(b) Add c times Row i to Row j. Let $\tilde{D}$ be the new determinant. Its entries in Row j are $a_{jk} + ca_{ik}$. If we expand $\tilde{D}$ by this Row j, we see that we can write it as $\tilde{D} = D_1 + cD_2$, where $D_1 = D$ has in Row j the a_{jk}, whereas D_2 has in that Row j the a_{ik} from the addition. Hence D_2 has a_{ik} in both Row i and Row j. Interchanging these two rows gives D_2 back, but on the other hand it gives $-D_2$ by (a). Together $D_2 = -D_2 = 0$, so that $\tilde{D} = D_1 = D$.

(c) Expand the determinant by the row that has been multiplied.

CAUTION! $\det(c\mathbf{A}) = c^n \det \mathbf{A}$ (not $c \det \mathbf{A}$). Explain why. ▩

EXAMPLE 4 **Evaluation of Determinants by Reduction to Triangular Form**

Because of Theorem 1 we may evaluate determinants by reduction to triangular form, as in the Gauss elimination for a matrix. For instance (with the blue explanations always referring to the ***preceding determinant***)

$$D = \begin{vmatrix} 2 & 0 & -4 & 6 \\ 4 & 5 & 1 & 0 \\ 0 & 2 & 6 & -1 \\ -3 & 8 & 9 & 1 \end{vmatrix}$$

$$= \begin{vmatrix} 2 & 0 & -4 & 6 \\ 0 & 5 & 9 & -12 \\ 0 & 2 & 6 & -1 \\ 0 & 8 & 3 & 10 \end{vmatrix} \quad \begin{matrix} \\ \text{Row 2} - 2\ \text{Row 1} \\ \\ \text{Row 4} + 1.5\ \text{Row 1} \end{matrix}$$

$$= \begin{vmatrix} 2 & 0 & -4 & 6 \\ 0 & 5 & 9 & -12 \\ 0 & 0 & 2.4 & 3.8 \\ 0 & 0 & -11.4 & 29.2 \end{vmatrix} \quad \begin{matrix} \\ \\ \text{Row 3} - 0.4\ \text{Row 2} \\ \text{Row 4} - 1.6\ \text{Row 2} \end{matrix}$$

$$= \begin{vmatrix} 2 & 0 & -4 & 6 \\ 0 & 5 & 9 & -12 \\ 0 & 0 & 2.4 & 3.8 \\ 0 & 0 & -0 & 47.25 \end{vmatrix} \quad \text{Row 4} + 4.75\ \text{Row 3}$$

$$= 2 \cdot 5 \cdot 2.4 \cdot 47.25 = 1134.$$

■

THEOREM 2

Further Properties of *n*th-Order Determinants

(a)–(c) *in Theorem* 1 *hold also for columns.*

(d) ***Transposition*** *leaves the value of a determinant unaltered.*

(e) ***A zero row or column*** *renders the value of a determinant zero.*

(f) ***Proportional rows or columns*** *render the value of a determinant zero. In particular, a determinant with two identical rows or columns has the value zero.*

PROOF **(a)–(e)** follow directly from the fact that a determinant can be expanded by any row column. In **(d),** transposition is defined as for matrices, that is, the jth row becomes the jth column of the transpose.

(f) If Row j = c times Row i, then $D = cD_1$, where D_1 has Row j = Row i. Hence an interchange of these rows reproduces D_1, but it also gives $-D_1$ by Theorem 1(a). Hence $D_1 = 0$ and $D = cD_1 = 0$. Similarly for columns. ■

It is quite remarkable that the important concept of the rank of a matrix **A,** which is the maximum number of linearly independent row or column vectors of **A** (see Sec. 7.4), can be related to determinants. Here we may assume that rank $\mathbf{A} > 0$ because the only matrices with rank 0 are the zero matrices (see Sec. 7.4).

THEOREM 3

Rank in Terms of Determinants

An $m \times n$ matrix $\mathbf{A} = [a_{jk}]$ has rank $r \geqq 1$ if and only if **A** *has an $r \times r$ submatrix with nonzero determinant, whereas every square submatrix with more than r rows that* **A** *has (or does not have!) has determinant equal to zero.*

In particular, if **A** *is square, $n \times n$, it has rank n if and only if*

$$\det \mathbf{A} \neq 0.$$

PROOF The key idea is that elementary row operations (Sec. 7.3) alter neither rank (by Theorem 1 in Sec. 7.4) nor the property of a determinant being nonzero (by Theorem 1 in this section). The echelon form $\hat{\mathbf{A}}$ of $\mathbf{A}$ (see Sec. 7.3) has r nonzero row vectors (which are the first r row vectors) if and only if rank $\mathbf{A} = r$. Let $\hat{\mathbf{R}}$ be the $r \times r$ submatrix in the left upper corner of $\hat{\mathbf{A}}$ (so that the entries of $\hat{\mathbf{R}}$ are in both the first r rows and r columns of $\hat{\mathbf{A}}$). Now $\hat{\mathbf{R}}$ is triangular, with all diagonal entries r_{jj} nonzero. Thus, $\det \hat{\mathbf{R}} = r_{11} \cdots r_{rr} \neq 0$. Also $\det \mathbf{R} \neq 0$ for the corresponding $r \times r$ submatrix $\mathbf{R}$ of $\mathbf{A}$ because $\hat{\mathbf{R}}$ results from $\mathbf{R}$ by elementary row operations. Similarly, $\det \mathbf{S} = 0$ for any square submatrix $\mathbf{S}$ of $r + 1$ or more rows perhaps contained in $\mathbf{A}$ because the corresponding submatrix $\hat{\mathbf{S}}$ of $\hat{\mathbf{A}}$ must contain a row of zeros (otherwise we would have rank $\mathbf{A} \geqq r + 1$), so that $\det \hat{\mathbf{S}} = 0$ by Theorem 2. This proves the theorem for an $m \times n$ matrix.

In particular, if $\mathbf{A}$ is square, $n \times n$, then rank $\mathbf{A} = n$ if and only if $\mathbf{A}$ contains an $n \times n$ submatrix with nonzero determinant. But the only such submatrix can be $\mathbf{A}$ itself, hence $\det \mathbf{A} \neq 0$. ■

Cramer's Rule

Theorem 3 opens the way to the classical solution formula for linear systems known as **Cramer's rule**[2], which gives solutions as quotients of determinants. ***Cramer's rule is not practical in computations*** (for which the methods in Secs. 7.3 and 20.1–20.3 are suitable), but is of ***theoretical interest*** in differential equations (Secs. 2.10, 3.3) and other theories that have engineering applications.

THEOREM 4

Cramer's Theorem (Solution of Linear Systems by Determinants)

(a) *If a linear system of n equations in the same number of unknowns* $x_1, \cdots, x_n$

$$(6)\qquad \begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ \cdots\cdots\cdots\cdots\cdots&\cdots\cdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n \end{aligned}$$

has a nonzero coefficient determinant $D = \det \mathbf{A}$, *the system has precisely one solution. This solution is given by the formulas*

$$(7)\qquad x_1 = \frac{D_1}{D}, \quad x_2 = \frac{D_2}{D}, \quad \cdots, \quad x_n = \frac{D_n}{D} \qquad \textbf{(Cramer's rule)}$$

where D_k *is the determinant obtained from D by replacing in D the kth column by the column with the entries* $b_1, \cdots, b_n$.

(b) *Hence if the system* (6) *is* ***homogeneous*** *and* $D \neq 0$, *it has only the trivial solution* $x_1 = 0, x_2 = 0, \cdots, x_n = 0$. *If* $D = 0$, *the homogeneous system also has nontrivial solutions.*

[2]GABRIEL CRAMER (1704–1752), Swiss mathematician.

PROOF The augmented matrix $\tilde{\mathbf{A}}$ of the system (6) is of size $n \times (n+1)$. Hence its rank can be at most n. Now if

$$D = \det \mathbf{A} = \begin{vmatrix} a_{11} & \cdots & a_{1n} \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ a_{n1} & \cdots & a_{nn} \end{vmatrix} \neq 0, \tag{8}$$

then rank $\mathbf{A} = n$ by Theorem 3. Thus rank $\tilde{\mathbf{A}}$ = rank $\mathbf{A}$. Hence, by the Fundamental Theorem in Sec. 7.5, the system (6) has a unique solution.

Let us now prove (7). Expanding D by its kth column, we obtain

$$D = a_{1k}C_{1k} + a_{2k}C_{2k} + \cdots + a_{nk}C_{nk}, \tag{9}$$

where C_{ik} is the cofactor of entry a_{ik} in D. If we replace the entries in the kth column of D by any other numbers, we obtain a new determinant, say, $\hat{D}$. Clearly, its expansion by the kth column will be of the form (9), with $a_{1k}, \cdots, a_{nk}$ replaced by those new numbers and the cofactors C_{ik} as before. In particular, if we choose as new numbers the entries $a_{1l}, \cdots, a_{nl}$ of the lth column of D (where $l \neq k$), we have a new determinant $\hat{D}$ which has twice the column $[a_{1l} \quad \cdots \quad a_{nl}]^T$, once as its lth column, and once as its kth because of the replacement. Hence $\hat{D} = 0$ by Theorem 2(f). If we now expand $\hat{D}$ by the column that has been replaced (the kth column), we thus obtain

$$a_{1l}C_{1k} + a_{2l}C_{2k} + \cdots + a_{nl}C_{nk} = 0 \qquad (l \neq k). \tag{10}$$

We now multiply the first equation in (6) by C_{1k} on both sides, the second by $C_{2k}, \cdots$, the last by C_{nk}, and add the resulting equations. This gives

$$\begin{aligned} &C_{1k}(a_{11}x_1 + \cdots + a_{1n}x_n) + \cdots + C_{nk}(a_{n1}x_1 + \cdots + a_{nn}x_n) \\ &\qquad = b_1C_{1k} + \cdots + b_nC_{nk}. \end{aligned} \tag{11}$$

Collecting terms with the same x_j, we can write the left side as

$$x_1(a_{11}C_{1k} + a_{21}C_{2k} + \cdots + a_{n1}C_{nk}) + \cdots + x_n(a_{1n}C_{1k} + a_{2n}C_{2k} + \cdots + a_{nn}C_{nk}).$$

From this we see that x_k is multiplied by

$$a_{1k}C_{1k} + a_{2k}C_{2k} + \cdots + a_{nk}C_{nk}.$$

Equation (9) shows that this equals D. Similarly, x_l is multiplied by

$$a_{1l}C_{1k} + a_{2l}C_{2k} + \cdots + a_{nl}C_{nk}.$$

Equation (10) shows that this is zero when $l \neq k$. Accordingly, the left side of (11) equals simply x_kD, so that (11) becomes

$$x_kD = b_1C_{1k} + b_2C_{2k} + \cdots + b_nC_{nk}.$$

Now the right side of this is D_k as defined in the theorem, expanded by its kth column, so that division by D gives (7). This proves Cramer's rule.

If (6) is homogeneous and $D \neq 0$, then each D_k has a column of zeros, so that $D_k = 0$ by Theorem 2(e), and (7) gives the trivial solution.

Finally, if (6) is homogeneous and $D = 0$, then rank $\mathbf{A} < n$ by Theorem 3, so that nontrivial solutions exist by Theorem 2 in Sec. 7.5. ■

Illustrations of Theorem 4 for $n = 2$ and 3 are given in Sec. 7.6, and an important application of the present formulas will follow in the next section.

PROBLEM SET 7.7

1. **(Second-order determinant)** Expand a general second-order determinant in four possible ways and show that the results agree.
2. **(Minors, cofactors)** Complete the list of minors and cofactors in Example 1.
3. **(Third-order determinant)** Do the task indicated in Example 2. Also evaluate D by reduction to triangular form.
4. **(Scalar multiplication)** Show that $\det(k\mathbf{A}) = k^n \det \mathbf{A}$ (not $k \det \mathbf{A}$), where $\mathbf{A}$ is any $n \times n$ matrix. Give an example.

5–16 EVALUATION OF DETERMINANTS

Evaluate, showing the details of your work.

5. $\begin{vmatrix} 13 & 8 \\ -2 & 7 \end{vmatrix}$

6. $\begin{vmatrix} \cos n\theta & \sin n\theta \\ -\sin n\theta & \cos n\theta \end{vmatrix}$

7. $\begin{vmatrix} \cos\alpha & \sin\alpha \\ \sin\beta & \cos\beta \end{vmatrix}$

8. $\begin{vmatrix} 14 & 2 & 5 \\ 2 & 0 & 8 \\ 5 & 8 & -2 \end{vmatrix}$

9. $\begin{vmatrix} 70.4 & 0.3 & 0.8 \\ 0 & 0.5 & 2.6 \\ 0 & 0 & -1.9 \end{vmatrix}$

10. $\begin{vmatrix} 2 & 1 & 2 \\ -2 & 2 & 1 \\ 1 & 2 & -2 \end{vmatrix}$

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

12. $\begin{vmatrix} 0 & a & b \\ -a & 0 & c \\ -b & -c & 0 \end{vmatrix}$

13. $\begin{vmatrix} u & v & w \\ w & u & v \\ v & w & u \end{vmatrix}$

14. $\begin{vmatrix} 1 & -2 & 0 & 0 \\ 4 & 3 & 5 & 0 \\ 0 & 2 & 7 & 5 \\ 0 & 0 & 2 & 4 \end{vmatrix}$

15. $\begin{vmatrix} 1 & 2 & 0 & 0 \\ 3 & 4 & 0 & 0 \\ 0 & 0 & 5 & 6 \\ 0 & 0 & 7 & 8 \end{vmatrix}$

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

17. **(Expansion numerically impractical)** Show that the computation of an nth-order determinant by expansion involves $n!$ multiplications, which if a multiplication takes 10^{-9} sec would take these times:

n	10	15	20	25
Time	0.004 sec	22 min	77 years	$0.5 \cdot 10^9$ years

18–20 CRAMER'S RULE

Solve by Cramer's rule and check by Gauss elimination and back substitution. (Show details.)

18. $2x - 5y = 23$
$4x + 6y = -2$

19. $3y + 4z = 14.8$
$4x + 2y - z = -6.3$
$x - y + 5z = 13.5$

20. $w + 2x - 3z = 30$
$4x - 5y + 2z = 13$
$2w + 8x - 4y + z = 42$
$3w + y - 5z = 35$

21–23 RANK BY DETERMINANTS

Find the rank by Theorem 3 (which is not a very practical way) and check by row reduction. (Show details.)

21. $\begin{bmatrix} 8 & 4 \\ -2 & -1 \\ 6 & 3 \end{bmatrix}$

22. $\begin{bmatrix} 2 & 1 & 0 \\ 13 & -13 & 12 \\ -3 & 5 & -4 \end{bmatrix}$

23. $\begin{bmatrix} 0.4 & 0 & -2.4 & 3.0 \\ 1.2 & 0.6 & 0 & 0.3 \\ 0 & 1.2 & 1.2 & 0 \end{bmatrix}$

24. **TEAM PROJECT. Geometrical Applications: Curves and Surfaces Through Given Points.** The idea is to get an equation from the vanishing of the determinant of a homogeneous linear system as the condition for a nontrivial solution in Cramer's theorem. We explain the trick for obtaining such a system for the case of a line L through two given points P_1: (x_1, y_1) and P_2: (x_2, y_2). The unknown line is $ax + by = -c$, say. We write it as $ax + by + c \cdot 1 = 0$. To get a nontrivial solution a, b, c, the determinant of the "coefficients" x, y, 1 must be zero. The system is

$$ax + by + c \cdot 1 = 0 \quad (\text{Line } L)$$

$$(12) \quad ax_1 + by_1 + c \cdot 1 = 0 \quad (P_1 \text{ on } L)$$

$$ax_2 + by_2 + c \cdot 1 = 0 \quad (P_2 \text{ on } L).$$

(a) Line through two points. Derive from $D = 0$ in (12) the familiar formula

$$\frac{x - x_1}{x_1 - x_2} = \frac{y - y_1}{y_1 - y_2}.$$

(b) Plane. Find the analog of (12) for a plane through three given points. Apply it when the points are (1, 1, 1), (3, 2, 6), (5, 0, 5).

(c) Circle. Find a similar formula for a circle in the plane through three given points. Find and sketch the circle through (2, 6), (6, 4), (7, 1).

(d) Sphere. Find the analog of the formula in (c) for a sphere through four given points. Find the sphere through (0, 0, 5), (4, 0, 1), (0, 4, 1), (0, 0, −3) by this formula or by inspection.

(e) General conic section. Find a formula for a general conic section (the vanishing of a determinant of 6th order). Try it out for a quadratic parabola and for a more general conic section of your own choice.

25. **WRITING PROJECT. General Properties of Determinants.** Illustrate each statement in Theorems 1 and 2 with an example of your choice.

26. **CAS EXPERIMENT. Determinant of Zeros and Ones.** Find the value of the determinant of the $n \times n$ matrix $\mathbf{A}_n$ with main diagonal entries all 0 and all others 1. Try to find a formula for this. Try to prove it by induction. Interpret $\mathbf{A}_3$ and $\mathbf{A}_4$ as *"incidence matrices"* (as in Problem Set 7.1 but without the minuses) of a triangle and a tetrahedron, respectively; similarly for an *"n-simplex"*, having n vertices and $n(n-1)/2$ edges (and spanning R^{n-1}, $n = 5, 6, \cdots$).

7.8 Inverse of a Matrix. Gauss–Jordan Elimination

In this section we consider* square *matrices exclusively.

The **inverse** of an $n \times n$ matrix $\mathbf{A} = [a_{jk}]$ is denoted by $\mathbf{A}^{-1}$ and is an $n \times n$ matrix such that

$$(1) \quad \mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$

where $\mathbf{I}$ is the $n \times n$ unit matrix (see Sec. 7.2).

If $\mathbf{A}$ has an inverse, then $\mathbf{A}$ is called a **nonsingular matrix.** If $\mathbf{A}$ has no inverse, then $\mathbf{A}$ is called a **singular matrix.**

If $\mathbf{A}$ has an inverse, the inverse is unique.

Indeed, if both $\mathbf{B}$ and $\mathbf{C}$ are inverses of $\mathbf{A}$, then $\mathbf{AB} = \mathbf{I}$ and $\mathbf{CA} = \mathbf{I}$, so that we obtain the uniqueness from

$$\mathbf{B} = \mathbf{IB} = (\mathbf{CA})\mathbf{B} = \mathbf{C}(\mathbf{AB}) = \mathbf{CI} = \mathbf{C}.$$

We prove next that $\mathbf{A}$ has an inverse (is nonsingular) if and only if it has maximum possible rank n. The proof will also show that $\mathbf{Ax} = \mathbf{b}$ implies $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$ provided $\mathbf{A}^{-1}$ exists, and will thus give a motivation for the inverse as well as a relation to linear systems. (But this will ***not*** give a good method of solving $\mathbf{Ax} = \mathbf{b}$ ***numerically*** because the Gauss elimination in Sec. 7.3 requires fewer computations.)

THEOREM 1

Existence of the Inverse

The inverse $\mathbf{A}^{-1}$ *of an* $n \times n$ *matrix* $\mathbf{A}$ *exists if and only if* rank $\mathbf{A} = n$, *thus* (by Theorem 3, Sec. 7.7) *if and only if* $\det \mathbf{A} \neq 0$. *Hence* $\mathbf{A}$ *is nonsingular if* rank $\mathbf{A} = n$, *and is singular if* rank $\mathbf{A} < n$.

PROOF Let $\mathbf{A}$ be a given $n \times n$ matrix and consider the linear system

$$\mathbf{Ax} = \mathbf{b}. \tag{2}$$

If the inverse $\mathbf{A}^{-1}$ exists, then multiplication from the left on both sides and use of (1) gives

$$\mathbf{A}^{-1}\mathbf{Ax} = \mathbf{x} = \mathbf{A}^{-1}\mathbf{b}.$$

This shows that (2) has a unique solution $\mathbf{x}$. Hence $\mathbf{A}$ must have rank n by the Fundamental Theorem in Sec. 7.5.

Conversely, let rank $\mathbf{A} = n$. Then by the same theorem, the system (2) has a unique solution $\mathbf{x}$ for any $\mathbf{b}$. Now the back substitution following the Gauss elimination (Sec. 7.3) shows that the components x_j of $\mathbf{x}$ are linear combinations of those of $\mathbf{b}$. Hence we can write

$$\mathbf{x} = \mathbf{Bb} \tag{3}$$

with $\mathbf{B}$ to be determined. Substitution into (2) gives

$$\mathbf{Ax} = \mathbf{A}(\mathbf{Bb}) = (\mathbf{AB})\mathbf{b} = \mathbf{Cb} = \mathbf{b} \qquad (\mathbf{C} = \mathbf{AB})$$

for any $\mathbf{b}$. Hence $\mathbf{C} = \mathbf{AB} = \mathbf{I}$, the unit matrix. Similarly, if we substitute (2) into (3) we get

$$\mathbf{x} = \mathbf{Bb} = \mathbf{B}(\mathbf{Ax}) = (\mathbf{BA})\mathbf{x}$$

for any $\mathbf{x}$ (and $\mathbf{b} = \mathbf{Ax}$). Hence $\mathbf{BA} = \mathbf{I}$. Together, $\mathbf{B} = \mathbf{A}^{-1}$ exists. ∎

---

[3]WILHELM JORDAN (1842–1899), German mathematician and geodesist. [See *American Mathematical Monthly* **94** (1987), 130–142.]

We do ***not recommend*** it as a method for solving systems of linear equations, since the number of operations in addition to those of the Gauss elimination is larger than that for back substitution, which the Gauss–Jordan elimination avoids. See also Sec. 20.1.

Determination of the Inverse by the Gauss–Jordan Method

For the practical determination of the inverse $\mathbf{A}^{-1}$ of a nonsingular $n \times n$ matrix $\mathbf{A}$ we can use the Gauss elimination (Sec. 7.3), actually a variant of it, called the **Gauss-Jordan elimination**[3] (footnote of p. 316). The idea of the method is as follows.

Using $\mathbf{A}$, we form n linear systems

$$\mathbf{A}\mathbf{x}_{(1)} = \mathbf{e}_{(1)}, \qquad \cdots, \qquad \mathbf{A}\mathbf{x}_{(n)} = \mathbf{e}_{(n)}$$

where $\mathbf{e}_{(1)}, \cdots, \mathbf{e}_{(n)}$ are the columns of the $n \times n$ unit matrix $\mathbf{I}$; thus, $\mathbf{e}_{(1)} = [1 \quad 0 \quad \cdots \quad 0]^T$, $\mathbf{e}_{(2)} = [0 \quad 1 \quad 0 \quad \cdots \quad 0]^T$, etc. These are n vector equations in the unknown vectors $\mathbf{x}_{(1)}, \cdots, \mathbf{x}_{(n)}$. We combine them into a single matrix equation $\mathbf{AX} = \mathbf{I}$, with the unknown matrix $\mathbf{X}$ having the columns $\mathbf{x}_{(1)}, \cdots, \mathbf{x}_{(n)}$. Correspondingly, we combine the n augmented matrices $[\mathbf{A} \quad \mathbf{e}_{(1)}], \cdots, [\mathbf{A} \quad \mathbf{e}_{(n)}]$ into one $n \times 2n$ "augmented matrix" $\tilde{\mathbf{A}} = [\mathbf{A} \quad \mathbf{I}]$. Now multiplication of $\mathbf{AX} = \mathbf{I}$ by $\mathbf{A}^{-1}$ from the left gives $\mathbf{X} = \mathbf{A}^{-1}\mathbf{I} = \mathbf{A}^{-1}$. Hence, to solve $\mathbf{AX} = \mathbf{I}$ for $\mathbf{X}$, we can apply the Gauss elimination to $\tilde{\mathbf{A}} = [\mathbf{A} \quad \mathbf{I}]$. This gives a matrix of the form $[\mathbf{U} \quad \mathbf{H}]$ with upper triangular $\mathbf{U}$ because the Gauss elimination triangularizes systems. The Gauss–Jordan method reduces $\mathbf{U}$ by further elementary row operations to diagonal form, in fact to the unit matrix $\mathbf{I}$. This is done by eliminating the entries of $\mathbf{U}$ above the main diagonal and making the diagonal entries all 1 by multiplication (see the example below). Of course, the method operates on the entire matrix $[\mathbf{U} \quad \mathbf{H}]$, transforming $\mathbf{H}$ into some matrix $\mathbf{K}$, hence the entire $[\mathbf{U} \quad \mathbf{H}]$ to $[\mathbf{I} \quad \mathbf{K}]$. This is the "augmented matrix" of $\mathbf{IX} = \mathbf{K}$. Now $\mathbf{IX} = \mathbf{X} = \mathbf{A}^{-1}$, as shown before. By comparison, $\mathbf{K} = \mathbf{A}^{-1}$, so that we can read $\mathbf{A}^{-1}$ directly from $[\mathbf{I} \quad \mathbf{K}]$.

The following example illustrates the practical details of the method.

EXAMPLE 1 **Inverse of a Matrix. Gauss–Jordan Elimination**

Determine the inverse $\mathbf{A}^{-1}$ of

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

Solution. We apply the Gauss elimination (Sec. 7.3) to the following $n \times 2n = 3 \times 6$ matrix, where BLUE always refers to the previous matrix.

$$[\mathbf{A} \quad \mathbf{I}] = \left[\begin{array}{ccc|ccc} -1 & 1 & 2 & 1 & 0 & 0 \\ 3 & -1 & 1 & 0 & 1 & 0 \\ -1 & 3 & 4 & 0 & 0 & 1 \end{array}\right]$$

$$\left[\begin{array}{ccc|ccc} -1 & 1 & 2 & 1 & 0 & 0 \\ 0 & 2 & 7 & 3 & 1 & 0 \\ 0 & 2 & 2 & -1 & 0 & 1 \end{array}\right] \quad \begin{array}{l} \\ \text{Row 2} + 3\ \text{Row 1} \\ \text{Row 3} - \text{Row 1} \end{array}$$

$$\left[\begin{array}{ccc|ccc} -1 & 1 & 2 & 1 & 0 & 0 \\ 0 & 2 & 7 & 3 & 1 & 0 \\ 0 & 0 & -5 & -4 & -1 & 1 \end{array}\right] \quad \begin{array}{l} \\ \\ \text{Row 3} - \text{Row 2} \end{array}$$

This is [U H] as produced by the Gauss elimination. Now follow the additional Gauss–Jordan steps, reducing **U** to **I**, that is, to diagonal form with entries 1 on the main diagonal.

$$\left[\begin{array}{ccc|ccc} 1 & -1 & -2 & -1 & 0 & 0 \\ 0 & 1 & 3.5 & 1.5 & 0.5 & 0 \\ 0 & 0 & 1 & 0.8 & 0.2 & -0.2 \end{array}\right] \quad -\text{Row 1}$$

$$\left[\begin{array}{ccc|ccc} 1 & -1 & 0 & 0.6 & 0.4 & -0.4 \\ 0 & 1 & 0 & -1.3 & -0.2 & 0.7 \\ 0 & 0 & 1 & 0.8 & 0.2 & -0.2 \end{array}\right] \quad \begin{array}{l} 0.5 \text{ Row 2} \\ -0.2 \text{ Row 3} \\ \text{Row 1} + 2 \text{ Row 3} \\ \text{Row 2} - 3.5 \text{ Row 3} \end{array}$$

$$\left[\begin{array}{ccc|ccc} 1 & 0 & 0 & -0.7 & 0.2 & 0.3 \\ 0 & 1 & 0 & -1.3 & -0.2 & 0.7 \\ 0 & 0 & 1 & 0.8 & 0.2 & -0.2 \end{array}\right] \quad \text{Row 1} + \text{Row 2}$$

The last three columns constitute $\mathbf{A}^{-1}$. Check:

$$\begin{bmatrix} -1 & 1 & 2 \\ 3 & -1 & 1 \\ -1 & 3 & 4 \end{bmatrix} \begin{bmatrix} -0.7 & 0.2 & 0.3 \\ -1.3 & -0.2 & 0.7 \\ 0.8 & 0.2 & -0.2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Hence $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$. Similarly, $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. ■

Useful Formulas for Inverses

The explicit formula (4) in the following theorem is often useful in theoretical studies (as opposed to *computing* inverses). In fact, the special case $n = 2$ occurs quite frequently in geometrical and other applications.

THEOREM 2

Inverse of a Matrix

The inverse of a nonsingular $n \times n$ matrix $\mathbf{A} = [a_{jk}]$ is given by

$$(4) \qquad \mathbf{A}^{-1} = \frac{1}{\det \mathbf{A}} [C_{jk}]^{\mathsf{T}} = \frac{1}{\det \mathbf{A}} \begin{bmatrix} C_{11} & C_{21} & \cdots & C_{n1} \\ C_{12} & C_{22} & \cdots & C_{n2} \\ \cdot & \cdot & \cdots & \cdot \\ C_{1n} & C_{2n} & \cdots & C_{nn} \end{bmatrix},$$

where C_{jk} is the cofactor of a_{jk} in det **A** (see Sec. 7.7). (CAUTION! Note well that in $\mathbf{A}^{-1}$, the cofactor C_{jk} occupies the same place as a_{kj} (not a_{jk}) does in **A**.)

In particular, the inverse of

$$(4^*) \qquad \mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \quad \textit{is} \quad \mathbf{A}^{-1} = \frac{1}{\det \mathbf{A}} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}.$$

PROOF We denote the right side of (4) by $\mathbf{B}$ and show that $\mathbf{BA} = \mathbf{I}$. We first write

$$\mathbf{BA} = \mathbf{G} = [g_{kl}] \tag{5}$$

and then show that $\mathbf{G} = \mathbf{I}$. Now by the definition of matrix multiplication and because of the form of $\mathbf{B}$ in (4), we obtain (CAUTION! C_{sk}, not C_{ks})

$$g_{kl} = \sum_{s=1}^{n} \frac{C_{sk}}{\det \mathbf{A}} a_{sl} = \frac{1}{\det \mathbf{A}} (a_{1l}C_{1k} + \cdots + a_{nl}C_{nk}). \tag{6}$$

Now (9) and (10) in Sec. 7.7 show that the sum $(\cdots)$ on the right is $D = \det \mathbf{A}$ when $l = k$, and is zero when $l \neq k$. Hence

$$g_{kk} = \frac{1}{\det \mathbf{A}} \det \mathbf{A} = 1,$$

$$g_{kl} = 0 \quad (l \neq k),$$

In particular, for $n = 2$ we have in (4) in the first row $C_{11} = a_{22}$, $C_{21} = -a_{12}$ and in the second row $C_{12} = -a_{21}$, $C_{22} = a_{11}$. This gives (4*). ∎

EXAMPLE 2 **Inverse of a 2 × 2 Matrix**

$$\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 2 & 4 \end{bmatrix}, \qquad \mathbf{A}^{-1} = \frac{1}{10} \begin{bmatrix} 4 & -1 \\ -2 & 3 \end{bmatrix} = \begin{bmatrix} 0.4 & -0.1 \\ -0.2 & 0.3 \end{bmatrix}$$ ∎

EXAMPLE 3 **Further Illustration of Theorem 2**

Using (4), find the inverse of

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

Solution. We obtain $\det \mathbf{A} = -1(-7) - 1 \cdot 13 + 2 \cdot 8 = 10$, and in (4),

$$C_{11} = \begin{vmatrix} -1 & 1 \\ 3 & 4 \end{vmatrix} = -7, \qquad C_{21} = -\begin{vmatrix} 1 & 2 \\ 3 & 4 \end{vmatrix} = 2, \qquad C_{31} = \begin{vmatrix} 1 & 2 \\ -1 & 1 \end{vmatrix} = 3,$$

$$C_{12} = -\begin{vmatrix} 3 & 1 \\ -1 & 4 \end{vmatrix} = -13, \qquad C_{22} = \begin{vmatrix} -1 & 2 \\ -1 & 4 \end{vmatrix} = -2, \qquad C_{32} = -\begin{vmatrix} -1 & 2 \\ 3 & 1 \end{vmatrix} = 7,$$

$$C_{13} = \begin{vmatrix} 3 & -1 \\ -1 & 3 \end{vmatrix} = 8, \qquad C_{23} = -\begin{vmatrix} -1 & 1 \\ -1 & 3 \end{vmatrix} = 2, \qquad C_{33} = \begin{vmatrix} -1 & 1 \\ 3 & -1 \end{vmatrix} = -2,$$

so that by (4), in agreement with Example 1,

$$\mathbf{A}^{-1} = \begin{bmatrix} -0.7 & 0.2 & 0.3 \\ -1.3 & -0.2 & 0.7 \\ 0.8 & 0.2 & -0.2 \end{bmatrix}.$$ ∎

Diagonal matrices $\mathbf{A} = [a_{jk}]$, $a_{jk} = 0$ when $j \neq k$, have an inverse if and only if all $a_{jj} \neq 0$. Then $\mathbf{A}^{-1}$ is diagonal, too, with entries $1/a_{11}, \cdots, 1/a_{nn}$.

PROOF For a diagonal matrix we have in (4)

$$\frac{C_{11}}{D} = \frac{a_{22} \cdots a_{nn}}{a_{11}a_{22} \cdots a_{nn}} = \frac{1}{a_{11}}, \qquad \text{etc.}$$

EXAMPLE 4 Inverse of a Diagonal Matrix

Let

$$\mathbf{A} = \begin{bmatrix} -0.5 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

Then the inverse is

$$\mathbf{A}^{-1} = \begin{bmatrix} -2 & 0 & 0 \\ 0 & 0.25 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Products can be inverted by taking the inverse of each factor and multiplying these inverses ***in reverse order,***

$$(\mathbf{AC})^{-1} = \mathbf{C}^{-1}\mathbf{A}^{-1}. \tag{7}$$

Hence for more than two factors,

$$(\mathbf{AC} \cdots \mathbf{PQ})^{-1} = \mathbf{Q}^{-1}\mathbf{P}^{-1} \cdots \mathbf{C}^{-1}\mathbf{A}^{-1}. \tag{8}$$

PROOF The idea is to start from (1) for $\mathbf{AC}$ instead of $\mathbf{A}$, that is, $\mathbf{AC}(\mathbf{AC})^{-1} = \mathbf{I}$, and multiply it on both sides from the left, first by $\mathbf{A}^{-1}$, which because of $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$ gives

$$\begin{aligned} \mathbf{A}^{-1}\mathbf{AC}(\mathbf{AC})^{-1} &= \mathbf{C}(\mathbf{AC})^{-1} \\ &= \mathbf{A}^{-1}\mathbf{I} = \mathbf{A}^{-1}, \end{aligned}$$

and then multiplying this on both sides from the left, this time by $\mathbf{C}^{-1}$ and by using $\mathbf{C}^{-1}\mathbf{C} = \mathbf{I}$,

$$\mathbf{C}^{-1}\mathbf{C}(\mathbf{AC})^{-1} = (\mathbf{AC})^{-1} = \mathbf{C}^{-1}\mathbf{A}^{-1}.$$

This proves (7), and from it, (8) follows by induction.

We also note that *the inverse of the inverse is the given matrix,* as you may prove,

$$(\mathbf{A}^{-1})^{-1} = \mathbf{A}. \tag{9}$$

Unusual Properties of Matrix Multiplication. Cancellation Laws

Section 7.2 contains warnings that some properties of matrix multiplication deviate from those for numbers, and we are now able to explain the restricted validity of the so-called **cancellation laws** [2.] and [3.] below, using rank and inverse, concepts that were not yet available in Sec. 7.2. The deviations from the usual are of great practical importance and must be carefully observed. They are as follows.

[1.] Matrix multiplication is not commutative, that is, in general we have

$$\mathbf{AB} \neq \mathbf{BA}.$$

[2.] $\mathbf{AB} = \mathbf{0}$ does not generally imply $\mathbf{A} = \mathbf{0}$ or $\mathbf{B} = \mathbf{0}$ (or $\mathbf{BA} = \mathbf{0}$); for example,

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

[3.] $\mathbf{AC} = \mathbf{AD}$ does not generally imply $\mathbf{C} = \mathbf{D}$ (even when $\mathbf{A} \neq \mathbf{0}$).

Complete answers to [2.] and [3.] are contained in the following theorem.

THEOREM 3

Cancellation Laws

Let $\mathbf{A}, \mathbf{B}, \mathbf{C}$ *be* $n \times n$ *matrices. Then:*

(a) *If* rank $\mathbf{A} = n$ *and* $\mathbf{AB} = \mathbf{AC}$, *then* $\mathbf{B} = \mathbf{C}$.

(b) *If* rank $\mathbf{A} = n$, *then* $\mathbf{AB} = \mathbf{0}$ *implies* $\mathbf{B} = \mathbf{0}$. *Hence if* $\mathbf{AB} = \mathbf{0}$, *but* $\mathbf{A} \neq \mathbf{0}$ *as well as* $\mathbf{B} \neq \mathbf{0}$, *then* rank $\mathbf{A} < n$ *and* rank $\mathbf{B} < n$.

(c) *If* $\mathbf{A}$ *is singular, so are* $\mathbf{BA}$ *and* $\mathbf{AB}$.

PROOF **(a)** The inverse of $\mathbf{A}$ exists by Theorem 1. Multiplication by $\mathbf{A}^{-1}$ from the left gives $\mathbf{A}^{-1}\mathbf{AB} = \mathbf{A}^{-1}\mathbf{AC}$, hence $\mathbf{B} = \mathbf{C}$.

(b) Let rank $\mathbf{A} = n$. Then $\mathbf{A}^{-1}$ exists, and $\mathbf{AB} = \mathbf{0}$ implies $\mathbf{A}^{-1}\mathbf{AB} = \mathbf{B} = \mathbf{0}$. Similarly when rank $\mathbf{B} = n$. This implies the second statement in (b).

(c_1) Rank $\mathbf{A} < n$ by Theorem 1. Hence $\mathbf{Ax} = \mathbf{0}$ has nontrivial solutions by Theorem 2 in Sec. 7.5. Multiplication by $\mathbf{B}$ shows that these solutions are also solutions of $\mathbf{BAx} = \mathbf{0}$, so that rank $(\mathbf{BA}) < n$ by Theorem 2 in Sec. 7.5 and $\mathbf{BA}$ is singular by Theorem 1.

(c_2) $\mathbf{A}^T$ is singular by Theorem 2(d) in Sec. 7.7. Hence $\mathbf{B}^T\mathbf{A}^T$ is singular by part (c_1), and is equal to $(\mathbf{AB})^T$ by (10d) in Sec. 7.2. Hence $\mathbf{AB}$ is singular by Theorem 2(d) in Sec. 7.7. ∎

Determinants of Matrix Products

The determinant of a matrix product $\mathbf{AB}$ or $\mathbf{BA}$ can be written as the product of the determinants of the factors, and it is interesting that $\det \mathbf{AB} = \det \mathbf{BA}$, although $\mathbf{AB} \neq \mathbf{BA}$ in general. The corresponding formula (10) is needed occasionally and can be obtained by Gauss-Jordan elimination (see Example 1) and from the theorem just proved.

THEOREM 4

Determinant of a Product of Matrices

For any $n \times n$ matrices **A** *and* **B,**

(10) $$\det(\mathbf{AB}) = \det(\mathbf{BA}) = \det\mathbf{A}\,\det\mathbf{B}.$$

PROOF If **A** or **B** is singular, so are **AB** and **BA** by Theorem 3(c), and (10) reduces to $0 = 0$ by Theorem 3 in Sec. 7.7.

Now let **A** and **B** be nonsingular. Then we can reduce **A** to a diagonal matrix $\hat{\mathbf{A}} = [a_{jk}]$ by Gauss–Jordan steps. Under these operations, det **A** retains its value, by Theorem 1 in Sec. 7.7, (a) and (b) [not (c)] except perhaps for a sign reversal in row interchanging when pivoting. But the same operations reduce **AB** to $\hat{\mathbf{A}}\mathbf{B}$ with the same effect on det (**AB**). Hence it remains to prove (10) for $\hat{\mathbf{A}}\mathbf{B}$; written out,

$$\hat{\mathbf{A}}\mathbf{B} = \begin{bmatrix} \hat{a}_{11} & 0 & \cdots & 0 \\ 0 & \hat{a}_{22} & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & \hat{a}_{nn} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ & & & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{bmatrix}$$

$$= \begin{bmatrix} \hat{a}_{11}b_{11} & \hat{a}_{11}b_{12} & \cdots & \hat{a}_{11}b_{1n} \\ \hat{a}_{22}b_{21} & \hat{a}_{22}b_{22} & \cdots & \hat{a}_{22}b_{2n} \\ & & \vdots & \\ \hat{a}_{nn}b_{n1} & \hat{a}_{nn}b_{n2} & \cdots & \hat{a}_{nn}b_{nn} \end{bmatrix}.$$

We now take the determinant det ($\hat{\mathbf{A}}\mathbf{B}$). On the right we can take out a factor $\hat{a}_{11}$ from the first row, $\hat{a}_{22}$ from the second, $\cdots$, $\hat{a}_{nn}$ from the nth. But this product $\hat{a}_{11}\,\hat{a}_{22}\cdots\hat{a}_{nn}$ equals det $\hat{\mathbf{A}}$ because $\hat{\mathbf{A}}$ is diagonal. The remaining determinant is det **B.** This proves (10) for det (**AB**), and the proof for det (**BA**) follows by the same idea. ∎

This completes our discussion of linear systems (Secs. 7.3–7.8). Section 7.9 on vector spaces and linear transformations is optional. ***Numeric methods*** are discussed in Secs. 20.1–20.4, which are independent of other sections on numerics.

PROBLEM SET 7.8

1–12 INVERSE

Find the inverse by Gauss–Jordan [or by (4*) if $n = 2$] or state that it does not exist. Check by using (1).

1. $\begin{bmatrix} 1.20 & 4.64 \\ 0.50 & 3.60 \end{bmatrix}$

2. $\begin{bmatrix} 0.6 & 0.8 \\ 0.8 & -0.6 \end{bmatrix}$

3. $\begin{bmatrix} \cos 2\theta & \sin 2\theta \\ -\sin 2\theta & \cos 2\theta \end{bmatrix}$

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

5. $\begin{bmatrix} 3 & -1 & 1 \\ -15 & 6 & -5 \\ 5 & -2 & 2 \end{bmatrix}$

6. $\begin{bmatrix} 29 & -11 & 10 \\ -160 & 61 & -55 \\ 55 & -21 & 19 \end{bmatrix}$

7. $\begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 5 & 4 & 1 \end{bmatrix}$

8. $\begin{bmatrix} 1 & 2 & 5 \\ 0 & -1 & 2 \\ 2 & 4 & 11 \end{bmatrix}$

9. $\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

10. $\begin{bmatrix} 0 & 8 & 0 \\ 0 & 0 & 4 \\ 2 & 0 & 0 \end{bmatrix}$

11. $\begin{bmatrix} 1 & 2 & 5 \\ 0 & -1 & 2 \\ 2 & 4 & 10 \end{bmatrix}$

12. $\begin{bmatrix} 1 & 2 & -9 \\ -2 & -4 & 19 \\ 0 & -1 & 2 \end{bmatrix}$

13. (Triangular matrix) Is the inverse of a triangular matrix always triangular (as in Prob. 7)? Give reason.

14. (Rotation) Give an application of the matrix in Prob. 3 that makes the form of its inverse obvious.

15. (Inverse of the square) Verify $(\mathbf{A}^2)^{-1} = (\mathbf{A}^{-1})^2$ for $\mathbf{A}$ in Prob. 5.

16. Prove the formula in Prob. 15.

17. (Inverse of the transpose) Verify $(\mathbf{A}^T)^{-1} = (\mathbf{A}^{-1})^T$ for $\mathbf{A}$ in Prob. 5.

18. Prove the formula in Prob. 17.

19. (Inverse of the inverse) Prove that $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$.

20. (Row interchange) Same question as in Prob. 14 for the matrix in Prob. 9.

21–23 EXPLICIT FORMULA (4) FOR THE INVERSE

Formula (4) is generally not very practical. To understand its use, apply it:

21. To Prob. 9. **22.** To Prob. 4. **23.** To Prob. 7.

7.9 Vector Spaces, Inner Product Spaces, Linear Transformations *Optional*

In Sec. 7.4 we have seen that ***special vector spaces*** arise quite naturally in connection with matrices and linear systems, that their elements, called *vectors,* satisfy rules quite similar to those for numbers [(3) and (4) in Sec. 7.1], and that they are often obtained as *spans* (sets of linear combinations) of finitely many given vectors. Each such vector has n real numbers as its *components.* Look this up before going on.

Now if we take *all* vectors with n real numbers as components (*"real vectors"*), we obtain the very important **real n-dimensional vector space** R^n. This is a standard name and notation. Thus, each vector in R^n is an ordered n-tuple of real numbers.

Particular cases are R^2, the space of all ordered pairs (**"vectors in the plane"**) and R^3, the space of all ordered triples (**"vectors in 3-space"**). These vectors have wide applications in mechanics, geometry, and calculus that are basic to the engineer and physicist.

Similarly, if we take all ordered n-tuples of *complex numbers* as vectors and complex numbers as scalars, we obtain the **complex vector space** C^n, which we shall consider in Sec. 8.5.

This is not all. There are other sets of practical interest (sets of matrices, functions, transformations, etc.) for which addition and scalar multiplication can be defined in a natural way so that they form a "vector space". This suggests to create from the ***"concrete model"*** R^n the ***"abstract concept"*** of a ***"real vector space"*** V by taking the basic properties (3) and (4) in Sec. 7.1 as axioms. These axioms guarantee that one obtains a useful and applicable theory of those more general situations. Note that each axiom expresses a simple property of R^n or, as a matter of fact, of R^3. Selecting good axioms needs experience and is a process of trial and error that often extends over a long period of time.

DEFINITION

Real Vector Space

A nonempty set V of elements $\mathbf{a}, \mathbf{b}, \cdots$ is called a **real vector space** (or *real linear space*), and these elements are called **vectors** (regardless of their nature, which will come out from the context or will be left arbitrary) if in V there are defined two algebraic operations (called *vector addition* and *scalar multiplication*) as follows.

I. Vector addition associates with every pair of vectors $\mathbf{a}$ and $\mathbf{b}$ of V a unique vector of V, called the *sum* of $\mathbf{a}$ and $\mathbf{b}$ and denoted by $\mathbf{a} + \mathbf{b}$, such that the following axioms are satisfied.

I.1 *Commutativity.* For any two vectors $\mathbf{a}$ and $\mathbf{b}$ of V,

$$\mathbf{a} + \mathbf{b} = \mathbf{b} + \mathbf{a}.$$

I.2 *Associativity.* For any three vectors $\mathbf{u}, \mathbf{v}, \mathbf{w}$ of V,

$$(\mathbf{u} + \mathbf{v}) + \mathbf{w} = \mathbf{u} + (\mathbf{v} + \mathbf{w}) \quad \text{(written } \mathbf{u} + \mathbf{v} + \mathbf{w}\text{)}.$$

I.3 There is a unique vector in V, called the *zero vector* and denoted by $\mathbf{0}$, such that for every $\mathbf{a}$ in V,

$$\mathbf{a} + \mathbf{0} = \mathbf{a}.$$

I.4 For every $\mathbf{a}$ in V there is a unique vector in V that is denoted by $-\mathbf{a}$ and is such that

$$\mathbf{a} + (-\mathbf{a}) = \mathbf{0}.$$

II. Scalar multiplication. The real numbers are called **scalars.** Scalar multiplication associates with every $\mathbf{a}$ in V and every scalar c a unique vector of V, called the *product* of c and $\mathbf{a}$ and denoted by $c\mathbf{a}$ (or $\mathbf{a}c$) such that the following axioms are satisfied.

II.1 *Distributivity.* For every scalar c and vectors $\mathbf{a}$ and $\mathbf{b}$ in V,

$$c(\mathbf{a} + \mathbf{b}) = c\mathbf{a} + c\mathbf{b}.$$

II.2 *Distributivity.* For all scalars c and k and every $\mathbf{a}$ in V,

$$(c + k)\mathbf{a} = c\mathbf{a} + k\mathbf{a}.$$

II.3 *Associativity.* For all scalars c and k and every $\mathbf{a}$ in V,

$$c(k\mathbf{a}) = (ck)\mathbf{a} \quad \text{(written } ck\mathbf{a}\text{)}.$$

II.4 For every $\mathbf{a}$ in V,

$$1\mathbf{a} = \mathbf{a}.$$

A **complex vector space** is obtained if, instead of real numbers, we take complex numbers as scalars.

Basic concepts related to the concept of a vector space are defined as in Sec. 7.4.

A **linear combination** of vectors $\mathbf{a}_{(1)}, \cdots, \mathbf{a}_{(m)}$ in a vector space V is an expression

$$c_1\mathbf{a}_{(1)} + \cdots + c_m\mathbf{a}_{(m)} \qquad (c_1, \cdots, c_m \text{ any scalars}).$$

These vectors form a **linearly independent set** (briefly, they are called **linearly independent**) if

$$c_1\mathbf{a}_{(1)} + \cdots + c_m\mathbf{a}_{(m)} = \mathbf{0} \tag{1}$$

implies that $c_1 = 0, \cdots, c_m = 0$. Otherwise, if (1) also holds with scalars not all zero, the vectors are called **linearly dependent.**

Note that (1) with $m = 1$ is $c\mathbf{a} = \mathbf{0}$ and shows that a single vector $\mathbf{a}$ is linearly independent if and only if $\mathbf{a} \neq \mathbf{0}$.

V has **dimension n,** or is **n-dimensional,** if it contains a linearly independent set of n vectors, whereas any set of more than n vectors in V is linearly dependent. That set of n linearly independent vectors is called a **basis** for V. Then every vector in V can be written as a linear combination of the basis vectors; for a given basis, this representation is unique (see Prob. 14).

EXAMPLE 1 **Vector Space of Matrices**

The real 2×2 matrices form a four-dimensional real vector space. A basis is

$$\mathbf{B}_{11} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{B}_{12} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{B}_{21} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad \mathbf{B}_{22} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

because any 2×2 matrix $\mathbf{A} = [a_{jk}]$ has a unique representation $\mathbf{A} = a_{11}\mathbf{B}_{11} + a_{12}\mathbf{B}_{12} + a_{21}\mathbf{B}_{21} + a_{22}\mathbf{B}_{22}$. Similarly, the real $m \times n$ matrices with fixed m and n form an mn-dimensional vector space. What is the dimension of the vector space of all 3×3 skew-symmetric matrices? Can you find a basis? ▩

EXAMPLE 2 **Vector Space of Polynomials**

The set of all constant, linear, and quadratic polynomials in x together is a vector space of dimension 3 with basis $\{1, x, x^2\}$ under the usual addition and multiplication by real numbers because these two operations give polynomials not exceeding degree 2. What is the dimension of the vector space of all polynomials of degree not exceeding a given fixed n? Can you find a basis? ▩

If a vector space V contains a linearly independent set of n vectors for every n, no matter how large, then V is called **infinite dimensional,** as opposed to a *finite dimensional* (n-dimensional) vector space just defined. An example of an infinite dimensional vector space is the space of all continuous functions on some interval $[a, b]$ of the x-axis, as we mention without proof.

Inner Product Spaces

If $\mathbf{a}$ and $\mathbf{b}$ are vectors in R^n, regarded as column vectors, we can form the product $\mathbf{a}^T\mathbf{b}$. This is a 1×1 matrix, which we can identify with its single entry, that is, with a number. This product is called the **inner product** or **dot product** of $\mathbf{a}$ and $\mathbf{b}$. Other notations for it are $(\mathbf{a}, \mathbf{b})$ and $\mathbf{a} \cdot \mathbf{b}$. Thus

$$\mathbf{a}^T\mathbf{b} = (\mathbf{a}, \mathbf{b}) = \mathbf{a} \cdot \mathbf{b} = [a_1 \cdots a_n] \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} = \sum_{l=1}^{n} a_l b_l = a_1 b_1 + \cdots + a_n b_n.$$

We now extend this concept to general real vector spaces by taking basic properties of $(\mathbf{a}, \mathbf{b})$ as axioms for an "abstract inner product" $(\mathbf{a}, \mathbf{b})$ as follows.

DEFINITION

Real Inner Product Space

A real vector space V is called a **real inner product space** (or *real pre-Hilbert*[4] *space*) if it has the following property. With every pair of vectors $\mathbf{a}$ and $\mathbf{b}$ in V there is associated a real number, which is denoted by $(\mathbf{a}, \mathbf{b})$ and is called the **inner product** of $\mathbf{a}$ and $\mathbf{b}$, such that the following axioms are satisfied.

I. For all scalars q_1 and q_2 and all vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}$ in V,

$$(q_1\mathbf{a} + q_2\mathbf{b}, \mathbf{c}) = q_1(\mathbf{a}, \mathbf{c}) + q_2(\mathbf{b}, \mathbf{c}) \qquad (\textit{Linearity}).$$

II. For all vectors $\mathbf{a}$ and $\mathbf{b}$ in V,

$$(\mathbf{a}, \mathbf{b}) = (\mathbf{b}, \mathbf{a}) \qquad (\textit{Symmetry}).$$

III. For every $\mathbf{a}$ in V,

$$\left.\begin{array}{l} (\mathbf{a}, \mathbf{a}) \geqq 0, \\ (\mathbf{a}, \mathbf{a}) = 0 \quad \text{if and only if} \quad \mathbf{a} = \mathbf{0} \end{array}\right\} \quad (\textit{Positive-definiteness}).$$

Vectors whose inner product is zero are called **orthogonal.**

The *length* or **norm** of a vector in V is defined by

$$\|\mathbf{a}\| = \sqrt{(\mathbf{a}, \mathbf{a})} \qquad (\geqq 0). \tag{2}$$

A vector of norm 1 is called a **unit vector.**

From these axioms and from (2) one can derive the basic inequality

$$|(\mathbf{a}, \mathbf{b})| \leqq \|\mathbf{a}\| \, \|\mathbf{b}\| \qquad (\textit{Cauchy–Schwarz}[5] \textit{ inequality}). \tag{3}$$

From this follows

$$\|\mathbf{a} + \mathbf{b}\| \leqq \|\mathbf{a}\| + \|\mathbf{b}\| \qquad (\textit{Triangle inequality}). \tag{4}$$

A simple direct calculation gives

$$\|\mathbf{a} + \mathbf{b}\|^2 + \|\mathbf{a} - \mathbf{b}\|^2 = 2(\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2) \qquad (\textit{Parallelogram equality}). \tag{5}$$

---

[4]DAVID HILBERT (1862–1943), great German mathematician, taught at Königsberg and Göttingen and was the creator of the famous Göttingen mathematical school. He is known for his basic work in algebra, the calculus of variations, integral equations, functional analysis, and mathematical logic. His "Foundations of Geometry" helped the axiomatic method to gain general recognition. His famous 23 problems (presented in 1900 at the International Congress of Mathematicians in Paris) considerably influenced the development of modern mathematics.

If V is finite dimensional, it is actually a so-called *Hilbert space;* see Ref. [GR7], p. 73, listed in App. I.

[5]HERMANN AMANDUS SCHWARZ (1843–1921). German mathematician, known by his work in complex analysis (conformal mapping) and differential geometry. For Cauchy see Sec. 2.5.

EXAMPLE 3 ***n*-Dimensional Euclidean Space**

R^n with the inner product

$$(\mathbf{a}, \mathbf{b}) = \mathbf{a}^T\mathbf{b} = a_1b_1 + \cdots + a_nb_n \tag{6}$$

(where both $\mathbf{a}$ and $\mathbf{b}$ are *column* vectors) is called the ***n*-dimensional Euclidean space** and is denoted by E^n or again simply by R^n. Axioms I–III hold, as direct calculation shows. Equation (2) gives the **"Euclidean norm"**

$$\|\mathbf{a}\| = \sqrt{(\mathbf{a}, \mathbf{a})} = \sqrt{\mathbf{a}^T\mathbf{a}} = \sqrt{a_1^2 + \cdots + a_n^2}. \tag{7}$$

■

EXAMPLE 4 **An Inner Product for Functions. Function Space**

The set of all real-valued continuous functions $f(x)$, $g(x)$, $\cdots$ on a given interval $\alpha \leqq x \leqq \beta$ is a real vector space under the usual addition of functions and multiplication by scalars (real numbers). On this **"function space"** we can define an inner product by the integral

$$(f, g) = \int_\alpha^\beta f(x)\, g(x)\, dx. \tag{8}$$

Axioms I–III can be verified by direct calculation. Equation (2) gives the norm

$$\|f\| = \sqrt{(f, f)} = \sqrt{\int_\alpha^\beta f(x)^2\, dx}. \tag{9}$$

■

Our examples give a first impression of the great generality of the abstract concepts of vector spaces and inner product spaces. Further details belong to more advanced courses (on functional analysis, meaning abstract modern analysis; see Ref. [GR7] listed in App. 1) and cannot be discussed here. Instead we now take up a related topic where matrices play a central role.

Linear Transformations

Let X and Y be any vector spaces. To each vector $\mathbf{x}$ in X we assign a unique vector $\mathbf{y}$ in Y. Then we say that a **mapping** (or **transformation** or **operator**) of X into Y is given. Such a mapping is denoted by a capital letter, say F. The vector $\mathbf{y}$ in Y assigned to a vector $\mathbf{x}$ in X is called the **image** of $\mathbf{x}$ under F and is denoted by $F(\mathbf{x})$ [or $F\mathbf{x}$, without parentheses].

F is called a **linear mapping** or **linear transformation** if for all vectors $\mathbf{v}$ and $\mathbf{x}$ in X and scalars c,

$$\begin{aligned} F(\mathbf{v} + \mathbf{x}) &= F(\mathbf{v}) + F(\mathbf{x}) \\ F(c\mathbf{x}) &= cF(\mathbf{x}). \end{aligned} \tag{10}$$

Linear Transformation of Space R^n into Space R^m

From now on we let $X = R^n$ and $Y = R^m$. Then any real $m \times n$ matrix $\mathbf{A} = [a_{jk}]$ gives a transformation of R^n into R^m,

$$\mathbf{y} = \mathbf{A}\mathbf{x}. \tag{11}$$

Since $\mathbf{A}(\mathbf{u} + \mathbf{x}) = \mathbf{A}\mathbf{u} + \mathbf{A}\mathbf{x}$ and $\mathbf{A}(c\mathbf{x}) = c\mathbf{A}\mathbf{x}$, this transformation is linear.

We show that, conversely, every linear transformation F of R^n into R^m can be given in terms of an $m \times n$ matrix $\mathbf{A}$, after a basis for R^n and a basis for R^m have been chosen. This can be proved as follows.

Let $\mathbf{e}_{(1)}, \cdots, \mathbf{e}_{(n)}$ be any basis for R^n. Then every $\mathbf{x}$ in R^n has a unique representation

$$\mathbf{x} = x_1\mathbf{e}_{(1)} + \cdots + x_n\mathbf{e}_{(n)}.$$

Since F is linear, this representation implies for the image $F(\mathbf{x})$:

$$F(\mathbf{x}) = F(x_1\mathbf{e}_{(1)} + \cdots + x_n\mathbf{e}_{(n)}) = x_1F(\mathbf{e}_{(1)}) + \cdots + x_nF(\mathbf{e}_{(n)}).$$

Hence F is uniquely determined by the images of the vectors of a basis for R^n. We now choose for R^n the **"standard basis"**

$$\mathbf{e}_{(1)} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \mathbf{e}_{(2)} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \cdots, \quad \mathbf{e}_{(n)} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \tag{12}$$

where $\mathbf{e}_{(j)}$ has its jth component equal to 1 and all others 0. We show that we can now determine an $m \times n$ matrix $\mathbf{A} = [a_{jk}]$ such that for every $\mathbf{x}$ in R^n and image $\mathbf{y} = F(\mathbf{x})$ in R^m,

$$\mathbf{y} = F(\mathbf{x}) = \mathbf{A}\mathbf{x}.$$

Indeed, from the image $\mathbf{y}^{(1)} = F(\mathbf{e}_{(1)})$ of $\mathbf{e}_{(1)}$ we get the condition

$$\mathbf{y}^{(1)} = \begin{bmatrix} y_1^{(1)} \\ y_2^{(1)} \\ \vdots \\ y_m^{(1)} \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

from which we can determine the first column of $\mathbf{A}$, namely $a_{11} = y_1^{(1)}$, $a_{21} = y_2^{(1)}, \cdots, a_{m1} = y_m^{(1)}$. Similarly, from the image of $\mathbf{e}_{(2)}$ we get the second column of $\mathbf{A}$, and so on. This completes the proof. ∎

We say that $\mathbf{A}$ **represents** F, or *is a representation of* F, with respect to the bases for R^n and R^m. Quite generally, the purpose of a **"representation"** is the replacement of one object of study by another object whose properties are more readily apparent.

In three-dimensional Euclidean space E^3 the standard basis is usually written $\mathbf{e}_{(1)} = \mathbf{i}$, $\mathbf{e}_{(2)} = \mathbf{j}$, $\mathbf{e}_{(3)} = \mathbf{k}$. Thus,

$$\mathbf{i} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{j} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \mathbf{k} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}. \tag{13}$$

These are the three unit vectors in the positive directions of the axes of the **Cartesian coordinate system in space,** that is, the usual coordinate system with the same scale of measurement on the three mutually perpendicular coordinate axes.

EXAMPLE 5 Linear Transformations

Interpreted as transformations of Cartesian coordinates in the plane, the matrices

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}, \quad \begin{bmatrix} a & 0 \\ 0 & 1 \end{bmatrix}$$

represent a reflection in the line $x_2 = x_1$, a reflection in the x_1-axis, a reflection in the origin, and a stretch (when $a > 1$, or a contraction when $0 < a < 1$) in the x_1-direction, respectively. ▩

EXAMPLE 6 Linear Transformations

Our discussion preceding Example 5 is simpler than it may look at first sight. To see this, find $\mathbf{A}$ representing the linear transformation that maps (x_1, x_2) onto $(2x_1 - 5x_2, 3x_1 + 4x_2)$.

Solution. Obviously, the transformation is

$$y_1 = 2x_1 - 5x_2$$
$$y_2 = 3x_1 + 4x_2.$$

From this we can directly see that the matrix is

$$\mathbf{A} = \begin{bmatrix} 2 & -5 \\ 3 & 4 \end{bmatrix}. \qquad \text{Check:} \quad \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 2 & -5 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2x_1 - 5x_2 \\ 3x_1 + 4x_2 \end{bmatrix}. \quad ▩$$

If $\mathbf{A}$ in (11) is square, $n \times n$, then (11) maps R^n into R^n. If this $\mathbf{A}$ is nonsingular, so that $\mathbf{A}^{-1}$ exists (see Sec. 7.8), then multiplication of (11) by $\mathbf{A}^{-1}$ from the left and use of $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$ gives the **inverse transformation**

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}. \tag{14}$$

It maps every $\mathbf{y} = \mathbf{y}_0$ onto that $\mathbf{x}$, which by (11) is mapped onto $\mathbf{y}_0$. *The inverse of a linear transformation is itself linear,* because it is given by a matrix, as (14) shows.

PROBLEM SET 7.9

1–12 VECTOR SPACES

(Additional problems in Problem Set 7.4.)

Is the given set (taken with the usual addition and scalar multiplication) a vector space? (Give a reason.) If your answer is yes, find the dimension and a basis.

1. All vectors in R^3 satisfying $5v_1 - 3v_2 + 2v_3 = 0$
2. All vectors in R^3 satisfying $2v_1 + 3v_2 - v_3 = 0$, $v_1 - 4v_2 + v_3 = 0$
3. All 2×3 matrices with all entries nonnegative
4. All symmetric 3×3 matrices
5. All vectors in R^5 with the first three components 0
6. All vectors in R^4 with $v_1 + v_2 = 0$, $v_3 - v_4 = 1$
7. All skew-symmetric 2×2 matrices
8. All $n \times n$ matrices $\mathbf{A}$ with fixed n and $\det \mathbf{A} = 0$
9. All polynomials with positive coefficients and degree 3 or less
10. All functions $f(x) = a \cos x + b \sin x$ with any constants a and b
11. All functions $f(x) = (ax + b)e^{-x}$ with any constants a and b
12. All 2×3 matrices with the second row any multiple of $[4 \quad 0 \quad -9]$

13. **(Different bases)** Find three bases for R^2.

14. **(Uniqueness)** Show that the representation $\mathbf{v} = c_1\mathbf{a}_{(1)} + \cdots + c_n\mathbf{a}_{(n)}$ of any given vector in an n-dimensional vector space V in terms of a given basis $\mathbf{a}_{(1)}, \cdots, \mathbf{a}_{(n)}$ for V is unique.

15–20 LINEAR TRANSFORMATIONS

Find the inverse transformation. (Show the details of your work.)

15. $y_1 = x_1 - 2x_2$
$y_2 = 4x_1 - 3x_2$

16. $y_1 = 5x_1 - x_2$
$y_2 = 3x_1 - x_2$

17. $y_1 = 3x_1 - x_2$
$y_2 = -5x_1 + 2x_2$

18. $y_1 = 0.25x_1 - 0.1x_3$
$y_2 = x_2 - 0.8x_3$
$y_3 = 0.2x_3$

19. $y_1 = 2x_1 - 3x_2$
$y_2 = -10x_1 + 16x_2 + x_3$
$y_3 = -7x_1 + 11x_2 + x_3$

20. $y_1 = x_1 + x_2 - 2x_3$
$y_2 = x_1 + x_2 + 2x_3$
$y_3 = -2x_1 + 2x_2 + 4x_3$

21–26 INNER PRODUCT. ORTHOGONALITY

Find the Euclidean norm of the vectors

21. $[4 \quad 2 \quad -6]^T$

22. $[0 \quad -3 \quad 3 \quad 0 \quad 5 \quad 1]^T$

23. $[16 \quad -32 \quad 0]^T$

24. $[\frac{1}{4} \quad \frac{3}{8} \quad \frac{1}{2} \quad 2]^T$

25. $[0 \quad 1 \quad 0 \quad 0 \quad -1 \quad 1 \quad -1]^T$

26. $[\frac{2}{3} \quad -\frac{2}{3} \quad \frac{1}{3}]^T$

27. **(Orthogonality)** Show that the vectors in Probs. 21 and 23 are orthogonal.

28. Find all vectors $\mathbf{v}$ in R^3 orthogonal to $[2 \quad 0 \quad 1]^T$.

29. **(Unit vectors)** Find all unit vectors orthogonal to $[4 \quad -3]^T$. Make a sketch.

30. **(Triangle inequality)** Verify (4) for the vectors in Probs. 21 and 23.

CHAPTER 7 REVIEW QUESTIONS AND PROBLEMS

1. What properties of matrix multiplication differ from those of the multiplication of numbers? What about division of matrices?
2. Let $\mathbf{A}$ be a 50×50 matrix and $\mathbf{B}$ a 50×20 matrix. Are the following expressions defined or not? $\mathbf{A} + \mathbf{B}$, $\mathbf{A}^2$, $\mathbf{B}^2$, $\mathbf{AB}$, $\mathbf{BA}$, $\mathbf{AA}^T$, $\mathbf{B}^T\mathbf{A}$, $\mathbf{B}^T\mathbf{B}$, $\mathbf{BB}^T$, $\mathbf{B}^T\mathbf{AB}$. (Give reasons.)
3. How is matrix multiplication motivated?
4. Are there any linear systems without solutions? With one solution? With more than one solution? Give simple examples.
5. How can you give the rank of a matrix in terms of row vectors? Of column vectors? Of determinants?
6. What is the role of rank in connection with solving linear systems?
7. What is the row space of a matrix? The column space? The null space?
8. What is the idea of Gauss elimination and back substitution?
9. What is the inverse of a matrix? When does it exist? How would you determine it?
10. What is Cramer's rule? When would you apply it?

11–19 LINEAR SYSTEMS

Find all solutions or indicate that no solution exists. (Show the details of your work.)

11. $9x - 3y = 15$
$5x + 4y = 48$

12. $-2x - 4y + 7z = -6$
$x + 2y + 16z = 3$

13. $3x + 5y - 8z = 18$
$x + 2y - 3z = 6$

14. $5x - 10y = 2$
$3x + y = 13$
$-x + 6y = 6$

15. $-8x + 2z = 1$
$6y + 4z = 3$
$12x + 2y = 2$

16. $2y + z = -1$
$2x + 3y - z = -12$
$5x - 4y + 3z = 32$

17. $3x + 7y = 0$
$5x - 4y = 47$
$6x + 9y = 15$

18. $-x + 4y - 2z = 1$
$3x + 4y + 6z = 1$
$x - 2y + 2z = -\frac{1}{2}$

19. $7x + 9y - 14z = 36$
$-x - 3y + 2z = -12$
$2x + y - 4z = 4$

20–30 CALCULATIONS WITH MATRICES AND VECTORS

Calculate the following expressions (showing the details of your work) or indicate why they do not exist, when

$$\mathbf{A} = \begin{bmatrix} 9 & 2 & 8 \\ 2 & 18 & 10 \\ 8 & 10 & 15 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 & 2 & 6 \\ -2 & 0 & -3 \\ -6 & 3 & 0 \end{bmatrix},$$

$$\mathbf{a} = \begin{bmatrix} 3 \\ 7 \\ 1 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 4 \\ 0 \\ 2 \end{bmatrix}.$$

20. $\mathbf{AB}$, $\mathbf{BA}$

21. $\mathbf{A} - \mathbf{A}^T$

22. $\mathbf{A}^2 + \mathbf{B}^2$

23. det $\mathbf{A}$, det $\mathbf{B}$, det $\mathbf{AB}$

24. $\mathbf{AA}^T$, $\mathbf{A}^T\mathbf{A}$

25. $0.2\mathbf{BB}^T$

26. $\mathbf{Aa}$, $\mathbf{a}^T\mathbf{A}$, $\mathbf{a}^T\mathbf{Aa}$

27. $\mathbf{a}^T\mathbf{b}$, $\mathbf{b}^T\mathbf{a}$, $\mathbf{ab}^T$

28. $\mathbf{b}^T\mathbf{Bb}$

29. $\mathbf{a}^T\mathbf{B}$, $\mathbf{B}^T\mathbf{a}$

30. $0.1(\mathbf{A} + \mathbf{A}^T)(\mathbf{B} - \mathbf{B}^T)$

31–36 RANK

Determine the ranks of the coefficient matrix and the augmented matrix and state how many solutions the linear system will have.

31. In Prob. 13 **32.** In Prob. 12 **33.** In Prob. 17

34. In Prob. 14 **35.** In Prob. 19 **36.** In Prob. 18

37–42 INVERSE

Find the inverse or state why it does not exist. (Show details.)

37. Of the coefficient matrix in Prob. 11

38. Of the coefficient matrix in Prob. 15

39. Of the coefficient matrix in Prob. 16

40. Of the coefficient matrix in Prob. 18

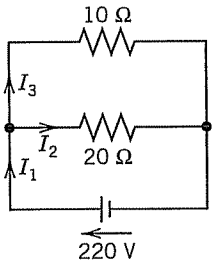
41. Of the augmented matrix in Prob. 14

42. Of the diagonal matrix with entries 3, −1, 5

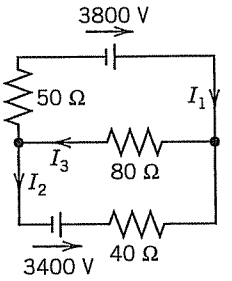
43–45 NETWORKS

Find the currents in the following networks.

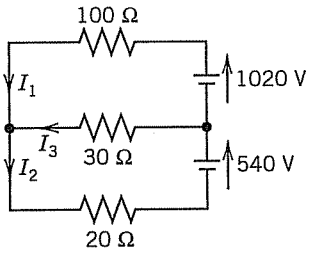
43.



44.



45.



SUMMARY OF CHAPTER 7

Linear Algebra: Matrices, Vectors, Determinants Linear Systems of Equations

An $m \times n$ **matrix** $\mathbf{A} = [a_{jk}]$ is a rectangular array of numbers or functions ("entries", "elements") arranged in m horizontal **rows** and n vertical **columns.** If $m = n$, the matrix is called **square.** A $1 \times n$ matrix is called a **row vector** and an $m \times 1$ matrix a **column vector** (Sec. 7.1).

The **sum** $\mathbf{A} + \mathbf{B}$ of matrices of the same **size** (i.e., both $m \times n$) is obtained by adding corresponding entries. The **product** of $\mathbf{A}$ by a scalar c is obtained by multiplying each a_{jk} by c (Sec. 7.1).

The **product** $\mathbf{C} = \mathbf{AB}$ of an $m \times n$ matrix $\mathbf{A}$ by an $r \times p$ matrix $\mathbf{B} = [b_{jk}]$ is defined only when $r = n$, and is the $m \times p$ matrix $\mathbf{C} = [c_{jk}]$ with entries

$$(1) \qquad c_{jk} = a_{j1}b_{1k} + a_{j2}b_{2k} + \cdots + a_{jn}b_{nk} \qquad \text{(row } j \text{ of } \mathbf{A} \text{ times column } k \text{ of } \mathbf{B}\text{)}.$$

This multiplication is motivated by the composition of **linear transformations** (Secs. 7.2, 7.9). It is associative, but is ***not commutative:*** if $\mathbf{AB}$ is defined, $\mathbf{BA}$ may not be defined, but even if $\mathbf{BA}$ is defined, $\mathbf{AB} \neq \mathbf{BA}$ in general. Also $\mathbf{AB} = \mathbf{0}$ may not imply $\mathbf{A} = \mathbf{0}$ or $\mathbf{B} = \mathbf{0}$ or $\mathbf{BA} = \mathbf{0}$ (Secs. 7.2, 7.8). Illustrations:

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

$$\begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix}$$

$$[1 \quad 2] \begin{bmatrix} 3 \\ 4 \end{bmatrix} = [11], \qquad \begin{bmatrix} 3 \\ 4 \end{bmatrix} [1 \quad 2] = \begin{bmatrix} 3 & 6 \\ 4 & 8 \end{bmatrix}.$$

The **transpose** $\mathbf{A}^T$ of a matrix $\mathbf{A} = [a_{jk}]$ is $\mathbf{A}^T = [a_{kj}]$; rows become columns and conversely (Sec. 7.2). Here, $\mathbf{A}$ need not be square. If it is and $\mathbf{A} = \mathbf{A}^T$, then $\mathbf{A}$ is called **symmetric;** if $\mathbf{A} = -\mathbf{A}^T$, it is called **skew-symmetric.** For a product, $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$ (Sec. 7.2).

A main application of matrices concerns **linear systems of equations**

$$\mathbf{Ax} = \mathbf{b} \tag{2}$$

(Sec. 7.3)

(m equations in n unknowns $x_1, \cdots, x_n$; $\mathbf{A}$ and $\mathbf{b}$ given). The most important method of solution is the **Gauss elimination** (Sec. 7.3), which reduces the system to "triangular" form by *elementary row operations,* which leave the set of solutions unchanged. (Numeric aspects and variants, such as *Doolittle's* and *Cholesky's methods,* are discussed in Secs. 20.1 and 20.2)

Cramer's rule (Secs. 7.6, 7.7) represents the unknowns in a system (2) of n equations in n unknowns as quotients of determinants; for numeric work it is impractical. **Determinants** (Sec. 7.7) have decreased in importance, but will retain their place in eigenvalue problems, elementary geometry, etc.

The **inverse** $\mathbf{A}^{-1}$ of a square matrix satisfies $\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. It exists if and only if $\det \mathbf{A} \neq 0$. It can be computed by the *Gauss–Jordan elimination* (Sec. 7.8).

The **rank** r of a matrix $\mathbf{A}$ is the maximum number of linearly independent rows or columns of $\mathbf{A}$ or, equivalently, the number of rows of the largest square submatrix of $\mathbf{A}$ with nonzero determinant (Secs. 7.4, 7.7).

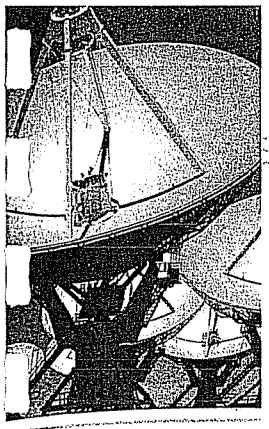
The system (2) has solutions if and only if rank $\mathbf{A}$ = rank $[\mathbf{A} \quad \mathbf{b}]$, where $[\mathbf{A} \quad \mathbf{b}]$ is the **augmented matrix** (Fundamental Theorem, Sec. 7.5).

The **homogeneous system**

$$\mathbf{Ax} = \mathbf{0} \tag{3}$$

has solutions $\mathbf{x} \neq \mathbf{0}$ ("nontrivial solutions") if and only if rank $\mathbf{A} < n$, in the case $m = n$ equivalently if and only if $\det \mathbf{A} = 0$ (Secs. 7.6, 7.7).

Vector spaces, inner product spaces, and linear transformations are discussed in Sec. 7.9. See also Sec. 7.4.



CHAPTER 8

Linear Algebra: Matrix Eigenvalue Problems

Matrix eigenvalue problems concern the solutions of vector equations

$$\mathbf{Ax} = \lambda\mathbf{x} \tag{1}$$

where $\mathbf{A}$ is a given ***square*** matrix and vector $\mathbf{x}$ and scalar λ are unknown. Clearly, $\mathbf{x} = \mathbf{0}$ is a solution of (1), giving $\mathbf{0} = \mathbf{0}$. But this of no interest, and we want to find solution vectors $\mathbf{x} \neq \mathbf{0}$ of (1), called **eigenvectors** of $\mathbf{A}$. We shall see that eigenvectors can be found only for certain values of the scalar λ; these values λ for which an eigenvector exists are called the **eigenvalues** of $\mathbf{A}$. Geometrically, solving (1) in this way means that we are looking for vectors $\mathbf{x}$ for which the multiplication of $\mathbf{x}$ by the matrix $\mathbf{A}$ has the same effect as the multiplication of $\mathbf{x}$ by a scalar λ, giving a vector $\lambda\mathbf{x}$ with components proportional to those of $\mathbf{x}$, and λ as the factor of proportionality.

Eigenvalue problems are of greatest practical interest to the engineer, physicist, and mathematician, and we shall see that their theory makes up a beautiful chapter in linear algebra that has found numerous applications.

We shall explain how to solve that vector equation (1) in Sec. 8.1, show a few typical applications in Sec. 8.2, and then discuss eigenvalue problems for symmetric, skew-symmetric, and orthogonal matrices in Sec. 8.3. In Sec. 8.4 we show how to obtain eigenvalues by diagonalization of a matrix. We also consider the complex counterparts of those matrices (Hermitian, skew-Hermitian, and unitary matrices, Sec. 8.5), which play a role in modern physics.

COMMENT. ***Numerics for eigenvalues* (Secs. 20.6–20.9) *can be studied immediately after this chapter.***

Prerequisite: Chap. 7.
Sections that may be omitted in a shorter course: 8.4, 8.5
References and Answers to Problems: App. 1 Part B, App. 2.

8.1 Eigenvalues, Eigenvectors

From the viewpoint of engineering applications, eigenvalue problems are among the most important problems in connection with matrices, and the student should follow the present discussion with particular attention. We begin by defining the basic concepts and show how to solve these problems, by examples as well as in general. Then we shall turn to applications.

Let $\mathbf{A} = [a_{jk}]$ be a given $n \times n$ matrix and consider the vector equation

$$\mathbf{Ax} = \lambda\mathbf{x}. \tag{1}$$

Here $\mathbf{x}$ is an unknown vector and λ an unknown scalar. Our task is to determine $\mathbf{x}$'s and λ's that satisfy (1). Geometrically, we are looking for vectors $\mathbf{x}$ for which the multiplication by $\mathbf{A}$ has the same effect as the multiplication by a scalar λ; in other words, $\mathbf{Ax}$ should be proportional to $\mathbf{x}$.

Clearly, the zero vector $\mathbf{x} = \mathbf{0}$ is a solution of (1) for any value of λ, because $\mathbf{A0} = \mathbf{0}$. This is of no interest. A value of λ for which (1) has a solution $\mathbf{x} \neq \mathbf{0}$ is called an **eigenvalue** or *characteristic value* (or *latent root*) of the matrix $\mathbf{A}$. ("Eigen" is German and means "proper" or "characteristic.") The corresponding solutions $\mathbf{x} \neq \mathbf{0}$ of (1) are called the **eigenvectors** or *characteristic vectors* of $\mathbf{A}$ corresponding to that eigenvalue λ. The set of all the eigenvalues of $\mathbf{A}$ is called the **spectrum** of $\mathbf{A}$. We shall see that the spectrum consists of at least one eigenvalue and at most of n numerically different eigenvalues. The largest of the absolute values of the eigenvalues of $\mathbf{A}$ is called the *spectral radius* of $\mathbf{A}$, a name to be motivated later.

How to Find Eigenvalues and Eigenvectors

The problem of determining the eigenvalues and eigenvectors of a matrix is called an **eigenvalue problem.** (More precisely: an *algebraic* eigenvalue problem, as opposed to an eigenvalue problem involving an ODE, PDE (see Secs. 5.7 and 12.3) or integral equation.) Such problems occur in physical, technical, geometric, and other applications, as we shall see. We show how to solve them, first by an example and then in general. Some typical applications will follow afterwards.

EXAMPLE 1 **Determination of Eigenvalues and Eigenvectors**

We illustrate all the steps in terms of the matrix

$$\mathbf{A} = \begin{bmatrix} -5 & 2 \\ 2 & -2 \end{bmatrix}.$$

Solution. (a) ***Eigenvalues.*** These must be determined ***first.*** Equation (1) is

$$\mathbf{Ax} = \begin{bmatrix} -5 & 2 \\ 2 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \lambda \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}; \qquad \text{in components,} \qquad \begin{aligned} -5x_1 + 2x_2 &= \lambda x_1 \\ 2x_1 - 2x_2 &= \lambda x_2. \end{aligned}$$

Transferring the terms on the right to the left, we get

$$\begin{aligned} (-5 - \lambda)x_1 + \quad 2x_2 \quad &= 0 \\ 2x_1 \quad + (-2 - \lambda)x_2 &= 0. \end{aligned} \tag{2*}$$

This can be written in matrix notation

(3*) $$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$$

because (1) is $\mathbf{Ax} - \lambda\mathbf{x} = \mathbf{Ax} - \lambda\mathbf{Ix} = (\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$, which gives (3*). We see that this is a ***homogeneous*** linear system. By Cramer's theorem in Sec. 7.7 it has a nontrivial solution $\mathbf{x} \neq \mathbf{0}$ (an eigenvector of $\mathbf{A}$ we are looking for) if and only if its coefficient determinant is zero, that is,

(4*) $$D(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} -5 - \lambda & 2 \\ 2 & -2 - \lambda \end{vmatrix} = (-5 - \lambda)(-2 - \lambda) - 4 = \lambda^2 + 7\lambda + 6 = 0.$$

We call $D(\lambda)$ the **characteristic determinant** or, if expanded, the **characteristic polynomial,** and $D(\lambda) = 0$ the **characteristic equation** of $\mathbf{A}$. The solutions of this quadratic equation are $\lambda_1 = -1$ and $\lambda_2 = -6$. These are the eigenvalues of $\mathbf{A}$.

($\mathbf{b_1}$) *Eigenvector of* $\mathbf{A}$ *corresponding to* $\boldsymbol{\lambda_1}$. This vector is obtained from (2*) with $\lambda = \lambda_1 = -1$, that is,

$$\begin{aligned} -4x_1 + 2x_2 &= 0 \\ 2x_1 - x_2 &= 0. \end{aligned}$$

A solution is $x_2 = 2x_1$, as we see from either of the two equations, so that we need only one of them. This determines an eigenvector corresponding to $\lambda_1 = -1$ up to a scalar multiple. If we choose $x_1 = 1$, we obtain the eigenvector

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}. \qquad \text{Check:} \qquad \mathbf{Ax}_1 = \begin{bmatrix} -5 & 2 \\ 2 & -2 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} -1 \\ -2 \end{bmatrix} = (-1)\mathbf{x}_1 = \lambda_1\mathbf{x}_1.$$

($\mathbf{b_2}$) *Eigenvector of* $\mathbf{A}$ *corresponding to* $\boldsymbol{\lambda_2}$. For $\lambda = \lambda_2 = -6$, equation (2*) becomes

$$\begin{aligned} x_1 + 2x_2 &= 0 \\ 2x_1 + 4x_2 &= 0. \end{aligned}$$

A solution is $x_2 = -x_1/2$ with arbitrary x_1. If we choose $x_1 = 2$, we get $x_2 = -1$. Thus an eigenvector of $\mathbf{A}$ corresponding to $\lambda_2 = -6$ is

$$\mathbf{x}_2 = \begin{bmatrix} 2 \\ -1 \end{bmatrix}. \qquad \text{Check:} \qquad \mathbf{Ax}_2 = \begin{bmatrix} -5 & 2 \\ 2 & -2 \end{bmatrix} \begin{bmatrix} 2 \\ -1 \end{bmatrix} = \begin{bmatrix} -12 \\ 6 \end{bmatrix} = (-6)\mathbf{x}_2 = \lambda_2\mathbf{x}_2. \quad \blacksquare$$

This example illustrates the general case as follows. Equation (1) written in components is

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= \lambda x_1 \\ a_{21}x_1 + \cdots + a_{2n}x_n &= \lambda x_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots & \\ a_{n1}x_1 + \cdots + a_{nn}x_n &= \lambda x_n. \end{aligned}$$

Transferring the terms on the right side to the left side, we have

(2)
$$\begin{aligned} (a_{11} - \lambda)x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= 0 \\ a_{21}x_1 + (a_{22} - \lambda)x_2 + \cdots + a_{2n}x_n &= 0 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + (a_{nn} - \lambda)x_n &= 0. \end{aligned}$$

In matrix notation,

(3) $$\boxed{(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}.}$$

By Cramer's theorem in Sec. 7.7, this homogeneous linear system of equations has a nontrivial solution if and only if the corresponding determinant of the coefficients is zero:

$$(4) \quad D(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} a_{11} - \lambda & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} - \lambda & \cdots & a_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} - \lambda \end{vmatrix} = 0.$$

$\mathbf{A} - \lambda\mathbf{I}$ is called the **characteristic matrix** and $D(\lambda)$ the **characteristic determinant** of **A.** Equation (4) is called the **characteristic equation** of **A.** By developing $D(\lambda)$ we obtain a polynomial of nth degree in λ. This is called the **characteristic polynomial** of **A.**

This proves the following important theorem.

THEOREM 1

Eigenvalues

The eigenvalues of a square matrix **A** *are the roots of the characteristic equation* (4) *of* **A.**

Hence an $n \times n$ *matrix has at least one eigenvalue and at most* n *numerically different eigenvalues.*

For larger n, the actual computation of eigenvalues will in general require the use of Newton's method (Sec. 19.2) or another numeric approximation method in Secs. 20.7–20.9.

The eigenvalues must be determined first. Once these are known, corresponding eigen*vectors* are obtained from the system (2), for instance, by the Gauss elimination, where λ is the eigenvalue for which an eigenvector is wanted. This is what we did in Example 1 and shall do again in the examples below. (To prevent misunderstandings: *numeric approximation methods* (Sec. 20.8) may determine eigen*vectors first.*)

Eigenvectors have the following properties.

THEOREM 2

Eigenvectors, Eigenspace

If $\mathbf{w}$ *and* $\mathbf{x}$ *are eigenvectors of a matrix* **A** *corresponding to* ***the same*** *eigenvalue* λ, *so are* $\mathbf{w} + \mathbf{x}$ (provided $\mathbf{x} \neq -\mathbf{w}$) *and* $k\mathbf{x}$ *for any* $k \neq 0$.

Hence the eigenvectors corresponding to one and the same eigenvalue λ *of* **A,** *together with* **0,** *form a vector space* (cf. Sec. 7.4), *called the* **eigenspace** *of* **A** *corresponding to that* λ.

PROOF $\mathbf{Aw} = \lambda\mathbf{w}$ and $\mathbf{Ax} = \lambda\mathbf{x}$ *imply* $\mathbf{A}(\mathbf{w} + \mathbf{x}) = \mathbf{Aw} + \mathbf{Ax} = \lambda\mathbf{w} + \lambda\mathbf{x} = \lambda(\mathbf{w} + \mathbf{x})$ and $\mathbf{A}(k\mathbf{w}) = k(\mathbf{Aw}) = k(\lambda\mathbf{w}) = \lambda(k\mathbf{w})$; hence $\mathbf{A}(k\mathbf{w} + \ell\mathbf{x}) = \lambda(k\mathbf{w} + \ell\mathbf{x})$. ∎

In particular, *an eigenvector* $\mathbf{x}$ *is determined only up to a constant factor.* Hence we can **normalize x,** that is, multiply it by a scalar to get a unit vector (see Sec. 7.9). For instance, $\mathbf{x}_1 = [1 \quad 2]^T$ in Example 1 has the length $\| \mathbf{x}_1 \| = \sqrt{1^2 + 2^2} = \sqrt{5}$; hence $[1/\sqrt{5} \quad 2/\sqrt{5}]^T$ is a normalized eigenvector (a unit eigenvector).

Examples 2 and 3 will illustrate that an $n \times n$ matrix may have n linearly independent eigenvectors, or it may have fewer than n. In Example 4 we shall see that a *real* matrix may have *complex* eigenvalues and eigenvectors.

EXAMPLE 2 Multiple Eigenvalues

Find the eigenvalues and eigenvectors of

$$\mathbf{A} = \begin{bmatrix} -2 & 2 & -3 \\ 2 & 1 & -6 \\ -1 & -2 & 0 \end{bmatrix}.$$

Solution. For our matrix, the characteristic determinant gives the characteristic equation

$$-\lambda^3 - \lambda^2 + 21\lambda + 45 = 0.$$

The roots (eigenvalues of $\mathbf{A}$) are $\lambda_1 = 5$, $\lambda_2 = \lambda_3 = -3$. To find eigenvectors, we apply the Gauss elimination (Sec. 7.3) to the system $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$, first with $\lambda = 5$ and then with $\lambda = -3$. For $\lambda = 5$ the characteristic matrix is

$$\mathbf{A} - \lambda\mathbf{I} = \mathbf{A} - 5\mathbf{I} = \begin{bmatrix} -7 & 2 & -3 \\ 2 & -4 & -6 \\ -1 & -2 & -5 \end{bmatrix}. \qquad \text{It row-reduces to} \qquad \begin{bmatrix} -7 & 2 & -3 \\ 0 & -\frac{24}{7} & -\frac{48}{7} \\ 0 & 0 & 0 \end{bmatrix}.$$

Hence it has rank 2. Choosing $x_3 = -1$ we have $x_2 = 2$ from $-\frac{24}{7}x_2 - \frac{48}{7}x_3 = 0$ and then $x_1 = 1$ from $-7x_1 + 2x_2 - 3x_3 = 0$. Hence an eigenvector of $\mathbf{A}$ coresponding to $\lambda = 5$ is $\mathbf{x}_1 = [1 \quad 2 \quad -1]^T$.

For $\lambda = -3$ the characteristic matrix

$$\mathbf{A} - \lambda\mathbf{I} = \mathbf{A} + 3\mathbf{I} = \begin{bmatrix} 1 & 2 & -3 \\ 2 & 4 & -6 \\ -1 & -2 & 3 \end{bmatrix} \qquad \text{row-reduces to} \qquad \begin{bmatrix} 1 & 2 & -3 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Hence it has rank 1. From $x_1 + 2x_2 - 3x_3 = 0$ we have $x_1 = -2x_2 + 3x_3$. Choosing $x_2 = 1$, $x_3 = 0$ and $x_2 = 0$, $x_3 = 1$, we obtain two linearly independent eigenvectors of $\mathbf{A}$ corresponding to $\lambda = -3$ [as they must exist by (5), Sec. 7.5, with rank $= 1$ and $n = 3$],

$$\mathbf{x}_2 = \begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix}$$

and

$$\mathbf{x}_3 = \begin{bmatrix} 3 \\ 0 \\ 1 \end{bmatrix}.$$

The order M_λ of an eigenvalue λ as a root of the characteristic polynomial is called the **algebraic multiplicity** of λ. The number m_λ of linearly independent eigenvectors corresponding to λ is called the **geometric multiplicity** of λ. Thus m_λ is the dimension of the eigenspace corresponding to this λ. Since the characteristic polynomial has degree n, the sum of all the algebraic multiplicities must equal n. In Example 2 for $\lambda = -3$ we have $m_\lambda = M_\lambda = 2$. In general, $m_\lambda \leqq M_\lambda$, as can be shown. The difference $\Delta_\lambda = M_\lambda - m_\lambda$ is called the **defect** of λ. Thus $\Delta_{-3} = 0$ in Example 2, but positive defects Δ_λ can easily occur:

EXAMPLE 3 **Algebraic Multiplicity, Geometric Multiplicity. Positive Defect**

The characteristic equation of the matrix

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \quad \text{is} \quad \det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} -\lambda & 1 \\ 0 & -\lambda \end{vmatrix} = \lambda^2 = 0.$$

Hence $\lambda = 0$ is an eigenvalue of algebraic multiplicity $M_0 = 2$. But its geometric multiplicity is only $m_0 = 1$, since eigenvectors result from $-0x_1 + x_2 = 0$, hence $x_2 = 0$, in the form $[x_1 \quad 0]^\mathsf{T}$. Hence for $\lambda = 0$ the defect is $\Delta_0 = 1$.

Similarly, the characteristic equation of the matrix

$$\mathbf{A} = \begin{bmatrix} 3 & 2 \\ 0 & 3 \end{bmatrix} \quad \text{is} \quad \det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} 3 - \lambda & 2 \\ 0 & 3 - \lambda \end{vmatrix} = (3 - \lambda)^2 = 0.$$

Hence $\lambda = 3$ is an eigenvalue of algebraic multiplicity $M_3 = 2$, but its geometric multiplicity is only $m_3 = 1$, since eigenvectors result from $0x_1 + 2x_2 = 0$ in the form $[x_1 \quad 0]^\mathsf{T}$. ■

EXAMPLE 4 **Real Matrices with Complex Eigenvalues and Eigenvectors**

Since real polynomials may have complex roots (which then occur in conjugate pairs), a real matrix may have complex eigenvalues and eigenvectors. For instance, the characteristic equation of the skew-symmetric matrix

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \quad \text{is} \quad \det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} -\lambda & 1 \\ -1 & -\lambda \end{vmatrix} = \lambda^2 + 1 = 0.$$

It gives the eigenvalues $\lambda_1 = i$ $(= \sqrt{-1})$, $\lambda_2 = -i$. Eigenvectors are obtained from $-ix_1 + x_2 = 0$ and $ix_1 + x_2 = 0$, respectively, and we can choose $x_1 = 1$ to get

$$\begin{bmatrix} 1 \\ i \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 \\ -i \end{bmatrix}. \qquad ■$$

In the next section we shall need the following simple theorem.

THEOREM 3

Eigenvalues of the Transpose

The transpose $\mathbf{A}^\mathsf{T}$ *of a square matrix* $\mathbf{A}$ *has the same eigenvalues as* $\mathbf{A}$.

PROOF Transposition does not change the value of the characteristic determinant, as follows from Theorem 2d in Sec. 7.7. ■

Having gained a first impression of matrix eigenvalue problems, in the next section we illustrate their importance with some typical applications.

PROBLEM SET 8.1

1–25 EIGENVALUES AND EIGENVECTORS

Find the eigenvalues and eigenvectors of the following matrices. (Use the given λ or factors.)

1. $\begin{bmatrix} -2 & 0 \\ 0 & 0.4 \end{bmatrix}$

2. $\begin{bmatrix} a & b \\ 0 & c \end{bmatrix}$

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

4. $\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$

5. $\begin{bmatrix} 5 & -2 \\ 9 & -6 \end{bmatrix}$

6. $\begin{bmatrix} a & b \\ -b & a \end{bmatrix}$

7. $\begin{bmatrix} 0.8 & -0.6 \\ 0.6 & 0.8 \end{bmatrix}$

8. $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

9. $\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$

10. $\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$

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

12. $\begin{bmatrix} 85 & -28 & -28 \\ -10 & -11 & -11 \\ -46 & -2 & -2 \end{bmatrix}$

13. $\begin{bmatrix} 6 & 2 & -2 \\ 2 & 5 & 0 \\ -2 & 0 & 7 \end{bmatrix}$, $\lambda = 3$

14. $\begin{bmatrix} 2 & 0 & -2 \\ 0 & 0 & -2 \\ -2 & -2 & 1 \end{bmatrix}$, $\lambda = 1$

15. $\begin{bmatrix} 1 & 0 & 0 \\ 2 & 4 & 0 \\ 6 & 4 & 2 \end{bmatrix}$

16. $\begin{bmatrix} 0.5 & 0.2 & 0.1 \\ 0 & 1.0 & 1.5 \\ 0 & 0 & 3.5 \end{bmatrix}$

17. $\begin{bmatrix} 4 & -2 & 3 \\ -2 & 1 & 6 \\ 1 & 2 & 2 \end{bmatrix}$, $\lambda = -3$

18. $\begin{bmatrix} 3 & 0 & 12 \\ -6 & 3 & 0 \\ 9 & 6 & 3 \end{bmatrix}$, $\lambda = 9$

19. $\begin{bmatrix} 13 & 5 & 2 \\ 2 & 7 & -8 \\ 5 & 4 & 7 \end{bmatrix}$

20. $\begin{bmatrix} 0 & 0 & -5 & 7 \\ 0 & 0 & 7 & -5 \\ 0 & 0 & 19 & -1 \\ 0 & 0 & -1 & 19 \end{bmatrix}$

21. $\begin{bmatrix} 0 & -2 & 2 & 0 \\ -4 & 2 & -2 & 4 \\ 0 & 2 & 2 & -4 \\ 0 & 2 & -6 & 4 \end{bmatrix}$, $\lambda = 4$

22. $\begin{bmatrix} -3 & 0 & 4 & 2 \\ 0 & 1 & -2 & 4 \\ 2 & 4 & -1 & -2 \\ 0 & 2 & -2 & 3 \end{bmatrix}$, $(\lambda - 3)^2$

23. $\begin{bmatrix} 2 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 2 & 0 & 3 & 0 \\ 1 & 4 & 2 & -6 \end{bmatrix}$

24. $\begin{bmatrix} -3 & 0 & -2 & 8 \\ 0 & 1 & 4 & -2 \\ -4 & 10 & -1 & -2 \\ 6 & -4 & -2 & 3 \end{bmatrix}$, $\lambda = 3$, $\lambda = -5$

25. $\begin{bmatrix} -1 & 0 & 12 & 0 \\ 0 & -1 & 0 & 12 \\ 0 & 0 & -1 & -4 \\ 0 & 0 & -4 & -1 \end{bmatrix}$, $(\lambda + 1)^2$

26. (Multiple eigenvalues) Find further 2×2 and 3×3 matrices with multiple eigenvalues. (See Example 2.)

27. (Nonzero defect) Find further 2×2 and 3×3 matrices with positive defect. (See Example 3.)

28. (Transpose) Illustrate Theorem 3 with examples of your own.

29. (Complex eigenvalues) Show that the eigenvalues of a real matrix are real or complex conjugate in pairs.

30. (Inverse) Show that the inverse $\mathbf{A}^{-1}$ exists if and only if none of the eigenvalues $\lambda_1, \cdots, \lambda_n$ of $\mathbf{A}$ is zero, and then $\mathbf{A}^{-1}$ has the eigenvalues $1/\lambda_1, \cdots, 1/\lambda_n$.

8.2 Some Applications of Eigenvalue Problems

In this section we discuss a few typical examples from the range of applications of matrix eigenvalue problems, which is incredibly large. Chapter 4 shows matrix eigenvalue problems related to ODEs governing mechanical systems and electrical networks. To keep our present discussion independent of Chap. 4, we include a typical application of that kind as our last example.

EXAMPLE 1 Stretching of an Elastic Membrane

An elastic membrane in the x_1x_2-plane with boundary circle ${x_1}^2 + {x_2}^2 = 1$ (Fig. 158) is stretched so that a point P: (x_1, x_2) goes over into the point Q: (y_1, y_2) given by

$$(1) \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \mathbf{A}\mathbf{x} = \begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}; \qquad \text{in components,} \quad \begin{aligned} y_1 &= 5x_1 + 3x_2 \\ y_2 &= 3x_1 + 5x_2. \end{aligned}$$

Find the **principal directions,** that is, the directions of the position vector $\mathbf{x}$ of P for which the direction of the position vector $\mathbf{y}$ of Q is the same or exactly opposite. What shape does the boundary circle take under this deformation?

Solution. We are looking for vectors $\mathbf{x}$ such that $\mathbf{y} = \lambda\mathbf{x}$. Since $\mathbf{y} = \mathbf{A}\mathbf{x}$, this gives $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$, the equation of an eigenvalue problem. In components, $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ is

$$(2) \qquad \begin{aligned} 5x_1 + 3x_2 &= \lambda x_1 \\ 3x_1 + 5x_2 &= \lambda x_2 \end{aligned} \qquad \text{or} \qquad \begin{aligned} (5 - \lambda)x_1 + 3x_2 &= 0 \\ 3x_1 + (5 - \lambda)x_2 &= 0. \end{aligned}$$

The characteristic equation is

$$(3) \qquad \begin{vmatrix} 5 - \lambda & 3 \\ 3 & 5 - \lambda \end{vmatrix} = (5 - \lambda)^2 - 9 = 0.$$

Its solutions are $\lambda_1 = 8$ and $\lambda_2 = 2$. These are the eigenvalues of our problem. For $\lambda = \lambda_1 = 8$, our system (2) becomes

$$\begin{aligned} -3x_1 + 3x_2 &= 0, \\ 3x_1 - 3x_2 &= 0. \end{aligned} \quad \Big| \quad \begin{aligned} &\text{Solution } x_2 = x_1, \quad x_1 \text{ arbitrary,} \\ &\text{for instance, } x_1 = x_2 = 1. \end{aligned}$$

For $\lambda_2 = 2$, our system (2) becomes

$$\begin{aligned} 3x_1 + 3x_2 &= 0, \\ 3x_1 + 3x_2 &= 0. \end{aligned} \quad \Big| \quad \begin{aligned} &\text{Solution } x_2 = -x_1, \quad x_1 \text{ arbitrary,} \\ &\text{for instance, } x_1 = 1, x_2 = -1. \end{aligned}$$

We thus obtain as eigenvectors of $\mathbf{A}$, for instance, $[1 \;\; 1]^T$ corresponding to λ_1 and $[1 \;\; -1]^T$ corresponding to λ_2 (or a nonzero scalar multiple of these). These vectors make 45° and 135° angles with the positive x_1-direction. They give the principal directions, the answer to our problem. The eigenvalues show that in the principal directions the membrane is stretched by factors 8 and 2, respectively; see Fig. 158.

Accordingly, if we choose the principal directions as directions of a new Cartesian u_1u_2-coordinate system, say, with the positive u_1-semi-axis in the first quadrant and the positive u_2-semi-axis in the second quadrant of the x_1x_2-system, and if we set $u_1 = r\cos\phi$, $u_2 = r\sin\phi$, then a boundary point of the unstretched circular membrane has coordinates $\cos\phi$, $\sin\phi$. Hence, after the stretch we have

$$z_1 = 8\cos\phi, \qquad z_2 = 2\sin\phi.$$

Since $\cos^2\phi + \sin^2\phi = 1$, this shows that the deformed boundary is an ellipse (Fig. 158)

$$(4) \qquad \frac{{z_1}^2}{8^2} + \frac{{z_2}^2}{2^2} = 1. \qquad \blacksquare$$

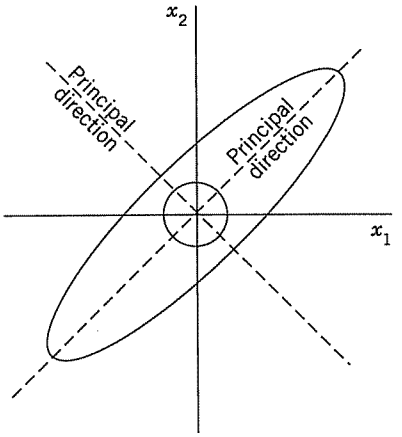


Fig. 158. Undeformed and deformed membrane in Example 1

EXAMPLE 2 **Eigenvalue Problems Arising from Markov Processes**

Markov processes as considered in Example 13 of Sec. 7.2 lead to eigenvalue problems if we ask for the limit state of the process in which the state vector $\mathbf{x}$ is reproduced under the multiplication by the stochastic matrix $\mathbf{A}$ governing the process, that is, $\mathbf{Ax} = \mathbf{x}$. Hence $\mathbf{A}$ should have the eigenvalue 1, and $\mathbf{x}$ should be a corresponding eigenvector. This is of practical interest because it shows the long-term tendency of the development modeled by the process.

In that example,

$$\mathbf{A} = \begin{bmatrix} 0.7 & 0.1 & 0 \\ 0.2 & 0.9 & 0.2 \\ 0.1 & 0 & 0.8 \end{bmatrix}. \qquad \text{For the transpose,} \qquad \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0.1 & 0.9 & 0 \\ 0 & 0.2 & 0.8 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

Hence $\mathbf{A}^T$ has the eigenvalue 1, and the same is true for $\mathbf{A}$ by Theorem 3 in Sec. 8.1. An eigenvector $\mathbf{x}$ of $\mathbf{A}$ for $\lambda = 1$ is obtained from

$$\mathbf{A} - \mathbf{I} = \begin{bmatrix} -0.3 & 0.1 & 0 \\ 0.2 & -0.1 & 0.2 \\ 0.1 & 0 & -0.2 \end{bmatrix}, \qquad \text{row-reduced to} \qquad \begin{bmatrix} -3/10 & 1/10 & 0 \\ 0 & -1/30 & 1/5 \\ 0 & 0 & 0 \end{bmatrix}.$$

Taking $x_3 = 1$, we get $x_2 = 6$ from $-x_2/30 + x_3/5 = 0$ and then $x_1 = 2$ from $-3x_1/10 + x_2/10 = 0$. This gives $\mathbf{x} = [2 \quad 6 \quad 1]^T$. It means that in the long run, the ratio Commercial : Industrial : Residential will approach 2:6:1, provided that the probabilities given by $\mathbf{A}$ remain (about) the same. (We switched to ordinary fractions to avoid rounding errors.)

EXAMPLE 3 **Eigenvalue Problems Arising from Population Models. Leslie Model**

The Leslie model describes age-specified population growth, as follows. Let the oldest age attained by the females in some animal population be 9 years. Divide the population into three age classes of 3 years each. Let the *"Leslie matrix"* be

$$\mathbf{L} = [l_{jk}] = \begin{bmatrix} 0 & 2.3 & 0.4 \\ 0.6 & 0 & 0 \\ 0 & 0.3 & 0 \end{bmatrix} \tag{5}$$

where l_{1k} is the average number of daughters born to a single female during the time she is in age class k, and $l_{j,j-1}$ $(j = 2, 3)$ is the fraction of females in age class $j - 1$ that will survive and pass into class j. (a) What is the number of females in each class after 3, 6, 9 years if each class initially consists of 400 females? (b) For what initial distribution will the number of females in each class change by the same proportion? What is this rate of change?

Solution. (a) Initially, $\mathbf{x}_{(0)}^{\mathsf{T}} = [400 \quad 400 \quad 400]$. After 3 years,

$$\mathbf{x}_{(3)} = \mathbf{L}\mathbf{x}_{(0)} = \begin{bmatrix} 0 & 2.3 & 0.4 \\ 0.6 & 0 & 0 \\ 0 & 0.3 & 0 \end{bmatrix} \begin{bmatrix} 400 \\ 400 \\ 400 \end{bmatrix} = \begin{bmatrix} 1080 \\ 240 \\ 120 \end{bmatrix}.$$

Similarly, after 6 years the number of females in each class is given by $\mathbf{x}_{(6)}^{\mathsf{T}} = (\mathbf{L}\mathbf{x}_{(3)})^{\mathsf{T}} = [600 \quad 648 \quad 72]$, and after 9 years we have $\mathbf{x}_{(9)}^{\mathsf{T}} = (\mathbf{L}\mathbf{x}_{(6)})^{\mathsf{T}} = [1519.2 \quad 360 \quad 194.4]$.

(b) Proportional change means that we are looking for a distribution vector $\mathbf{x}$ such that $\mathbf{L}\mathbf{x} = \lambda\mathbf{x}$, where λ is the rate of change (growth if $\lambda > 1$, decrease if $\lambda < 1$). The characteristic equation is (develop the characteristic determinant by the first column)

$$\det(\mathbf{L} - \lambda\mathbf{I}) = -\lambda^3 - 0.6(-2.3\lambda - 0.3 \cdot 0.4) = -\lambda^3 + 1.38\lambda + 0.072 = 0.$$

A positive root is found to be (for instance, by Newton's method, Sec. 19.2) $\lambda = 1.2$. A corresponding eigenvector $\mathbf{x}$ can be determined from the characteristic matrix

$$\mathbf{A} - 1.2\mathbf{I} = \begin{bmatrix} -1.2 & 2.3 & 0.4 \\ 0.6 & -1.2 & 0 \\ 0 & 0.3 & -1.2 \end{bmatrix}, \qquad \text{say,} \qquad \mathbf{x} = \begin{bmatrix} 1 \\ 0.5 \\ 0.125 \end{bmatrix}$$

where $x_3 = 0.125$ is chosen, $x_2 = 0.5$ then follows from $0.3x_2 - 1.2x_3 = 0$, and $x_1 = 1$ from $-1.2x_1 + 2.3x_2 + 0.4x_3 = 0$. To get an initial population of 1200 as before, we multiply $\mathbf{x}$ by $1200/(1 + 0.5 + 0.125) = 738$. *Answer:* Proportional growth of the numbers of females in the three classes will occur if the initial values are 738, 369, 92 in classes 1, 2, 3, respectively. The growth rate will be 1.2 per 3 years. ■

EXAMPLE 4 **Vibrating System of Two Masses on Two Springs (Fig. 159)**

Mass–spring systems involving several masses and springs can be treated as eigenvalue problems. For instance, the mechanical system in Fig. 159 is governed by the system of ODEs

$$\begin{aligned} y_1'' &= -5y_1 + 2y_2 \\ y_2'' &= \quad 2y_1 - 2y_2 \end{aligned} \tag{6}$$

where y_1 and y_2 are the displacements of the masses from rest, as shown in the figure, and primes denote derivatives with respect to time t. In vector form, this becomes

$$\mathbf{y}'' = \begin{bmatrix} y_1'' \\ y_2'' \end{bmatrix} = \mathbf{A}\mathbf{y} = \begin{bmatrix} -5 & 2 \\ 2 & -2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}. \tag{7}$$

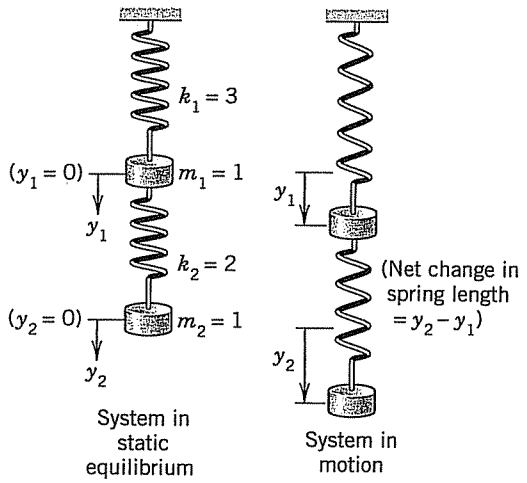


Fig. 159. Masses on springs in Example 4

We try a vector solution of the form

$$\mathbf{y} = \mathbf{x}e^{\omega t}. \tag{8}$$

This is suggested by a mechanical system of a single mass on a spring (Sec. 2.4), whose motion is given by exponential functions (and sines and cosines). Substitution into (7) gives

$$\omega^2 \mathbf{x}e^{\omega t} = \mathbf{A}\mathbf{x}e^{\omega t}.$$

Dividing by $e^{\omega t}$ and writing $\omega^2 = \lambda$, we see that our mechanical system leads to the eigenvalue problem

$$\mathbf{A}\mathbf{x} = \lambda \mathbf{x} \qquad \text{where } \lambda = \omega^2. \tag{9}$$

From Example 1 in Sec. 8.1 we see that $\mathbf{A}$ has the eigenvalues $\lambda_1 = -1$ and $\lambda_2 = -6$. Consequently, $\omega = \sqrt{-1} = \pm i$ and $\sqrt{-6} = \pm i\sqrt{6}$, respectively. Corresponding eigenvectors are

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad \text{and} \quad \mathbf{x}_2 = \begin{bmatrix} 2 \\ -1 \end{bmatrix}. \tag{10}$$

From (8) we thus obtain the four complex solutions [see (10), Sec. 2.2]

$$\mathbf{x}_1 e^{\pm it} = \mathbf{x}_1(\cos t \pm i \sin t),$$
$$\mathbf{x}_2 e^{\pm i\sqrt{6}t} = \mathbf{x}_2(\cos \sqrt{6}\, t \pm i \sin \sqrt{6}\, t).$$

By addition and subtraction (see Sec. 2.2) we get the four real solutions

$$\mathbf{x}_1 \cos t, \quad \mathbf{x}_1 \sin t, \quad \mathbf{x}_2 \cos \sqrt{6}\, t, \quad \mathbf{x}_2 \sin \sqrt{6}\, t.$$

A general solution is obtained by taking a linear combination of these,

$$\mathbf{y} = \mathbf{x}_1(a_1 \cos t + b_1 \sin t) + \mathbf{x}_2(a_2 \cos \sqrt{6}\, t + b_2 \sin \sqrt{6}\, t)$$

with arbitrary constants a_1, b_1, a_2, b_2 (to which values can be assigned by prescribing initial displacement and initial velocity of each of the two masses). By (10), the components of $\mathbf{y}$ are

$$y_1 = a_1 \cos t + b_1 \sin t + 2a_2 \cos \sqrt{6}\, t + 2b_2 \sin \sqrt{6}\, t$$
$$y_2 = 2a_1 \cos t + 2b_1 \sin t - a_2 \cos \sqrt{6}\, t - b_2 \sin \sqrt{6}\, t.$$

These functions describe harmonic oscillations of the two masses. Physically, this had to be expected because we have neglected damping. ■

PROBLEM SET 8.2

1–6 LINEAR TRANSFORMATIONS

Find the matrix $\mathbf{A}$ in the indicated linear transformation $\mathbf{y} = \mathbf{A}\mathbf{x}$. Explain the geometric significance of the eigenvalues and eigenvectors of $\mathbf{A}$. Show the details.

1. Reflection about the y-axis in R^2
2. Reflection about the xy-plane in R^3
3. Orthogonal projection (perpendicular projection) of R^2 onto the x-axis
4. Orthogonal projection of R^3 onto the plane $y = x$
5. Dilatation (uniform stretching) in R^2 by a factor 5
6. Counterclockwise rotation through the angle $\pi/2$ about the origin in R^2

7–14 ELASTIC DEFORMATIONS

Given $\mathbf{A}$ in a deformation $\mathbf{y} = \mathbf{A}\mathbf{x}$, find the principal directions and corresponding factors of extension or contraction. Show the details.

7. $\begin{bmatrix} 3 & 5 \\ 5 & 3 \end{bmatrix}$

8. $\begin{bmatrix} 0.4 & 0.8 \\ 0.8 & 0.4 \end{bmatrix}$

9. $\begin{bmatrix} 2.5 & 1.5 \\ 1.5 & 6.5 \end{bmatrix}$

10. $\begin{bmatrix} 5 & 4 \\ 4 & 11 \end{bmatrix}$

11. $\begin{bmatrix} 7 & \sqrt{6} \\ \sqrt{6} & 2 \end{bmatrix}$

12. $\begin{bmatrix} 5 & 2 \\ 2 & 13 \end{bmatrix}$

13. $\begin{bmatrix} -2 & 3 \\ 3 & -2 \end{bmatrix}$

14. $\begin{bmatrix} 10.5 & 1/\sqrt{2} \\ 1/\sqrt{2} & 10.0 \end{bmatrix}$

15. (Leontief[1] input–output model) Suppose that three industries are interrelated so that their outputs are used as inputs by themselves, according to the 3×3 **consumption matrix**

$$\mathbf{A} = [a_{jk}] = \begin{bmatrix} 0.2 & 0.5 & 0 \\ 0.6 & 0 & 0.3 \\ 0.2 & 0.5 & 0.7 \end{bmatrix}$$

where a_{jk} is the fraction of the output of industry k consumed (purchased) by industry j. Let p_j be the price charged by industry j for its total output. A problem is to find prices so that for each industry, total expenditures equal total income. Show that this leads to $\mathbf{Ap} = \mathbf{p}$, where $\mathbf{p} = [p_1 \quad p_2 \quad p_3]^T$, and find a solution $\mathbf{p}$ with nonnegative p_1, p_2, p_3.

16. Show that a consumption matrix as considered in Prob. 15 must have column sums 1 and always has the eigenvalue 1.

17. (Open Leontief input–output model) If not the whole output but only a portion of it is consumed by the industries themselves, then instead of $\mathbf{Ax} = \mathbf{x}$ (as in Prob. 15), we have $\mathbf{x} - \mathbf{Ax} = \mathbf{y}$, where $\mathbf{x} = [x_1 \quad x_2 \quad x_3]^T$ is produced, $\mathbf{Ax}$ is consumed by the industries, and, thus, $\mathbf{y}$ is the net production available for other consumers. Find for what production $\mathbf{x}$ a given **demand vector** $\mathbf{y} = [0.136 \quad 0.272 \quad 0.136]^T$ can be achieved if the consumption matrix is

$$\mathbf{A} = \begin{bmatrix} 0.2 & 0.4 & 0.2 \\ 0.3 & 0 & 0.1 \\ 0.2 & 0.4 & 0.5 \end{bmatrix}.$$

18–20 MARKOV PROCESSES

Find limit states of the Markov processes modeled by the following matrices. (Show the details.)

18. $\begin{bmatrix} 0.1 & 0.4 \\ 0.9 & 0.6 \end{bmatrix}$

19. $\begin{bmatrix} 0.5 & 0.3 & 0.2 \\ 0.3 & 0.5 & 0.2 \\ 0.2 & 0.2 & 0.6 \end{bmatrix}$

20. $\begin{bmatrix} 0.6 & 0.1 & 0.2 \\ 0.4 & 0.1 & 0.4 \\ 0 & 0.8 & 0.4 \end{bmatrix}$

21–23 POPULATION MODEL WITH AGE SPECIFICATION

Find the growth rate in the Leslie model (see Example 3) with the matrix as given. (Show details.)

21. $\begin{bmatrix} 0 & 3.45 & 0.60 \\ 0.90 & 0 & 0 \\ 0 & 0.45 & 0 \end{bmatrix}$

22. $\begin{bmatrix} 0 & 12.0 & 0 \\ 0.75 & 0 & 0 \\ 0 & 0.30 & 0 \end{bmatrix}$

23. $\begin{bmatrix} 0 & 7.280 & 2.975 \\ 0.560 & 0 & 0 \\ 0 & 0.420 & 0 \end{bmatrix}$

24. TEAM PROJECT. General Properties of Eigenvalues and Eigenvectors. Prove the following statements and illustrate them with examples of your own choice. Here, $\lambda_1, \cdots, \lambda_n$ are the (not necessarily distinct) eigenvalues of a given $n \times n$ matrix $\mathbf{A} = [a_{jk}]$.

(a) Trace. The sum of the main diagonal entries is called the **trace** of $\mathbf{A}$. It equals the sum of the eigenvalues.

(b) "Spectral shift." $\mathbf{A} - k\mathbf{I}$ has the eigenvalues $\lambda_1 - k, \cdots, \lambda_n - k$ and the same eigenvectors as $\mathbf{A}$.

(c) Scalar multiples, powers. $k\mathbf{A}$ has the eigenvalues $k\lambda_1, \cdots, k\lambda_n$. $\mathbf{A}^m$ $(m = 1, 2, \cdots)$ has the eigenvalues $\lambda_1^m, \cdots, \lambda_n^m$. The eigenvectors are those of $\mathbf{A}$.

(d) Spectral mapping theorem. The **"polynomial matrix"**

$$p(\mathbf{A}) = k_m\mathbf{A}^m + k_{m-1}\mathbf{A}^{m-1} + \cdots + k_1\mathbf{A} + k_0\mathbf{I}$$

has the eigenvalues

$$p(\lambda_j) = k_m\lambda_j^m + k_{m-1}\lambda_j^{m-1} + \cdots + k_1\lambda_j + k_0$$

where $j = 1, \cdots, n$, and the same eigenvectors as $\mathbf{A}$.

(e) Perron's theorem. Show that a Leslie matrix $\mathbf{L}$ with positive $l_{12}, l_{13}, l_{21}, l_{32}$ has a positive eigenvalue. (This is a special case of the famous Perron–Frobenius theorem in Sec. 20.7, which is difficult to prove in its general form.)

---

[1]WASSILY LEONTIEF (1906–1999). American economist at New York University. For his input–output analysis he was awarded the Nobel Prize in 1973.

8.3 Symmetric, Skew-Symmetric, and Orthogonal Matrices

We consider three classes of real square matrices that occur quite frequently in applications because they have several remarkable properties which we shall now discuss. The first two of these classes have already been mentioned in Sec. 7.2.

DEFINITIONS

Symmetric, Skew-Symmetric, and Orthogonal Matrices

A *real* square matrix $\mathbf{A} = [a_{jk}]$ is called

symmetric if transposition leaves it unchanged,

$$\text{(1)} \qquad \mathbf{A}^T = \mathbf{A}, \qquad \text{thus} \qquad a_{kj} = a_{jk},$$

skew-symmetric if transposition gives the negative of $\mathbf{A}$,

$$\text{(2)} \qquad \mathbf{A}^T = -\mathbf{A}, \qquad \text{thus} \qquad a_{kj} = -a_{jk},$$

orthogonal if transposition gives the inverse of $\mathbf{A}$,

$$\text{(3)} \qquad \mathbf{A}^T = \mathbf{A}^{-1}.$$

EXAMPLE 1 **Symmetric, Skew-Symmetric, and Orthogonal Matrices**

The matrices

$$\begin{bmatrix} -3 & 1 & 5 \\ 1 & 0 & -2 \\ 5 & -2 & 4 \end{bmatrix}, \quad \begin{bmatrix} 0 & 9 & -12 \\ -9 & 0 & 20 \\ 12 & -20 & 0 \end{bmatrix}, \quad \begin{bmatrix} \frac{2}{3} & \frac{1}{3} & \frac{2}{3} \\ -\frac{2}{3} & \frac{2}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{2}{3} & -\frac{2}{3} \end{bmatrix}$$

are symmetric, skew-symmetric, and orthogonal, respectively, as you should verify. Every skew-symmetric matrix has all main diagonal entries zero. (Can you prove this?) ■

Any real square matrix $\mathbf{A}$ may be written as the sum of a symmetric matrix $\mathbf{R}$ and a skew-symmetric matrix $\mathbf{S}$, where

$$\text{(4)} \qquad \mathbf{R} = \tfrac{1}{2}(\mathbf{A} + \mathbf{A}^T) \qquad \text{and} \qquad \mathbf{S} = \tfrac{1}{2}(\mathbf{A} - \mathbf{A}^T).$$

EXAMPLE 2 **Illustration of Formula (4)**

$$\mathbf{A} = \begin{bmatrix} 9 & 5 & 2 \\ 2 & 3 & -8 \\ 5 & 4 & 3 \end{bmatrix} = \mathbf{R} + \mathbf{S} = \begin{bmatrix} 9.0 & 3.5 & 3.5 \\ 3.5 & 3.0 & -2.0 \\ 3.5 & -2.0 & 3.0 \end{bmatrix} + \begin{bmatrix} 0 & 1.5 & -1.5 \\ -1.5 & 0 & -6.0 \\ 1.5 & 6.0 & 0 \end{bmatrix}$$

■

THEOREM 1

Eigenvalues of Symmetric and Skew-Symmetric Matrices

(a) *The eigenvalues of a symmetric matrix are real.*

(b) *The eigenvalues of a skew-symmetric matrix are pure imaginary or zero.*

This basic theorem (and an extension of it) will be proved in Sec. 8.5.

EXAMPLE 3 **Eigenvalues of Symmetric and Skew-Symmetric Matrices**

The matrices in (1) and (7) of Sec. 8.2 are symmetric and have real eigenvalues. The skew-symmetric matrix in Example 1 has the eigenvalues 0, $-25i$, and $25i$. (Verify this.) The following matrix has the real eigenvalues 1 and 5 but is not symmetric. Does this contradict Theorem 1?

$$\begin{bmatrix} 3 & 4 \\ 1 & 3 \end{bmatrix}$$

Orthogonal Transformations and Orthogonal Matrices

Orthogonal transformations are transformations

$$\mathbf{y} = \mathbf{A}\mathbf{x} \qquad \text{where } \mathbf{A} \text{ is an orthogonal matrix.} \tag{5}$$

With each vector $\mathbf{x}$ in R^n such a transformation assigns a vector $\mathbf{y}$ in R^n. For instance, the plane rotation through an angle θ

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{6}$$

is an orthogonal transformation. It can be shown that any orthogonal transformation in the plane or in three-dimensional space is a **rotation** (possibly combined with a reflection in a straight line or a plane, respectively).

The main reason for the importance of orthogonal matrices is as follows.

THEOREM 2

Invariance of Inner Product

An orthogonal transformation preserves the value of the **inner product** *of vectors* $\mathbf{a}$ *and* $\mathbf{b}$ *in* R^n, *defined by*

$$\mathbf{a} \cdot \mathbf{b} = \mathbf{a}^T\mathbf{b} = [a_1 \quad \cdots \quad a_n] \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}. \tag{7}$$

That is, for any $\mathbf{a}$ *and* $\mathbf{b}$ *in* R^n, *orthogonal* $n \times n$ *matrix* $\mathbf{A}$, *and* $\mathbf{u} = \mathbf{A}\mathbf{a}$, $\mathbf{v} = \mathbf{A}\mathbf{b}$ *we have* $\mathbf{u} \cdot \mathbf{v} = \mathbf{a} \cdot \mathbf{b}$.

Hence the transformation also preserves the **length** *or* **norm** *of any vector* $\mathbf{a}$ *in* R^n *given by*

$$\|\mathbf{a}\| = \sqrt{\mathbf{a} \cdot \mathbf{a}} = \sqrt{\mathbf{a}^T\mathbf{a}}. \tag{8}$$

PROOF Let $\mathbf{A}$ be orthogonal. Let $\mathbf{u} = \mathbf{Aa}$ and $\mathbf{v} = \mathbf{Ab}$. We must show that $\mathbf{u} \cdot \mathbf{v} = \mathbf{a} \cdot \mathbf{b}$. Now $(\mathbf{Aa})^T = \mathbf{a}^T\mathbf{A}^T$ by (10d) in Sec. 7.2 and $\mathbf{A}^T\mathbf{A} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$ by (3). Hence

$$\mathbf{u} \cdot \mathbf{v} = \mathbf{u}^T\mathbf{v} = (\mathbf{Aa})^T\mathbf{Ab} = \mathbf{a}^T\mathbf{A}^T\mathbf{Ab} = \mathbf{a}^T\mathbf{I}\mathbf{b} = \mathbf{a}^T\mathbf{b} = \mathbf{a} \cdot \mathbf{b}. \tag{9}$$

From this the invariance of $\| \mathbf{a} \|$ follows if we set $\mathbf{b} = \mathbf{a}$. ∎

Orthogonal matrices have further interesting properties as follows.

THEOREM 3

Orthonormality of Column and Row Vectors

A real square matrix is orthogonal if and only if its column vectors $\mathbf{a}_1, \cdots, \mathbf{a}_n$ *(and also its row vectors) form an* **orthonormal system,** *that is,*

$$\mathbf{a}_j \cdot \mathbf{a}_k = \mathbf{a}_j^T\mathbf{a}_k = \begin{cases} 0 & \text{if } j \neq k \\ 1 & \text{if } j = k. \end{cases} \tag{10}$$

PROOF **(a)** Let $\mathbf{A}$ be orthogonal. Then $\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}^T\mathbf{A} = \mathbf{I}$, in terms of column vectors $\mathbf{a}_1, \cdots, \mathbf{a}_n$,

$$\mathbf{I} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{A}^T\mathbf{A} = \begin{bmatrix} \mathbf{a}_1^T \\ \vdots \\ \mathbf{a}_n^T \end{bmatrix} [\mathbf{a}_1 \cdots \mathbf{a}_n] = \begin{bmatrix} \mathbf{a}_1^T\mathbf{a}_1 & \mathbf{a}_1^T\mathbf{a}_2 \cdots \mathbf{a}_1^T\mathbf{a}_n \\ \cdot & \cdot \quad \cdots \quad \cdot \\ \mathbf{a}_n^T\mathbf{a}_1 & \mathbf{a}_n^T\mathbf{a}_2 \cdots \mathbf{a}_n^T\mathbf{a}_n \end{bmatrix}. \tag{11}$$

The last equality implies (10), by the definition of the $n \times n$ unit matrix $\mathbf{I}$. From (3) it follows that the inverse of an orthogonal matrix is orthogonal (see CAS Experiment 20). Now the column vectors of $\mathbf{A}^{-1}$ $(= \mathbf{A}^T)$ are the row vectors of $\mathbf{A}$. Hence the row vectors of $\mathbf{A}$ also form an orthonormal system.

(b) Conversely, if the column vectors of $\mathbf{A}$ satisfy (10), the off-diagonal entries in (11) must be 0 and the diagonal entries 1. Hence $\mathbf{A}^T\mathbf{A} = \mathbf{I}$, as (11) shows. Similarly, $\mathbf{AA}^T = \mathbf{I}$. This implies $\mathbf{A}^T = \mathbf{A}^{-1}$ because also $\mathbf{A}^{-1}\mathbf{A} = \mathbf{AA}^{-1} = \mathbf{I}$ and the inverse is unique. Hence $\mathbf{A}$ is orthogonal. Similarly when the row vectors of $\mathbf{A}$ form an orthonormal system, by what has been said at the end of part **(a)**. ∎

THEOREM 4

Determinant of an Orthogonal Matrix

The determinant of an orthogonal matrix has the value $+1$ *or* -1.

PROOF From $\det \mathbf{AB} = \det \mathbf{A} \det \mathbf{B}$ (Sec. 7.8, Theorem 4) and $\det \mathbf{A}^T = \det \mathbf{A}$ (Sec. 7.7, Theorem 2d), we get for an orthogonal matrix

$$1 = \det \mathbf{I} = \det (\mathbf{AA}^{-1}) = \det (\mathbf{AA}^T) = \det \mathbf{A} \det \mathbf{A}^T = (\det \mathbf{A})^2. \quad ∎$$

EXAMPLE 4 **Illustration of Theorems 3 and 4**

The last matrix in Example 1 and the matrix in (6) illustrate Theorems 3 and 4 because their determinants are -1 and $+1$, as you should verify. ∎

THEOREM 5

Eigenvalues of an Orthogonal Matrix

The eigenvalues of an orthogonal matrix **A** *are real or complex conjugates in pairs and have absolute value* 1.

PROOF The first part of the statement holds for any real matrix **A** because its characteristic polynomial has real coefficients, so that its zeros (the eigenvalues of **A**) must be as indicated. The claim that $|\lambda| = 1$ will be proved in Sec. 8.5. ∎

EXAMPLE 5 **Eigenvalues of an Orthogonal Matrix**

The orthogonal matrix in Example 1 has the characteristic equation

$$-\lambda^3 + \tfrac{2}{3}\lambda^2 + \tfrac{2}{3}\lambda - 1 = 0.$$

Now one of the eigenvalues must be real (why?), hence $+1$ or -1. Trying, we find -1. Division by $\lambda + 1$ gives $-(\lambda^2 - 5\lambda/3 + 1) = 0$ and the two eigenvalues $(5 + i\sqrt{11})/6$ and $(5 - i\sqrt{11})/6$, which have absolute value 1. Verify all of this. ∎

Looking back at this section, you will find that the numerous basic results it contains have relatively short, straightforward proofs. This is typical of large portions of matrix eigenvalue theory.

PROBLEM SET 8.3

1. **(Verification)** Verify the statements in Example 1.
2. Verify the statements in Examples 3 and 4.
3. Are the eigenvalues of $\mathbf{A} + \mathbf{B}$ of the form $\lambda_j + \mu_j$, where λ_j and μ_j are the eigenvalues of **A** and **B**, respectively?
4. **(Orthogonality)** Prove that eigenvectors of a symmetric matrix corresponding to different eigenvalues are orthogonal. Give an example.
5. **(Skew-symmetric matrix)** Show that the inverse of a skew-symmetric matrix is skew-symmetric.
6. Do there exist nonsingular skew-symmetric $n \times n$ matrices with odd n?
7. **(Orthogonal matrix)** Do there exist skew-symmetric orthogonal 3×3 matrices?
8. **(Symmetric matrix)** Do there exist nondiagonal symmetric 3×3 matrices that are orthogonal?

9–17 EIGENVALUES OF SYMMETRIC, SKEW-SYMMETRIC, AND ORTHOGONAL MATRICES

Are the following matrices symmetric, skew-symmetric, or orthogonal? Find their spectrum (thereby illustrating Theorems 1 and 5). (Show the details of your work.)

9. $\begin{bmatrix} 0.96 & -0.28 \\ 0.28 & 0.96 \end{bmatrix}$

10. $\begin{bmatrix} a & b \\ -b & a \end{bmatrix}$

11. $\begin{bmatrix} 3 & 1 \\ -1 & 1 \end{bmatrix}$

12. $\begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}$

13. $\begin{bmatrix} 14 & 4 & -2 \\ 4 & 14 & 2 \\ -2 & 2 & 17 \end{bmatrix}$

14. $\begin{bmatrix} 0 & -6 & -12 \\ 6 & 0 & -12 \\ 12 & 12 & 0 \end{bmatrix}$

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

16. $\begin{bmatrix} \frac{4}{9} & \frac{8}{9} & \frac{1}{9} \\ -\frac{7}{9} & \frac{4}{9} & -\frac{4}{9} \\ -\frac{4}{9} & \frac{1}{9} & \frac{8}{9} \end{bmatrix}$

17. $\begin{bmatrix} a & b & b \\ b & a & b \\ b & b & a \end{bmatrix}$

18. **(Rotation in space)** Give a geometric interpretation of the transformation $\mathbf{y} = \mathbf{A}\mathbf{x}$ with **A** as in Prob. 12 and **x** and **y** referred to a Cartesian coordinate system.
19. **WRITING PROJECT. Section Summary.** Summarize the main concepts and facts in this section, with illustrative examples of your own.

20. CAS EXPERIMENT. Orthogonal Matrices.

(a) Products. Inverse. Prove that the product of two orthogonal matrices is orthogonal, and so is the inverse of an orthogonal matrix. What does this mean in terms of rotations?

(b) Rotation. Show that (6) is an orthogonal transformation. Verify that it satisfies Theorem 3. Find the inverse transformation.

(c) Powers. Write a program for computing powers $\mathbf{A}^m$ ($m = 1, 2, \cdots$) of a 2×2 matrix $\mathbf{A}$ and their spectra. Apply it to the matrix in Prob. 9 (call it $\mathbf{A}$). To what rotation does $\mathbf{A}$ correspond? Do the eigenvalues of $\mathbf{A}^m$ have a limit as $m \to \infty$?

(d) Compute the eigenvalues of $(0.9\mathbf{A})^m$, where $\mathbf{A}$ is the matrix in Prob. 9. Plot them as points. What is their limit? Along what kind of curve do these points approach the limit?

(e) Find $\mathbf{A}$ such that $\mathbf{y} = \mathbf{Ax}$ is a counterclockwise rotation through 30° in the plane.

8.4 Eigenbases. Diagonalization. Quadratic Forms

So far we have emphasized properties of eigen*values*. We now turn to general properties of eigen*vectors*. Eigenvectors of an $n \times n$ matrix $\mathbf{A}$ may (or may not!) form a basis for R^n. If we are interested in a transformation $\mathbf{y} = \mathbf{Ax}$, such an **"eigenbasis"** (basis of eigenvectors)—if it exists—is of great advantage because then we can represent any $\mathbf{x}$ in R^n uniquely as a linear combination of the eigenvectors $\mathbf{x}_1, \cdots, \mathbf{x}_n$, say,

$$\mathbf{x} = c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \cdots + c_n\mathbf{x}_n.$$

And, denoting the corresponding (not necessarily distinct) eigenvalues of the matrix $\mathbf{A}$ by $\lambda_1, \cdots, \lambda_n$, we have $\mathbf{Ax}_j = \lambda_j\mathbf{x}_j$, so that we simply obtain

$$\begin{aligned} \mathbf{y} = \mathbf{Ax} &= \mathbf{A}(c_1\mathbf{x}_1 + \cdots + c_n\mathbf{x}_n) \\ &= c_1\mathbf{Ax}_1 + \cdots + c_n\mathbf{Ax}_n \\ &= c_1\lambda_1\mathbf{x}_1 + \cdots + c_n\lambda_n\mathbf{x}_n. \end{aligned} \tag{1}$$

This shows that we have decomposed the complicated action of $\mathbf{A}$ on an arbitrary vector $\mathbf{x}$ into a sum of simple actions (multiplication by scalars) on the eigenvectors of $\mathbf{A}$. This is the point of an eigenbasis.

Now if the n eigenvalues are all different, we do obtain a basis:

THEOREM 1

Basis of Eigenvectors

*If an $n \times n$ matrix $\mathbf{A}$ has n **distinct** eigenvalues, then $\mathbf{A}$ has a basis of eigenvectors $\mathbf{x}_1, \cdots, \mathbf{x}_n$ for R^n.*

PROOF All we have to show is that $\mathbf{x}_1, \cdots, \mathbf{x}_n$ are linearly independent. Suppose they are not. Let r be the largest integer such that $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ is a linearly independent set. Then $r < n$ and the set $\{\mathbf{x}_1, \cdots, \mathbf{x}_r, \mathbf{x}_{r+1}\}$ is linearly dependent. Thus there are scalars $c_1, \cdots, c_{r+1}$, not all zero, such that

$$c_1\mathbf{x}_1 + \cdots + c_{r+1}\mathbf{x}_{r+1} = \mathbf{0} \tag{2}$$

(see Sec. 7.4). Multiplying both sides by $\mathbf{A}$ and using $\mathbf{Ax}_j = \lambda_j\mathbf{x}_j$, we obtain

$$c_1\lambda_1\mathbf{x}_1 + \cdots + c_{r+1}\lambda_{r+1}\mathbf{x}_{r+1} = \mathbf{0}. \tag{3}$$

To get rid of the last term, we subtract λ_{r+1} times (2) from this, obtaining

$$c_1(\lambda_1 - \lambda_{r+1})\mathbf{x}_1 + \cdots + c_r(\lambda_r - \lambda_{r+1})\mathbf{x}_r = \mathbf{0}.$$

Here $c_1(\lambda_1 - \lambda_{r+1}) = 0, \cdots, c_r(\lambda_r - \lambda_{r+1}) = 0$ since $\{x_1, \cdots, x_r\}$ is linearly independent. Hence $c_1 = \cdots = c_r = 0$, since all the eigenvalues are distinct. But with this, (2) reduces to $c_{r+1}\mathbf{x}_{r+1} = \mathbf{0}$, hence $c_{r+1} = 0$, since $\mathbf{x}_{r+1} \neq \mathbf{0}$ (an eigenvector!). This contradicts the fact that not all scalars in (2) are zero. Hence the conclusion of the theorem must hold. ▩

EXAMPLE 1 **Eigenbasis. Nondistinct Eigenvalues. Nonexistence**

The matrix $\mathbf{A} = \begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix}$ has a basis of eigenvectors $\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \end{bmatrix}$ corresponding to the eigenvalues $\lambda_1 = 8, \lambda_2 = 2$. (See Example 1 in Sec. 8.2.)

Even if not all n eigenvalues are different, a matrix $\mathbf{A}$ may still provide an eigenbasis for R^n. See Example 2 in Sec. 8.1, where $n = 3$.

On the other hand, $\mathbf{A}$ ***may not have enough linearly independent eigenvectors to make up a basis.*** For instance, $\mathbf{A}$ in Example 3 of Sec. 8.1 is

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \quad \text{and has only one eigenvector} \quad \begin{bmatrix} k \\ 0 \end{bmatrix} \quad (k \neq 0, \text{ arbitrary}). \quad ▩$$

Actually, eigenbases exist under much more general conditions than those in Theorem 1. An important case is the following.

THEOREM 2

Symmetric Matrices

A symmetric matrix has an orthonormal basis of eigenvectors for R^n.

For a proof (which is involved) see Ref. [B3], vol. 1, pp. 270–272.

EXAMPLE 2 **Orthonormal Basis of Eigenvectors**

The first matrix in Example 1 is symmetric, and an orthonormal basis of eigenvectors is $[1/\sqrt{2} \quad 1/\sqrt{2}]^\mathsf{T}$, $[1/\sqrt{2} \quad -1/\sqrt{2}]^\mathsf{T}$. ▩

Diagonalization of Matrices

Eigenbases also play a role in reducing a matrix $\mathbf{A}$ to a diagonal matrix whose entries are the eigenvalues of $\mathbf{A}$. This is done by a "similarity transformation," which is defined as follows (and will have various applications in numerics in Chap. 20).

DEFINITION

Similar Matrices. Similarity Transformation

An $n \times n$ matrix $\hat{\mathbf{A}}$ is called **similar** to an $n \times n$ matrix $\mathbf{A}$ if

$$\hat{\mathbf{A}} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P} \tag{4}$$

for some (nonsingular!) $n \times n$ matrix $\mathbf{P}$. This transformation, which gives $\hat{\mathbf{A}}$ from $\mathbf{A}$, is called a **similarity transformation.**

The key property of this transformation is that it preserves the eigenvalues of $\mathbf{A}$:

THEOREM 3

Eigenvalues and Eigenvectors of Similar Matrices

If $\hat{\mathbf{A}}$ *is similar to* $\mathbf{A}$, then $\hat{\mathbf{A}}$ *has the same eigenvalues as* $\mathbf{A}$.

Furthermore, if $\mathbf{x}$ *is an eigenvector of* $\mathbf{A}$, then $\mathbf{y} = \mathbf{P}^{-1}\mathbf{x}$ *is an eigenvector of* $\hat{\mathbf{A}}$ *corresponding to the same eigenvalue.*

PROOF From $\mathbf{Ax} = \lambda\mathbf{x}$ (λ an eigenvalue, $\mathbf{x} \neq \mathbf{0}$) we get $\mathbf{P}^{-1}\mathbf{Ax} = \lambda\mathbf{P}^{-1}\mathbf{x}$. Now $\mathbf{I} = \mathbf{PP}^{-1}$. By this ***"identity trick"*** the previous equation gives

$$\mathbf{P}^{-1}\mathbf{Ax} = \mathbf{P}^{-1}\mathbf{AIx} = \mathbf{P}^{-1}\mathbf{APP}^{-1}\mathbf{x} = \hat{\mathbf{A}}(\mathbf{P}^{-1}\mathbf{x}) = \lambda\mathbf{P}^{-1}\mathbf{x}.$$

Hence λ is an eigenvalue of $\hat{\mathbf{A}}$ and $\mathbf{P}^{-1}\mathbf{x}$ a corresponding eigenvector. Indeed, $\mathbf{P}^{-1}\mathbf{x} = \mathbf{0}$ would give $\mathbf{x} = \mathbf{Ix} = \mathbf{PP}^{-1}\mathbf{x} = \mathbf{P0} = \mathbf{0}$, contradicting $\mathbf{x} \neq \mathbf{0}$. ▩

EXAMPLE 3 **Eigenvalues and Vectors of Similar Matrices**

Let $$\mathbf{A} = \begin{bmatrix} 6 & -3 \\ 4 & -1 \end{bmatrix} \quad \text{and} \quad \mathbf{P} = \begin{bmatrix} 1 & 3 \\ 1 & 4 \end{bmatrix}.$$

Then $$\hat{\mathbf{A}} = \begin{bmatrix} 4 & -3 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 6 & -3 \\ 4 & -1 \end{bmatrix} \begin{bmatrix} 1 & 3 \\ 1 & 4 \end{bmatrix} = \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}.$$

Here $\mathbf{P}^{-1}$ was obtained from (4*) in Sec. 7.8 with $\det \mathbf{P} = 1$. We see that $\hat{\mathbf{A}}$ has the eigenvalues $\lambda_1 = 3$, $\lambda_2 = 2$. The characteristic equation of $\mathbf{A}$ is $(6 - \lambda)(-1 - \lambda) + 12 = \lambda^2 - 5\lambda + 6 = 0$. It has the roots (the eigenvalues of $\mathbf{A}$) $\lambda_1 = 3$, $\lambda_2 = 2$, confirming the first part of Theorem 3.

We confirm the second part. From the first component of $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$ we have $(6 - \lambda)x_1 - 3x_2 = 0$. For $\lambda = 3$ this gives $3x_1 - 3x_2 = 0$, say, $\mathbf{x}_1 = [1 \quad 1]^T$. For $\lambda = 2$ it gives $4x_1 - 3x_2 = 0$, say, $\mathbf{x}_2 = [3 \quad 4]^T$. In Theorem 3 we thus have

$$\mathbf{y}_1 = \mathbf{P}^{-1}\mathbf{x}_1 = \begin{bmatrix} 4 & -3 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad \mathbf{y}_2 = \mathbf{P}^{-1}\mathbf{x}_2 = \begin{bmatrix} 4 & -3 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Indeed, these are eigenvectors of the diagonal matrix $\hat{\mathbf{A}}$.

Perhaps we see that $\mathbf{x}_1$ and $\mathbf{x}_2$ are the columns of $\mathbf{P}$. This suggests the general method of transforming a matrix $\mathbf{A}$ to diagonal form $\mathbf{D}$ by using $\mathbf{P} = \mathbf{X}$, the matrix with eigenvectors as columns: ▩

THEOREM 4

Diagonalization of a Matrix

If an $n \times n$ *matrix* $\mathbf{A}$ *has a basis of eigenvectors, then*

$$\mathbf{D} = \mathbf{X}^{-1}\mathbf{AX} \tag{5}$$

is diagonal, with the eigenvalues of $\mathbf{A}$ *as the entries on the main diagonal. Here* $\mathbf{X}$ *is the matrix with these eigenvectors as column vectors. Also,*

$$\mathbf{D}^m = \mathbf{X}^{-1}\mathbf{A}^m\mathbf{X} \qquad (m = 2, 3, \cdots). \tag{5*}$$

PROOF Let $\mathbf{x}_1, \cdots, \mathbf{x}_n$ constitute a basis of eigenvectors of $\mathbf{A}$ for R^n. Let the corresponding eigenvalues of $\mathbf{A}$ be $\lambda_1, \cdots, \lambda_n$, respectively, so that $\mathbf{A}\mathbf{x}_1 = \lambda_1\mathbf{x}_1, \cdots, \mathbf{A}\mathbf{x}_n = \lambda_n\mathbf{x}_n$. Then $\mathbf{X} = [\mathbf{x}_1 \quad \cdots \quad \mathbf{x}_n]$ has rank n, by Theorem 3 in Sec. 7.4. Hence $\mathbf{X}^{-1}$ exists by Theorem 1 in Sec. 7.8. We claim that

$$(6) \quad \mathbf{AX} = \mathbf{A}[\mathbf{x}_1 \quad \cdots \quad \mathbf{x}_n] = [\mathbf{A}\mathbf{x}_1 \quad \cdots \quad \mathbf{A}\mathbf{x}_n] = [\lambda_1\mathbf{x}_1 \quad \cdots \quad \lambda_n\mathbf{x}_n] = \mathbf{XD}$$

where $\mathbf{D}$ is the diagonal matrix as in (5). The fourth equality in (6) follows by direct calculation. (Try it for $n = 2$ and then for general n.) The third equality uses $\mathbf{A}\mathbf{x}_k = \lambda_k\mathbf{x}_k$. The second equality results if we note that the first column of $\mathbf{AX}$ is $\mathbf{A}$ times the first column of $\mathbf{X}$, and so on. For instance, when $n = 2$ and we write $\mathbf{x}_1 = [x_{11} \quad x_{21}]^\mathsf{T}$, $\mathbf{x}_2 = [x_{12} \quad x_{22}]^\mathsf{T}$, we have

$$\mathbf{AX} = \mathbf{A}[x_1 \quad \mathbf{x}_2] = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix}$$

$$= \begin{bmatrix} a_{11}x_{11} + a_{12}x_{21} & a_{11}x_{12} + a_{12}x_{22} \\ a_{21}x_{11} + a_{22}x_{21} & a_{21}x_{12} + a_{22}x_{22} \end{bmatrix} = [\mathbf{A}\mathbf{x}_1 \quad \mathbf{A}\mathbf{x}_2].$$

Column 1 Column 2

If we multiply (6) by $\mathbf{X}^{-1}$ from the left, we obtain (5). Since (5) is a similarity transformation, Theorem 3 implies that $\mathbf{D}$ has the same eigenvalues as $\mathbf{A}$. Equation (5*) follows if we note that

$$\mathbf{D}^2 = \mathbf{DD} = \mathbf{X}^{-1}\mathbf{AXX}^{-1}\mathbf{AX} = \mathbf{X}^{-1}\mathbf{AAX} = \mathbf{X}^{-1}\mathbf{A}^2\mathbf{X}, \qquad \text{etc.}$$

■

EXAMPLE 4 Diagonalization

Diagonalize

$$\mathbf{A} = \begin{bmatrix} 7.3 & 0.2 & -3.7 \\ -11.5 & 1.0 & 5.5 \\ 17.7 & 1.8 & -9.3 \end{bmatrix}.$$

Solution. The characteristic determinant gives the characteristic equation $-\lambda^3 - \lambda^2 + 12\lambda = 0$. The roots (eigenvalues of $\mathbf{A}$) are $\lambda_1 = 3$, $\lambda_2 = -4$, $\lambda_3 = 0$. By the Gauss elimination applied to $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$ with $\lambda = \lambda_1, \lambda_2, \lambda_3$ we find eigenvectors and then $\mathbf{X}^{-1}$ by the Gauss–Jordan elimination (Sec. 7.8, Example 1). The results are

$$\begin{bmatrix} -1 \\ 3 \\ -1 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ -1 \\ 3 \end{bmatrix}, \quad \begin{bmatrix} 2 \\ 1 \\ 4 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} -1 & 1 & 2 \\ 3 & -1 & 1 \\ -1 & 3 & 4 \end{bmatrix}, \quad \mathbf{X}^{-1} = \begin{bmatrix} -0.7 & 0.2 & 0.3 \\ -1.3 & -0.2 & 0.7 \\ 0.8 & 0.2 & -0.2 \end{bmatrix}.$$

Calculating $\mathbf{AX}$ and multiplying by $\mathbf{X}^{-1}$ from the left, we thus obtain

$$\mathbf{D} = \mathbf{X}^{-1}\mathbf{AX} = \begin{bmatrix} -0.7 & 0.2 & 0.3 \\ -1.3 & -0.2 & 0.7 \\ 0.8 & 0.2 & -0.2 \end{bmatrix} \begin{bmatrix} -3 & -4 & 0 \\ 9 & 4 & 0 \\ -3 & -12 & 0 \end{bmatrix} = \begin{bmatrix} 3 & 0 & 0 \\ 0 & -4 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

■

Quadratic Forms. Transformation to Principal Axes

By definition, a **quadratic form** Q in the components $x_1, \cdots, x_n$ of a vector $\mathbf{x}$ is a sum of n^2 terms, namely,

$$
\begin{aligned}
Q = \mathbf{x}^T\mathbf{A}\mathbf{x} &= \sum_{j=1}^{n}\sum_{k=1}^{n} a_{jk}x_jx_k \\
&= a_{11}x_1^2 + a_{12}x_1x_2 + \cdots + a_{1n}x_1x_n \\
&\quad + a_{21}x_2x_1 + a_{22}x_2^2 + \cdots + a_{2n}x_2x_n \\
&\quad + \cdots\cdots\cdots\cdots\cdots\cdots \\
&\quad + a_{n1}x_nx_1 + a_{n2}x_nx_2 + \cdots + a_{nn}x_n^2.
\end{aligned} \tag{7}
$$

$\mathbf{A} = [a_{jk}]$ is called the **coefficient matrix** of the form. We may assume that $\mathbf{A}$ is ***symmetric,*** because we can take off-diagonal terms together in pairs and write the result as a sum of two equal terms; see the following example.

EXAMPLE 5 **Quadratic Form. Symmetric Coefficient Matrix**

Let

$$
\mathbf{x}^T\mathbf{A}\mathbf{x} = [x_1 \quad x_2]\begin{bmatrix} 3 & 4 \\ 6 & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 3x_1^2 + 4x_1x_2 + 6x_2x_1 + 2x_2^2 = 3x_1^2 + 10x_1x_2 + 2x_2^2.
$$

Here $4 + 6 = 10 = 5 + 5$. From the corresponding ***symmetric*** matrix $\mathbf{C} = [c_{jk}]$, where $c_{jk} = \frac{1}{2}(a_{jk} + a_{kj})$, thus $c_{11} = 3$, $c_{12} = c_{21} = 5$, $c_{22} = 2$, we get the same result; indeed,

$$
\mathbf{x}^T\mathbf{C}\mathbf{x} = [x_1 \quad x_2]\begin{bmatrix} 3 & 5 \\ 5 & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 3x_1^2 + 5x_1x_2 + 5x_2x_1 + 2x_2^2 = 3x_1^2 + 10x_1x_2 + 2x_2^2. \qquad \blacksquare
$$

Quadratic forms occur in physics and geometry, for instance, in connection with conic sections (ellipses $x_1^2/a^2 + x_2^2/b^2 = 1$, etc.) and quadratic surfaces (cones, etc.). Their transformation to principal axes is an important practical task related to the diagonalization of matrices, as follows.

By Theorem 2 the *symmetric* coefficient matrix $\mathbf{A}$ of (7) has an orthonormal basis of eigenvectors. Hence if we take these as column vectors, we obtain a matrix $\mathbf{X}$ that is orthogonal, so that $\mathbf{X}^{-1} = \mathbf{X}^T$. From (5) we thus have $\mathbf{A} = \mathbf{X}\mathbf{D}\mathbf{X}^{-1} = \mathbf{X}\mathbf{D}\mathbf{X}^T$. Substitution into (7) gives

$$
Q = \mathbf{x}^T\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{x}. \tag{8}
$$

If we set $\mathbf{X}^T\mathbf{x} = \mathbf{y}$, then, since $\mathbf{X}^T = \mathbf{X}^{-1}$, we get

$$
\mathbf{x} = \mathbf{X}\mathbf{y}. \tag{9}
$$

Furthermore, in (8) we have $\mathbf{x}^T\mathbf{X} = (\mathbf{X}^T\mathbf{x})^T = \mathbf{y}^T$ and $\mathbf{X}^T\mathbf{x} = \mathbf{y}$, so that Q becomes simply

$$
Q = \mathbf{y}^T\mathbf{D}\mathbf{y} = \lambda_1y_1^2 + \lambda_2y_2^2 + \cdots + \lambda_ny_n^2. \tag{10}
$$

This proves the following basic theorem.

THEOREM 5

Principal Axes Theorem

The substitution (9) *transforms a quadratic form*

$$Q = \mathbf{x}^T\mathbf{A}\mathbf{x} = \sum_{j=1}^{n}\sum_{k=1}^{n} a_{jk}x_jx_k \qquad (a_{kj} = a_{jk})$$

to the principal axes form or **canonical form** (10), *where* $\lambda_1, \cdots, \lambda_n$ *are the (not necessarily distinct) eigenvalues of the (symmetric!) matrix* $\mathbf{A}$, *and* $\mathbf{X}$ *is an orthogonal matrix with corresponding eigenvectors* $\mathbf{x}_1, \cdots, \mathbf{x}_n$, *respectively, as column vectors.*

EXAMPLE 6 **Transformation to Principal Axes. Conic Sections**

Find out what type of conic section the following quadratic form represents and transform it to principal axes:

$$Q = 17x_1^2 - 30x_1x_2 + 17x_2^2 = 128.$$

Solution. We have $Q = \mathbf{x}^T\mathbf{A}\mathbf{x}$, where

$$\mathbf{A} = \begin{bmatrix} 17 & -15 \\ -15 & 17 \end{bmatrix}, \qquad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

This gives the characteristic equation $(17 - \lambda)^2 - 15^2 = 0$. It has the roots $\lambda_1 = 2$, $\lambda_2 = 32$. Hence (10) becomes

$$Q = 2y_1^2 + 32y_2^2.$$

We see that $Q = 128$ represents the ellipse $2y_1^2 + 32y_2^2 = 128$, that is,

$$\frac{y_1^2}{8^2} + \frac{y_2^2}{2^2} = 1.$$

If we want to know the direction of the principal axes in the x_1x_2-coordinates, we have to determine normalized eigenvectors from $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$ with $\lambda = \lambda_1 = 2$ and $\lambda = \lambda_2 = 32$ and then use (9). We get

$$\begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix},$$

hence

$$\mathbf{x} = \mathbf{X}\mathbf{y} = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \qquad \begin{aligned} x_1 &= y_1/\sqrt{2} - y_2/\sqrt{2} \\ x_2 &= y_1/\sqrt{2} + y_2/\sqrt{2}. \end{aligned}$$

This is a 45° rotation. Our results agree with those in Sec. 8.2, Example 1, except for the notations. See also Fig. 158 in that example. ■

PROBLEM SET 8.4

1–9 DIAGONALIZATION OF MATRICES

Find an eigenbasis (a basis of eigenvectors) and diagonalize. (Show the details.)

1. $\begin{bmatrix} 3 & 2 \\ 2 & 6 \end{bmatrix}$

2. $\begin{bmatrix} 0 & 16 \\ 4 & 0 \end{bmatrix}$

3. $\begin{bmatrix} 5 & 1 \\ 1 & 5 \end{bmatrix}$

4. $\begin{bmatrix} 3 & 2 \\ -5 & -4 \end{bmatrix}$

5. $\begin{bmatrix} 1.0 & 6.0 \\ 1.5 & 1.0 \end{bmatrix}$

6. $\begin{bmatrix} 2 & 7 \\ 6 & -9 \end{bmatrix}$

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

8. $\begin{bmatrix} -6 & -6 & 10 \\ -5 & -5 & 5 \\ -9 & -9 & 13 \end{bmatrix}$

9. $\begin{bmatrix} 3 & 10 & -15 \\ -18 & 39 & 9 \\ -24 & 40 & -15 \end{bmatrix}$

10. **(Orthonormal basis)** Illustrate Theorem 2 with further examples.

11. **(No basis)** Find further 2×2 and 3×3 matrices without eigenbases.

12. **PROJECT. Similarity of Matrices.** Similarity is basic, for instance in designing numeric methods.

(a) Trace. By definition, the **trace** of an $n \times n$ matrix $\mathbf{A} = [a_{jk}]$ is the sum of the diagonal entries,

$$\text{trace } \mathbf{A} = a_{11} + a_{22} + \cdots + a_{nn}.$$

Show that the trace equals the sum of the eigenvalues, each counted as often as its algebraic multiplicity indicates. Illustrate this with the matrices in Probs. 1, 3, 5, 7, 9.

(b) Trace of product. Let $\mathbf{B} = [b_{jk}]$ be $n \times n$. Show that similar matrices have equal traces, by first proving

$$\text{trace } \mathbf{AB} = \sum_{i=1}^{n} \sum_{l=1}^{n} a_{il} b_{li} = \text{trace } \mathbf{BA}.$$

(c) Find a relationship between $\hat{\mathbf{A}}$ in (4) and $\tilde{\mathbf{A}} = \mathbf{PAP}^{-1}$.

(d) Diagonalization. What can you do in (5) if you want to change the order of the eigenvalues in $\mathbf{D}$, for instance, interchange $d_{11} = \lambda_1$ and $d_{22} = \lambda_2$?

13–18 SIMILAR MATRICES HAVE EQUAL SPECTRA

Verify this for $\mathbf{A}$ and $\hat{\mathbf{A}} = \mathbf{P}^{-1}\mathbf{AP}$. Find eigenvectors $\mathbf{y}$ of $\hat{\mathbf{A}}$. Show that $\mathbf{x} = \mathbf{Py}$ are eigenvectors of $\mathbf{A}$. (Show the details of your work.)

13. $\mathbf{A} = \begin{bmatrix} -5 & 0 \\ 0 & 2 \end{bmatrix}$, $\mathbf{P} = \begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix}$

14. $\mathbf{A} = \begin{bmatrix} 3 & 4 \\ 4 & -3 \end{bmatrix}$, $\mathbf{P} = \begin{bmatrix} 5 & 2 \\ 2 & 1 \end{bmatrix}$

15. $\mathbf{A} = \begin{bmatrix} 4 & 2 \\ -4 & -2 \end{bmatrix}$, $\mathbf{P} = \begin{bmatrix} 1 & 3 \\ 3 & 6 \end{bmatrix}$

16. $\mathbf{A} = \begin{bmatrix} 3 & 0 \\ 7 & -2 \end{bmatrix}$, $\mathbf{P} = \begin{bmatrix} 8 & 2 \\ 1 & 4 \end{bmatrix}$

17. $\mathbf{A} = \begin{bmatrix} 4 & 0 & 0 \\ 12 & -2 & 0 \\ 21 & -6 & 1 \end{bmatrix}$, $\mathbf{P} = \begin{bmatrix} 4 & 0 & 6 \\ 0 & 2 & 0 \\ 6 & 0 & 10 \end{bmatrix}$

18. $\mathbf{A} = \begin{bmatrix} -5 & 0 & 15 \\ 3 & 4 & -9 \\ -5 & 0 & 15 \end{bmatrix}$, $\mathbf{P} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

19–28 TRANSFORMATION TO PRINCIPAL AXES. CONIC SECTIONS

What kind of conic section (or pair of straight lines) is given by the quadratic form? Transform it to principal axes. Express $\mathbf{x}^\mathsf{T} = [x_1 \quad x_2]$ in terms of the new coordinate vector $\mathbf{y}^\mathsf{T} = [y_1 \quad y_2]$, as in Example 6.

19. $x_1^2 + 24x_1x_2 - 6x_2^2 = 5$

20. $3x_1^2 + 4\sqrt{3}x_1x_2 + 7x_2^2 = 9$

21. $3x_1^2 - 8x_1x_2 - 3x_2^2 = 0$

22. $6x_1^2 + 16x_1x_2 - 6x_2^2 = 20$

23. $4x_1^2 + 2\sqrt{3}x_1x_2 + 2x_2^2 = 10$

24. $7x_1^2 - 24x_1x_2 = 144$

25. $x_1^2 - 12x_1x_2 + x_2^2 = 35$

26. $3x_1^2 + 22x_1x_2 + 3x_2^2 = 0$

27. $12x_1^2 + 32x_1x_2 + 12x_2^2 = 112$

28. $6.5x_1^2 + 5.0x_1x_2 + 6.5x_2^2 = 36$

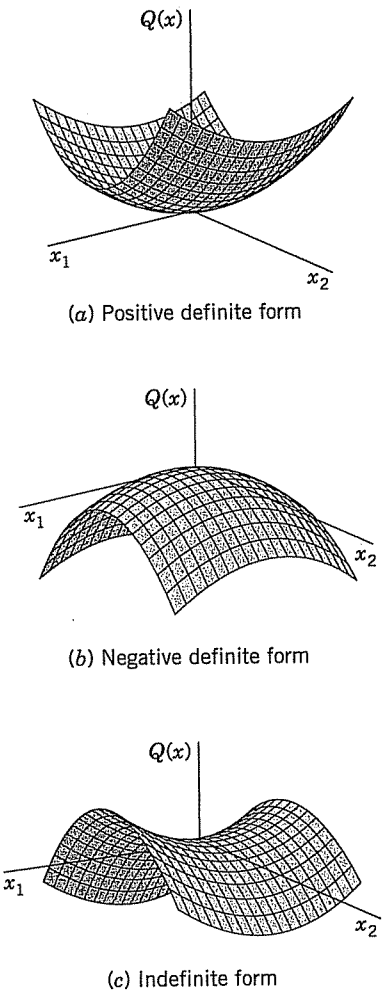
29. (Definiteness) A quadratic form $Q(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x}$ and its (symmetric!) matrix $\mathbf{A}$ are called **(a) positive definite** if $Q(\mathbf{x}) > 0$ for all $\mathbf{x} \neq \mathbf{0}$, **(b) negative definite** if $Q(\mathbf{x}) < 0$ for all $\mathbf{x} \neq 0$, **(c) indefinite** if $Q(\mathbf{x})$ takes both positive and negative values. (See Fig. 160.) [$Q(\mathbf{x})$ and $\mathbf{A}$ are called *positive semidefinite* (*negative semidefinite*) if $Q(\mathbf{x}) \geqq 0$ ($Q(\mathbf{x}) \leqq 0$) for all $\mathbf{x}$.] A necessary and sufficient condition for positive definiteness is that all the **"principal minors"** are positive (see Ref. [B3], vol. 1, p. 306), that is,

$$a_{11} > 0, \qquad \begin{vmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{vmatrix} > 0,$$

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{vmatrix} > 0, \qquad \cdots, \qquad \det \mathbf{A} > 0.$$

Show that the form in Prob. 23 is positive definite, whereas that in Prob. 19 is indefinite.

30. (Definiteness) Show that necessary and sufficient for (a), (b), (c) in Prob. 29 is that the eigenvalues of $\mathbf{A}$ are (a) all positive, (b) all negative, (c) both positive and negative. *Hint.* Use Theorem 5.



(*a*) Positive definite form

(*b*) Negative definite form

(*c*) Indefinite form

Fig. 160. Quadratic forms in two variables

8.5 Complex Matrices and Forms. *Optional*

The three classes of real matrices in Sec. 8.3 have complex counterparts that are of practical interest in certain applications, mainly because of their spectra (see Theorem 1 in this section), for instance, in quantum mechanics. To define these classes, we need the following standard

Notations

$\bar{\mathbf{A}} = [\bar{a}_{jk}]$ is obtained from $\mathbf{A} = [a_{jk}]$ by replacing each entry $a_{jk} = \alpha + i\beta$ (α, β real) with its complex conjugate $\bar{a}_{jk} = \alpha - i\beta$. Also, $\bar{\mathbf{A}}^T = [\bar{a}_{kj}]$ is the transpose of $\bar{\mathbf{A}}$, hence the conjugate transpose of $\mathbf{A}$.

EXAMPLE 1 Notations

$$\text{If } \mathbf{A} = \begin{bmatrix} 3 + 4i & 1 - i \\ 6 & 2 - 5i \end{bmatrix}, \text{ then } \bar{\mathbf{A}} = \begin{bmatrix} 3 - 4i & 1 + i \\ 6 & 2 + 5i \end{bmatrix} \text{ and } \bar{\mathbf{A}}^T = \begin{bmatrix} 3 - 4i & 6 \\ 1 + i & 2 + 5i \end{bmatrix}. \blacksquare$$

DEFINITION

Hermitian, Skew-Hermitian, and Unitary Matrices

A square matrix $\mathbf{A} = [a_{kj}]$ is called

Hermitian	if	$\bar{\mathbf{A}}^{\mathsf{T}} = \mathbf{A}$,	that is,	$\bar{a}_{kj} = a_{jk}$
skew-Hermitian	if	$\bar{\mathbf{A}}^{\mathsf{T}} = -\mathbf{A}$,	that is,	$\bar{a}_{kj} = -a_{jk}$
unitary	if	$\bar{\mathbf{A}}^{\mathsf{T}} = \mathbf{A}^{-1}$.		

The first two classes are named after Hermite (see footnote 13 in Problem Set 5.8).

From the definitions we see the following. If $\mathbf{A}$ is Hermitian, the entries on the main diagonal must satisfy $\bar{a}_{jj} = a_{jj}$; that is, they are real. Similarly, if $\mathbf{A}$ is skew-Hermitian, then $\bar{a}_{jj} = -a_{jj}$. If we set $a_{jj} = \alpha + i\beta$, this becomes $\alpha - i\beta = -(\alpha + i\beta)$. Hence $\alpha = 0$, so that a_{jj} must be pure imaginary or 0.

EXAMPLE 2 **Hermitian, Skew-Hermitian, and Unitary Matrices**

$$\mathbf{A} = \begin{bmatrix} 4 & 1 - 3i \\ 1 + 3i & 7 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 3i & 2 + i \\ -2 + i & -i \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} \frac{1}{2}i & \frac{1}{2}\sqrt{3} \\ \frac{1}{2}\sqrt{3} & \frac{1}{2}i \end{bmatrix}$$

are Hermitian, skew-Hermitian, and unitary matrices, respectively, as you may verify by using the definitions. ■

If a Hermitian matrix is real, then $\bar{\mathbf{A}}^{\mathsf{T}} = \mathbf{A}^{\mathsf{T}} = \mathbf{A}$. Hence a real Hermitian matrix is a symmetric matrix (Sec. 8.3.).

Similarly, if a skew-Hermitian matrix is real, then $\bar{\mathbf{A}}^{\mathsf{T}} = \mathbf{A}^{\mathsf{T}} = -\mathbf{A}$. Hence a real skew-Hermitian matrix is a skew-symmetric matrix.

Finally, if a unitary matrix is real, then $\bar{\mathbf{A}}^{\mathsf{T}} = \mathbf{A}^{\mathsf{T}} = \mathbf{A}^{-1}$. Hence a real unitary matrix is an orthogonal matrix.

This shows that *Hermitian, skew-Hermitian, and unitary matrices generalize symmetric, skew-symmetric, and orthogonal matrices, respectively.*

Eigenvalues

It is quite remarkable that the matrices under consideration have spectra (sets of eigenvalues; see Sec. 8.1) that can be characterized in a general way as follows (see Fig. 161).

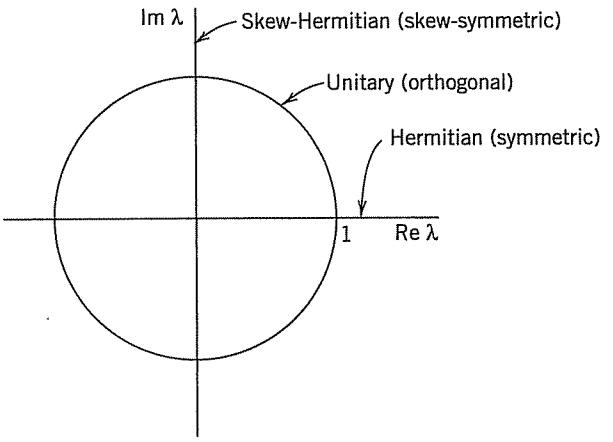


Fig. 161. Location of the eigenvalues of Hermitian, skew-Hermitian, and unitary matrices in the complex λ-plane

THEOREM 1

Eigenvalues

(a) *The eigenvalues of a Hermitian matrix (and thus of a symmetric matrix) are real.*

(b) *The eigenvalues of a skew-Hermitian matrix (and thus of a skew-symmetric matrix) are pure imaginary or zero.*

(c) *The eigenvalues of a unitary matrix (and thus of an orthogonal matrix) have absolute value* 1.

EXAMPLE 3 **Illustration of Theorem 1**

For the matrices in Example 2 we find by direct calculation

	Matrix	Characteristic Equation	Eigenvalues
A	Hermitian	$\lambda^2 - 11\lambda + 18 = 0$	$9, \quad 2$
B	Skew-Hermitian	$\lambda^2 - 2i\lambda + 8 = 0$	$4i, \quad -2i$
C	Unitary	$\lambda^2 - i\lambda - 1 = 0$	$\frac{1}{2}\sqrt{3} + \frac{1}{2}i, \quad -\frac{1}{2}\sqrt{3} + \frac{1}{2}i$

and $|\pm\frac{1}{2}\sqrt{3} + \frac{1}{2}i|^2 = \frac{3}{4} + \frac{1}{4} = 1$. ■

PROOF We prove Theorem 1. Let λ be an eigenvalue and $\mathbf{x}$ an eigenvector of $\mathbf{A}$. Multiply $\mathbf{Ax} = \lambda\mathbf{x}$ from the left by $\bar{\mathbf{x}}^\mathsf{T}$, thus $\bar{\mathbf{x}}^\mathsf{T}\mathbf{Ax} = \lambda\bar{\mathbf{x}}^\mathsf{T}\mathbf{x}$, and divide by $\bar{\mathbf{x}}^\mathsf{T}\mathbf{x} = \bar{x}_1 x_1 + \cdots + \bar{x}_n x_n = |x_1|^2 + \cdots + |x_n|^2$, which is real and not 0 because $\mathbf{x} \neq \mathbf{0}$. This gives

$$\lambda = \frac{\bar{\mathbf{x}}^\mathsf{T}\mathbf{Ax}}{\bar{\mathbf{x}}^\mathsf{T}\mathbf{x}}. \tag{1}$$

(a) If $\mathbf{A}$ is Hermitian, $\bar{\mathbf{A}}^\mathsf{T} = \mathbf{A}$ or $\mathbf{A}^\mathsf{T} = \bar{\mathbf{A}}$ and we show that then the numerator in (1) is real, which makes λ real. $\bar{\mathbf{x}}^\mathsf{T}\mathbf{Ax}$ is a scalar; hence taking the transpose has no effect. Thus

$$\bar{\mathbf{x}}^\mathsf{T}\mathbf{Ax} = (\bar{\mathbf{x}}^\mathsf{T}\mathbf{Ax})^\mathsf{T} = \mathbf{x}^\mathsf{T}\mathbf{A}^\mathsf{T}\bar{\mathbf{x}} = \mathbf{x}^\mathsf{T}\bar{\mathbf{A}}\bar{\mathbf{x}} = \overline{(\bar{\mathbf{x}}^\mathsf{T}\mathbf{Ax})}. \tag{2}$$

Hence, $\bar{\mathbf{x}}^\mathsf{T}\mathbf{Ax}$ equals its complex conjugate, so that it must be real. ($a + ib = a - ib$ implies $b = 0$.)

(b) If $\mathbf{A}$ is skew-Hermitian, $\mathbf{A}^\mathsf{T} = -\bar{\mathbf{A}}$ and instead of (2) we obtain

$$\bar{\mathbf{x}}^\mathsf{T}\mathbf{Ax} = -\overline{(\bar{\mathbf{x}}^\mathsf{T}\mathbf{Ax})} \tag{3}$$

so that $\bar{\mathbf{x}}^\mathsf{T}\mathbf{Ax}$ equals minus its complex conjugate and is pure imaginary or 0. ($a + ib = -(a - ib)$ implies $a = 0$.)

(c) Let $\mathbf{A}$ be unitary. We take $\mathbf{Ax} = \lambda\mathbf{x}$ and its conjugate transpose

$$(\overline{\mathbf{Ax}})^\mathsf{T} = (\bar{\lambda}\bar{\mathbf{x}})^\mathsf{T} = \bar{\lambda}\bar{\mathbf{x}}^\mathsf{T}$$

and multiply the two left sides and the two right sides,

$$(\overline{\mathbf{Ax}})^\mathsf{T}\mathbf{Ax} = \bar{\lambda}\lambda\bar{\mathbf{x}}^\mathsf{T}\mathbf{x} = |\lambda|^2\bar{\mathbf{x}}^\mathsf{T}\mathbf{x}.$$

But $\mathbf{A}$ is unitary, $\bar{\mathbf{A}}^{\mathsf{T}} = \mathbf{A}^{-1}$, so that on the left we obtain

$$(\overline{\mathbf{A}\mathbf{x}})^{\mathsf{T}}\mathbf{A}\mathbf{x} = \bar{\mathbf{x}}^{\mathsf{T}}\bar{\mathbf{A}}^{\mathsf{T}}\mathbf{A}\mathbf{x} = \bar{\mathbf{x}}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{A}\mathbf{x} = \bar{\mathbf{x}}^{\mathsf{T}}\mathbf{I}\,\mathbf{x} = \bar{\mathbf{x}}^{\mathsf{T}}\mathbf{x}.$$

Together, $\bar{\mathbf{x}}^{\mathsf{T}}\mathbf{x} = |\lambda|^2\bar{\mathbf{x}}^{\mathsf{T}}\mathbf{x}$. We now divide by $\bar{\mathbf{x}}^{\mathsf{T}}\mathbf{x}$ ($\neq 0$) to get $|\lambda|^2 = 1$. Hence $|\lambda| = 1$.

This proves Theorem 1 as well as Theorems 1 and 5 in Sec. 8.3. ∎

Key properties of orthogonal matrices (invariance of the inner product, orthonormality of rows and columns; see Sec. 8.3) generalize to unitary matrices in a remarkable way.

To see this, instead of R^n we now use the **complex vector space** C^n of all complex vectors with n complex numbers as components, and complex numbers as scalars. For such complex vectors the **inner product** is defined by (note the overbar for the complex conjugate)

$$\mathbf{a}\bullet\mathbf{b} = \bar{\mathbf{a}}^{\mathsf{T}}\mathbf{b}. \tag{4}$$

The **length** or **norm** of such a complex vector is a *real* number defined by

$$\|\mathbf{a}\| = \sqrt{\mathbf{a}\bullet\mathbf{a}} = \sqrt{\mathbf{a}^{-\mathsf{T}}\mathbf{a}} = \sqrt{\bar{a}_1a_1 + \cdots + \bar{a}_na_n} = \sqrt{|a_1|^2 + \cdots + |a_n|^2}. \tag{5}$$

THEOREM 2

Invariance of Inner Product

A **unitary transformation,** *that is,* $\mathbf{y} = \mathbf{A}\mathbf{x}$ *with a unitary matrix* $\mathbf{A}$, *preserves the value of the inner product* (4), *hence also the norm* (5).

PROOF The proof is the same as that of Theorem 2 in Sec. 8.3, which the theorem generalizes. In the analog of (9), Sec. 8.3, we now have bars,

$$\mathbf{u}\bullet\mathbf{v} = \bar{\mathbf{u}}^{\mathsf{T}}\mathbf{v} = (\overline{\mathbf{A}\mathbf{a}})^{\mathsf{T}}\mathbf{A}\mathbf{b} = \bar{\mathbf{a}}^{\mathsf{T}}\bar{\mathbf{A}}^{\mathsf{T}}\mathbf{A}\mathbf{b} = \bar{\mathbf{a}}^{\mathsf{T}}\mathbf{I}\,\mathbf{b} = \bar{\mathbf{a}}^{\mathsf{T}}\mathbf{b} = \mathbf{a}\bullet\mathbf{b}.$$ ∎

The complex analog of an orthonormal systems of real vectors (see Sec. 8.3) is defined as follows.

DEFINITION

Unitary System

A *unitary system* is a set of complex vectors satisfying the relationships

$$\mathbf{a}_j\bullet\mathbf{a}_k = \bar{\mathbf{a}}_j^{\mathsf{T}}\mathbf{a}_k = \begin{cases} 0 & \text{if } j \neq k \\ 1 & \text{if } j = k. \end{cases} \tag{6}$$

Theorem 3 in Sec. 8.3 extends to complex as follows.

THEOREM 3

Unitary Systems of Column and Row Vectors

A complex square matrix is unitary if and only if its column vectors (and also its row vectors) form a unitary system.

PROOF The proof is the same as that of Theorem 3 in Sec. 8.3, except for the bars required in $\bar{\mathbf{A}}^{\mathsf{T}} = \mathbf{A}^{-1}$ and in (4) and (6) of the present section. ∎

THEOREM 4

Determinant of a Unitary Matrix

Let $\mathbf{A}$ *be a unitary matrix. Then its determinant has absolute value one, that is,* $|\det \mathbf{A}| = 1$.

PROOF Similarly as in Sec. 8.3 we obtain

$$1 = \det(\mathbf{A}\mathbf{A}^{-1}) = \det(\mathbf{A}\bar{\mathbf{A}}^{\mathsf{T}}) = \det \mathbf{A} \det \bar{\mathbf{A}}^{\mathsf{T}} = \det \mathbf{A} \det \bar{\mathbf{A}}$$
$$= \det \mathbf{A} \overline{\det \mathbf{A}} = |\det \mathbf{A}|^2.$$

Hence $|\det \mathbf{A}| = 1$ (where $\det \mathbf{A}$ may now be complex). ∎

EXAMPLE 4 **Unitary Matrix Illustrating Theorems 1c and 2–4**

For the vectors $\mathbf{a}^{\mathsf{T}} = [2 \quad -i]$ and $\mathbf{b}^{\mathsf{T}} = [1 + i \quad 4i]$ we get $\bar{\mathbf{a}}^{\mathsf{T}} = [2 \quad i]^{\mathsf{T}}$ and $\bar{\mathbf{a}}^{\mathsf{T}}\mathbf{b} = 2(1 + i) - 4 = -2 + 2i$ and with

$$\mathbf{A} = \begin{bmatrix} 0.8i & 0.6 \\ 0.6 & 0.8i \end{bmatrix} \quad \text{also} \quad \mathbf{A}\mathbf{a} = \begin{bmatrix} i \\ 2 \end{bmatrix} \quad \text{and} \quad \mathbf{A}\mathbf{b} = \begin{bmatrix} -0.8 + 3.2i \\ -2.6 + 0.6i \end{bmatrix},$$

as one can readily verify. This gives $(\overline{\mathbf{A}}\overline{\mathbf{a}})^{\mathsf{T}}\mathbf{A}\mathbf{b} = -2 + 2i$, illustrating Theorem 2. The matrix is unitary. Its columns form a unitary system,

$$\bar{\mathbf{a}}_1^{\mathsf{T}}\mathbf{a}_1 = -0.8i \cdot 0.8i + 0.6^2 = 1, \qquad \bar{\mathbf{a}}_1^{\mathsf{T}}\mathbf{a}_2 = -0.8i \cdot 0.6 + 0.6 \cdot 0.8i = 0,$$
$$\bar{\mathbf{a}}_2^{\mathsf{T}}\mathbf{a}_2 = 0.6^2 + (-0.8i)0.8i = 1$$

and so do its rows. Also, $\det \mathbf{A} = -1$. The eigenvalues are $0.6 + 0.8i$ and $-0.6 + 0.8i$, with eigenvectors $[1 \quad 1]^{\mathsf{T}}$ and $[1 \quad -1]^{\mathsf{T}}$, respectively. ∎

Theorem 2 in Sec. 8.4 on the existence of an eigenbasis extends to complex matrices as follows.

THEOREM 5

Basis of Eigenvectors

A Hermitian, skew-Hermitian, or unitary matrix has a basis of eigenvectors for C^n *that is a unitary system.*

For a proof see Ref. [B3], vol. 1, pp. 270–272 and p. 244 (Definition 2).

EXAMPLE 5 **Unitary Eigenbases**

The matrices **A, B, C** in Example 2 have the following unitary systems of eigenvectors, as you should verify.

A: $\frac{1}{\sqrt{35}}[1 - 3i \quad 5]^{\mathsf{T}} \quad (\lambda = 9), \qquad \frac{1}{\sqrt{14}}[1 - 3i \quad -2]^{\mathsf{T}} \quad (\lambda = 2)$

B: $\frac{1}{\sqrt{30}}[1 - 2i \quad -5]^{\mathsf{T}} \quad (\lambda = -2i), \qquad \frac{1}{\sqrt{30}}[5 \quad 1 + 2i]^{\mathsf{T}} \quad (\lambda = 4i)$

C: $\frac{1}{\sqrt{2}}[1 \quad 1]^{\mathsf{T}} \quad (\lambda = \frac{1}{2}(i + \sqrt{3})), \qquad \frac{1}{\sqrt{2}}[1 \quad -1]^{\mathsf{T}} \quad (\lambda = \frac{1}{2}(i - \sqrt{3}))$. ∎

Hermitian and Skew-Hermitian Forms

The concept of a quadratic form (Sec. 8.4) can be extended to complex. We call the numerator $\bar{\mathbf{x}}^T\mathbf{A}\mathbf{x}$ in (1) a **form** in the components $x_1, \cdots, x_n$ of $\mathbf{x}$, which may now be complex. This form is again a sum of n^2 terms

$$\bar{\mathbf{x}}^T\mathbf{A}\mathbf{x} = \sum_{j=1}^{n}\sum_{k=1}^{n} a_{jk}\bar{x}_j x_k$$

$$= a_{11}\bar{x}_1x_1 + \cdots + a_{1n}\bar{x}_1x_n$$

$$+ a_{21}\bar{x}_2x_1 + \cdots + a_{2n}\bar{x}_2x_n \tag{7}$$

$$+ \cdots\cdots\cdots\cdots\cdots$$

$$+ a_{n1}\bar{x}_nx_1 + \cdots + a_{nn}\bar{x}_nx_n.$$

$\mathbf{A}$ is called its **coefficient matrix.** The form is called a **Hermitian** or **skew-Hermitian form** if $\mathbf{A}$ is Hermitian or skew-Hermitian, respectively. *The value of a Hermitian form is real, and that of a skew-Hermitian form is pure imaginary or zero.* This can be seen directly from (2) and (3) and accounts for the importance of these forms in physics. Note that (2) and (3) are valid for any vectors because in the proof of (2) and (3) we did not use that $\mathbf{x}$ is an eigenvector but only that $\bar{\mathbf{x}}^T\mathbf{x}$ is real and not 0.

EXAMPLE 6 Hermitian Form

For $\mathbf{A}$ in Example 2 and, say, $\mathbf{x} = [1 + i \quad 5i]^T$ we get

$$\bar{\mathbf{x}}^T\mathbf{A}\mathbf{x} = [1 - i \quad -5i]\begin{bmatrix} 4 & 1 - 3i \\ 1 + 3i & 7 \end{bmatrix}\begin{bmatrix} 1 + i \\ 5i \end{bmatrix} = [1 - i \quad -5i]\begin{bmatrix} 4(1 + i) + (1 - 3i)\cdot 5i \\ (1 + 3i)(1 + i) + 7\cdot 5i \end{bmatrix} = 223. \quad \blacksquare$$

Clearly, if $\mathbf{A}$ and $\mathbf{x}$ in (4) are real, then (7) reduces to a quadratic form, as discussed in the last section.

PROBLEM SET 8.5

1. **(Verification)** Verify the statements in Examples 2 and 3.
2. **(Product)** Show $(\overline{\mathbf{BA}})^T = -\mathbf{AB}$ for $\mathbf{A}$ and $\mathbf{B}$ in Example 2. For any $n \times n$ Hermitian $\mathbf{A}$ and skew-Hermitian $\mathbf{B}$.
3. Show that $(\overline{\mathbf{ABC}})^T = -\mathbf{C}^{-1}\mathbf{BA}$ for any $n \times n$ Hermitian $\mathbf{A}$, skew-Hermitian $\mathbf{B}$, and unitary $\mathbf{C}$.
4. **(Eigenvectors)** Find eigenvectors of $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ in Examples 2 and 3.

5–11 EIGENVALUES AND EIGENVECTORS

Are the matrices in Probs. 5–11 Hermitian? Skew-Hermitian? Unitary? Find their eigenvalues (thereby verifying Theorem 1) and eigenvectors.

5. $\begin{bmatrix} 4 & i \\ -i & 2 \end{bmatrix}$

6. $\begin{bmatrix} 0 & 2i \\ 2i & 0 \end{bmatrix}$

7. $\begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{i}{\sqrt{2}} \\ -\frac{i}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}$

8. $\begin{bmatrix} 0 & i \\ i & 0 \end{bmatrix}$

9. $\begin{bmatrix} 5i & 0 & 0 \\ 0 & 0 & 5i \\ 0 & 5i & 0 \end{bmatrix}$

10. $\begin{bmatrix} 0 & 1 + i & 0 \\ 1 - i & 0 & 1 + i \\ 0 & 1 - i & 0 \end{bmatrix}$

11. $\begin{bmatrix} 0 & 0 & i \\ 0 & i & 0 \\ i & 0 & 0 \end{bmatrix}$

12. PROJECT. Complex Matrices

(a) Decomposition. Show that any square matrix may be written as the sum of a Hermitian and a skew-Hermitian matrix. Give examples.

(b) Normal matrix. This important concept denotes a matrix that commutes with its conjugate transpose, $\mathbf{A}\overline{\mathbf{A}}^{\mathsf{T}} = \overline{\mathbf{A}}^{\mathsf{T}}\mathbf{A}$. Prove that Hermitian, skew-Hermitian, and unitary matrices are normal. Give corresponding examples of your own.

(c) Normality criterion. Prove that $\mathbf{A}$ is normal if and only if the Hermitian and skew-Hermitian matrices in (a) commute.

(d) Find a simple matrix that is not normal. Find a normal matrix that is not Hermitian, skew-Hermitian, or unitary.

(e) Unitary matrices. Prove that the product of two unitary $n \times n$ matrices and the inverse of a unitary matrix are unitary. Give examples.

(f) Powers of unitary matrices in applications may sometimes be very simple. Show that $\mathbf{C}^{12} = \mathbf{I}$ in Example 2. Find further examples.

13–15 COMPLEX FORMS

Is the given matrix (call it $\mathbf{A}$) Hermitian or skew-Hermitian? Find $\overline{\mathbf{x}}^{\mathsf{T}}\mathbf{A}\mathbf{x}$. (Show all the details.) a, b, c, k are real.

13. $\begin{bmatrix} 0 & -3i \\ -3i & 0 \end{bmatrix}$, $\mathbf{x} = \begin{bmatrix} 4+i \\ 3-i \end{bmatrix}$

14. $\begin{bmatrix} a & b+ic \\ b-ic & k \end{bmatrix}$, $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

15. $\begin{bmatrix} 2 & 1+i \\ 1-i & 1 \end{bmatrix}$, $\mathbf{x} = \begin{bmatrix} i \\ 2i \end{bmatrix}$

16. (Pauli spin matrices) Find the eigenvalues and eigenvectors of the so-called *Pauli spin matrices* and show that $\mathbf{S}_x\mathbf{S}_y = i\mathbf{S}_z$, $\mathbf{S}_y\mathbf{S}_x = -i\mathbf{S}_z$, $\mathbf{S}_x^2 = \mathbf{S}_y^2 = \mathbf{S}_z^2 = \mathbf{I}$, where

$$\mathbf{S}_x = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \mathbf{S}_y = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix},$$

$$\mathbf{S}_z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

CHAPTER 8 REVIEW QUESTIONS AND PROBLEMS

1. In solving an eigenvalue problem, what is given and what is sought?
2. Do there exist square matrices without eigenvalues? Eigenvectors corresponding to more than one eigenvalue of a given matrix?
3. What is the defect? Why is it important? Give examples.
4. Can a complex matrix have real eigenvalues? Real eigenvectors? Give reasons.
5. What is diagonalization of a matrix? Transformation of a form to principal axes?
6. What is an eigenbasis? When does it exist? Why is it important?
7. Does a 3×3 matrix always have a real eigenvalue?
8. Give a few typical applications in which eigenvalue problems occur.

9–14 DIAGONALIZATION

Find an eigenbasis and diagonalize. (Show the details.)

9. $\begin{bmatrix} 101 & 72 \\ -144 & -103 \end{bmatrix}$

10. $\begin{bmatrix} 14.4 & -11.2 \\ -11.2 & 102.6 \end{bmatrix}$

11. $\begin{bmatrix} -14 & 10 \\ -10 & 11 \end{bmatrix}$

12. $\begin{bmatrix} 15 & 4 & -4 \\ 6 & 10 & 8 \\ -12 & -2 & -7 \end{bmatrix}$, $\lambda = 18$

13. $\begin{bmatrix} 5 & \frac{8}{3} & -\frac{2}{3} \\ 2 & \frac{2}{3} & \frac{4}{3} \\ -4 & -\frac{4}{3} & -\frac{8}{3} \end{bmatrix}$

14. $\begin{bmatrix} -6 & 11 & 3 \\ 4 & 1 & 3 \\ -4 & 10 & 8 \end{bmatrix}$, $\lambda = 2$

15–17 SIMILARITY

Verify that $\mathbf{A}$ and $\hat{\mathbf{A}} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ have the same spectrum. Here, $\mathbf{A}$, $\mathbf{P}$ are:

15. $\begin{bmatrix} 3.8 & 2.4 \\ 2.4 & 0.2 \end{bmatrix}, \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$

16. $\begin{bmatrix} -22 & 20 & 10 \\ -4 & 20 & -8 \\ 28 & -14 & 29 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 2 \\ 0 & 2 & 4 \\ 2 & 8 & 0 \end{bmatrix}$

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

Transformation to Canonical Form. Reduce the quadratic form to principal axes.

18. $11.56x_1^2 + 20.16x_1x_2 + 17.44x_2^2 = 100$

19. $1.09x_1^2 - 0.06x_1x_2 + 1.01x_2^2 = 1$

20. $14x_1^2 + 24x_1x_2 - 4x_2^2 = 20$

SUMMARY OF CHAPTER 8

Linear Algebra: Matrix Eigenvalue Problems

The practical importance of matrix eigenvalue problems can hardly be overrated. The problems are defined by the vector equation

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}. \tag{1}$$

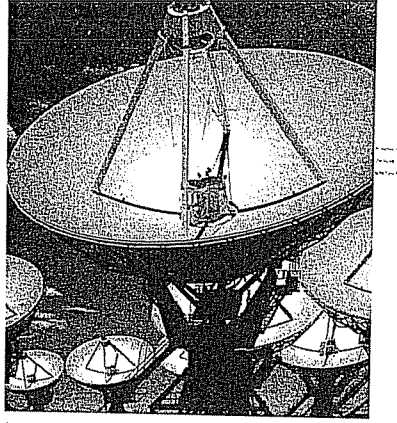
$\mathbf{A}$ is a given square matrix. All matrices in this chapter are ***square.*** λ is a scalar. To *solve* the problem (1) means to determine values of λ, called **eigenvalues** (or **characteristic values**) of $\mathbf{A}$, such that (1) has a nontrivial solution $\mathbf{x}$ (that is, $\mathbf{x} \neq \mathbf{0}$), called an **eigenvector** of $\mathbf{A}$ corresponding to that λ. An $n \times n$ matrix has at least one and at most n numerically different eigenvalues. These are the solutions of the **characteristic equation** (Sec. 8.1)

$$D(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} a_{11} - \lambda & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} - \lambda & \cdots & a_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} - \lambda \end{vmatrix} = 0. \tag{2}$$

$D(\lambda)$ is called the **characteristic determinant** of $\mathbf{A}$. By expanding it we get the **characteristic polynomial** of $\mathbf{A}$, which is of degree n in λ. Some typical applications are shown in Sec. 8.2.

Section 8.3 is devoted to eigenvalue problems for **symmetric** ($\mathbf{A}^\mathsf{T} = \mathbf{A}$), **skew-symmetric** ($\mathbf{A}^\mathsf{T} = -\mathbf{A}$), and **orthogonal matrices** ($\mathbf{A}^\mathsf{T} = \mathbf{A}^{-1}$). Section 8.4 concerns the diagonalization of matrices and the transformation of quadratic forms to principal axes and its relation to eigenvalues.

Section 8.5 extends Sec. 8.3 to the complex analogs of those real matrices, called **Hermitian** ($\bar{\mathbf{A}}^\mathsf{T} = \mathbf{A}$), **skew-Hermitian** ($\bar{\mathbf{A}}^\mathsf{T} = -\mathbf{A}$), and **unitary matrices** ($\bar{\mathbf{A}}^\mathsf{T} = \mathbf{A}^{-1}$). All the eigenvalues of a Hermitian matrix (and a symmetric one) are real. For a skew-Hermitian (and a skew-symmetric) matrix they are pure imaginary or zero. For a unitary (and an orthogonal) matrix they have absolute value 1.



CHAPTER 9

Vector Differential Calculus. Grad, Div, Curl

This chapter deals with vectors and vector functions in **3-space,** the space of three dimensions with the usual measurement of distance (given by the Pythagorean theorem). This includes **2-space** (the plane) as a special case. It extends the differential calculus to those vector functions and the vector fields they represent. Forces, velocities, and various other quantities are vectors. This makes the algebra, geometry, and calculus of these vector functions the natural instrument for the engineer and physicist in solid mechanics, fluid flow, heat flow, electrostatics, and so on. The engineer must understand these vector functions and fields as the basis of the design and construction of systems, such as airplanes, laser generators, and robots.

In Secs. 9.1–9.3 we explain the basic *algebraic* operations with vectors in 3-space. *Calculus* begins in Sec. 9.4 with the extension of differentiation to vector functions in a simple and natural fashion. Application to curves and their use in mechanics follows in Sec. 9.5.

We finally discuss three physically important concepts related to scalar and vector fields, namely, the gradient (Sec. 9.7), divergence (Sec. 9.8), and curl (Sec. 9.9). (The use of these concepts in integral theorems follows in the next chapter. Their form in **curvilinear coordinates** is given in App. A3.4.)

We shall keep this chapter ***independent of Chaps. 7 and 8.*** Our present approach is in harmony with Chap. 7, with the restriction to two and three dimensions providing for a richer theory with basic physical, engineering, and geometric applications.

Prerequisite: Elementary use of second- and third-order determinants in Sec. 9.3.
Sections that may be omitted in a shorter course: 9.5, 9.6.
References and Answers to Problems: App. 1 Part B, App. 2.

9.1 Vectors in 2-Space and 3-Space

In physics and geometry and its engineering applications we use two kinds of quantities: scalars and vectors. A **scalar** is a quantity that is determined by its magnitude; this is the number of units measured on a suitable scale. For instance, length, voltage, and temperature are scalars.

A **vector** is a quantity that is determined by both its magnitude and its direction. Thus it is an **arrow** or **directed line segment.** For instance, a force is a vector, and so is a velocity, giving the speed and direction of motion (Fig. 162).

We denote vectors by lowercase boldface letters **a, b, v,** etc. In handwriting you may use arrows, for instance $\vec{a}$ (in place of **a**), $\vec{b}$, etc.

A vector (arrow) has a tail, called its **initial point,** and a tip, called its **terminal point.** This is motivated in the **translation** (displacement without rotation) of the triangle in Fig. 163, where the initial point P of the vector **a** is the original position of a point, and the terminal point Q is the terminal position of that point, its position *after* the translation. The length of the arrow equals the distance between P and Q. This is called the **length** (or *magnitude*) of the vector **a** and is denoted by $|\mathbf{a}|$. Another name for *length* is **norm** (or *Euclidean norm*).

A vector of length 1 is called a **unit vector.**

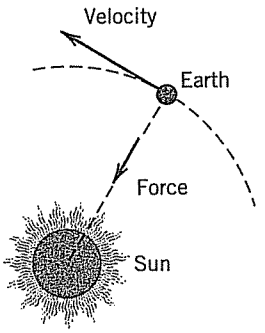


Fig. 162. Force and velocity

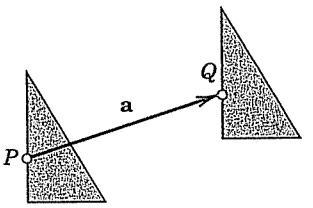


Fig. 163. Translation

Of course, we would like to calculate with vectors. For instance, we want to find the resultant of forces or compare parallel forces of different magnitude. This motivates our next ideas: to define *components* of a vector, and then the two basic algebraic operations of *vector addition* and *scalar multiplication.*

For this we must first define *equality of vectors* in a way that is practical in connection with forces and other applications.

DEFINITION

Equality of Vectors

Two vectors **a** and **b** are equal, written $\mathbf{a} = \mathbf{b}$, if they have the same length and the same direction [as explained in Fig. 164; in particular, note (B)]. Hence a vector can be arbitrarily translated; that is, its initial point can be chosen arbitrarily.

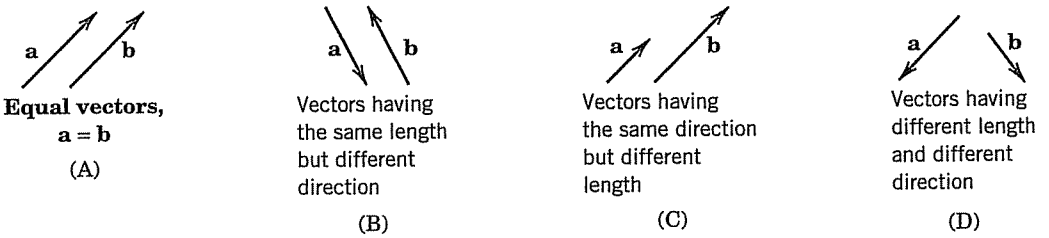


Fig. 164. (A) Equal vectors. (B)–(D) Different vectors

Components of a Vector

We choose an xyz **Cartesian coordinate system**[1] in space (Fig. 165), that is, a usual rectangular coordinate system with the same scale of measurement on the three mutually perpendicular coordinate axes. Let **a** be a given vector with initial point P: (x_1, y_1, z_1) and terminal point Q: (x_2, y_2, z_2). Then the three coordinate differences

$$(1) \qquad a_1 = x_2 - x_1, \qquad a_2 = y_2 - y_1, \qquad a_3 = z_2 - z_1$$

are called the **components** of the vector **a** with respect to that coordinate system, and we write simply $\mathbf{a} = [a_1, a_2, a_3]$. See Fig. 166.

The **length** $|\mathbf{a}|$ of **a** can now readily be expressed in terms of components because from (1) and the Pythagorean theorem we have

$$(2) \qquad |\mathbf{a}| = \sqrt{a_1^2 + a_2^2 + a_3^2}.$$

EXAMPLE 1 **Components and Length of a Vector**

The vector **a** with initial point P: (4, 0, 2) and terminal point Q: (6, −1, 2) has the components

$$a_1 = 6 - 4 = 2, \qquad a_2 = -1 - 0 = -1, \qquad a_3 = 2 - 2 = 0.$$

Hence $\mathbf{a} = [2, -1, 0]$. (Can you sketch **a**, as in Fig. 166?) Equation (2) gives the length

$$|\mathbf{a}| = \sqrt{2^2 + (-1)^2 + 0^2} = \sqrt{5}.$$

If we choose (−1, 5, 8) as the initial point of **a**, the corresponding terminal point is (1, 4, 8).

If we choose the origin (0, 0, 0) as the initial point of **a**, the corresponding terminal point is (2, −1, 0); its coordinates equal the components of **a**. This suggests that we can determine each point in space by a vector, called the *position vector* of the point, as follows. ■

A Cartesian coordinate system being given, the **position vector r** of a point A: (x, y, z) is the vector with the origin (0, 0, 0) as the initial point and A as the terminal point (see Fig. 167). Thus in components, $\mathbf{r} = [x, y, z]$. This can be seen directly from (1) with $x_1 = y_1 = z_1 = 0$.

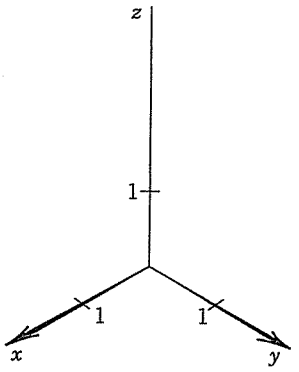


Fig. 165. Cartesian coordinate system

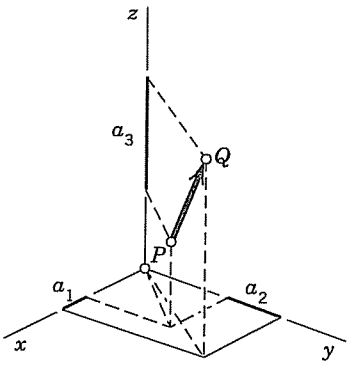


Fig. 166. Components of a vector

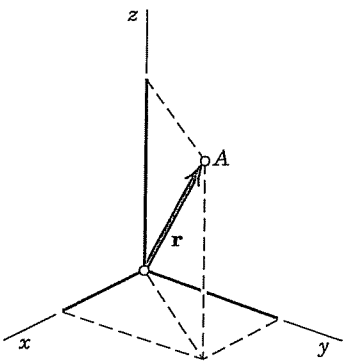


Fig. 167. Position vector **r** of a point A: (x, y, z)

[1]Named after the French philosopher and mathematician RENATUS CARTESIUS, latinized for RENÉ DESCARTES (1596–1650), who invented analytic geometry. His basic work *Géométrie* appeared in 1637, as an appendix to his *Discours de la méthode.*

Furthermore, if we translate a vector **a**, with initial point P and terminal point Q, then corresponding coordinates of P and Q change by the same amount, so that the differences in (1) remain unchanged. This proves

THEOREM 1

Vectors as Ordered Triples of Real Numbers

A fixed Cartesian coordinate system being given, each vector is uniquely determined by its ordered triple of corresponding components. Conversely, to each ordered triple of real numbers (a_1, a_2, a_3) *there corresponds precisely one vector* $\mathbf{a} = [a_1, a_2, a_3]$, *with* $(0, 0, 0)$ *corresponding to the* **zero vector 0,** *which has length* 0 *and no direction.*

Hence a vector equation $\mathbf{a} = \mathbf{b}$ *is equivalent to the three equations* $a_1 = b_1$, $a_2 = b_2$, $a_3 = b_3$ *for the components.*

We now see that from our "geometric" definition of a vector as an arrow we have arrived at an "algebraic" characterization of a vector by Theorem 1. We could have started from the latter and reversed our process. This shows that the two approaches are equivalent.

Vector Addition, Scalar Multiplication

Applications suggest calculation with vectors that are practically useful and are almost as simple as the arithmetics for real numbers. The first is addition and the second is multiplication by a number.

DEFINITION

Addition of Vectors

The **sum** $\mathbf{a} + \mathbf{b}$ of two vectors $\mathbf{a} = [a_1, a_2, a_3]$ and $\mathbf{b} = [b_1, b_2, b_3]$ is obtained by adding the corresponding components,

$$\mathbf{a} + \mathbf{b} = [a_1 + b_1, \quad a_2 + b_2, \quad a_3 + b_3]. \tag{3}$$

Geometrically, place the vectors as in Fig. 168 (the initial point of **b** at the terminal point of **a**); then $\mathbf{a} + \mathbf{b}$ is the vector drawn from the initial point of **a** to the terminal point of **b.**

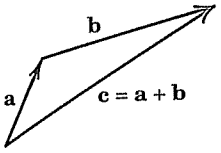


Fig. 168. Vector addition

For forces, this addition is the parallelogram law by which we obtain the **resultant** of two forces in mechanics. See Fig. 169.

Figure 170 shows (for the plane) that the "algebraic" way and the "geometric way" of vector addition give the same vector.

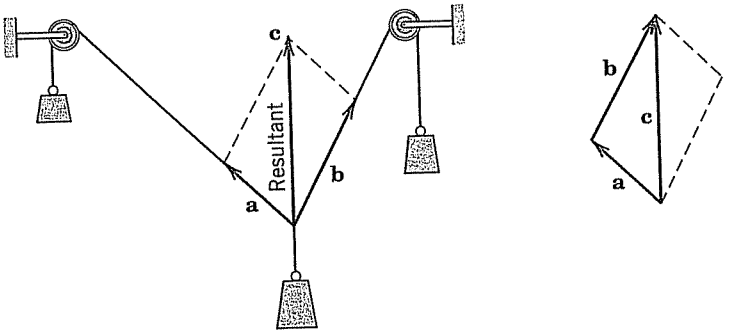


Fig. 169. Resultant of two forces **(parallelogram law)**

Basic Properties of Vector Addition. Familiar laws for real numbers give immediately (see also Figs. 171 and 172)

$$
\begin{aligned}
&\text{(a)} && \mathbf{a} + \mathbf{b} = \mathbf{b} + \mathbf{a} && \textit{(Commutativity)}\\
&\text{(b)} && (\mathbf{u} + \mathbf{v}) + \mathbf{w} = \mathbf{u} + (\mathbf{v} + \mathbf{w}) && \textit{(Associativity)}\\
&\text{(c)} && \mathbf{a} + \mathbf{0} = \mathbf{0} + \mathbf{a} = \mathbf{a}\\
&\text{(d)} && \mathbf{a} + (-\mathbf{a}) = \mathbf{0}.
\end{aligned}
\tag{4}
$$

Here $-\mathbf{a}$ denotes the vector having the length $|\mathbf{a}|$ and the direction opposite to that of $\mathbf{a}$.

In (4b) we may simply write $\mathbf{u} + \mathbf{v} + \mathbf{w}$, and similarly for sums of more than three vectors. Instead of $\mathbf{a} + \mathbf{a}$ we also write $2\mathbf{a}$, and so on. This (and the notation $-\mathbf{a}$ used just before) motivates defining the second algebraic operation for vectors as follows.

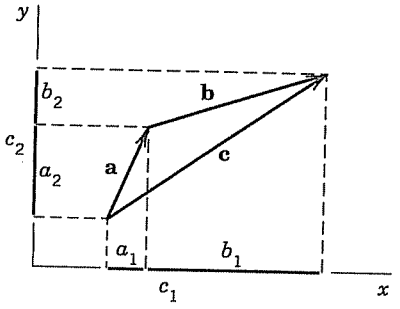


Fig. 170. Vector addition

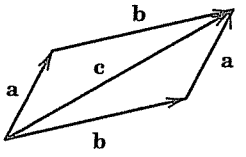


Fig. 171. Cummutativity of vector addition

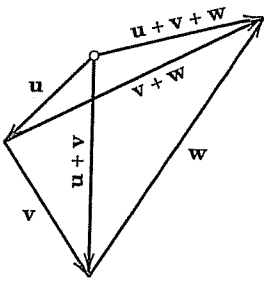


Fig. 172. Associativity of vector addition

DEFINITION

Scalar Multiplication (Multiplication by a Number)

The product $c\mathbf{a}$ of any vector $\mathbf{a} = [a_1, a_2, a_3]$ and any scalar c (real number c) is the vector obtained by multiplying each component of $\mathbf{a}$ by c,

$$c\mathbf{a} = [ca_1, ca_2, ca_3]. \tag{5}$$

Geometrically, if $\mathbf{a} \neq \mathbf{0}$, then $c\mathbf{a}$ with $c > 0$ has the direction of $\mathbf{a}$ and with $c < 0$ the direction opposite to $\mathbf{a}$. In any case, the length of $c\mathbf{a}$ is $|c\mathbf{a}| = |c||\mathbf{a}|$, and $c\mathbf{a} = \mathbf{0}$ if $\mathbf{a} = \mathbf{0}$ or $c = 0$ (or both). (See Fig. 173.)

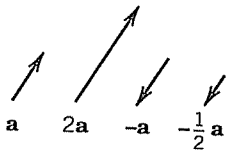


Fig. 173. Scalar multiplication [multiplication of vectors by scalars (numbers)]

Basic Properties of Scalar Multiplication. From the definitions we obtain directly

$$
\begin{aligned}
&\text{(a)} && c(\mathbf{a} + \mathbf{b}) = c\mathbf{a} + c\mathbf{b}\\
&\text{(b)} && (c + k)\mathbf{a} = c\mathbf{a} + k\mathbf{a}\\
&\text{(c)} && c(k\mathbf{a}) = (ck)\mathbf{a} && \text{(written } ck\mathbf{a}\text{)}\\
&\text{(d)} && 1\mathbf{a} = \mathbf{a}.
\end{aligned}
\tag{6}
$$

You may prove that (4) and (6) imply for any vector $\mathbf{a}$

$$(7)\quad \text{(a)}\quad 0\mathbf{a} = \mathbf{0} \qquad \text{(b)}\quad (-1)\mathbf{a} = -\mathbf{a}.$$

Instead of $\mathbf{b} + (-\mathbf{a})$ we simply write $\mathbf{b} - \mathbf{a}$ (Fig. 174).

EXAMPLE 2 **Vector Addition. Multiplication by Scalars**

With respect to a given coordinate system, let

$$\mathbf{a} = [4, 0, 1] \quad \text{and} \quad \mathbf{b} = [2, -5, \tfrac{1}{3}].$$

Then $-\mathbf{a} = [-4, 0, -1]$, $7\mathbf{a} = [28, 0, 7]$, $\mathbf{a} + \mathbf{b} = [6, -5, \frac{4}{3}]$, and

$$2(\mathbf{a} - \mathbf{b}) = 2[2, 5, \tfrac{2}{3}] = [4, 10, \tfrac{4}{3}] = 2\mathbf{a} - 2\mathbf{b}.$$

Unit Vectors i, j, k. Besides $\mathbf{a} = [a_1, a_2, a_3]$ another popular way of writing vectors is

$$(8)\quad \mathbf{a} = a_1\mathbf{i} + a_2\mathbf{j} + a_3\mathbf{k}$$

In this representation, **i, j, k** are the unit vectors in the positive directions of the axes of a Cartesian coordinate system (Fig. 175). Hence, in components,

$$(9)\quad \mathbf{i} = [1, 0, 0], \qquad \mathbf{j} = [0, 1, 0], \qquad \mathbf{k} = [0, 0, 1]$$

and the right side of (8) is a sum of three vectors parallel to the three axes.

EXAMPLE 3 **ijk Notation for Vectors**

In Example 2 we have $\mathbf{a} = 4\mathbf{i} + \mathbf{k}$, $\mathbf{b} = 2\mathbf{i} - 5\mathbf{j} + \frac{1}{3}\mathbf{k}$, and so on.

All the vectors $\mathbf{a} = [a_1, a_2, a_3] = a_1\mathbf{i} + a_2\mathbf{j} + a_3\mathbf{k}$ (with real numbers as components) form the **real vector space** R^3 with the two *algebraic operations* of vector addition and scalar multiplication as just defined. R^3 has **dimension** 3. The triple of vectors **i, j, k** is called a **standard basis** of R^3. A Cartesian coordinate system being given, the representation (8) of a given vector is unique.

Vector space R^3 is a model of a general vector space, as discussed in Sec. 7.9, but is not needed in this chapter.

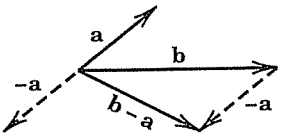


Fig. 174. Difference of vectors

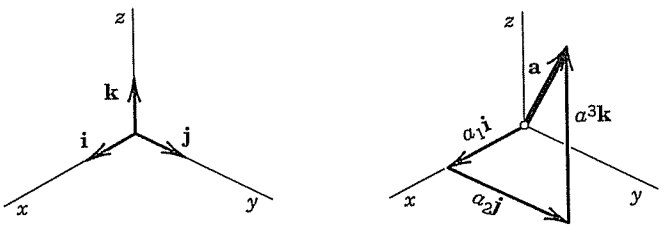


Fig. 175. The unit vectors **i, j, k** and the representation (8)

PROBLEM SET 9.1

1–6 COMPONENTS AND LENGTH

Find the components of the vector $\mathbf{v}$ with given initial point P and terminal point Q. Find $|\mathbf{v}|$. Sketch $|\mathbf{v}|$. Find the **unit vector** in the direction of $\mathbf{v}$.

1. P: (3, 2, 0), Q: (5, −2, 0)
2. P: (1, 1, 1), Q: (−4, −4, −4)
3. P: (1, 0, 1.2), Q: (0, 0, 6.2)
4. P: (2, −2, 0), Q: (0, 4, 6)
5. P: (4, 3, 2), Q: (−4, −3, 2)
6. P: (0, 0, 0), Q: (6, 8, 10)

7–12 Given the components v_1, v_2, v_3 of a vector $\mathbf{v}$ and a particular initial point P, find the corresponding terminal point Q and the length of $\mathbf{v}$.

7. 3, −1, 0; P: (4, 6, 0)
8. 8, 4, −2; P: (−8, −4, 2)
9. $\frac{1}{4}$, 2, $\frac{3}{4}$; P: $(0, -\frac{1}{2}, \frac{3}{2})$;
10. 3, 2, 6; P: (0, 0, 0)
11. 4, $\frac{1}{2}$, $-\frac{3}{2}$; P: $(-4, \frac{1}{2}, 2)$
12. 3, −3, 3; P: (1, 3, −3)

13–20 VECTOR ADDITION AND SCALAR MULTIPLICATION

Let $\mathbf{a} = [2, -1, 0] = 2\mathbf{i} - \mathbf{j}$, $\mathbf{b} = [-4, 2, 5] = -4\mathbf{i} + 2\mathbf{j} + 5\mathbf{k}$, $\mathbf{c} = [0, 0, 3] = 3\mathbf{k}$. Find:

13. $2\mathbf{a}$, $-\mathbf{a}$, $-\frac{1}{2}\mathbf{a}$
14. $\mathbf{a} + 2\mathbf{b}$, $2\mathbf{b} + \mathbf{a}$
15. $5(\mathbf{a} - \mathbf{c})$, $5\mathbf{a} - 5\mathbf{c}$
16. $(3\mathbf{a} - 5\mathbf{b}) + 2\mathbf{c}$, $3\mathbf{a} + (-5\mathbf{b} + 2\mathbf{c})$
17. $6\mathbf{a} - 4\mathbf{b} + 2\mathbf{c}$, $2(3\mathbf{a} - 2\mathbf{b} + \mathbf{c})$
18. $(1/|\mathbf{a}|)\mathbf{a}$, $(1/|\mathbf{c}|)\mathbf{c}$
19. $\mathbf{a} + \mathbf{b} + \mathbf{c}$, $-3\mathbf{a} - 3\mathbf{b} - 3\mathbf{c}$
20. $|\mathbf{a} + \mathbf{b}|$, $|\mathbf{a}| + |\mathbf{b}|$

21. What laws do Probs. 14–17 illustrate?
22. Prove (4) and (6).
23. Find the midpoint of the segment PQ in Probs. 7 and 9.

24–28 FORCES

Find the resultant (in components) and its magnitude.

24. $\mathbf{p} = [1, 2, 0]$, $\mathbf{q} = [0, 4, -1]$, $\mathbf{u} = [4, 0, -3]$, $\mathbf{v} = [6, 2, 4]$
25. $\mathbf{p} = [2, 2, 2]$, $\mathbf{q} = [-4, -4, 0]$, $\mathbf{u} = [2, 2, 7]$
26. $\mathbf{p} = [-1, -3, -5]$, $\mathbf{q} = [6, 4, 2]$, $\mathbf{u} = [-5, -1, 3]$
27. $\mathbf{p} = [8, 2, -4]$, $\mathbf{q} = 3\mathbf{p}$, $\mathbf{u} = -5\mathbf{p}$
28. $\mathbf{p} = [3, 0, -2]$, $\mathbf{q} = [2, 5, 1]$, $\mathbf{u} = 4\mathbf{q}$

29. Find $\mathbf{v}$ so that $\mathbf{v}$, $\mathbf{p}$, $\mathbf{q}$, $\mathbf{u}$ in Prob. 25 are in equilibrium.
30. For what c is the resultant of [3, 1, 7], [4, 4, 5], and [3, 2, c] parallel to the xy-plane?
31. Find forces $\mathbf{p}$, $\mathbf{q}$, $\mathbf{u}$ in the direction of the coordinate axes such that $\mathbf{p}$, $\mathbf{q}$, $\mathbf{u}$, $\mathbf{v} = [2, 3, 0]$, $\mathbf{w} = [7, -1, 11]$ are in equilibrium. Are $\mathbf{p}$, $\mathbf{q}$, $\mathbf{u}$ uniquely determined?
32. If $|\mathbf{p}| = 1$ and $|\mathbf{q}| = 2$, what can be said about the magnitude and direction of the resultant? Can you think of an application where this matters?
33. Same question as in Prob. 32 if $|\mathbf{p}| = 3$, $|\mathbf{q}| = 2$, $|\mathbf{u}| = 1$.
34. **(Relative velocity)** If airplanes A and B are moving southwest with speed $|\mathbf{v}_A| = 500$ mph and northwest with speed $|\mathbf{v}_B| = 400$ mph, respectively, what is the relative velocity $\mathbf{v} = \mathbf{v}_B - \mathbf{v}_A$ of B with respect to A?
35. **(Relative velocity)** Same question as in Prob. 34 for two ships moving northwest with speed $|\mathbf{v}_A| = 20$ knots and northeast with speed $|\mathbf{v}_B| = 25$ knots.
36. **(Reflection)** If a ray of light is reflected once in each of two mutually perpendicular mirrors, what can you say about the reflected ray?
37. **(Rope)** Find the magnitude of the force in each rope in the figure for any weight $\mathbf{w}$ and angle α.
38. **TEAM PROJECT. Geometric Applications.** To increase your skill in dealing with vectors, use vectors to prove the following (see the figures).

(a) The diagonals of a parallelogram bisect each other.

(b) The line through the midpoints of adjacent sides of a parallelogram bisects one of the diagonals in the ratio 1:3.

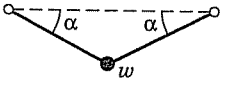
(c) Obtain (b) from (a).

(d) The three medians of a triangle (the segments from a vertex to the midpoint of the opposite side) meet at a single point, which divides the medians in the ratio 2:1.

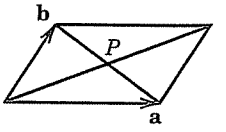
(e) The quadrilateral whose vertices are the midpoints of the sides of an arbitrary quadrilateral is a parallelogram.

(f) The four space diagonals of a parallelepiped meet and bisect each other.

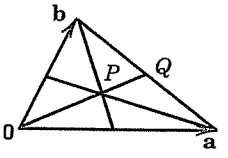
(g) The sum of the vectors drawn from the center of a regular polygon to its vertices is the zero vector.



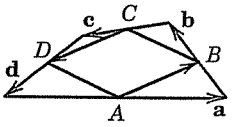
Problem 37



Team Project 38(a)



Team Project 38(d)



Team Project 38(e)

9.2 Inner Product (Dot Product)

We shall now define a multiplication of two vectors that gives a scalar as the product and is suggested by various applications, in particular when angles between vectors and lengths of vectors are involved.

DEFINITION

Inner Product (Dot Product) of Vectors

The **inner product** or **dot product** $\mathbf{a} \cdot \mathbf{b}$ (read "**a** dot **b**") of two vectors **a** and **b** is the product of their lengths times the cosine of their angle (see Fig. 176),

$$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}||\mathbf{b}| \cos \gamma \quad \text{if} \quad \mathbf{a} \neq \mathbf{0}, \mathbf{b} \neq \mathbf{0} \tag{1}$$

$$\mathbf{a} \cdot \mathbf{b} = 0 \quad \text{if} \quad \mathbf{a} = \mathbf{0} \text{ or } \mathbf{b} = \mathbf{0}.$$

The angle γ, $0 \leqq \gamma \leqq \pi$, between **a** and **b** is measured when the initial points of the vectors coincide, as in Fig. 176. In components, $\mathbf{a} = [a_1, a_2, a_3]$, $\mathbf{b} = [b_1, b_2, b_3]$, and

$$\mathbf{a} \cdot \mathbf{b} = a_1 b_1 + a_2 b_2 + a_3 b_3. \tag{2}$$

The second line in (1) is needed because γ is undefined when $\mathbf{a} = \mathbf{0}$ or $\mathbf{b} = \mathbf{0}$. The derivation of (2) from (1) is shown below.

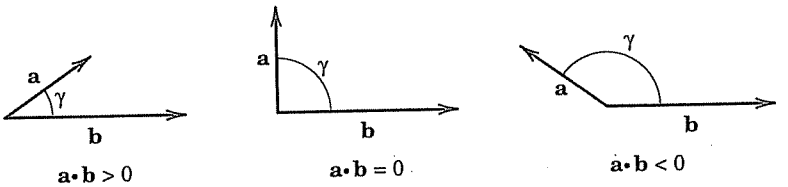


Fig. 176. Angle between vectors and value of inner product

Orthogonality. Since the cosine in (1) may be positive, 0, or negative, so may be the inner product (Fig. 176). The case that the inner product is zero is of particular practical interest and suggests the following concept.

A vector **a** is called **orthogonal** to a vector **b** if $\mathbf{a} \cdot \mathbf{b} = 0$. Then **b** is also orthogonal to **a**, and we call **a** and **b** **orthogonal vectors.** Clearly, this happens for nonzero vectors if and only if $\cos \gamma = 0$; thus $\gamma = \pi/2$ (90°). This proves the important

THEOREM 1

Orthogonality

The inner product of two nonzero vectors is 0 *if and only if these vectors are perpendicular.*

Length and Angle. Equation (1) with $\mathbf{b} = \mathbf{a}$ gives $\mathbf{a} \cdot \mathbf{a} = |\mathbf{a}|^2$. Hence

$$|\mathbf{a}| = \sqrt{\mathbf{a} \cdot \mathbf{a}}. \tag{3}$$

From (3) and (1) we obtain for the angle γ between two nonzero vectors

$$\cos \gamma = \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}||\mathbf{b}|} = \frac{\mathbf{a} \cdot \mathbf{b}}{\sqrt{\mathbf{a} \cdot \mathbf{a}}\sqrt{\mathbf{b} \cdot \mathbf{b}}}. \tag{4}$$

EXAMPLE 1 Inner Product. Angle Between Vectors

Find the inner product and the lengths of $\mathbf{a} = [1, 2, 0]$ and $\mathbf{b} = [3, -2, 1]$ as well as the angle between these vectors.

Solution. $\mathbf{a} \cdot \mathbf{b} = 1 \cdot 3 + 2 \cdot (-2) + 0 \cdot 1 = -1$, $|\mathbf{a}| = \sqrt{\mathbf{a} \cdot \mathbf{a}} = \sqrt{5}$, $|\mathbf{b}| = \sqrt{\mathbf{b} \cdot \mathbf{b}} = \sqrt{14}$, and (4) gives the angle

$$\gamma = \arccos \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}||\mathbf{b}|} = \arccos (-0.11952) = 1.69061 = 96.865°.$$

From the definition we see that the inner product has the following properties. For any vectors **a, b, c** and scalars q_1, q_2,

$$\text{(a)} \quad (q_1\mathbf{a} + q_2\mathbf{b}) \cdot \mathbf{c} = q_1\mathbf{a} \cdot \mathbf{c} + q_2\mathbf{b} \cdot \mathbf{c} \qquad \textit{(Linearity)}$$

$$\text{(b)} \quad \mathbf{a} \cdot \mathbf{b} = \mathbf{b} \cdot \mathbf{a} \qquad \textit{(Symmetry)} \tag{5}$$

$$\text{(c)} \quad \left.\begin{array}{c} \mathbf{a} \cdot \mathbf{a} \geqq 0 \\ \mathbf{a} \cdot \mathbf{a} = 0 \quad \text{if and only if} \quad \mathbf{a} = \mathbf{0} \end{array}\right\} \textit{(Positive-definiteness).}$$

Hence *dot multiplication is commutative* [see (5b)] *and is distributive with respect to vector addition;* in fact, from (5a) with $q_1 = 1$ and $q_2 = 1$ we have

$$(\mathbf{a} + \mathbf{b}) \cdot \mathbf{c} = \mathbf{a} \cdot \mathbf{c} + \mathbf{b} \cdot \mathbf{c} \qquad \textit{(Distributivity).} \tag{5a*}$$

Furthermore, from (1) and $|\cos \gamma| \leqq 1$ we see that

$$|\mathbf{a} \cdot \mathbf{b}| \leqq |\mathbf{a}||\mathbf{b}| \qquad \textit{(Cauchy–Schwarz inequality).} \tag{6}$$

Using this and (3), you may prove (see Prob. 18)

$$|\mathbf{a} + \mathbf{b}| \leqq |\mathbf{a}| + |\mathbf{b}| \qquad \textbf{(Triangle inequality).} \tag{7}$$

Geometrically, (7) with $<$ says that one side of a triangle must be shorter than the other two sides together; this motivates the name of (7).

A simple direct calculation with inner products shows that

$$|\mathbf{a} + \mathbf{b}|^2 + |\mathbf{a} - \mathbf{b}|^2 = 2(|\mathbf{a}|^2 + |\mathbf{b}|^2) \qquad \textit{(Parallelogram equality).} \tag{8}$$

Equations (6)–(8) play a basic role in so-called *Hilbert spaces* (abstract inner product spaces), which form the basis of quantum mechanics (see Ref. [GR7] listed in App. 1).

Derivation of (2) from (1). We write $\mathbf{a} = a_1\mathbf{i} + a_2\mathbf{j} + a_3\mathbf{k}$ and $\mathbf{b} = b_1\mathbf{i} + b_2\mathbf{j} + b_3\mathbf{k}$, as in (8) of Sec. 9.1. If we substitute this into $\mathbf{a} \bullet \mathbf{b}$ and use (5a*), we first have a sum of $3 \times 3 = 9$ products

$$\mathbf{a} \bullet \mathbf{b} = a_1b_1\mathbf{i} \bullet \mathbf{i} + a_1b_2\mathbf{i} \bullet \mathbf{j} + \cdots + a_3b_3\mathbf{k} \bullet \mathbf{k}.$$

Now **i, j, k** are unit vectors, so that $\mathbf{i} \bullet \mathbf{i} = \mathbf{j} \bullet \mathbf{j} = \mathbf{k} \bullet \mathbf{k} = 1$ by (3). Since the coordinate axes are perpendicular, so are **i, j, k,** and Theorem 1 implies that the other six of those nine products are 0, namely, $\mathbf{i} \bullet \mathbf{j} = \mathbf{j} \bullet \mathbf{i} = \mathbf{j} \bullet \mathbf{k} = \mathbf{k} \bullet \mathbf{j} = \mathbf{k} \bullet \mathbf{i} = \mathbf{i} \bullet \mathbf{k} = 0$. But this reduces our sum for $\mathbf{a} \bullet \mathbf{b}$ to (2). ∎

Applications of Inner Products

Typical applications of inner products are shown in the following examples and in Problem Set 9.2.

EXAMPLE 2 **Work Done by a Force Expressed as an Inner Product**

This is a major application. It concerns a body on which a *constant* force **p** acts. (For a *variable* force, see Sec. 10.1.) Let the body be given a displacement **d.** Then the work done by **p** in the displacement is defined as

(9) $$W = |\mathbf{p}||\mathbf{d}| \cos \alpha = \mathbf{p} \bullet \mathbf{d},$$

that is, magnitude $|\mathbf{p}|$ of the force times length $|\mathbf{d}|$ of the displacement times the cosine of the angle α between **p** and **d** (Fig. 177). If $\alpha < 90°$, as in Fig. 177, then $W > 0$. If **p** and **d** are orthogonal, then the work is zero (why?). If $\alpha > 90°$, then $W < 0$, which means that in the displacement one has to do work against the force. (Think of swimming across a river at some angle α against the current.) ∎

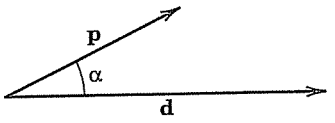


Fig. 177. Work done by a force

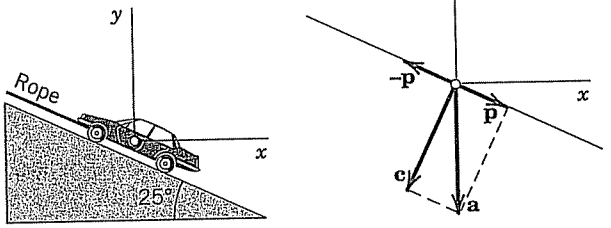


Fig. 178. Example 3

EXAMPLE 3 **Component of a Force in a Given Direction**

What force in the rope in Fig. 178 will hold a car of 5000 lb in equilibrium if the ramp makes an angle of 25° with the horizontal?

Solution. Introducing coordinates as shown, the weight is $\mathbf{a} = [0, -5000]$ because this force points downward, in the negative y-direction. We have to represent **a** as a sum (resultant) of two forces, $\mathbf{a} = \mathbf{c} + \mathbf{p}$, where **c** is the force the car exerts on the ramp, which is of no interest to us, and **p** is parallel to the rope, of magnitude (see Fig. 178)

$$|\mathbf{p}| = |\mathbf{a}| \cos \gamma = 5000 \cos 65° = 2113 \text{ [lb]}$$

and direction of the unit vector **u** opposite to the direction of the rope; here $\gamma = 90° - 25° = 65°$ is the angle between **a** and **p.** Now a vector in the direction of the rope is

$$\mathbf{b} = [-1, \tan 25°] = [-1, 0.46631], \quad \text{thus} \quad |\mathbf{b}| = 1.10338,$$

so that

$$\mathbf{u} = -\frac{1}{|\mathbf{b}|}\,\mathbf{b} = [0.90631,\ -0.42262].$$

Since $|\mathbf{u}| = 1$ and $\cos\gamma > 0$, we see that we can also write our result as

$$|\mathbf{p}| = (|\mathbf{a}|\cos\gamma)|\mathbf{u}| = \mathbf{a}\cdot\mathbf{u} = -\frac{\mathbf{a}\cdot\mathbf{b}}{|\mathbf{b}|} = \frac{5000\cdot 0.46631}{1.10338} = 2113\ [\text{lb}].$$

Answer: About 2100 lb.

Example 3 is typical of applications in which one uses the concept of the **component** or **projection** *of a vector* **a** *in the direction of a vector* **b** ($\neq \mathbf{0}$), defined by (see Fig. 179)

(10) $$p = |\mathbf{a}|\cos\gamma.$$

Thus p is the length of the orthogonal projection of **a** on a straight line l parallel to **b,** taken with the plus sign if $p\mathbf{b}$ has the direction of **b** and with the minus sign if $p\mathbf{b}$ has the direction opposite to **b;** see Fig. 179.

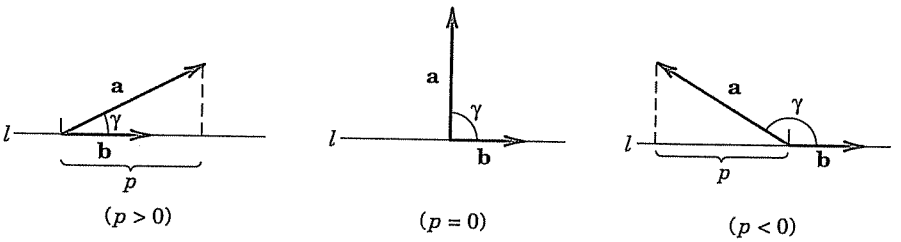


Fig. 179. Component of a vector **a** in the direction of a vector **b**

Multiplying (10) by $|\mathbf{b}|/|\mathbf{b}| = 1$, we have $\mathbf{a}\cdot\mathbf{b}$ in the numerator and thus

(11) $$p = \frac{\mathbf{a}\cdot\mathbf{b}}{|\mathbf{b}|} \qquad (\mathbf{b} \neq \mathbf{0}).$$

If **b** is a unit vector, as it is often used for fixing a direction, then (11) simply gives

(12) $$p = \mathbf{a}\cdot\mathbf{b} \qquad (|\mathbf{b}| = 1).$$

Figure 180 shows the projection p of **a** in the direction of **b** (as in Fig. 179) and the projection $q = |\mathbf{b}|\cos\gamma$ of **b** in the direction of **a.**

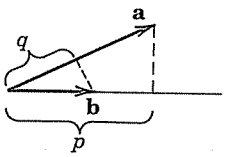


Fig. 180. Projections p of **a** on **b** and q of **b** on **a**

EXAMPLE 4 Orthonormal Basis

By definition, an *orthonormal basis* for 3-space is a basis $\{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$ consisting of orthogonal unit vectors. It has the great advantage that the determination of the coefficients in representations $\mathbf{v} = l_1\mathbf{a} + l_2\mathbf{b} + l_3\mathbf{c}$ of a given vector $\mathbf{v}$ is very simple. We claim that $l_1 = \mathbf{a} \cdot \mathbf{v}$, $l_2 = \mathbf{b} \cdot \mathbf{v}$, $l_3 = \mathbf{c} \cdot \mathbf{v}$. Indeed, this follows simply by taking the inner products of the representation with $\mathbf{a}, \mathbf{b}, \mathbf{c}$, respectively, and using the orthonormality of the basis, $\mathbf{a} \cdot \mathbf{v} = l_1\mathbf{a} \cdot \mathbf{a} + l_2\mathbf{a} \cdot \mathbf{b} + l_3\mathbf{a} \cdot \mathbf{c} = l_1$, etc.

For example, the unit vectors $\mathbf{i}, \mathbf{j}, \mathbf{k}$ in (8), Sec. 9.1, associated with a Cartesian coordinate system form an orthonormal basis, called the **standard basis** with respect to the given coordinate system. ▩

EXAMPLE 5 Orthogonal Straight Lines in the Plane

Find the straight line L_1 through the point P: (1, 3) in the xy-plane and perpendicular to the straight line L_2: $x - 2y + 2 = 0$; see Fig. 181.

Solution. The idea is to write a general straight line L_1: $a_1x + a_2y = c$ as $\mathbf{a} \cdot \mathbf{r} = c$ with $\mathbf{a} = [a_1, a_2] \neq \mathbf{0}$ and $\mathbf{r} = [x, y]$, according to (2). Now the line L_1^* through the origin and parallel to L_1 is $\mathbf{a} \cdot \mathbf{r} = 0$. Hence, by Theorem 1, the vector $\mathbf{a}$ is perpendicular to $\mathbf{r}$. Hence it is perpendicular to L_1^* and also to L_1 because L_1 and L_1^* are parallel. $\mathbf{a}$ is called a **normal vector** of L_1 (and of L_1^*).

Now a normal vector of the given line $x - 2y + 2 = 0$ is $\mathbf{b} = [1, -2]$. Thus L_1 is perpendicular to L_2 if $\mathbf{b} \cdot \mathbf{a} = a_1 - 2a_2 = 0$, for instance, if $\mathbf{a} = [2, 1]$. Hence L_1 is given by $2x + y = c$. It passes through P: (1, 3) when $2 \cdot 1 + 3 = c = 5$. *Answer:* $y = -2x + 5$. Show that the point of intersection is $(x, y) = (1.6, 1.8)$. ▩

EXAMPLE 6 Normal Vector to a Plane

Find a unit vector perpendicular to the plane $4x + 2y + 4z = -7$.

Solution. Using (2), we may write any plane in space as

$$\mathbf{a} \cdot \mathbf{r} = a_1x + a_2y + a_3z = c \tag{13}$$

where $\mathbf{a} = [a_1, a_2, a_3] \neq \mathbf{0}$ and $\mathbf{r} = [x, y, z]$. The unit vector in the direction of $\mathbf{a}$ is (Fig. 182)

$$\mathbf{n} = \frac{1}{|\mathbf{a}|}\,\mathbf{a}.$$

Dividing by $|\mathbf{a}|$, we obtain from (13)

$$\mathbf{n} \cdot \mathbf{r} = p \qquad \text{where} \qquad p = \frac{c}{|\mathbf{a}|}. \tag{14}$$

From (12) we see that p is the projection of $\mathbf{r}$ in the direction of $\mathbf{n}$. This projection has the same constant value $c/|\mathbf{a}|$ for the position vector $\mathbf{r}$ of any point in the plane. Clearly this holds if and only if $\mathbf{n}$ is perpendicular to the plane. $\mathbf{n}$ is called a **unit normal vector** of the plane (the other being $-\mathbf{n}$).

Furthermore, from this and the definition of projection it follows that $|p|$ is the distance of the plane from the origin. Representation (14) is called **Hesse's[2] normal form** of a plane. In our case, $\mathbf{a} = [4, 2, 4]$, $c = -7$, $|\mathbf{a}| = 6$, $\mathbf{n} = \frac{1}{6}\mathbf{a} = [\frac{2}{3}, \frac{1}{3}, \frac{2}{3}]$, and the plane has the distance 7/6 from the origin. ▩

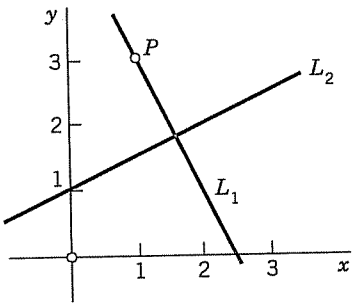


Fig. 181. Example 5

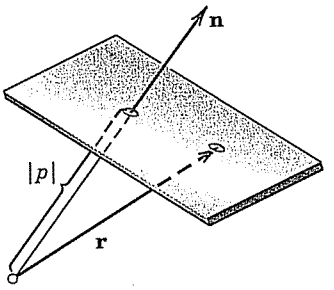


Fig. 182. Normal vector to a plane

[2]LUDWIG OTTO HESSE (1811–1874), German mathematician who contributed to the theory of curves and surfaces.

PROBLEM SET 9.2

1–12 INNER PRODUCT

Let $\mathbf{a} = [2, 1, 4]$, $\mathbf{b} = [-4, 0, 3]$, $\mathbf{c} = [3, -2, 1]$. Find

1. $\mathbf{a} \cdot \mathbf{b}$, $\mathbf{b} \cdot \mathbf{a}$
2. $|\mathbf{a}|$, $|\mathbf{b}|$, $|\mathbf{c}|$
3. $|3\mathbf{a} - 2\mathbf{b}|$, $|2\mathbf{b} - 3\mathbf{a}|$
4. $\mathbf{a} \cdot (\mathbf{b} + \mathbf{c})$, $\mathbf{a} \cdot \mathbf{b} + \mathbf{a} \cdot \mathbf{c}$
5. $(\mathbf{a} \cdot \mathbf{b})\mathbf{c}$, $\mathbf{a}(\mathbf{b} \cdot \mathbf{c})$
6. $\mathbf{a} \cdot \mathbf{b} + \mathbf{b} \cdot \mathbf{c} + \mathbf{c} \cdot \mathbf{a}$
7. $(\mathbf{a} - \mathbf{b}) \cdot \mathbf{c}$, $\mathbf{a} \cdot \mathbf{c} - \mathbf{b} \cdot \mathbf{c}$
8. $4\mathbf{a} \cdot 3\mathbf{c}$, $12\mathbf{a} \cdot \mathbf{c}$
9. $\mathbf{a} \cdot (\mathbf{b} - \mathbf{c})$, $\mathbf{a} \cdot (\mathbf{c} - \mathbf{b})$
10. $|\mathbf{b} + \mathbf{c}|$, $|\mathbf{b}| + |\mathbf{c}|$
11. $6(\mathbf{a} + \mathbf{b}) \cdot (\mathbf{a} - \mathbf{b})$
12. $|\mathbf{a} \cdot \mathbf{c}|$, $|\mathbf{a}||\mathbf{c}|$

13. What laws do Probs. 1, 3, 4, 7, 8 illustrate?
14. Does $\mathbf{u} \cdot \mathbf{v} = \mathbf{u} \cdot \mathbf{w}$ with $\mathbf{u} \neq \mathbf{0}$ imply that $\mathbf{v} = \mathbf{w}$?
15. Prove the Cauchy–Schwarz inequality.
16. Verify the Cauchy–Schwarz inequality, the triangle inequality, and the parallelogram equality for the above $\mathbf{a}$ and $\mathbf{b}$.
17. Prove the parallelogram equality.
18. **(Triangle inequality)** Prove (7). *Hint.* Use (3) for $|\mathbf{a} + \mathbf{b}|$ and (6) to prove the square of (7), then take roots.

19–22 WORK

Find the work done by a force $\mathbf{p}$ acting on a body if the body is displaced from a point A to a point B along the straight segment $\overrightarrow{AB}$. Sketch $\mathbf{p}$ and $\overrightarrow{AB}$. (Show the details of your work.)

19. $\mathbf{p} = [8, -4, 11]$, A: (1, 2, 0), B: (3, 6, 0)
20. $\mathbf{p} = [2, 7, -4]$, A: (3, 1, 0), B: (0, 2, 0)
21. $\mathbf{p} = [5, -2, 1]$, A: (4, 0, 3), B: (6, 0, 8)
22. $\mathbf{p} = [4, 3, 6]$, A: (5, 2, 10), B: (1, 3, 1)

23. Why is the work in Prob. 19 zero? Can work be negative? Explain.
24. Show that the work done by the resultant of $\mathbf{p}$ and $\mathbf{q}$ in a displacement from A to B is the sum of the work done by each force in that displacement.
25. Find the work $W = \mathbf{p} \cdot \mathbf{d}$ if $\mathbf{d} = 2\mathbf{i}$ and $\mathbf{p} = \mathbf{i}$, $\mathbf{i} + \mathbf{j}$, $\mathbf{j}$, $-\mathbf{i} + \mathbf{j}$ and sketch a figure similar to Fig. 177.

26–30 ANGLE BETWEEN VECTORS. ORTHOGONALITY

Let $\mathbf{a} = [1, 1, 1]$, $\mathbf{b} = [2, 3, 1]$, $\mathbf{c} = [-1, 1, 0]$. Find the angle between:

26. $\mathbf{a}$, $\mathbf{b}$
27. $\mathbf{b}$, $\mathbf{c}$
28. $\mathbf{a} - \mathbf{c}$, $\mathbf{b} - \mathbf{c}$
29. $\mathbf{a} + \mathbf{b}$, $\mathbf{c}$
30. $\mathbf{a}$, $\mathbf{b} + \mathbf{c}$

31. **(Planes)** Find the angle between the planes $x + y + z = 1$ and $2x - y + 2z = 0$.
32. **(Cosine law)** Deduce the law of cosines by using vectors $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{a} - \mathbf{b}$.
33. **(Triangle)** Find the angles of the triangle with vertices [0, 0, 0], [1, 2, 3], [4, −1, 3].
34. **(Addition law)** Obtain
$$\cos(\alpha - \beta) = \cos\alpha\cos\beta + \sin\alpha\sin\beta$$
by using $\mathbf{a} = [\cos\alpha, \sin\alpha]$, $\mathbf{b} = [\cos\beta, \sin\beta]$, where $0 \leqq \alpha \leqq \beta = 2\pi$.
35. **(Parallelogram)** Find the angles if the sides are [5, 0] and [1, 2].
36. **(Distance)** Find the distance of the plane $5x + 2y + z = 10$ from the origin.

37–40 COMPONENTS IN THE DIRECTION OF A VECTOR

Find the component of $\mathbf{a}$ in the direction of $\mathbf{b}$.

37. $\mathbf{a} = [1, 1, 3]$, $\mathbf{b} = [0, 0, 5]$
38. $\mathbf{a} = [2, 0, 6]$, $\mathbf{b} = [3, 4, -1]$
39. $\mathbf{a} = [0, 4, -3]$, $\mathbf{b} = [0, 4, 3]$
40. $\mathbf{a} = [-1, 2, 0]$, $\mathbf{b} = [1, -2, 0]$

41. Under what condition will the projection of $\mathbf{a}$ in the direction of $\mathbf{b}$ equal the projection of $\mathbf{b}$ in the direction of $\mathbf{a}$?
42. **TEAM PROJECT. Orthogonality** is particularly important, mainly because of the use of orthogonal coordinates, such as **Cartesian coordinates,** whose **"natural basis"** (9), Sec. 9.1, consists of three orthogonal unit vectors.

(a) Show that $\mathbf{a} = [2, -2, 4]$, $\mathbf{b} = [0, 8, 4]$, $\mathbf{c} = [-20, -4, 8]$ are orthogonal.

(b) For what values of a_1 are $\mathbf{a} = [a_1, 2, 0]$ and $\mathbf{b} = [3, 4, -1]$ orthogonal?

(c) Show that the straight lines $4x + 2y = 1$ and $5x - 10y = 7$ are orthogonal.

(d) Find all unit vectors $\mathbf{a} = [a_1, a_2]$ in the plane orthogonal to [4, 3].

(e) Find all vectors orthogonal to $\mathbf{a} = [2, 1, 0]$. Do they form a vector space?

(f) For what c are the planes $4x - 2y + 3z = 6$ and $2x - cy + 5z = 1$ orthogonal?

(g) Under what condition will the diagonals of a parallelogram be orthogonal? (Prove your answer.)

(h) What is the angle between a light ray and its reflection in three orthogonal plane mirrors (known as a **"corner reflector"**)?

(i) Discuss further applications in physics and geometry in which orthogonality plays a role.

9.3 Vector Product (Cross Product)

The dot product in Sec. 9.2 is a *scalar*. We shall see that in some applications, for instance, in connection with rotations, we shall need a product that is again a *vector*:

DEFINITION

Vector Product (Cross Product, Outer Product) of Vectors

The **vector product** (also called **cross product** or **outer product**) $\mathbf{a} \times \mathbf{b}$ (read "**a** cross **b**") of two vectors **a** and **b** is the vector

$$\mathbf{v} = \mathbf{a} \times \mathbf{b}$$

as follows. If **a** and **b** have the same or opposite direction, or if $\mathbf{a} = \mathbf{0}$ or $\mathbf{b} = \mathbf{0}$, then $\mathbf{v} = \mathbf{a} \times \mathbf{b} = \mathbf{0}$. In any other case $\mathbf{v} = \mathbf{a} \times \mathbf{b}$ has the length

$$|\mathbf{v}| = |\mathbf{a} \times \mathbf{b}| = |\mathbf{a}||\mathbf{b}| \sin \gamma. \tag{1}$$

This is the area of the blue parallelogram in Fig. 183. γ is the angle between **a** and **b** (as in Sec. 9.2). The direction of $\mathbf{v} = \mathbf{a} \times \mathbf{b}$ is perpendicular to both **a** and **b** and such that **a, b, v,** in this order, form a *right-handed triple* as in Figs. 183–185 (explanation below).

In components, let $\mathbf{a} = [a_1, a_2, a_3]$ and $\mathbf{b} = [b_1, b_2, b_3]$. Then $\mathbf{v} = [v_1, v_2, v_3] = \mathbf{a} \times \mathbf{b}$ has the components

$$v_1 = a_2b_3 - a_3b_2, \qquad v_2 = a_3b_1 - a_1b_3, \qquad v_3 = a_1b_2 - a_2b_1. \tag{2}$$

Here the Cartesian coordinate system is *right-handed*, as explained below (see also Fig. 186). (For a left-handed system, each component of **v** must be multiplied by -1. Derivation of (2) in App. 4.)

Right-Handed Triple. A triple of vectors **a, b, v** is *right-handed* if the vectors in the given order assume the same sort of orientation as the thumb, index finger, and middle finger of the right hand when these are held as in Fig. 184. We may also say that if **a** is rotated into the direction of **b** through the angle γ ($< \pi$), then **v** advances in the same direction as a right-handed screw would if turned in the same way (Fig. 185).

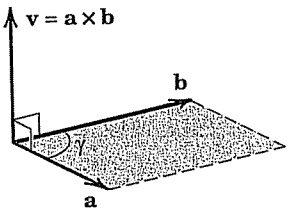


Fig. 183. Vector product

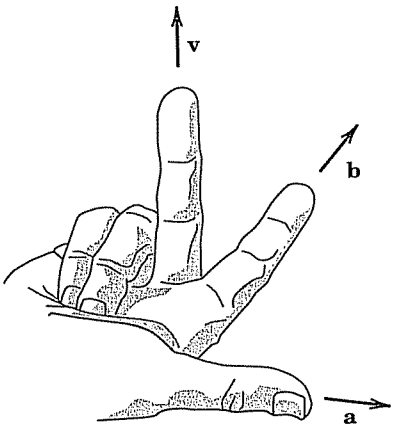


Fig. 184. Right-handed triple of vectors **a, b, v**

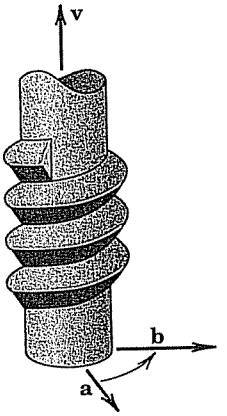
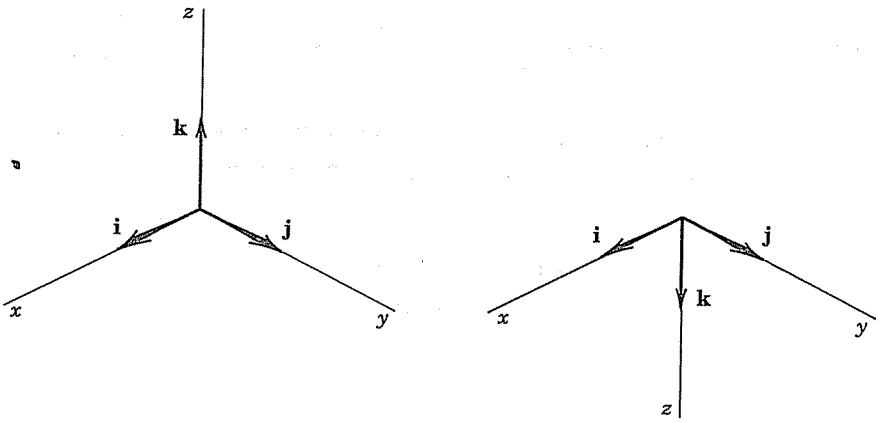


Fig. 185. Right-handed screw



(a) Right-handed (b) Left-handed

Fig. 186. The two types of Cartesian coordinate systems

Right-Handed Cartesian Coordinate System. The system is called **right-handed** if the corresponding unit vectors **i, j, k** in the positive directions of the axes (see Sec. 9.1) form a right-handed triple as in Fig. 186a. The system is called **left-handed** if the sense of **k** is reversed, as in Fig. 186b. In applications, we prefer right-handed systems.

How to Memorize (2). If you know second- and third-order determinants, you see that (2) can be written

$$(2^*) \qquad v_1 = \begin{vmatrix} a_2 & a_3 \\ b_2 & b_3 \end{vmatrix}, \qquad v_2 = -\begin{vmatrix} a_1 & a_3 \\ b_1 & b_3 \end{vmatrix} = +\begin{vmatrix} a_3 & a_1 \\ b_3 & b_1 \end{vmatrix}, \qquad v_3 = \begin{vmatrix} a_1 & a_2 \\ b_1 & b_2 \end{vmatrix}$$

and $\mathbf{v} = [v_1, v_2, v_3] = v_1\mathbf{i} + v_2\mathbf{j} + v_3\mathbf{k}$ is the expansion of the following symbolic determinant by its first row. (We call the determinant "symbolic" because the first row consists of vectors rather than of numbers.)

$$(2^{**}) \qquad \mathbf{v} = \mathbf{a} \times \mathbf{b} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix} = \begin{vmatrix} a_2 & a_3 \\ b_2 & b_3 \end{vmatrix}\mathbf{i} - \begin{vmatrix} a_1 & a_3 \\ b_1 & b_3 \end{vmatrix}\mathbf{j} + \begin{vmatrix} a_1 & a_2 \\ b_1 & b_2 \end{vmatrix}\mathbf{k}.$$

For a left-handed system the determinant has a minus sign in front.

EXAMPLE 1 **Vector Product**

For the vector product $\mathbf{v} = \mathbf{a} \times \mathbf{b}$ of $\mathbf{a} = [1, 1, 0]$ and $\mathbf{b} = [3, 0, 0]$ in right-handed coordinates we obtain from (2)

$$v_1 = 0, \qquad v_2 = 0, \qquad v_3 = 1 \cdot 0 - 1 \cdot 3 = -3.$$

We confirm this by (2**):

$$\mathbf{v} = \mathbf{a} \times \mathbf{b} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 1 & 1 & 0 \\ 3 & 0 & 0 \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ 0 & 0 \end{vmatrix}\mathbf{i} - \begin{vmatrix} 1 & 0 \\ 3 & 0 \end{vmatrix}\mathbf{j} + \begin{vmatrix} 1 & 1 \\ 3 & 0 \end{vmatrix}\mathbf{k} = -3\mathbf{k} = [0, 0, -3].$$

To check the result in this simple case, sketch **a, b,** and **v.** Can you see that two vectors in the xy-plane must always have their vector product parallel to the z-axis (or equal to the zero vector)? ■

EXAMPLE 2 **Vector Products of the Standard Basis Vectors**

$$\mathbf{i} \times \mathbf{j} = \mathbf{k}, \qquad \mathbf{j} \times \mathbf{k} = \mathbf{i}, \qquad \mathbf{k} \times \mathbf{i} = \mathbf{j}$$
$$\mathbf{j} \times \mathbf{i} = -\mathbf{k}, \qquad \mathbf{k} \times \mathbf{j} = -\mathbf{i}, \qquad \mathbf{i} \times \mathbf{k} = -\mathbf{j}. \tag{3}$$

We shall use this in the next proof. ■

THEOREM 1 **General Properties of Vector Products**

(a) *For every scalar l,*

$$(l\mathbf{a}) \times \mathbf{b} = l(\mathbf{a} \times \mathbf{b}) = \mathbf{a} \times (l\mathbf{b}). \tag{4}$$

(b) *Cross multiplication is distributive with respect to vector addition; that is,*

$$(\alpha) \quad \mathbf{a} \times (\mathbf{b} + \mathbf{c}) = (\mathbf{a} \times \mathbf{b}) + (\mathbf{a} \times \mathbf{c}),$$
$$(\beta) \quad (\mathbf{a} + \mathbf{b}) \times \mathbf{c} = (\mathbf{a} \times \mathbf{c}) + (\mathbf{b} \times \mathbf{c}). \tag{5}$$

(c) *Cross multiplication is* ***not commutative*** *but* ***anticommutative;*** *that is,*

$$\mathbf{b} \times \mathbf{a} = -(\mathbf{a} \times \mathbf{b}) \tag{6}$$

(Fig. 187).

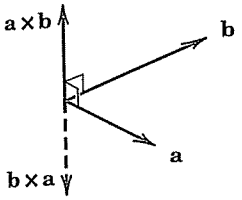


Fig. 187. Anticommutativity of cross multiplication

(d) *Cross multiplication is* ***not associative;*** *that is, in general,*

$$\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) \neq (\mathbf{a} \times \mathbf{b}) \times \mathbf{c} \tag{7}$$

so that the parentheses cannot be omitted.

PROOF (4) follows directly from the definition. In (5α), formula (2*) gives for the first component on the left

$$\begin{vmatrix} a_2 & a_3 \\ b_2 + c_2 & b_3 + c_3 \end{vmatrix} = a_2(b_3 + c_3) - a_3(b_2 + c_2)$$
$$= (a_2b_3 - a_3b_2) + (a_2c_3 - a_3c_2)$$
$$= \begin{vmatrix} a_2 & a_3 \\ b_2 & b_3 \end{vmatrix} + \begin{vmatrix} a_2 & a_3 \\ c_2 & c_3 \end{vmatrix}.$$

By (2*) the sum of the two determinants is the first component of $(\mathbf{a} \times \mathbf{b}) + (\mathbf{a} \times \mathbf{c})$, the right side of (5a). For the other components in (5α) and in (5β), equality follows by the same idea.

Anticommutativity (6) follows from (2**) by noting that the interchange of Rows 2 and 3 multiplies the determinant by -1. We can confirm this geometrically if we set $\mathbf{a} \times \mathbf{b} = \mathbf{v}$ and $\mathbf{b} \times \mathbf{a} = \mathbf{w}$; then $|\mathbf{v}| = |\mathbf{w}|$ by (1), and for $\mathbf{b}, \mathbf{a}, \mathbf{w}$ to form a *right-handed* triple, we must have $\mathbf{w} = -\mathbf{v}$.

Finally, $\mathbf{i} \times (\mathbf{i} \times \mathbf{j}) = \mathbf{i} \times \mathbf{k} = -\mathbf{j}$, whereas $(\mathbf{i} \times \mathbf{i}) \times \mathbf{j} = \mathbf{0} \times \mathbf{j} = \mathbf{0}$ (see Example 2). This proves (7). ■

Typical Applications of Vector Products

EXAMPLE 3 **Moment of a Force**

In mechanics the moment m of a force $\mathbf{p}$ about a point Q is defined as the product $m = |\mathbf{p}|d$, where d is the (perpendicular) distance between Q and the line of action L of $\mathbf{p}$ (Fig. 188). If $\mathbf{r}$ is the vector from Q to any point A on L, then $d = |\mathbf{r}| \sin \gamma$ (Fig. 188) and

$$m = |\mathbf{r}|\,|\mathbf{p}| \sin \gamma.$$

Since γ is the angle between $\mathbf{r}$ and $\mathbf{p}$, we see from (1) that $m = |\mathbf{r} \times \mathbf{p}|$. The vector

(8) $$\mathbf{m} = \mathbf{r} \times \mathbf{p}$$

is called the **moment vector** or **vector moment** of $\mathbf{p}$ about Q. Its magnitude is m. If $\mathbf{m} \neq \mathbf{0}$, its direction is that of the axis of the rotation about Q that $\mathbf{p}$ has the tendency to produce. This axis is perpendicular to both $\mathbf{r}$ and $\mathbf{p}$. ■

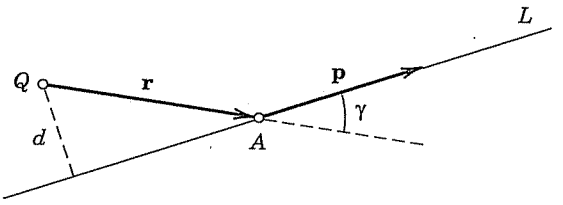


Fig. 188. Moment of a force $\mathbf{p}$

EXAMPLE 4 **Moment of a Force**

Find the moment of the force $\mathbf{p}$ in Fig. 189 about the center Q of the wheel.

Solution. Introducing coordinates as shown in Fig. 189, we have

$$\mathbf{p} = [1000 \cos 30°, \quad 1000 \sin 30°, \quad 0] = [866, \quad 500, \quad 0], \qquad \mathbf{r} = [0, \quad 1.5, \quad 0].$$

(Note that the center of the wheel is at $y = -1.5$ on the y-axis.) Hence (8) and (2**) give

$$\mathbf{m} = \mathbf{r} \times \mathbf{p} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 0 & 1.5 & 0 \\ 866 & 500 & 0 \end{vmatrix} = 0\mathbf{i} - 0\mathbf{j} + \begin{vmatrix} 0 & 1.5 \\ 866 & 500 \end{vmatrix}\mathbf{k} = [0, 0, -1299].$$

This moment vector is normal (perpendicular) to the plane of the wheel; hence it has the direction of the axis of rotation about the center of the wheel that the force has the tendency to produce. $\mathbf{m}$ points in the negative z-direction, the direction in which a right-handed screw would advance if turned in that way. ■

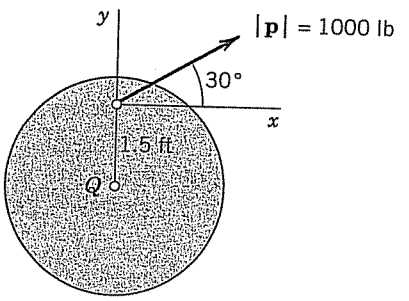


Fig. 189. Moment of a force $\mathbf{p}$

EXAMPLE 5 Velocity of a Rotating Body

A rotation of a rigid body B in space can be simply and uniquely described by a vector $\mathbf{w}$ as follows. The direction of $\mathbf{w}$ is that of the axis of rotation and such that the rotation appears clockwise if one looks from the initial point of $\mathbf{w}$ to its terminal point. The length of $\mathbf{w}$ is equal to the **angular speed** ω (> 0) of the rotation, that is, the linear (or tangential) speed of a point of B divided by its distance from the axis of rotation.

Let P be any point of B and d its distance from the axis. Then P has the speed ωd. Let $\mathbf{r}$ be the position vector of P referred to a coordinate system with origin 0 on the axis of rotation. Then $d = |\mathbf{r}| \sin \gamma$, where γ is the angle between $\mathbf{w}$ and $\mathbf{r}$. Therefore,

$$\omega d = |\mathbf{w}|\,|\mathbf{r}| \sin \gamma = |\mathbf{w} \times \mathbf{r}|.$$

From this and the definition of vector product we see that the velocity vector $\mathbf{v}$ of P can be represented in the form (Fig. 190)

$$\mathbf{v} = \mathbf{w} \times \mathbf{r}. \tag{9}$$

This simple formula is useful for determining $\mathbf{v}$ at any point of B.

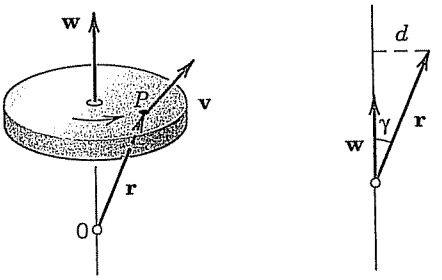


Fig. 190. Rotation of a rigid body

Scalar Triple Product

The most important product of vectors with more than two factors is the **scalar triple product** or **mixed triple product** of three vectors **a, b, c.** It is denoted by $(\mathbf{a}\ \ \mathbf{b}\ \ \mathbf{c})$ and defined by

$$(\mathbf{a}\ \ \mathbf{b}\ \ \mathbf{c}) = \mathbf{a} \bullet (\mathbf{b} \times \mathbf{c}). \tag{10*}$$

Because of the dot product it is a scalar. In terms of components $\mathbf{a} = [a_1, a_2, a_3]$, $\mathbf{b} = [b_1, b_2, b_3]$, $\mathbf{c} = [c_1, c_2, c_3]$ we can write it as a third-order determinant. For this we set $\mathbf{b} \times \mathbf{c} = \mathbf{v} = [v_1, v_2, v_3]$. Then from the dot product in components [formula (2) in Sec. 9.2] and from (2*) with $\mathbf{b}$ and $\mathbf{c}$ instead of $\mathbf{a}$ and $\mathbf{b}$ we first obtain

$$\begin{aligned}\mathbf{a} \bullet (\mathbf{b} \times \mathbf{c}) = \mathbf{a} \bullet \mathbf{v} &= a_1 v_1 + a_2 v_2 + a_3 v_3 \\ &= a_1 \begin{vmatrix} b_2 & b_3 \\ c_2 & c_3 \end{vmatrix} + a_2 \begin{vmatrix} b_3 & b_1 \\ c_3 & c_1 \end{vmatrix} + a_3 \begin{vmatrix} b_1 & b_2 \\ c_1 & c_2 \end{vmatrix}.\end{aligned}$$

The sum on the right is the expansion of a third-order determinant by its first row. Thus

$$(\mathbf{a}\ \ \mathbf{b}\ \ \mathbf{c}) = \mathbf{a} \bullet (\mathbf{b} \times \mathbf{c}) = \begin{vmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{vmatrix}. \tag{10}$$

The most important properties of the scalar triple product are as follows.

THEOREM 2

Properties and Applications of Scalar Triple Products

(a) *In* (10) *the dot and cross can be interchanged:*

$$(\mathbf{a}\ \ \mathbf{b}\ \ \mathbf{c}) = \mathbf{a}\cdot(\mathbf{b}\times\mathbf{c}) = (\mathbf{a}\times\mathbf{b})\cdot\mathbf{c}. \tag{11}$$

(b) Geometric interpretation. *The absolute value* $|(\mathbf{a}\ \ \mathbf{b}\ \ \mathbf{c})|$ *of* (10) *is the volume of the parallelepiped* (oblique box) *with* **a, b, c** *as edge vectors* (Fig. 191).

(c) Linear independence. *Three vectors in* R^3 *are linearly independent if and only if their scalar triple product is not zero.*

PROOF **(a)** Dot multiplication is commutative, so that by (10)

$$(\mathbf{a}\times\mathbf{b})\cdot\mathbf{c} = \mathbf{c}\cdot(\mathbf{a}\times\mathbf{b}) = \begin{vmatrix} c_1 & c_2 & c_3 \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix}.$$

From this we obtain the determinant in (10) by interchanging Rows 1 and 2 and in the result Rows 2 and 3. But this does not change the value of the determinant because each interchange produces a factor -1, and $(-1)(-1) = 1$. This proves (11).

(b) The volume of that box equals the height $h = |\mathbf{a}||\cos\gamma|$ (Fig. 191) times the area of the base, which is the area $|\mathbf{b}\times\mathbf{c}|$ of the parallelogram with sides **b** and **c.** Hence the volume is

$$|\mathbf{a}||\mathbf{b}\times\mathbf{c}|\,|\cos\gamma| = |\mathbf{a}\cdot(\mathbf{b}\times\mathbf{c})| \qquad \text{(Fig. 191)}$$

as given by the absolute value of (11).

(c) Three nonzero vectors, if we let their initial points coincide, are linearly independent if and only if they do not lie in the same plane (or do not lie on the same straight line). This happens if and only if the triple product in (b) is not zero, so that the independence criterion follows. (The case that one of the vectors is the zero vector is trivial.) ∎

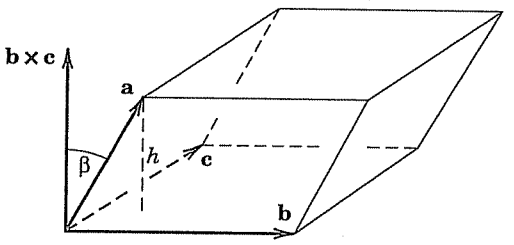


Fig. 191. Geometric interpretation of a scalar triple product

EXAMPLE 6 **Tetrahedron**

A tetrahedron is determined by three edge vectors **a, b, c,** as indicated in Fig. 192. Find its volume when $\mathbf{a} = [2, 0, 3]$, $\mathbf{b} = [0, 4, 1]$, $\mathbf{c} = [5, 6, 0]$.

Solution. The volume V of the parallelepiped with these vectors as edge vectors is the absolute value of the scalar triple product

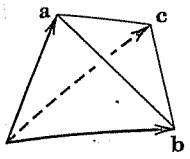


Fig. 192. Tetrahedron

$$(\mathbf{a}\ \ \mathbf{b}\ \ \mathbf{c}) = \begin{vmatrix} 2 & 0 & 3 \\ 0 & 4 & 1 \\ 5 & 6 & 0 \end{vmatrix} = 2\begin{vmatrix} 4 & 1 \\ 6 & 0 \end{vmatrix} + 3\begin{vmatrix} 0 & 4 \\ 5 & 6 \end{vmatrix} = -12 - 60 = -72.$$

Hence $V = 72$. The minus sign indicates that if the coordinates are right-handed, the triple **a, b, c** is left-handed. The volume of a tetrahedron is $\frac{1}{6}$ of that of the parallelepiped (can you prove it?), hence 12.

Can you sketch the tetrahedron, choosing the origin as the common initial point of the vectors? What are the coordinates of the four vertices? ■

This is the end of vector *algebra* (in space R^3 and in the plane). Vector *calculus* (differentiation) begins in the next section.

PROBLEM SET 9.3

1–20 VECTOR PRODUCT, SCALAR TRIPLE PRODUCT

With respect to right-handed Cartesian coordinates, let $\mathbf{a} = [1, 2, 0]$, $\mathbf{b} = [3, -4, 0]$, $\mathbf{c} = [3, 5, 2]$, $\mathbf{d} = [6, 2, -3]$. Showing details, find:

1. $\mathbf{a} \times \mathbf{b}$, $\mathbf{b} \times \mathbf{a}$
2. $\mathbf{a} \times \mathbf{c}$, $|\mathbf{a} \times \mathbf{c}|$, $\mathbf{a} \cdot \mathbf{c}$
3. $(\mathbf{a} + \mathbf{b}) \times \mathbf{c}$, $\mathbf{a} \times \mathbf{c} + \mathbf{b} \times \mathbf{c}$
4. $(\mathbf{c} + \mathbf{d}) \times \mathbf{d}$, $\mathbf{c} \times \mathbf{d}$
5. $2\mathbf{a} \times 3\mathbf{b}$, $3\mathbf{a} \times 2\mathbf{b}$, $6\mathbf{a} \times \mathbf{b}$
6. $\mathbf{b} \times \mathbf{c} + \mathbf{c} \times \mathbf{b}$
7. $\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})$, $(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c}$
8. $(\mathbf{a} + \mathbf{b}) \times (\mathbf{b} + \mathbf{a})$
9. $(\mathbf{a} \times \mathbf{b}) \cdot (\mathbf{c} \times \mathbf{d})$, $(\mathbf{b} \times \mathbf{a}) \cdot (\mathbf{d} \times \mathbf{c})$
10. $(\mathbf{a} \times \mathbf{b}) \times \mathbf{c}$, $\mathbf{a} \times (\mathbf{b} \times \mathbf{c})$
11. $\mathbf{d} \times \mathbf{c}$, $|\mathbf{d} \times \mathbf{c}|$, $|\mathbf{c} \times \mathbf{d}|$
12. $(\mathbf{a} + \mathbf{b}) \times (\mathbf{c} + \mathbf{d})$
13. $\mathbf{a} \times (\mathbf{b} + \mathbf{c} - \mathbf{d})$
14. $(\mathbf{i}\ \ \mathbf{j}\ \ \mathbf{k})$, $(\mathbf{i}\ \ \mathbf{k}\ \ \mathbf{j})$
15. $(\mathbf{i} + \mathbf{j}\ \ \mathbf{j} + \mathbf{k}\ \ \mathbf{k} + \mathbf{i})$
16. $(\mathbf{b} \times \mathbf{c}) \cdot \mathbf{d}$, $\mathbf{b} \cdot (\mathbf{c} \times \mathbf{d})$
17. $(\mathbf{a}\ \ \mathbf{b}\ \ \mathbf{d})$, $|(\mathbf{a}\ \ \mathbf{b}\ \ \mathbf{d})|$, $(\mathbf{b}\ \ \mathbf{a}\ \ \mathbf{d})$
18. $(\mathbf{a} + \mathbf{b}\ \ \mathbf{b} + \mathbf{c}\ \ \mathbf{c} + \mathbf{d})$
19. $(\mathbf{a} - \mathbf{c}\ \ \mathbf{b} - \mathbf{c}\ \ \mathbf{c})$, $(\mathbf{a}\ \ \mathbf{b}\ \ \mathbf{c})$
20. $(4\mathbf{a}\ \ 3\mathbf{b}\ \ 2\mathbf{c})$, $24(\mathbf{b}\ \ \mathbf{c}\ \ \mathbf{a})$

21. What properties of cross multiplication do Probs. 1, 3, 8, 10 illustrate?
22. Give the details of the proofs of (4) and (5).
23. Give the details of the proofs of (6) and (11).
24. **TEAM PROJECT. Useful Formulas for Two and More Vectors.** Prove (12)–(16), which are often useful in practical work, and illustrate each formula with two examples. *Hints.* For (13) choose Cartesian coordinates such that $\mathbf{d} = [d_1, 0, 0]$ and $\mathbf{c} = [c_1, c_2, 0]$. Show that each side of (13) then equals $[-b_2c_2d_1, b_1c_2d_1, 0]$, and give reasons why the two sides are then equal in any Cartesian coordinate system. For (14) and (15) use (13).

Formula (15) is called **Lagrange's identity.**

(12) $$|\mathbf{a} \times \mathbf{b}| = \sqrt{(\mathbf{a} \cdot \mathbf{a})(\mathbf{b} \cdot \mathbf{b}) - (\mathbf{a} \cdot \mathbf{b})^2}$$

(13) $$\mathbf{b} \times (\mathbf{c} \times \mathbf{d}) = (\mathbf{b} \cdot \mathbf{d})\mathbf{c} - (\mathbf{b} \cdot \mathbf{c})\mathbf{d}$$

(14) $$(\mathbf{a} \times \mathbf{b}) \times (\mathbf{c} \times \mathbf{d}) = (\mathbf{a}\ \ \mathbf{b}\ \ \mathbf{d})\mathbf{c} - (\mathbf{a}\ \ \mathbf{b}\ \ \mathbf{c})\mathbf{d}$$

(15) $$(\mathbf{a} \times \mathbf{b}) \cdot (\mathbf{c} \times \mathbf{d}) = (\mathbf{a} \cdot \mathbf{c})(\mathbf{b} \cdot \mathbf{d}) - (\mathbf{a} \cdot \mathbf{d})(\mathbf{b} \cdot \mathbf{c})$$

(16) $$(\mathbf{a}\ \ \mathbf{b}\ \ \mathbf{c}) = (\mathbf{b}\ \ \mathbf{c}\ \ \mathbf{a}) = (\mathbf{c}\ \ \mathbf{a}\ \ \mathbf{b}) = -(\mathbf{c}\ \ \mathbf{b}\ \ \mathbf{a}) = -(\mathbf{a}\ \ \mathbf{c}\ \ \mathbf{b})$$

25–28 MOMENT OF A FORCE

Find the moment vector $\mathbf{m}$ and the moment m of a force $\mathbf{p}$ about a point Q when $\mathbf{p}$ acts on a line through A.

25. $\mathbf{p} = [4, 4, 0]$, *Q:* (2, 1, 0), *A:* (0, 3, 0)
26. $\mathbf{p} = [0, 0, 5]$, *Q:* (3, 3, 0), *A:* (0, 0, 0)
27. $\mathbf{p} = [1, 2, 3]$, *Q:* (0, 1, 1), *A:* (1, 0, 3)
28. $\mathbf{p} = [4, 12, 8]$, *Q:* (3, 0, 5), *A:* (4, 3, 7)

29. **(Rotation)** A wheel is rotating about the y-axis with angular speed $\omega = 10\ \text{sec}^{-1}$. The rotation appears clockwise if one looks from the origin in the positive y-direction. Find the velocity and speed at the point (4, 3, 0).
30. **(Rotation)** What are the velocity and speed in Prob. 29 at the point (4, 2, −2) if the wheel rotates about the line $y = x$, $z = 0$ with $\omega = 5\ \text{sec}^{-1}$.

GEOMETRIC APPLICATIONS

31. **(Parallelogram)** Find the area if the vertices are (2, 2), (9, 2), (10, 3), (3, 3).

32. (Parallelogram) Find the area if the vertices are (3, 9, 8), (0, 5, 1), (−1, −3, −3), (2, 1, 4).

33. (Triangle) Find the area if the vertices are (1, 0, 0), (0, 1, 0), (0, 0, 1).

34. (Triangle) Find the area if the vertices are (4, 6, 5), (4, 9, 5), (8, 6, 7).

35. (Plane) Find a normal vector and a representation of the plane through the points (4, 8, 0), (0, 2, 6), (3, 0, 5).

36. (Plane) Find the plane through (2, 1, 3), (4, 4, 5), (1, 6, 0).

37. (Parallelepiped) Find the volume of the parallelepiped determined by the vertices (1, 1, 1), (4, 7, 2), (3, 2, 1), (5, 4, 3).

38. (Tetrahedron) Find the volume of the tetrahedron with vertices (0, 2, 1), (4, 3, 0), (6, 6, 5), (4, 7, 8).

39. (Linear dependence) For what c are the vectors [9, 1, 2], [−1, c, 5], [4, c, 5] linearly dependent?

40. WRITING PROJECT. Applications of Cross Products. Summarize the most important applications we have discussed in this section and give a few simple examples. No proofs.

9.4 Vector and Scalar Functions and Fields. Derivatives

We now begin with vector calculus. This calculus concerns two kinds of functions, namely, **vector functions,** whose values are vectors

$$\mathbf{v} = \mathbf{v}(P) = [v_1(P), v_2(P), v_3(P)]$$

depending on the points P in space, and **scalar functions,** whose values are scalars

$$f = f(P)$$

depending on P. Here, P is a point in the domain of definition, which in applications is a (three-dimensional) domain or a surface or a curve in space. We say that a vector function defines a **vector field,** and a scalar function defines a **scalar field** in that domain or on that surface or curve. Examples of vector functions are shown in Figs. 193–196. Examples of scalar fields are the temperature field in a body or the pressure field of the air in the earth's atmosphere. Vector and scalar functions may also depend on time t or on some other parameters.

Notation. If we introduce Cartesian coordinates x, y, z, then instead of $\mathbf{v}(P)$ we can also write

$$\mathbf{v}(x, y, z) = [v_1(x, y, z), \quad v_2(x, y, z), \quad v_3(x, y, z)],$$

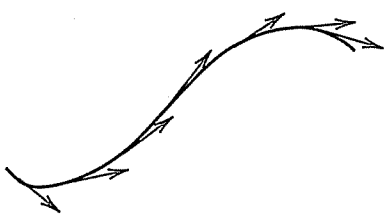


Fig. 193. Field of tangent vectors of a curve

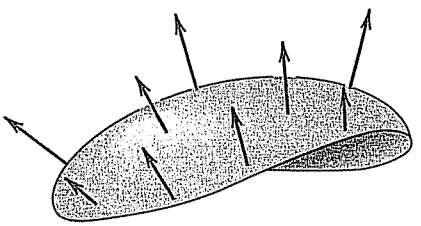


Fig. 194. Field of normal vectors of a surface

but we keep in mind that components depend on the choice of a coordinate system, whereas a vector field that has a physical or a geometric meaning should have magnitude and direction depending only on P, not on that choice. Similarly for the value of a scalar field $f(P) = f(x, y, z)$.

EXAMPLE 1 **Scalar Function (Euclidean Distance in Space)**

The distance $f(P)$ of any point P from a fixed point P_0 in space is a scalar function whose domain of definition is the whole space. $f(P)$ defines a scalar field in space. If we introduce a Cartesian coordinate system and P_0 has the coordinates x_0, y_0, z_0, then f is given by the well-known formula

$$f(P) = f(x, y, z) = \sqrt{(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2}$$

where x, y, z are the coordinates of P. If we replace the given Cartesian coordinate system with another such system by translating and rotating the given system, then the values of the coordinates of P and P_0 will in general change, but $f(P)$ will have the same value as before. Hence $f(P)$ is a scalar function. The direction cosines of the straight line through P and P_0 are not scalars because their values depend on the choice of the coordinate system. ▩

EXAMPLE 2 **Vector Field (Velocity Field)**

At any instant the velocity vectors $\mathbf{v}(P)$ of a rotating body B constitute a vector field, called the **velocity field** of the rotation. If we introduce a Cartesian coordinate system having the origin on the axis of rotation, then (see Example 5 in Sec. 9.3)

$$\mathbf{v}(x, y, z) = \mathbf{w} \times \mathbf{r} = \mathbf{w} \times [x, y, z] = \mathbf{w} \times (x\mathbf{i} + y\mathbf{j} + z\mathbf{k}) \tag{1}$$

where x, y, z are the coordinates of any point P of B at the instant under consideration. If the coordinates are such that the z-axis is the axis of rotation and $\mathbf{w}$ points in the positive z-direction, then $\mathbf{w} = \omega\mathbf{k}$ and

$$\mathbf{v} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 0 & 0 & \omega \\ x & y & z \end{vmatrix} = \omega[-y, x, 0] = \omega(-y\mathbf{i} + x\mathbf{j}).$$

An example of a rotating body and the corresponding velocity field are shown in Fig. 195. ▩

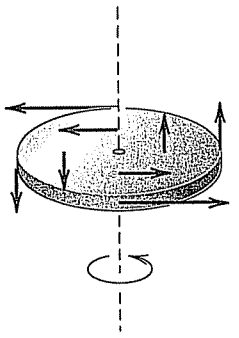


Fig. 195. Velocity field of a rotating body

EXAMPLE 3 **Vector Field (Field of Force, Gravitational Field)**

Let a particle A of mass M be fixed at a point P_0 and let a particle B of mass m be free to take up various positions P in space. Then A attracts B. According to **Newton's law of gravitation** the corresponding gravitational force $\mathbf{p}$ is directed from P to P_0, and its magnitude is proportional to $1/r^2$, where r is the distance between P and P_0, say,

$$|\mathbf{p}| = \frac{c}{r^2}, \qquad c = GMm. \tag{2}$$

Here $G = 6.67 \cdot 10^{-8}$ cm^3/(gm $\cdot$ sec^2) is the gravitational constant. Hence $\mathbf{p}$ defines a vector field in space. If we introduce Cartesian coordinates such that P_0 has the coordinates x_0, y_0, z_0 and P has the coordinates x, y, z, then by the Pythagorean theorem,

$$r = \sqrt{(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2} \qquad (\geqq 0).$$

Assuming that $r > 0$ and introducing the vector

$$\mathbf{r} = [x - x_0, \quad y - y_0, \quad z - z_0] = (x - x_0)\mathbf{i} + (y - y_0)\mathbf{j} + (z - z_0)\mathbf{k},$$

we have $|\mathbf{r}| = r$, and $(-1/r)\mathbf{r}$ is a unit vector in the direction of $\mathbf{p}$; the minus sign indicates that $\mathbf{p}$ is directed from P to P_0 (Fig. 196). From this and (2) we obtain

(3)
$$\mathbf{p} = |\mathbf{p}| \left(-\frac{1}{r}\,\mathbf{r}\right) = -\frac{c}{r^3}\,\mathbf{r} = \left[-c\,\frac{x - x_0}{r^3}, \quad -c\,\frac{y - y_0}{r^3}, \quad -c\,\frac{z - z_0}{r^3}\right]$$
$$= -c\,\frac{x - x_0}{r^3}\,\mathbf{i} - c\,\frac{y - y_0}{r^3}\,\mathbf{j} - c\,\frac{z - z_0}{r^3}\,\mathbf{k}.$$

This vector function describes the gravitational force acting on B. ■

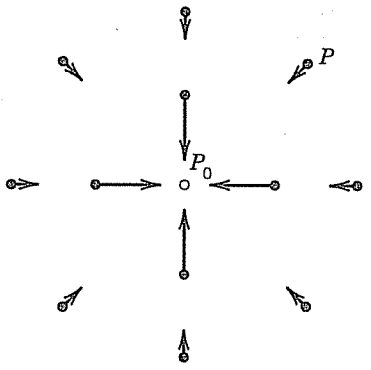


Fig. 196. Gravitational field in Example 3

Vector Calculus

We show next that the basic concepts of calculus, namely, convergence, continuity, and differentiability, can be defined for vector functions in a simple and natural way. Most important here is the derivative.

Convergence. An infinite sequence of vectors $\mathbf{a}_{(n)}$, $n = 1, 2, \cdots$, is said to **converge** if there is a vector $\mathbf{a}$ such that

(4)
$$\lim_{n \to \infty} |\mathbf{a}_{(n)} - \mathbf{a}| = 0.$$

$\mathbf{a}$ is called the **limit vector** of that sequence, and we write

(5)
$$\lim_{n \to \infty} \mathbf{a}_{(n)} = \mathbf{a}.$$

Cartesian coordinates being given, this sequence of vectors converges to $\mathbf{a}$ if and only if the three sequences of components of the vectors converge to the corresponding components of $\mathbf{a}$. We leave the simple proof to the student.

Similarly, a vector function $\mathbf{v}(t)$ of a real variable t is said to have the **limit** $\boldsymbol{l}$ as t approaches t_0, if $\mathbf{v}(t)$ is defined in some neighborhood of t_0 (possibly except at t_0) and

$$\lim_{t \to t_0} |\mathbf{v}(t) - \boldsymbol{l}| = 0. \tag{6}$$

Then we write

$$\lim_{t \to t_0} \mathbf{v}(t) = \boldsymbol{l}. \tag{7}$$

Here, a *neighborhood* of t_0 is an interval (segment) on the t-axis containing t_0 as an interior point (not as an endpoint).

Continuity. A vector function $\mathbf{v}(t)$ is said to be **continuous** at $t = t_0$ if it is defined in some neighborhood of t_0 (including at t_0 itself!) and

$$\lim_{t \to t_0} \mathbf{v}(t) = \mathbf{v}(t_0). \tag{8}$$

If we introduce a Cartesian coordinate system, we may write

$$\mathbf{v}(t) = [v_1(t), v_2(t), v_3(t)] = v_1(t)\mathbf{i} + v_2(t)\mathbf{j} + v_3(t)\mathbf{k}.$$

Then $\mathbf{v}(t)$ is continuous at t_0 if and only if its three components are continuous at t_0.

We now state the most important of these definitions.

DEFINITION

Derivative of a Vector Function

A vector function $\mathbf{v}(t)$ is said to be **differentiable** at a point t if the following limit exists:

$$\mathbf{v}'(t) = \lim_{\Delta t \to 0} \frac{\mathbf{v}(t + \Delta t) - \mathbf{v}(t)}{\Delta t}. \tag{9}$$

This vector $\mathbf{v}'(t)$ is called the **derivative** of $\mathbf{v}(t)$. See Fig. 197.

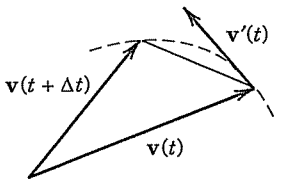


Fig. 197. Derivative of a vector function

In components with respect to a given Cartesian coordinate system,

$$\mathbf{v}'(t) = [v_1'(t), \quad v_2'(t), \quad v_3'(t)]. \tag{10}$$

Hence the derivative $\mathbf{v}'(t)$ ***is obtained by differentiating each component separately.*** For instance, if $\mathbf{v} = [t, t^2, 0]$, then $\mathbf{v}' = [1, 2t, 0]$.

Equation (10) follows from (9) and conversely because (9) is a "vector form" of the usual formula of calculus by which the derivative of a function of a single variable is defined. [The curve in Fig. 197 is the locus of the terminal points representing $\mathbf{v}(t)$ for values of the independent variable in some interval containing t and $t + \Delta t$ in (9)]. It follows that the familiar differentiation rules continue to hold for differentiating vector functions, for instance,

$$(c\mathbf{v})' = c\mathbf{v}' \qquad (c \text{ constant}),$$

$$(\mathbf{u} + \mathbf{v})' = \mathbf{u}' + \mathbf{v}'$$

and in particular

$$(\mathbf{u}\cdot\mathbf{v})' = \mathbf{u}'\cdot\mathbf{v} + \mathbf{u}\cdot\mathbf{v}' \tag{11}$$

$$(\mathbf{u}\times\mathbf{v})' = \mathbf{u}'\times\mathbf{v} + \mathbf{u}\times\mathbf{v}' \tag{12}$$

$$(\mathbf{u}\ \ \mathbf{v}\ \ \mathbf{w})' = (\mathbf{u}'\ \ \mathbf{v}\ \ \mathbf{w}) + (\mathbf{u}\ \ \mathbf{v}'\ \ \mathbf{w}) + (\mathbf{u}\ \ \mathbf{v}\ \ \mathbf{w}'). \tag{13}$$

The simple proofs are left to the student. In (12), note the order of the vectors carefully because cross multiplication is not commutative.

EXAMPLE 4 Derivative of a Vector Function of Constant Length

Let $\mathbf{v}(t)$ be a vector function whose length is constant, say, $|\mathbf{v}(t)| = c$. Then $|\mathbf{v}|^2 = \mathbf{v}\cdot\mathbf{v} = c^2$, and $(\mathbf{v}\cdot\mathbf{v})' = 2\mathbf{v}\cdot\mathbf{v}' = 0$, by differentiation [see (11)]. This yields the following result. *The derivative of a vector function* $\mathbf{v}(t)$ *of constant length is either the zero vector or is perpendicular to* $\mathbf{v}(t)$. ■

Partial Derivatives of a Vector Function

Our present discussion shows that partial differentiation of vector functions of two or more variables can be introduced as follows. Suppose that the components of a vector function

$$\mathbf{v} = [v_1,\ \ v_2,\ \ v_3] = v_1\mathbf{i} + v_2\mathbf{j} + v_3\mathbf{k}$$

are differentiable functions of n variables $t_1, \cdots, t_n$. Then the **partial derivative** of $\mathbf{v}$ with respect to t_m is denoted by $\partial\mathbf{v}/\partial t_m$ and is defined as the vector function

$$\frac{\partial\mathbf{v}}{\partial t_m} = \frac{\partial v_1}{\partial t_m}\mathbf{i} + \frac{\partial v_2}{\partial t_m}\mathbf{j} + \frac{\partial v_3}{\partial t_m}\mathbf{k}.$$

Similarly, second partial derivatives are

$$\frac{\partial^2\mathbf{v}}{\partial t_l\,\partial t_m} = \frac{\partial^2 v_1}{\partial t_l\,\partial t_m}\mathbf{i} + \frac{\partial^2 v_2}{\partial t_l\,\partial t_m}\mathbf{j} + \frac{\partial^2 v_3}{\partial t_l\,\partial t_m}\mathbf{k},$$

and so on.

EXAMPLE 5 Partial Derivatives

Let $\mathbf{r}(t_1, t_2) = a\cos t_1\,\mathbf{i} + a\sin t_1\,\mathbf{j} + t_2\mathbf{k}$. Then $\dfrac{\partial\mathbf{r}}{\partial t_1} = -a\sin t_1\,\mathbf{i} + a\cos t_1\,\mathbf{j}$ and $\dfrac{\partial\mathbf{r}}{\partial t_2} = \mathbf{k}$. ■

Various physical and geometric applications of derivatives of vector functions will be discussed in the next sections and in Chap. 10.

PROBLEM SET 9.4

1–6 SCALAR FIELDS

Determine the **isotherms** (curves of constant temperature T) of the temperature fields in the plane given by the following scalar functions. Sketch some isotherms.

1. $T = xy$
2. $T = 4x - 3y$
3. $T = y^2 - x^2$
4. $T = x/(x^2 + y^2)$
5. $T = y/(x^2 + y^2)$
6. $T = x^2 - y^2 + 8y$

7. **(Isobars)** For the pressure field $f(x, y) = 9x^2 + 16y^2$ find the *isobars* $f(x, y) = const$, the pressure at (4, 3), (−2, 2), (1, 5), and the region in which the pressure is between 4 and 16.

8. **CAS PROJECT. Scalar Fields in the Plane.** Sketch or graph isotherms of the following fields and describe what they look like.
   (a) $x^2 - 4x - y^2$ (b) $x^2y - y^3/3$
   (c) $\cos x \sinh y$ (d) $\sin x \sinh y$
   (e) $e^x \sin y$ (f) $e^{2x} \cos 2y$
   (g) $x^4 - 6x^2y^2 + y^4$ (h) $x^2 - 2x - y^2$

9–14 SCALAR FIELDS IN SPACE

What kind of surfaces are the level surfaces $f(x, y, z) = const$?

9. $f = x^2 + y^2 + 4z^2$
10. $f = x^2 + 4y^2$
11. $f = z - \sqrt{x^2 + y^2}$
12. $f = 4y^2 - z$
13. $f = 4x + 3y - 5z$
14. $f = z - x^2 - 4y^2$

15–20 VECTOR FIELDS

Sketch figures similar to Fig. 196.

15. $\mathbf{v} = \mathbf{i} - \mathbf{j}$
16. $\mathbf{v} = y\mathbf{i} + x\mathbf{j}$
17. $\mathbf{v} = \mathbf{i} + x^2\mathbf{j}$
18. $\mathbf{v} = x\mathbf{i} + y\mathbf{j}$
19. $\mathbf{v} = y\mathbf{i} - x\mathbf{j}$
20. $\mathbf{v} = (x - y)\mathbf{i} + (x + y)\mathbf{j}$

21–25 DIFFERENTIATION

21. Prove (11)–(13). Give two examples for each formula.
22. Find the first and second derivatives of $[4 \cos t, 4 \sin t, 2t]$.
23. Find the first partial derivatives of $[4x^2, 9z^2, xyz]$ and $[yz, zx, xy]$.
24. Find the first partial derivatives of $[\sin x \cosh y, \cos x \sinh y]$ and $[e^x \cos y, e^x \sin y]$.
25. **WRITING PROJECT. Differentiation of Vector Functions.** Summarize the essential ideas and facts and give examples of your own.

9.5 Curves. Arc Length. Curvature. Torsion

A major application of vector calculus concerns curves (this section) and surfaces (Sec. 10.5) and their use in physics and geometry. This field is called **differential geometry.** It plays a role in mechanics, computer-aided and traditional engineering design, geodesy and geography, space travel, and relativity theory (see Refs. [GR8], [GR9] in App. 1).

Curves C in space may occur as paths of moving bodies. This and other applications motivate **parametric representations** with **parameter** t, which may be time or something else (see Fig. 198)

(1) $$\mathbf{r}(t) = [x(t), \ y(t), \ z(t)] = x(t)\mathbf{i} + y(t)\mathbf{j} + z(t)\mathbf{k}.$$

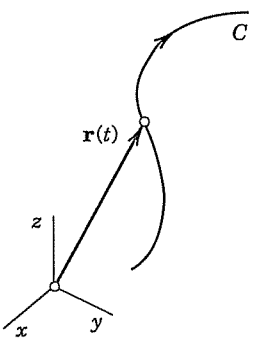


Fig. 198. Parametric representation of a curve

Here x, y, z are Cartesian coordinates (the usual rectangular coordinates; see Sec. 9.1). To each value $t = t_0$ there corresponds a point of C with position vector $\mathbf{r}(t_0)$, that is, with coordinates $x(t_0)$, $y(t_0)$, $z(t_0)$.

Parametric representations (1) have a key advantage over representations of a curve C in terms of its projections into the xy-plane and into the xz-plane, that is,

$$y = f(x), \qquad z = g(x) \tag{2}$$

(or by a pair of equations with y or with z as the independent variable). The advantage is that in (1) the coordinates x, y, z play the same role: all three are *dependent* variables. Moreover, the sense of increasing t, called the **positive sense** on C, induces an **orientation** of C, a direction of travel along C. The sense of decreasing t is then called the *negative sense* on C, given by (1).

EXAMPLE 1 **Circle**

The circle $x^2 + y^2 = 4$, $z = 0$ in the xy-plane with center 0 and radius 2 can be represented parametrically by

$$\mathbf{r}(t) = [2\cos t,\ 2\sin t,\ 0] \qquad \text{or simply by} \qquad \mathbf{r}(t) = [2\cos t,\ 2\sin t] \qquad \text{(Fig. 199)}$$

where $0 \leqq t \leqq 2\pi$. Indeed, $x^2 + y^2 = (2\cos t)^2 + (2\sin t)^2 = 4(\cos^2 t + \sin^2 t) = 4$. For $t = 0$ we have $\mathbf{r}(0) = [2, 0]$, for $t = \frac{1}{2}\pi$ we get $\mathbf{r}(\frac{1}{2}\pi) = [0, 2]$, and so on. The positive sense induced by this representation is the counterclockwise sense.

If we replace t with $t^* = -t$, we have $t = -t^*$ and get

$$\mathbf{r}^*(t^*) = [2\cos(-t^*),\ 2\sin(-t^*)] = [2\cos t^*,\ -2\sin t^*].$$

This has reversed the orientation, and the circle is now oriented clockwise. ■

EXAMPLE 2 **Ellipse**

The vector function

$$\mathbf{r}(t) = [a\cos t,\quad b\sin t,\quad 0] = a\cos t\ \mathbf{i} + b\sin t\ \mathbf{j} \tag{3}$$

(Fig. 200)

represents an ellipse in the xy-plane with center at the origin and principal axes in the direction of the x and y axes. In fact, since $\cos^2 t + \sin^2 t = 1$, we obtain from (3)

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1, \qquad z = 0.$$

If $b = a$, then (3) represents a *circle* of radius a. ■

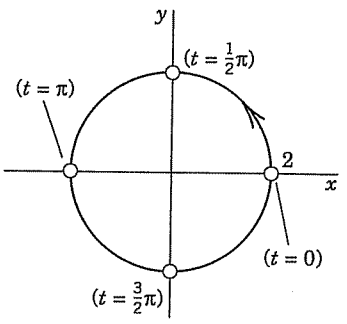


Fig. 199. Circle in Example 1

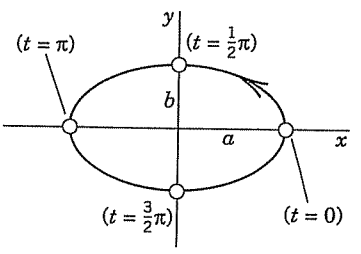


Fig. 200. Ellipse in Example 2

EXAMPLE 3 **Straight Line**

A straight line L through a point A with position vector $\mathbf{a}$ in the direction of a constant vector $\mathbf{b}$ (see Fig. 201) can be represented parametrically in the form

$$\mathbf{r}(t) = \mathbf{a} + t\mathbf{b} = [a_1 + tb_1, \quad a_2 + tb_2, \quad a_3 + tb_3]. \tag{4}$$

If $\mathbf{b}$ is a unit vector, its components are the **direction cosines** of L. In this case, $|t|$ measures the distance of the points of L from A. For instance, the straight line in the xy-plane through A: (3, 2) having slope 1 is (sketch it)

$$\mathbf{r}(t) = [3, \quad 2, \quad 0] + t[1, \quad 1, \quad 0] = [3 + t, \quad 2 + t, \quad 0].$$

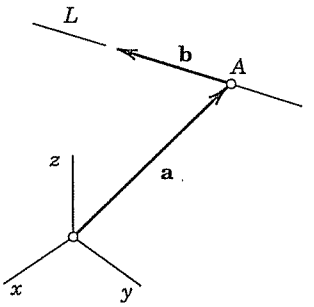


Fig. 201. Parametric representation of a straight line

A **plane curve** is a curve that lies in a plane in space. A curve that is not plane is called a **twisted curve.** A standard example of a twisted curve is the following.

EXAMPLE 4 **Circular Helix**

The twisted curve C represented by the vector function

$$\mathbf{r}(t) = [a\cos t, \quad a\sin t, \quad ct] = a\cos t\,\mathbf{i} + a\sin t\,\mathbf{j} + ct\,\mathbf{k} \qquad (c \neq 0) \tag{5}$$

is called a *circular helix.* It lies on the cylinder $x^2 + y^2 = a^2$. If $c > 0$, the helix is shaped like a right-handed screw (Fig. 202). If $c < 0$, it looks like a left-handed screw (Fig. 203). If $c = 0$, then (5) is a circle.

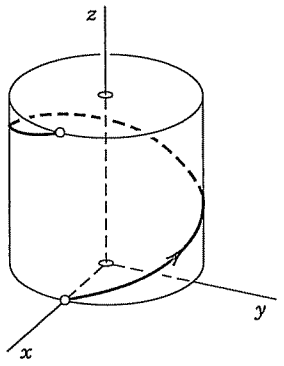


Fig. 202. Right-handed circular helix

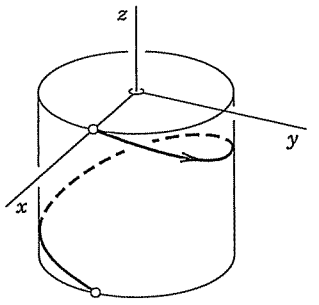


Fig. 203. Left-handed circular helix

A **simple curve** is a curve without **multiple points,** that is, without points at which the curve intersects or touches itself. Circle and helix are simple. Figure 204 shows curves that are not simple. An example is $[\sin 2t, \quad \cos t, \quad 0]$. Can you sketch it?

An **arc** of a curve is the portion between any two points of the curve. For simplicity, we say "curve" for curves as well as for arcs.

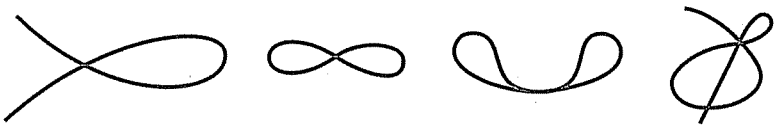


Fig. 204. Curves with multiple points

Tangent to a Curve

The next idea is the approximation of a curve by straight lines, leading to tangents and to a definition of length. Tangents are straight lines touching a curve. The **tangent** to a simple curve C at a point P of C is the limiting position of a straight line L through P and a point Q of C as Q approaches P along C. See Fig. 205.

If C is given by $\mathbf{r}(t)$, and P and Q correspond to t and $t + \Delta t$, then a vector in the direction of L is

$$\frac{1}{\Delta t}\,[\mathbf{r}(t + \Delta t) - \mathbf{r}(t)]. \tag{6}$$

In the limit this vector becomes the derivative

$$\mathbf{r}'(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t}\,[\mathbf{r}(t + \Delta t) - \mathbf{r}(t)], \tag{7}$$

provided $\mathbf{r}(t)$ is differentiable, as we shall assume from now on. If $\mathbf{r}'(t) \neq \mathbf{0}$, we call $\mathbf{r}'(t)$ a **tangent vector** of C at P because it has the direction of the tangent. The corresponding unit vector is the **unit tangent vector** (see Fig. 205)

$$\mathbf{u} = \frac{1}{|\mathbf{r}'|}\,\mathbf{r}'. \tag{8}$$

Note that both $\mathbf{r}'$ and $\mathbf{u}$ point in the direction of increasing t. Hence their sense depends on the orientation of C. It is reversed if we reverse the orientation.

It is now easy to see that the **tangent** to C at P is given by

$$\mathbf{q}(w) = \mathbf{r} + w\mathbf{r}' \qquad \text{(Fig. 206).} \tag{9}$$

This is the sum of the position vector $\mathbf{r}$ of P and a multiple of the tangent vector $\mathbf{r}'$ of C at P. Both vectors depend on P. The variable w is the parameter in (9).

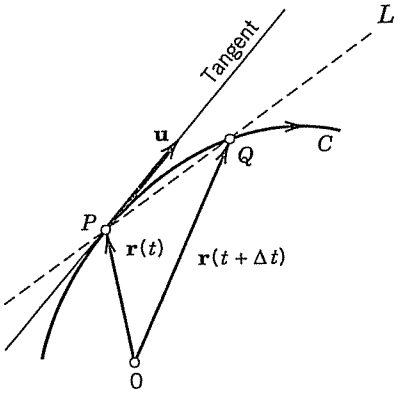


Fig. 205. Tangent to a curve

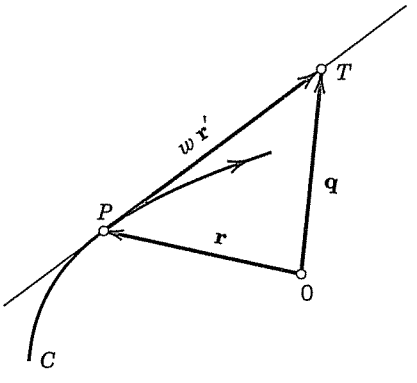


Fig. 206. Formula (9) for the tangent to a curve

EXAMPLE 5 Tangent to an Ellipse

Find the tangent to the ellipse $\frac{1}{4}x^2 + y^2 = 1$ at P: $(\sqrt{2}, 1/\sqrt{2})$.

Solution. Equation (3) with semi-axes $a = 2$ and $b = 1$ gives $\mathbf{r}(t) = [2\cos t, \sin t]$. The derivative is $\mathbf{r}'(t) = [-2\sin t, \cos t]$. Now P corresponds to $t = \pi/4$ because

$$\mathbf{r}(\pi/4) = [2\cos(\pi/4), \sin(\pi/4)] = [\sqrt{2}, \quad 1/\sqrt{2}].$$

Hence $\mathbf{r}'(\pi/4) = [-\sqrt{2}, \quad 1/\sqrt{2}]$. From (9) we thus get the *answer*

$$\mathbf{q}(w) = [\sqrt{2}, \quad 1/\sqrt{2}] + w[-\sqrt{2}, \quad 1/\sqrt{2}] = [\sqrt{2}(1 - w), \quad (1/\sqrt{2})(1 + w)].$$

To check the result, sketch or graph the ellipse and the tangent. ▩

Length of a Curve

We are now ready to define the length l of a curve. l will be the limit of the lengths of broken lines of n chords (see Fig. 207, where $n = 5$) with larger and larger n. For this, let $\mathbf{r}(t)$, $a \leqq t \leqq b$, represent C. For each $n = 1, 2, \cdots$ we subdivide ("partition") the interval $a \leqq t \leqq b$ by points

$$t_0 (= a), \quad t_1, \quad \cdots, \quad t_{n-1}, \quad t_n (= b), \qquad \text{where} \qquad t_0 < t_1 < \cdots < t_n.$$

This gives a broken line of chords with endpoints $\mathbf{r}(t_0), \cdots, \mathbf{r}(t_n)$. We do this arbitrarily but so that the greatest $|\Delta t_m| = |t_m - t_{m-1}|$ approaches 0 as $n \to \infty$. The lengths $l_1, l_2, \cdots$ of these chords can be obtained from the Pythagorean theorem. If $\mathbf{r}(t)$ has a continuous derivative $\mathbf{r}'(t)$, it can be shown that the sequence $l_1, l_2, \cdots$ has a limit, which is independent of the particular choice of the representation of C and of the choice of subdivisions. This limit is given by the integral

(10)
$$l = \int_a^b \sqrt{\mathbf{r}' \cdot \mathbf{r}'}\, dt \qquad \left(\mathbf{r}' = \frac{d\mathbf{r}}{dt}\right).$$

l is called the **length** of C, and C is called **rectifiable.** Formula (10) is made plausible in calculus for plane curves and is proved for curves in space in [GR8] listed in App. 1. The practical evaluation of the integral (10) will be difficult in general. Some simple cases are given in the problem set.

Arc Length s of a Curve

The length (10) of a curve C is a constant, a positive number. But if we replace the fixed b in (10) with a variable t, the integral becomes a function of t, denoted by $s(t)$ and called the *arc length function* or simply the **arc length** of C. Thus

(11)
$$s(t) = \int_a^t \sqrt{\mathbf{r}' \cdot \mathbf{r}'}\, d\tilde{t} \qquad \left(\mathbf{r}' = \frac{d\mathbf{r}}{d\tilde{t}}\right).$$

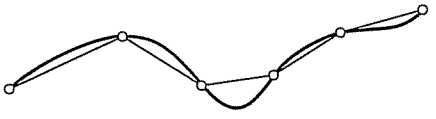


Fig. 207. Length of a curve.

Here the variable of integration is denoted by $\tilde{t}$ because t is now used in the upper limit.

Geometrically, $s(t_0)$ with some $t_0 > a$ is the length of the arc of C between the points with parametric values a and t_0. The choice of a (the point $s = 0$) is arbitrary; changing a means changing s by a constant.

Linear Element *ds*. If we differentiate (11) and square, we have

$$\text{(12)} \qquad \left(\frac{ds}{dt}\right)^2 = \frac{d\mathbf{r}}{dt} \cdot \frac{d\mathbf{r}}{dt} = |\mathbf{r}'(t)|^2 = \left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2.$$

It is customary to write

$$\text{(13*)} \qquad d\mathbf{r} = [dx, dy, dz] = dx\,\mathbf{i} + dy\,\mathbf{j} + dz\,\mathbf{k}$$

and

$$\text{(13)} \qquad ds^2 = d\mathbf{r} \cdot d\mathbf{r} = dx^2 + dy^2 + dz^2.$$

ds is called the **linear element** of C.

Arc Length as Parameter. The use of s in (1) instead of an arbitrary t simplifies various formulas. For the unit tangent vector (8) we simply obtain

$$\text{(14)} \qquad \mathbf{u}(s) = \mathbf{r}'(s).$$

Indeed, $|\mathbf{r}'(s)| = (ds/ds) = 1$ in (12) shows that $\mathbf{r}'(s)$ is a unit vector. Even greater simplifications due to the use of s will occur in curvature and torsion (below).

EXAMPLE 6 **Circular Helix. Circle. Arc Length as Parameter**

The **helix** $\mathbf{r}(t) = [a\cos t, a\sin t, ct]$ in (5) has the derivative $\mathbf{r}'(t) = [-a\sin t, a\cos t, c]$. Hence $\mathbf{r}' \cdot \mathbf{r}' = a^2 + c^2$, a constant, which we denote by K^2. Hence the integrand in (11) is constant, equal to K, and the integral is $s = Kt$. Thus $t = s/K$, so that a representation of the helix with the arc length s as parameter is

$$\text{(15)} \qquad \mathbf{r}^*(s) = \mathbf{r}\left(\frac{s}{K}\right) = \left[a\cos\frac{s}{K},\ a\sin\frac{s}{K},\ \frac{cs}{K}\right], \qquad K = \sqrt{a^2 + c^2}.$$

A **circle** is obtained if we set $c = 0$. Then $K = a$, $t = s/a$, and a representation with arc length s as parameter is

$$\mathbf{r}^*(s) = \mathbf{r}\left(\frac{s}{a}\right) = \left[a\cos\frac{s}{a},\ a\sin\frac{s}{a}\right].$$

Curves in Mechanics. Velocity. Acceleration

Curves play a basic role in mechanics, where they may serve as paths of moving bodies. Then such a curve C should be represented by a parametric representation $\mathbf{r}(t)$ with **time** t as parameter. The tangent vector (7) of C is then called the **velocity vector** $\mathbf{v}$ because, being tangent, it points in the instantaneous direction of motion and its length gives the **speed** $|\mathbf{v}| = |\mathbf{r}'| = \sqrt{\mathbf{r}' \cdot \mathbf{r}'} = ds/dt$; see (12). The second derivative of $\mathbf{r}(t)$ is called the

acceleration vector and is denoted by **a.** Its length $|\mathbf{a}|$ is called the **acceleration** of the motion. Thus

$$\mathbf{v}(t) = \mathbf{r}'(t), \qquad \mathbf{a}(t) = \mathbf{v}'(t) = \mathbf{r}''(t). \tag{16}$$

Tangential and Normal Acceleration. Whereas the velocity vector is always tangent to the path of motion, the acceleration vector will generally have another direction, so that it will be of the form

$$\mathbf{a} = \mathbf{a}_{\text{tan}} + \mathbf{a}_{\text{norm}}, \tag{17}$$

where the **tangential acceleration vector** $\mathbf{a}_{\text{tan}}$ is tangent to the path (or, sometimes, **0**) and the **normal acceleration vector** $\mathbf{a}_{\text{norm}}$ is normal (perpendicular) to the path (or, sometimes, **0**).

Expressions for the vectors in (17) are obtained from (16) by the chain rule. We first have

$$\mathbf{v}(t) = \frac{d\mathbf{r}}{dt} = \frac{d\mathbf{r}}{ds}\frac{ds}{dt} = \mathbf{u}(s)\frac{ds}{dt}$$

where $\mathbf{u}(s)$ is the unit tangent vector (14). Another differentiation gives

$$\mathbf{a}(t) = \frac{d\mathbf{v}}{dt} = \frac{d}{dt}\left(\mathbf{u}(s)\frac{ds}{dt}\right) = \frac{d\mathbf{u}}{ds}\left(\frac{ds}{dt}\right)^2 + \mathbf{u}(s)\frac{d^2s}{dt^2}. \tag{18}$$

Since the tangent vector $\mathbf{u}(s)$ has constant length (one), its derivative $d\mathbf{u}/ds$ is perpendicular to $\mathbf{u}(s)$ (by Example 4 in Sec. 9.4). Hence the first term on the right of (18) is the normal acceleration vector, and the second term on the right is the tangential acceleration vector, so that (18) is of the form (17).

Now the length of $\mathbf{a}_{\text{tan}}$ is the projection of $\mathbf{a}$ in the direction of $\mathbf{v}$, given by (11) in Sec. 9.2 with $\mathbf{b} = \mathbf{v}$; that is, $|\mathbf{a}_{\text{tan}}| = \mathbf{a}\cdot\mathbf{v}/|\mathbf{v}|$. Hence $\mathbf{a}_{\text{tan}}$ is this expression times the unit vector $(1/|\mathbf{v}|)\mathbf{v}$ in the direction of $\mathbf{v}$; that is,

$$\mathbf{a}_{\text{tan}} = \frac{\mathbf{a}\cdot\mathbf{v}}{\mathbf{v}\cdot\mathbf{v}}\mathbf{v}. \qquad \text{Also,} \qquad \mathbf{a}_{\text{norm}} = \mathbf{a} - \mathbf{a}_{\text{tan}}. \tag{18*}$$

Let us consider two basic examples, involving *centripetal* and *centrifugal accelerations* and *Coriolis acceleration,* as it occurs, for instance, in space travel.

EXAMPLE 7 **Centripetal Acceleration. Centrifugal Force**

The vector function

$$\mathbf{r}(t) = [R\cos\omega t, \quad R\sin\omega t] = R\cos\omega t\,\mathbf{i} + R\sin\omega t\,\mathbf{j} \qquad \text{(Fig. 208)}$$

(with fixed $\mathbf{i}$ and $\mathbf{j}$) represents a circle C of radius R with center at the origin of the xy-plane and describes the motion of a small body B counterclockwise around the circle. Differentiation gives the velocity vector

$$\mathbf{v} = \mathbf{r}' = [-R\omega\sin\omega t, \quad R\omega\cos\omega t] = -R\omega\sin\omega t\,\mathbf{i} + R\omega\cos\omega t\,\mathbf{j} \qquad \text{(Fig. 208).}$$

$\mathbf{v}$ is tangent to C. Its magnitude, the speed, is

$$|\mathbf{v}| = |\mathbf{r}'| = \sqrt{\mathbf{r}'\cdot\mathbf{r}'} = R\omega.$$

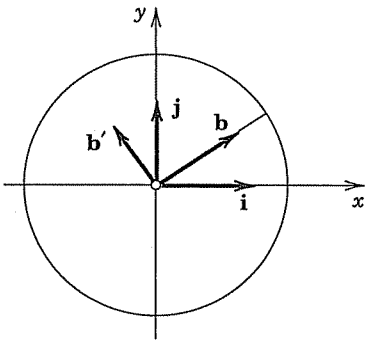


Fig. 208. Centripetal acceleration **a**

Hence it is constant. The speed divided by the distance R from the center is called the **angular speed.** It equals ω, so that it is constant, too. Differentiating the velocity vector, we obtain the acceleration vector

$$\mathbf{a} = \mathbf{v}' = [-R\omega^2 \cos \omega t, \quad -R\omega^2 \sin \omega t] = -R\omega^2 \cos \omega t\, \mathbf{i} - R\omega^2 \sin \omega t\, \mathbf{j}. \tag{19}$$

This shows that $\mathbf{a} = -\omega^2 \mathbf{r}$ (Fig. 208), so that there is an acceleration toward the center, called the **centripetal acceleration** of the motion. It occurs because the velocity vector is changing direction at a constant rate. Its magnitude is constant, $|\mathbf{a}| = \omega^2 |\mathbf{r}| = \omega^2 R$. Multiplying **a** by the mass m of B, we get the **centripetal force** $m\mathbf{a}$. The opposite vector $-m\mathbf{a}$ is called the **centrifugal force.** At each instant these two forces are in equilibrium.

We see that in this motion the acceleration vector is normal (perpendicular) to C; hence there is no tangential acceleration. ■

EXAMPLE 8 Superposition of Rotations. Coriolis Acceleration

A projectile is moving with constant speed along a meridian of the rotating earth in Fig. 209. Find its acceleration.

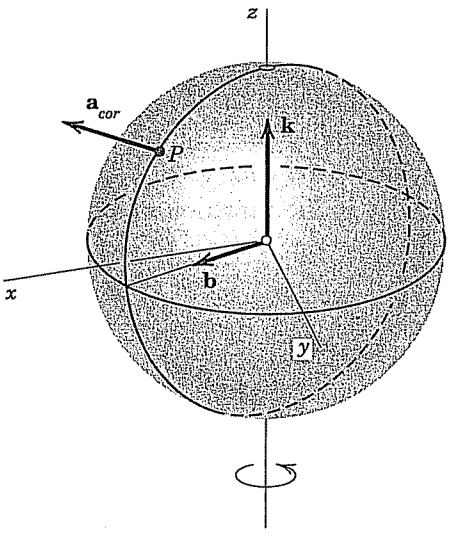


Fig. 209. Example 8. Superposition of two rotations

Solution. Let x, y, z be a fixed Cartesian coordinate system in space, with unit vectors **i, j, k** in the directions of the axes. Let the earth, together with a unit vector **b,** be rotating about the z-axis with angular speed $\omega > 0$ (see Example 7). Since **b** is rotaing together with the earth, it is of the form

$$\mathbf{b}(t) = \cos \omega t\, \mathbf{i} + \sin \omega t\, \mathbf{j}.$$

Let the projectile be moving on the meridian whose plane is spanned by **b** and **k** (Fig. 209) with constant angular speed $\gamma > 0$. Then its position vector in terms of **b** and **k** is

$$\mathbf{r}(t) = R \cos \gamma t\, \mathbf{b}(t) + R \sin \gamma t\, \mathbf{k} \qquad (R = \text{Radius of the earth}).$$

This is the model. The rest is calculation. The result will be unexpected and highly relevant for air and space travel. The first and second derivatives of $\mathbf{b}$ with respect to t are

$$\mathbf{b}'(t) = -\omega \sin \omega t\, \mathbf{i} + \omega \cos \omega t\, \mathbf{j}$$
$$\mathbf{b}''(t) = -\omega^2 \cos \omega t\, \mathbf{i} - \omega^2 \sin \omega t\, \mathbf{j} = -\omega^2 \mathbf{b}(t). \tag{20}$$

The first and second derivatives of $\mathbf{r}(t)$ with respect to t are

$$\mathbf{v} = \mathbf{r}'(t) = R \cos \gamma t\, \mathbf{b}' - \gamma R \sin \gamma t\, \mathbf{b} + \gamma R \cos \gamma t\, \mathbf{k}$$
$$\mathbf{a} = \mathbf{v}' = R \cos \gamma t\, \mathbf{b}'' - 2\gamma R \sin \gamma t\, \mathbf{b}' - \gamma^2 R \cos \gamma t\, \mathbf{b} - \gamma^2 R \sin \gamma t\, \mathbf{k} \tag{21}$$
$$= R \cos \gamma t\, \mathbf{b}'' - 2\gamma R \sin \gamma t\, \mathbf{b}' - \gamma^2 \mathbf{r}.$$

By analogy with Example 7 and because of $\mathbf{b}'' = -\omega^2 \mathbf{b}$ in (20) we conclude that the first term in $\mathbf{a}$ (involving ω in $\mathbf{b}''$!) is the centripetal acceleration due to the rotation of the earth. Similarly, the third term in the last line (involving γ!) is the centripetal acceleration due to the motion of the projectile on the meridian M of the rotating earth.

The second, unexpected term $-2\gamma R \sin \gamma t\, \mathbf{b}'$ in $\mathbf{a}$ is called the **Coriolis acceleration**[3] (Fig. 209) and is due to the interaction of the two rotations. On the Northern Hemisphere, $\sin \gamma t > 0$ (for $t > 0$; also $\gamma > 0$ by assumption), so that $\mathbf{a}_{\text{cor}}$ has the direction of $-\mathbf{b}'$, that is, opposite to the rotation of the earth. $|\mathbf{a}_{\text{cor}}|$ is maximum at the North Pole and zero at the equator. The projectile B of mass m_0 experiences a force $-m_0 \mathbf{a}_{\text{cor}}$ opposite to $m_0 \mathbf{a}_{\text{cor}}$, which tends to let B deviate from M to the right (and in the Southern Hemisphere, where $\sin \gamma t < 0$, to the left). This deviation has been observed for missiles, rockets, shells, and atmospheric air flow. ■

Curvature and Torsion. *Optional*

This optional portion of the section completes our discussion of curves from the viewpoint of vector calculus.

The **curvature** $\kappa(s)$ of a curve C: $\mathbf{r}(s)$ (s the arc length) at a point P of C measures the rate of change $|\mathbf{u}'(s)|$ of the unit tangent vector $\mathbf{u}(s)$ at P. Hence $\kappa(s)$ measures the deviation of C at P from a straight line (its tangent at P). Since $\mathbf{u}(s) = \mathbf{r}'(s)$, the definition is

$$\kappa(s) = |\mathbf{u}'(s)| = |\mathbf{r}''(s)| \qquad (' = d/ds). \tag{22}$$

The **torsion** $\tau(s)$ of C at P measures the rate of change of the **osculating plane** O (the plane spanned by $\mathbf{u}$ and $\mathbf{u}'$, see Fig. 210) of C at P. Hence $\tau(s)$ measures the deviation

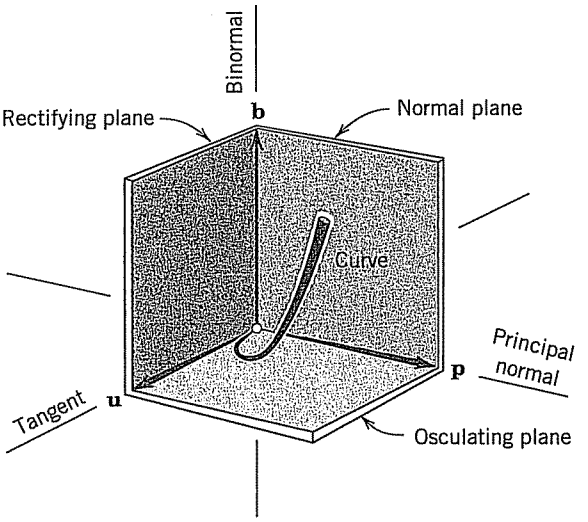


Fig. 210. Trihedron. Unit vectors **u, p, b** and planes

[3]GUSTAVE GASPARD CORIOLIS (1792–1843), French engineer who did research in mechanics.

of C at P from a plane (from O at P). Now the rate of change is also measured by the derivative $\mathbf{b}'$ of a normal vector $\mathbf{b}$ at O. By the definition of vector product, a unit normal vector of O is $\mathbf{b} = \mathbf{u} \times (1/\kappa)\mathbf{u}' = \mathbf{u} \times \mathbf{p}$, where $\mathbf{p} = (1/\kappa)\mathbf{u}'$ is called the **unit principal normal vector** and $\mathbf{b}$ is called the **unit binormal vector** of C at P; see Fig. 210. Here we must assume that $\kappa \neq 0$; hence $\kappa > 0$. The absolute value of the torsion is now defined by

$$|\tau(s)| = |\mathbf{b}'(s)|. \tag{23*}$$

Whereas $\kappa(s)$ is nonnegative, it is practical to give the torsion a sign, motivated by "right-handed" and "left-handed" (see Figs. 202, 203). This needs a little further calculation. Since $\mathbf{b}$ is a unit vector, it has constant length. Hence $\mathbf{b}'$ is perpendicular to $\mathbf{b}$ (see Example 4 in Sec. 9.4). Now $\mathbf{b}'$ is also perpendicular to $\mathbf{u}$ because by the definition of vector product we have $\mathbf{b} \bullet \mathbf{u} = 0$, $\mathbf{b} \bullet \mathbf{u}' = 0$. This implies

$$(\mathbf{b} \bullet \mathbf{u})' = 0; \qquad \text{that is,} \qquad \mathbf{b}' \bullet \mathbf{u} + \mathbf{b} \bullet \mathbf{u}' = \mathbf{b}' \bullet \mathbf{u} + 0 = 0.$$

Hence if $\mathbf{b}' \neq \mathbf{0}$ at P, it must have the direction of $\mathbf{p}$ or $-\mathbf{p}$, so that it must be of the form $\mathbf{b}' = -\tau\mathbf{p}$. Taking the dot product of this by $\mathbf{p}$ and using $\mathbf{p} \bullet \mathbf{p} = 1$ gives

$$\tau(s) = -\mathbf{p}(s) \bullet \mathbf{b}'(s). \tag{23}$$

The minus sign is chosen to make the torsion of a *right-handed* helix *positive* and that of a *left-handed* helix *negative* (Figs. 202, 203). The orthonormal vector triple **u, p, b** is called the **trihedron** of C. Figure 210 also shows the names of the three straight lines in the directions of **u, p, b,** which are the intersections of the **osculating plane,** the **normal plane,** and the **rectifying plane.**

PROBLEM SET 9.5

1–10 PARAMETRIC REPRESENTATIONS

Find a parametric representation of the following curves.

1. Circle of radius 3, center (4, 6)
2. Straight line through (5, 1, 2) and (11, 3, 0)
3. Straight line through (2, 0, 4) and (−3, 0, 9)
4. Straight line $y = 2x + 3$, $z = 7x$
5. Circle $y^2 + 4y + z^2 = 5$, $x = 3$
6. Ellipse $x^2 + y^2 = 1$, $z = y$
7. Straight line through (a, b, c) and $(a + 3, b - 2, c + 5)$
8. Intersection of $x + y - z = 2$, $2x - 5y + z = 3$
9. Circle $\frac{1}{2}x^2 + y^2 = 1$, $z = y$
10. Helix $x^2 + y^2 = 9$, $z = 4 \arctan(y/x)$

11–18 What curves are represented as follows?

11. $[2 + r\cos 4t, 6 + r\sin 4t, 2t]$
12. $[4 - 2t, 8t, -3 + 5t]$
13. $[2 + \cos 3t, -2 + \sin 3t, 5]$
14. $[t, t^2, t^3]$
15. $[\sqrt{\cos t}, \sqrt{\sin t}, 0]$ (*"Lamé curve"*)
16. $[\cosh t, \sinh t, 0]$
17. $[t, 1/t, 0]$
18. $[1, 5 + t, -5 + 1/t]$
19. Show that setting $t = -t^*$ reverses the orientation of $[a\cos t, a\sin t, 0]$.
20. If we set $t = e^{\tilde{t}}$ in Prob. 12, do we get the entire line? Explain.
21. **CAS PROJECT. Curves.** Graph the following more complicated curves.

    (a) $\mathbf{r}(t) = [2\cos t + \cos 2t, \quad 2\sin t - \sin 2t]$ (*Steiner's hypocycloid*)

    (b) $\mathbf{r}(t) = [\cos t + k\cos 2t, \quad \sin t - k\sin 2t]$ with $k = 10, 2, 1, \frac{1}{2}, 0, -\frac{1}{2}, -1$

    (c) $\mathbf{r}(t) = [\cos t, \quad \sin 5t]$ (a *Lissajous curve*)

    (d) $\mathbf{r}(t) = [\cos t, \quad \sin kt]$. For what k's will it be closed?

    (e) $\mathbf{r}(t) = [R\sin \omega t + \omega R t, \quad R\cos \omega t + R]$ (*cycloid*).

22–25 TANGENT

Given a curve C: $\mathbf{r}(t)$, find a tangent vector $\mathbf{r}'(t)$, a unit tangent vector $\mathbf{u}'(t)$, and the tangent of C at P. Sketch the curve and the tangent.

22. $\mathbf{r}(t) = [t, \quad t^2, \quad 0]$, P: (2, 4, 0)

23. $\mathbf{r}(t) = [5\cos t, \quad 5\sin t, \quad 0]$, P: (4, 3, 0)

24. $\mathbf{r}(t) = [3\cos t, \quad 3\sin t, \quad 4t]$, P: $(3, 0, 8\pi)$

25. $\mathbf{r}(t) = [\cosh t, \quad \sinh t]$, P: $(\frac{5}{3}, \frac{4}{3})$

26–28 LENGTH

Find the length and sketch the curve.

26. Circular helix $\mathbf{r}(t) = [2\cos t, \quad 2\sin t, \quad 6t]$ from (2, 0, 0) to $(2, 0, 24\pi)$

27. Catenary $\mathbf{r}(t) = [t, \quad \cosh t]$ from $t = 0$ to $t = 1$

28. Hypocycloid $\mathbf{r}(t) = [a\cos^3 t, \quad a\sin^3 t]$, total length

29. Show that (10) implies $\ell = \int_a^b \sqrt{1 + y'^2}\, dx$ for the length of a plane curve C: $y = f(x)$, $z = 0$, $a \leqq x \leqq b$.

30. Polar coordinates $\rho = \sqrt{x^2 + y^2}$, $\theta = \arctan(y/x)$ give $\ell = \int_\alpha^\beta \sqrt{\rho^2 + \rho'^2}\, d\theta$, where $\rho' = d\rho/d\theta$. Derive this. Use it to find the total length of the **cardioid** $\rho = a(1 - \cos\theta)$. Sketch this curve. *Hint.* Use (10) in App. 3.1.

31. CAS PROJECT. Polar Representations. Use your CAS to graph the following famous curves[4] and investigate their form depending on parameters a and b.

$\rho = a\theta$ *Spiral of Archimedes*

$\rho = ae^{b\theta}$ *Logarithmic spiral*

$\rho = \dfrac{2a\sin^2\theta}{\cos\theta}$ *Cissoid of Diocles*

$\rho = \dfrac{a}{\cos\theta} + b$ *Conchoid of Nicomedes*

$\rho = a/\theta$ *Hyperbolic spiral*

$\rho = \dfrac{3a\sin 2\theta}{\cos^3\theta + \sin^3\theta}$ *Folium of Descartes*

$\rho = 2a\dfrac{\sin 3\theta}{\sin 2\theta}$ *Maclaurin's trisectrix*

$\rho = 2a\cos\theta + b$ *Pascal's snail*

32–34 CURVES IN MECHANICS

Velocity and Acceleration. Forces on moving objects (cars, airplanes, etc.) require that the engineer knows corresponding **tangential** and **normal accelerations.** Find them, along with the **velocity** and **speed,** for the following motions. Sketch the path.

32. $\mathbf{r}(t) = [4t, \quad -3t, \quad 0]$

33. $\mathbf{r}(t) = [t, \quad t^2, \quad 0]$

34. $\mathbf{r}(t) = [\cos t, \quad 2\sin t, \quad 0]$

35. (Cycloid) Given

$$\mathbf{r}(t) = (R\sin\omega t + \omega R t)\mathbf{i} + (R\cos\omega t + R)\mathbf{j}.$$

This cycloid is the path of a point on the rim of a wheel of radius R that rolls without slipping along the x-axis. Find $\mathbf{v}$ and $\mathbf{a}$ at the maximum y-values of the curve.

36. CAS PROJECT. Paths of Motions. Gear transmissions and other engineering constructions often involve complicated paths whose study is greatly facilitated by the use of a CAS. To grasp the idea, graph the following paths and find the velocity, the speed, and the tangential and normal accelerations.

(a) $\mathbf{r}(t) = [2\cos t + \cos 2t, \quad 2\sin t - \sin 2t]$ *(Steiner's hypocycloid)*

(b) $\mathbf{r}(t) = [\cos t + \cos 2t, \quad \sin t - \sin 2t]$

(c) $\mathbf{r}(t) = [\cos t, \quad \sin 2t, \quad \cos 2t]$

(d) $\mathbf{r}(t) = [ct\cos t, \quad ct\sin t, \quad ct]$ $(c \neq 0)$

37. (Sun and earth) Find the acceleration of the earth toward the sun from (19) and the fact that the earth revolves about the sun in a nearly circular orbit with an almost constant speed of 30 km/sec.

38. (Earth and moon) Find the centripetal acceleration of the moon toward the earth, assuming that the orbit of the moon is a circle of radius 239,000 miles $= 3.85 \cdot 10^8$ m, and the time for one complete revolution is 27.3 days $= 2.36 \cdot 10^6$ sec.

39. (Satellite) Find the speed of an artificial earth satellite traveling at an altitude of 80 miles above the earth's surface, where $g = 31$ ft/sec^2. (The radius of the earth is 3960 miles.)

40. (Satellite) A satellite moves in a circular orbit 450 miles above the earth's surface and completes 1 revolution in 100 min. Find the acceleration of gravity at the orbit from these data and from the radius of the earth (3960 miles).

---

[4]Named after ARCHIMEDES (c. 287–212 B.C.), DESCARTES (Sec. 9.1), DIOCLES (200 B.C.), MACLAURIN (Sec. 15.4), NICOMEDES (250? B.C.) ÉTIENNE PASCAL (1588–1651), father of BLAISE PASCAL (1623–1662).

41–50 CURVATURE AND TORSION

41. Show that a circle of radius a has curvature $1/a$.

42. Using (22), show that if C is represented by $\mathbf{r}(t)$ with arbitrary t, then

$$(22^*) \quad \kappa(t) = \frac{\sqrt{(\mathbf{r}' \cdot \mathbf{r}')(\mathbf{r}'' \cdot \mathbf{r}'') - (\mathbf{r}' \cdot \mathbf{r}'')^2}}{(\mathbf{r}' \cdot \mathbf{r}')^{3/2}}.$$

43. Using (22*), show that for a curve $y = f(x)$ in the xy-plane,

$$(22^{**}) \quad \kappa(x) = \frac{|y''|}{(1 + y'^2)^{3/2}} \qquad \left(y' = \frac{dy}{dx}, \text{ etc.}\right).$$

44. Using $\mathbf{b} = \mathbf{u} \times \mathbf{p}$ and (23), show that

$$(23^{**}) \quad \tau(s) = (\mathbf{u} \quad \mathbf{p} \quad \mathbf{p}') = (\mathbf{r}' \quad \mathbf{r}'' \quad \mathbf{r}''')/\kappa^2 \qquad (\kappa > 0).$$

45. Show that the torsion of a plane curve (with $\kappa > 0$) is identically zero.

46. Show that if C is represented by $\mathbf{r}(t)$ with arbitrary parameter t, then, assuming $\kappa > 0$ as before,

$$(23^{***}) \quad \tau(t) = \frac{(\mathbf{r}' \quad \mathbf{r}'' \quad \mathbf{r}''')}{(\mathbf{r}' \cdot \mathbf{r}')(\mathbf{r}'' \cdot \mathbf{r}'') - (\mathbf{r}' \cdot \mathbf{r}'')^2}.$$

47. Find the torsion of C: $\mathbf{r}(t) = [t, t^2, t^3]$ (which looks similar to the curve in Fig. 210).

48. (Helix) Show that the helix $[a \cos t, \quad a \sin t, \quad ct]$ can be represented by $[a \cos (s/K), \quad a \sin (s/K), \quad cs/K]$, where $K = \sqrt{a^2 + c^2}$ and s is the arc length. Show that it has constant curvature $\kappa = a/K^2$ and torsion $\tau = c/K^2$.

49. Obtain κ and τ in Prob. 48 from (22*) and (23***) and the original representation in Prob. 48 with parameter t.

50. (Frenet[5] formulas) Show that

$$\mathbf{u}' = \kappa\mathbf{p}, \quad \mathbf{p}' = -\kappa\mathbf{u} + \tau\mathbf{b}, \quad \mathbf{b}' = -\tau\mathbf{p}.$$

9.6 Calculus Review: Functions of Several Variables. *Optional*

Curves required vector functions of a *single* variable x or s, and we now proceed to vector functions of *several* variables, beginning with a review from calculus. ***Go on to the next section, consulting this material only when needed.*** (We include this short section to keep the book reasonably self-contained. For partial derivatives see App. A3.2.)

Chain Rules

Figure 211 shows the notations in the following basic theorem.

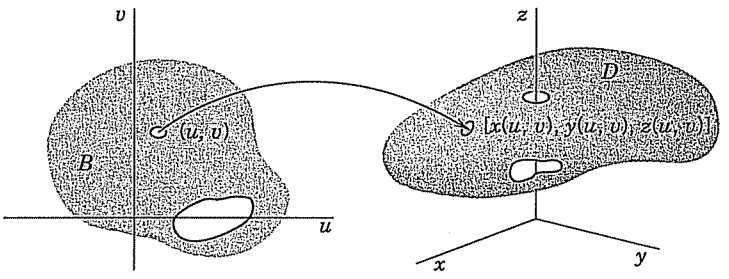


Fig. 211. Notations in Theorem 1

[5]JEAN-FRÉDÉRIC FRENET (1816–1900), French mathematician.

THEOREM 1

Chain Rule

Let $w = f(x, y, z)$ be continuous and have continuous first partial derivatives in a domain D in xyz-space. Let $x = x(u, v)$, $y = y(u, v)$, $z = z(u, v)$ be functions that are continuous and have first partial derivatives in a domain B in the uv-plane, where B is such that for every point (u, v) in B, the corresponding point $[x(u, v), y(u, v), z(u, v)]$ lies in D. See Fig. 211. *Then the function*

$$w = f(x(u, v), y(u, v), z(u, v))$$

is defined in B, has first partial derivatives with respect to u and v in B, and

(1)
$$\frac{\partial w}{\partial u} = \frac{\partial w}{\partial x}\frac{\partial x}{\partial u} + \frac{\partial w}{\partial y}\frac{\partial y}{\partial u} + \frac{\partial w}{\partial z}\frac{\partial z}{\partial u}$$
$$\frac{\partial w}{\partial v} = \frac{\partial w}{\partial x}\frac{\partial x}{\partial v} + \frac{\partial w}{\partial y}\frac{\partial y}{\partial v} + \frac{\partial w}{\partial z}\frac{\partial z}{\partial v}.$$

In this theorem, a **domain** D is an open connected point set in xyz-space, where "connected" means that any two points of D can be joined by a broken line of finitely many linear segments all of whose points belong to D. "Open" means that every point P of D has a neighborhood (a little ball with center P) all of whose points belong to D. For example, the interior of a cube or of an ellipsoid (the solid without the boundary surface) is a domain.

In calculus, x, y, z are often called the **intermediate variables,** in contrast with the **independent variables** u, v and the **dependent variable** w.

Special Cases of Practical Interest

If $w = f(x, y)$ and $x = x(u, v)$, $y = y(u, v)$ as before, then (1) becomes

(2)
$$\frac{\partial w}{\partial u} = \frac{\partial w}{\partial x}\frac{\partial x}{\partial u} + \frac{\partial w}{\partial y}\frac{\partial y}{\partial u}$$
$$\frac{\partial w}{\partial v} = \frac{\partial w}{\partial x}\frac{\partial x}{\partial v} + \frac{\partial w}{\partial y}\frac{\partial y}{\partial v}.$$

If $w = f(x, y, z)$ and $x = x(t)$, $y = y(t)$, $z = z(t)$, then (1) gives

(3)
$$\frac{dw}{dt} = \frac{\partial w}{\partial x}\frac{dx}{dt} + \frac{\partial w}{\partial y}\frac{dy}{dt} + \frac{\partial w}{\partial z}\frac{dz}{dt}.$$

If $w = f(x, y)$ and $x = x(t)$, $y = y(t)$, then (3) reduces to

(4)
$$\frac{dw}{dt} = \frac{\partial w}{\partial x}\frac{dx}{dt} + \frac{\partial w}{\partial y}\frac{dy}{dt}.$$

Finally, the simplest case $w = f(x)$, $x = x(t)$ gives

$$\frac{dw}{dt} = \frac{dw}{dx}\frac{dx}{dt}. \tag{5}$$

EXAMPLE 1 **Chain Rule**

If $w = x^2 - y^2$ and we define polar coordinates r, θ by $x = r\cos\theta$, $y = r\sin\theta$, then (2) gives

$$\frac{\partial w}{\partial r} = 2x\cos\theta - 2y\sin\theta = 2r\cos^2\theta - 2r\sin^2\theta = 2r\cos 2\theta$$
$$\frac{\partial w}{\partial \theta} = 2x(-r\sin\theta) - 2y(r\cos\theta) = -2r^2\cos\theta\sin\theta - 2r^2\sin\theta\cos\theta = -2r^2\sin 2\theta.$$

Mean Value Theorems

THEOREM 2 **Mean Value Theorem**

Let $f(x, y, z)$ be continuous and have continuous first partial derivatives in a domain D in xyz-space. Let P_0: (x_0, y_0, z_0) and P: $(x_0 + h, y_0 + k, z_0 + l)$ be points in D such that the straight line segment P_0P joining these points lies entirely in D. Then

$$f(x_0 + h, y_0 + k, z_0 + l) - f(x_0, y_0, z_0) = h\frac{\partial f}{\partial x} + k\frac{\partial f}{\partial y} + l\frac{\partial f}{\partial z}, \tag{6}$$

the partial derivatives being evaluated at a suitable point of that segment.

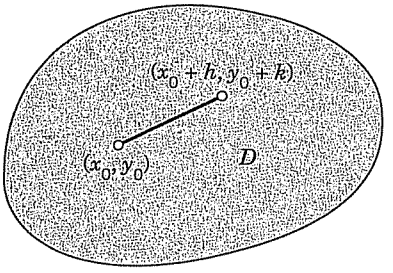


Fig. 212. Mean value theorem for a function of two variables [Formula (7)]

Special Cases

For a function $f(x, y)$ of two variables (satisfying assumptions as in the theorem), formula (6) reduces to (Fig. 212)

$$f(x_0 + h, y_0 + k) - f(x_0, y_0) = h\frac{\partial f}{\partial x} + k\frac{\partial f}{\partial y}, \tag{7}$$

and for a function $f(x)$ of a single variable, (6) becomes

$$f(x_0 + h) - f(x_0) = h\frac{df}{dx}, \tag{8}$$

where in (8), the domain D is a segment of the x-axis and the derivative is taken at a suitable point between x_0 and $x_0 + h$.

PROBLEM SET 9.6

1–5 DERIVATIVE

Find dw/dt by (3) or (4). Check the result by substitution and differentiation. (Show the details.)

1. $w = \sqrt{x^2 + y^2}$, $x = e^{2t}$, $y = e^{-2t}$
2. $w = y/x$, $x = g(t)$, $y = h(t)$
3. $w = x^y$, $x = \cosh t$, $y = \sinh t$
4. $w = xy + yz + zx$, $x = 2\cos t$, $y = 2\sin t$, $z = 5t$
5. $w = (x^2 + y^2 + z^2)^3$, $x = t^2$, $y = t^4$, $z = t^2$

6–9 PARTIAL DERIVATIVES

Find $\partial w/\partial u$ and $\partial w/\partial v$ by (1) and (2). Check the result by substitution and differentiation. (Show the details.)

6. $w = 4x^2 - 4y^2$, $x = u + 2v$, $y = 2u - v$
7. $w = x^2y^2$, $x = e^u \cos v$, $y = e^u \sin v$
8. $w = x^4 - 4x^2y^2 + y^4$, $x = uv$, $y = u/v$
9. $w = 1/(x^2 + y^2 + z^2)$, $x = u^2 + v^2$, $y = u^2 - v^2$, $z = 2uv$
10. **(Partial derivatives on a surface)** Let $w = f(x, y, z)$, and let $z = g(x, y)$ represent a surface S in space. Then on S, the function becomes

$$\tilde{w}(x, y) = f[x, y, g(x, y)].$$

Show that its partial derivatives are obtained from

$$\frac{\partial \tilde{w}}{\partial x} = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial z}\frac{\partial g}{\partial x}, \qquad \frac{\partial \tilde{w}}{\partial y} = \frac{\partial f}{\partial y} + \frac{\partial f}{\partial z}\frac{\partial g}{\partial y}$$

$[z = g(x, y)]$.

Apply this to $f = x^3 + y^3 + z^2$, $g = x^2 + y^2$ and check by substitution and direct differentiation. (The general formula will be needed in Sec. 10.9.)

9.7 Gradient of a Scalar Field. Directional Derivative

We shall see that some of the vector fields in applications—not all of them!—can be obtained from scalar fields. This is a considerable advantage because scalar fields can be handled more easily. The relation between these two kinds of fields is obtained by the "gradient," which is thus of great practical importance.

DEFINITION 1

Gradient

The **gradient** of a given scalar function $f(x, y, z)$ is denoted by grad f or ∇f (read **nabla** f) and is the vector function defined by

$$\text{grad } f = \nabla f = \left[\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}\right] = \frac{\partial f}{\partial x}\mathbf{i} + \frac{\partial f}{\partial y}\mathbf{j} + \frac{\partial f}{\partial z}\mathbf{k}. \tag{1}$$

Here x, y, z are Cartesian coordinates in a domain in 3-space in which f is defined and differentiable. (For curvilinear coordinates see App. 3.4.)

For instance, if $f(x, y, z) = 2y^3 + 4xz + 3x$, then grad $f = [4z + 3, \quad 6y^2, \quad 4x]$.

The notation ∇f is suggested by the *differential operator* ∇ (read *nabla*) defined by

$$\nabla = \frac{\partial}{\partial x}\mathbf{i} + \frac{\partial}{\partial y}\mathbf{j} + \frac{\partial}{\partial z}\mathbf{k}. \tag{1*}$$

Gradients are useful in several ways, notably in giving the rate of change of $f(x, y, z)$ in any direction in space, in obtaining surface normal vectors, and in deriving vector fields from scalar fields, as we are going to show in this section.

Directional Derivative

From calculus we know that the partial derivatives in (1) give the rates of change of $f(x, y, z)$ in the directions of the three coordinate axes. It seems natural to extend this and ask for the rate of change of f *in an arbitrary direction* in space. This leads to the following concept.

DEFINITION 2

Directional Derivative

The directional derivative $D_{\mathbf{b}}f$ or df/ds of a function $f(x, y, z)$ at a point P in the direction of a vector $\mathbf{b}$ is defined by (see Fig. 213)

$$D_{\mathbf{b}}f = \frac{df}{ds} = \lim_{s \to 0} \frac{f(Q) - f(P)}{s}. \tag{2}$$

Here Q is a variable point on the straight line L in the direction of $\mathbf{b}$, and $|s|$ is the distance between P and Q. Also, $s > 0$ if Q lies in the direction of $\mathbf{b}$ (as in Fig. 213), $s < 0$ if Q lies in the direction of $-\mathbf{b}$, and $s = 0$ if $Q = P$.

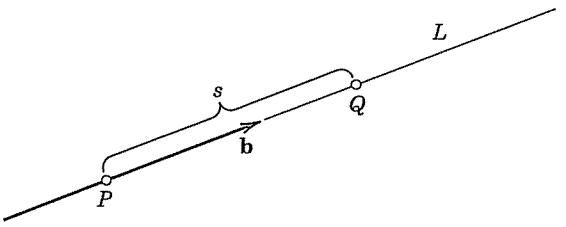


Fig. 213. Directional derivative

The next idea is to use Cartesian xyz-coordinates and for $\mathbf{b}$ a unit vector. Then the line L is given by

$$\mathbf{r}(s) = x(s)\mathbf{i} + y(s)\mathbf{j} + z(s)\mathbf{k} = \mathbf{p}_0 + s\mathbf{b} \qquad (|\mathbf{b}| = 1) \tag{3}$$

where $\mathbf{p}_0$ the position vector of P. Equation (2) now shows that $D_{\mathbf{b}}f = df/ds$ is the derivative of the function $f(x(s), y(s), z(s))$ with respect to the arc length s of L. Hence, assuming that f has continuous partial derivatives and applying the chain rule [formula (3) in the previous section], we obtain

$$D_{\mathbf{b}}f = \frac{df}{ds} = \frac{\partial f}{\partial x}x' + \frac{\partial f}{\partial y}y' + \frac{\partial f}{\partial z}z' \tag{4}$$

where primes denote derivatives with respect to s (which are taken at $s = 0$). But here, differentiating (3) gives $\mathbf{r}' = x'\mathbf{i} + y'\mathbf{j} + z'\mathbf{k} = \mathbf{b}$. Hence (4) is simply the inner product of grad f and $\mathbf{b}$ [see (2), Sec. 9.2]; that is,

$$(5) \qquad D_{\mathbf{b}}f = \frac{df}{ds} = \mathbf{b} \cdot \operatorname{grad} f \qquad (|\mathbf{b}| = 1).$$

ATTENTION! If the direction is given by a vector $\mathbf{a}$ of any length ($\neq 0$), then

$$(5^*) \qquad D_{\mathbf{a}}f = \frac{df}{ds} = \frac{1}{|\mathbf{a}|}\,\mathbf{a} \cdot \operatorname{grad} f.$$

EXAMPLE 1 **Gradient. Directional Derivative**

Find the directional derivative of $f(x, y, z) = 2x^2 + 3y^2 + z^2$ at P: (2, 1, 3) in the direction of $\mathbf{a} = [1, 0, -2]$.

Solution. grad $f = [4x, 6y, 2z]$ gives at P the vector grad $f(P) = [8, 6, 6]$. From this and (5*) we obtain, since $|\mathbf{a}| = \sqrt{5}$,

$$D_{\mathbf{a}}f(P) = \frac{1}{\sqrt{5}}\,[1, 0, -2] \cdot [8, 6, 6] = \frac{1}{\sqrt{5}}\,(8 + 0 - 12) = -\frac{4}{\sqrt{5}} = -1.789.$$

The minus sign indicates that at P the function f is decreasing in the direction of $\mathbf{a}$. ■

Gradient Is a Vector. Maximum Increase

grad f in (1) *looks* like a vector—after all, it has three components! But to prove that it *actually is* a vector, since it is defined in terms of components depending on the Cartesian coordinates, we must show that grad f has a length and direction independent of the choice of those coordinates. In contrast, $[\partial f/\partial x, \quad 2\partial f/\partial y, \quad \partial f/\partial z]$ also looks like a vector but does not have a length and direction independent of the choice of Cartesian coordinates.

Incidentally, the direction makes the gradient eminently useful: grad f *points in the direction of maximum increase of* f.

THEOREM 1

Vector Character of Gradient. Maximum Increase

Let $f(P) = f(x, y, z)$ be a scalar function having continuous first partial derivatives in some domain B in space. Then grad f *exists in B and is a vector, that is, its length and direction are independent of the particular choice of Cartesian coordinates. If* grad $f(P) \neq \mathbf{0}$ *at some point P, it has the direction of maximum increase of f at P.*

PROOF From (5) and the definition of inner product [(1) in Sec. 9.2] we have

$$(6) \qquad D_{\mathbf{b}}f = |\mathbf{b}|\,|\operatorname{grad} f| \cos\gamma = |\operatorname{grad} f| \cos\gamma$$

where γ is the angle between $\mathbf{b}$ and grad f. Now f is a scalar function. Hence its value at a point P depends on P but not on the particular choice of coordinates. The same holds for the arc length s of the line L in Fig. 213, hence also for $D_{\mathbf{b}}f$. Now (6) shows that $D_{\mathbf{b}}f$ is maximum when $\cos\gamma = 1$, $\gamma = 0$, and then $D_{\mathbf{b}}f = |\operatorname{grad} f|$. It follows that the length and direction of grad f are independent of the choice of coordinates. Since $\gamma = 0$ if and only if $\mathbf{b}$ has the direction of grad f, the latter is the direction of maximum increase of f at P, provided grad $f \neq \mathbf{0}$ at P. ■

Gradient as Surface Normal Vector

Gradients have an important application in connection with surfaces, namely, as surface normal vectors, as follows. Let S be a surface represented by $f(x, y, z) = c = const$, where f is differentiable. Such a surface is called a **level surface** of f, and for different c we get different level surfaces. Now let C be a curve on S through a point P of S. As a curve in space, C has a representation $\mathbf{r}(t) = [x(t), y(t), z(t)]$. For C to lie on the surface S, the components of $\mathbf{r}(t)$ must satisfy $f(x, y, z) = c$, that is,

$$f(x(t), y(t), z(t)) = c. \tag{7}$$

Now a tangent vector of C is $\mathbf{r}'(t) = [x'(t), y'(t), z'(t)]$. And the tangent vectors of all curves on S passing through P will generally form a plane, called the **tangent plane** of S at P. (Exceptions occur at edges or cusps of S, for instance, for the cone in Fig. 215 at the apex.) The normal of this plane (the straight line through P perpendicular to the tangent plane) is called the **surface normal** to S at P. A vector in the direction of the surface normal is called a **surface normal vector** of S at P. We can obtain such a vector quite simply by differentiating (7) with respect to t. By the chain rule,

$$\frac{\partial f}{\partial x}x' + \frac{\partial f}{\partial y}y' + \frac{\partial f}{\partial z}z' = (\text{grad } f) \cdot \mathbf{r}' = 0.$$

Hence grad f is orthogonal to all the vectors $\mathbf{r}'$ in the tangent plane, so that it is a normal vector of S at P. Our result is as follows (see Fig. 214).

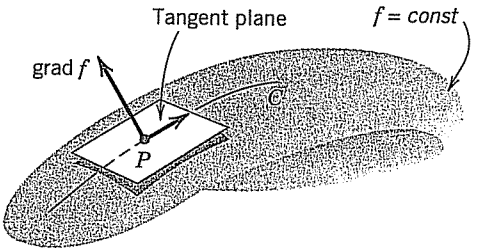


Fig. 214. Gradient as surface normal vector

THEOREM 2

Gradient as Surface Normal Vector

Let f be a differentiable scalar function in space. Let $f(x, y, z) = c = const$ represent a surface S. Then if the gradient of f at a point P of S is not the zero vector, it is a normal vector of S at P.

EXAMPLE 2 **Gradient as Surface Normal Vector. Cone**

Find a unit normal vector $\mathbf{n}$ of the cone of revolution $z^2 = 4(x^2 + y^2)$ at the point P: (1, 0, 2).

Solution. The cone is the level surface $f = 0$ of $f(x, y, z) = 4(x^2 + y^2) - z^2$. Thus (Fig. 215),

$$\text{grad } f = [8x, \quad 8y, \quad -2z], \qquad \text{grad } f(P) = [8, \quad 0, \quad -4]$$

$$\mathbf{n} = \frac{1}{|\text{grad } f(P)|}\,\text{grad } f(P) = \left[\frac{2}{\sqrt{5}}, \quad 0, \quad -\frac{1}{\sqrt{5}}\right].$$

$\mathbf{n}$ points downward since it has a negative z-component. The other unit normal vector of the cone at P is $-\mathbf{n}$. ■

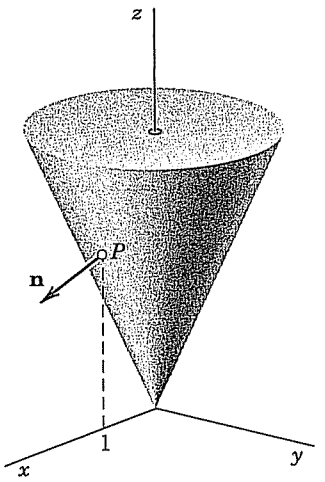


Fig. 215. Cone and unit normal vector **n**

Vector Fields That Are Gradients of Scalar Fields ("Potentials")

At the beginning of this section we mentioned that some vector fields have the advantage that they can be obtained from scalar fields, which can be handled more easily. Such a vector field is given by a vector function $\mathbf{v}(P)$, which is obtained as the gradient of a scalar function, say, $\mathbf{v}(P) = \text{grad } f(P)$. The function $f(P)$ is called a *potential function* or a **potential** of $\mathbf{v}(P)$. Such a $\mathbf{v}(P)$ and the corresponding vector field are called **conservative** because in such a vector field, energy is conserved; that is, no energy is lost (or gained) in displacing a body (or a charge in the case of an electrical field) from a point P to another point in the field and back to P. We show this in Sec. 10.2.

Conservative fields play a central role in physics and engineering. A basic application concerns the gravitational force (see Example 3 in Sec. 9.4) and we show that it has a potential which satisfies Laplace's equation, the most important partial differential equation in physics and its applications.

THEOREM 3

Gravitational Field. Laplace's Equation

The force of attraction

$$\mathbf{p} = -\frac{c}{r^3}\mathbf{r} = -c\left[\frac{x - x_0}{r^3}, \frac{y - y_0}{r^3}, \frac{z - z_0}{r^3}\right] \tag{8}$$

between two particles at points P_0: (x_0, y_0, z_0) *and* P: (x, y, z) (as given by Newton's law of gravitation) *has the potential* $f(x, y, z) = c/r$, *where* $r\ (> 0)$ *is the distance between* P_0 *and* P.

Thus $\mathbf{p} = \text{grad } f = \text{grad } (c/r)$. *This potential* f *is a solution of* **Laplace's equation**

$$\nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2} = 0. \tag{9}$$

[$\nabla^2 f$ (read *nabla squared* f) is called the **Laplacian** of f.]

PROOF That distance is $r = ((x - x_0)^2 + (y - y_0)^2 + (z - z_2)^2)^{1/2}$. The key observation now is that for the components of $\mathbf{p} = [p_1, p_2, p_3]$ we obtain by partial differentiation

$$\text{(10a)} \qquad \frac{\partial}{\partial x}\left(\frac{1}{r}\right) = \frac{-2(x - x_0)}{2[(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2]^{3/2}} = -\frac{x - x_0}{r^3}$$

and similarly

$$\text{(10b)} \qquad \frac{\partial}{\partial y}\left(\frac{1}{r}\right) = -\frac{y - y_0}{r^3}, \qquad \frac{\partial}{\partial z}\left(\frac{1}{r}\right) = -\frac{z - z_0}{r^3}.$$

From this we see that, indeed, $\mathbf{p}$ is the gradient of the scalar function $f = c/r$. The second statement of the theorem follows by partially differentiating (10), that is,

$$\frac{\partial^2}{\partial x^2}\left(\frac{1}{r}\right) = -\frac{1}{r^3} + \frac{3(x - x_0)^2}{r^5},$$

$$\frac{\partial^2}{\partial y^2}\left(\frac{1}{r}\right) = -\frac{1}{r^3} + \frac{3(y - y_0)^2}{r^5},$$

$$\frac{\partial^2}{\partial z^2}\left(\frac{1}{r}\right) = -\frac{1}{r^3} + \frac{3(z - z_0)^2}{r^5},$$

and then adding these three expressions. Their common denominator is r^5. Hence the three terms $-1/r^3$ contribute $-3r^2$ to the numerator, and the three other terms give the sum

$$3(x - x_0)^2 + 3(y - y_0)^2 + 3(z - z_0)^2 = 3r^2,$$

so that the numerator is 0, and we obtain (9). ∎

$\nabla^2 f$ is also denoted by Δf. The differential operator

$$\text{(11)} \qquad \nabla^2 = \Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$$

(read "nabla squared" or "delta") is called the **Laplace operator.** It can be shown that the field of force produced by any distribution of masses is given by a vector function that is the gradient of a scalar function f, and f satisfies (9) in any region that is free of matter.

The great importance of the Laplace equation also results from the fact that there are other laws in physics that are of the same form as Newton's law of gravitation. For instance, in electrostatics the force of attraction (or repulsion) between two particles of opposite (or

like) charge Q_1 and Q_2 is

(12) $$\mathbf{p} = \frac{k}{r^3}\mathbf{r}$$ **(Coulomb's law[6]).**

Laplace's equation will be discussed in detail in Chaps. 12 and 18.

A method for finding out whether a given vector field has a potential will be explained in Sec. 9.9.

PROBLEM SET 9.7

1–6 CALCULATION OF GRADIENTS

Find ∇f. Graph some level curves $f = const.$ Indicate ∇f by arrows at some points of these curves.

1. $f = x^2 + y^2$
2. $f = x^2 + \frac{1}{9}y^2$
3. $f = \dfrac{x}{y}$
4. $f = x^4 + y^4$
5. $f = (x - 2)(y + 2)$
6. $f = (x - 3)^2 + (y - 1)^2$

7–12 USE OF GRADIENTS. VELOCITY FIELDS

Given the velocity potential f of a flow, find the velocity $\mathbf{v} = \nabla f$ of the flow and its value at P. Make a sketch of $\mathbf{v}(P)$.

7. $f = x^2 + y^2 + z^2$, P: (3, 2, 2)
8. $f = \ln(x^2 + y^2)$, P: (4, 3)
9. $f = \cos x \cosh y$, P: $(\frac{1}{2}\pi, \ln 2)$
10. $f = x^2 + 4y^2 + 9z^2$, P: (3, 2, 1)
11. $f = e^x \sin y$, P: $(1, \pi)$
12. $f = (x^2 + y^2 + z^2)^{-1/2}$, P: (2, 1, 2)

13–18 HEAT FLOW

Experiments show that in a temperature field, heat flows in the direction of maximum decrease of temperature T. Find this direction in general and at a given point P. Sketch that direction at P as an arrow.

13. $T = x^2 - y^2$, P: (2, 1)
14. $T = \arctan \dfrac{y}{x}$, P: (2, 2)
15. $T = x^3 - 3xy^2$, P: $(\sqrt{8}, \sqrt{2})$
16. $T = x/(x^2 + y^2)$, P: (4, 0)
17. $T = 3x^2y - y^3$, P: (4, −2)
18. $T = \sin x \cosh y$, P: $(\frac{1}{4}\pi, \ln 5)$

19–24 ELECTRIC FORCE

The force in an electrostatic field $f(x, y, z)$ has the direction of the gradient of f. Find ∇f and its value at P.

19. $f = (x - 1)^2 - (y + 1)^2$, P: (4, −3)
20. $f = y/(x^2 + y^2)$, P: (5, 3)
21. $f = x^2 - 2x - y^2$, P: (−2, 6)
22. $f = \ln(x^2 + y^2)$, P: (3, 3)
23. $f = (x^2 + y^2 + z^2)^{-1/2}$, P: (12, 0, 16)
24. $f = x^2y - \frac{1}{3}y^3$, P: (2, 3)
25. **(Gradient)** What does it mean if $|\text{grad } f(P)| < |\text{grad } f(Q)|$ at two points P and Q in a scalar field?
26. **(Landscape)** If $z(x, y) = 2000 - 4x^2 - y^2$ [meters] gives the elevation of a mountain above sea level, what is the direction of steepest ascent at P: (3, −6)? What does the mountain look like?

27–32 SURFACE NORMAL

Find a normal vector of the surface at the given point P.

27. $ax + by + cz = d$, any P
28. $x^2 + 3y^2 + z^2 = 28$, P: (4, 1, 3)
29. $x^2 + y^2 = 25$, P: (4, 3, 8)
30. $x^2 - y^2 + 4z^2 = 67$, P: (−2, 1, 4)
31. $x^4 + y^4 + z^4 = 243$, P: (3, 3, 3)
32. $z = x^2 + y^2$, P: (3, 4, 25)

33–38 DIRECTIONAL DERIVATIVE

Find the directional derivative of f at P in the direction of $\mathbf{a}$.

33. $f = x^2 + y^2 - z$, P: (1, 1, −2), $\mathbf{a} = [1, 1, 2]$
34. $f = x^2 + y^2 + z^2$, P: (2, −2, 1), $\mathbf{a} = [-1, -1, 0]$
35. $f = xyz$, P: (−1, 1, 3), $\mathbf{a} = [1, -2, 2]$

[6]CHARLES AUGUSTIN DE COULOMB (1736–1806), French physicist and engineer. Coulomb's law was derived by him from his own very precise measurements.

36. $f = (x^2 + y^2 + z^2)^{-1/2}$, *P:* (4, 2, −4), $\mathbf{a} = [1, 2, -2]$

37. $f = e^x \sin y$, *P:* $(2, \frac{1}{2}\pi, 0)$, $\mathbf{a} = [2, 3, 0]$

38. $f = 4x^2 + y^2 + 9z^2$, *P:* (2, 4, 0), $\mathbf{a} = [-2, -4, 3]$

39–41 POTENTIALS

for a given vector field—if they exist!—can be obtained by a method to be discussed in Sec. 9.9. In simpler cases, use inspection. Find a potential $f = \text{grad } \mathbf{v}$ for given $\mathbf{v}(x, y, z)$.

39. $\mathbf{v} = [3x, \quad 5y, \quad -4z]$

40. $\mathbf{v} = [ye^x, e^x, 2z]$

41. $\mathbf{v} = [4x^3, 3y^2, -6z]$

42. Project. Useful Formulas for Gradients and Laplacians. Prove the following formulas and give for each of them two examples showing when they are advantageous.

$$\nabla(fg) = f\nabla g + g\nabla f$$

$$\nabla(f^n) = nf^{n-1}\nabla f$$

$$\nabla(f/g) = (1/g^2)(g\nabla f - f\nabla g)$$

$$\nabla^2(fg) = g\nabla^2 f + 2\nabla f \cdot \nabla g + f\nabla^2 g$$

43. CAS PROJECT. Equipotential Curves. Graph some **isotherms** (curves of constant temperature) and indicate directions of heat flow by arrows when the temperature $T(x, y)$ equals:

(a) $x^3 - 3xy^2$ (b) $\sin x \sinh y$ (c) $e^x \sin y$.

9.8 Divergence of a Vector Field

Vector calculus owes much of its importance in engineering and physics to the gradient, divergence, and curl. From a scalar field we can obtain a vector field by the gradient (Sec. 9.7). Conversely, from a vector field we can obtain a scalar field by the divergence, or another vector field by the curl (to be discussed in Sec. 9.9). These concepts were suggested by basic physical applications, as we shall see.

To begin, let $\mathbf{v}(x, y, z)$ be a differentiable vector function, where x, y, z are Cartesian coordinates, and let v_1, v_2, v_3 be the components of $\mathbf{v}$. Then the function

(1)
$$\text{div } \mathbf{v} = \frac{\partial v_1}{\partial x} + \frac{\partial v_2}{\partial y} + \frac{\partial v_3}{\partial z}$$

is called the **divergence** of $\mathbf{v}$ or the *divergence of the vector field defined by* $\mathbf{v}$. For example, if

$$\mathbf{v} = [3xz, 2xy, -yz^2] = 3xz\mathbf{i} + 2xy\mathbf{j} - yz^2\mathbf{k}, \quad \text{then} \quad \text{div } \mathbf{v} = 3z + 2x - 2yz.$$

Another common notation for the divergence is

$$\text{div } \mathbf{v} = \nabla \cdot \mathbf{v} = \left[\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}\right] \cdot [v_1, v_2, v_3]$$

$$= \left(\frac{\partial}{\partial x}\mathbf{i} + \frac{\partial}{\partial y}\mathbf{j} + \frac{\partial}{\partial z}\mathbf{k}\right) \cdot (v_1\mathbf{i} + v_2\mathbf{j} + v_3\mathbf{k})$$

$$= \frac{\partial v_1}{\partial x} + \frac{\partial v_2}{\partial y} + \frac{\partial v_3}{\partial z},$$

with the understanding that the "product" $(\partial/\partial x)v_1$ in the dot product means the partial derivative $\partial v_1/\partial x$, etc. This is a convenient notation, but nothing more. Note that $\nabla \cdot \mathbf{v}$ means the scalar div $\mathbf{v}$, whereas ∇f means the vector grad f defined in Sec. 9.7.

In Example 2 we shall see that the divergence has an important physical meaning. Clearly, the values of a function that characterizes a physical or geometric property must be independent of the particular choice of coordinates; that is, those values must be invariant with respect to coordinate transformations. Accordingly, the following theorem should hold.

THEOREM 1

Invariance of the Divergence

The divergence div $\mathbf{v}$ *is a scalar function, that is, its values depend only on the points in space* (and, of course, on $\mathbf{v}$) *but not on the choice of the coordinates in* (1), *so that with respect to other Cartesian coordinates* x^*, y^*, z^* *and corresponding components* v_1^*, v_2^*, v_3^* *of* $\mathbf{v}$,

(2) $$\operatorname{div} \mathbf{v} = \frac{\partial v_1^*}{\partial x^*} + \frac{\partial v_2^*}{\partial y^*} + \frac{\partial v_3^*}{\partial z^*}.$$

We shall prove this theorem in Sec. 10.7, using integrals.

Presently, let us turn to the more immediate practical task of gaining a feel for the significance of the divergence as follows. Let $f(x, y, z)$ be a twice differentiable scalar function. Then its gradient exists,

$$\mathbf{v} = \operatorname{grad} f = \left[\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}\right] = \frac{\partial f}{\partial x}\mathbf{i} + \frac{\partial f}{\partial y}\mathbf{j} + \frac{\partial f}{\partial z}\mathbf{k}$$

and we can differentiate once more, the first component with respect to x, the second with respect to y, the third with respect to z, and then form the divergence,

$$\operatorname{div} \mathbf{v} = \operatorname{div} (\operatorname{grad} f) = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2}.$$

Hence we have the basic result that *the divergence of the gradient is the Laplacian* (Sec. 9.7),

(3) $$\operatorname{div} (\operatorname{grad} f) = \nabla^2 f.$$

EXAMPLE 1 **Gravitational Force. Laplace's Equation**

The gravitational force $\mathbf{p}$ in Theorem 3 of the last section is the gradient of the scalar function $f(x, y, z) = c/r$, which satisfies Laplaces equation $\nabla^2 f = 0$. According to (3) this implies that div $\mathbf{p} = 0$ $(r > 0)$. ■

The following example from hydrodynamics shows the physical significance of the divergence of a vector field. (More physical details on this significance will be added in Sec. 10.8.)

EXAMPLE 2 **Flow of a Compressible Fluid. Physical Meaning of the Divergence**

We consider the motion of a fluid in a region R having no **sources** or **sinks** in R, that is, no points at which fluid is produced or disappears. The concept of **fluid state** is meant to cover also gases and vapors. Fluids in the restricted sense, or **liquids** (water or oil, for instance), have very small compressibility, which can be neglected in many problems. Gases and vapors have large compressibility; that is, their density ρ (= mass per unit volume) depends on the coordinates x, y, z in space (and may depend on time t). We assume that our fluid is compressible.

We consider the flow through a rectangular box B of small edges Δx, Δy, Δz parallel to the coordinate axes (Fig. 216), (Δ is a standard notation for small quantities; of course, it has nothing to do with the notation for the Laplacian in (11) of Sec. 9.7.) The box B has the volume $\Delta V = \Delta x\, \Delta y\, \Delta z$. Let $\mathbf{v} = [v_1, v_2, v_3] = v_1\mathbf{i} + v_2\mathbf{j} + v_3\mathbf{k}$ be the velocity vector of the motion. We set

$$\mathbf{u} = \rho\mathbf{v} = [u_1, u_2, u_3] = u_1\mathbf{i} + u_2\mathbf{j} + u_3\mathbf{k} \tag{4}$$

and assume that $\mathbf{u}$ and $\mathbf{v}$ are continuously differentiable vector functions of x, y, z, and t (that is, they have first partial derivatives which are continuous). Let us calculate the change in the mass included in B by considering the **flux** across the boundary, that is, the total loss of mass leaving B per unit time. Consider the flow through the left of the three faces of B that are visible in Fig. 216, whose area is $\Delta x\, \Delta z$. Since the vectors $v_1\,\mathbf{i}$ and $v_3\,\mathbf{k}$ are parallel to that face, the components v_1 and v_3 of $\mathbf{v}$ contribute nothing to this flow. Hence the mass of fluid entering through that face during a short time interval Δt is given approximately by

$$(\rho v_2)_y\, \Delta x\, \Delta z\, \Delta t = (u_2)_y\, \Delta x\, \Delta z\, \Delta t,$$

where the subscript y indicates that this expression refers to the left face. The mass of fluid leaving the box B through the opposite face during the same time interval is approximately $(u_2)_{y+\Delta y}\, \Delta x\, \Delta z\, \Delta t$, where the subscript $y + \Delta y$ indicates that this expression refers to the right face (which is not visible in Fig. 216). The difference

$$\Delta u_2\, \Delta x\, \Delta z\, \Delta t = \frac{\Delta u_2}{\Delta y}\, \Delta V\, \Delta t \qquad [\Delta u_2 = (u_2)_{y+\Delta y} - (u_2)_y]$$

is the approximate loss of mass. Two similar expressions are obtained by considering the other two pairs of parallel faces of B. If we add these three expressions, we find that the total loss of mass in B during the time interval Δt is approximately

$$\left(\frac{\Delta u_1}{\Delta x} + \frac{\Delta u_2}{\Delta y} + \frac{\Delta u_3}{\Delta z}\right) \Delta V\, \Delta t,$$

where

$$\Delta u_1 = (u_1)_{x+\Delta x} - (u_1)_x \qquad \text{and} \qquad \Delta u_3 = (u_3)_{z+\Delta z} - (u_3)_z.$$

This loss of mass in B is caused by the time rate of change of the density and is thus equal to

$$-\frac{\partial \rho}{\partial t}\, \Delta V\, \Delta t.$$

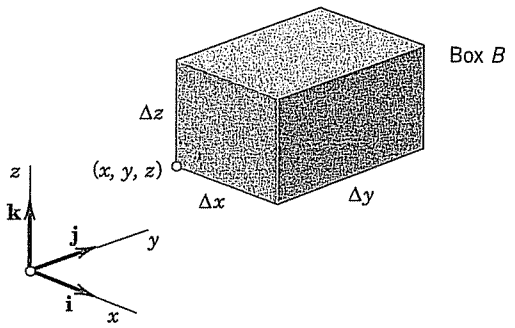


Fig. 216. Physical interpretation of the divergence

If we equate both expressions, divide the resulting equation by $\Delta V\,\Delta t$, and let Δx, Δy, Δz, and Δt approach zero, then we obtain

$$\operatorname{div}\mathbf{u} = \operatorname{div}(\rho\mathbf{v}) = -\frac{\partial\rho}{\partial t}$$

or

$$\frac{\partial\rho}{\partial t} + \operatorname{div}(\rho\mathbf{v}) = 0. \tag{5}$$

This important relation is called the *condition for the conservation of mass* or the **continuity equation** *of a compressible fluid flow.*

If the flow is **steady,** that is, independent of time, then $\partial\rho/\partial t = 0$ and the continuity equation is

$$\operatorname{div}(\rho\mathbf{v}) = 0. \tag{6}$$

If the density ρ is constant, so that the fluid is incompressible, then equation (6) becomes

$$\operatorname{div}\mathbf{v} = 0. \tag{7}$$

This relation is known as the **condition of incompressibility.** It expresses the fact that the balance of outflow and inflow for a given volume element is zero at any time. Clearly, the assumption that the flow has no sources or sinks in R is essential to our argument.

From this discussion you should conclude and remember that, roughly speaking, ***the divergence measures outflow minus inflow.*** ■

Comment. The **divergence theorem** of Gauss, an integral theorem involving the divergence, follows in the next chapter (Sec. 10.7).

PROBLEM SET 9.8

1–7 CALCULATION OF THE DIVERGENCE

Find the divergence of the following vector functions.

1. $[x^3 + y^3,\quad 3xy^2,\quad 3zy^2]$
2. $[e^{2x}\cos 2y,\quad e^{2x}\sin 2y,\quad 5e^{2z}]$
3. $[x^2 + y^2,\quad 2xyz,\quad z^2 + x^2]$
4. $(x^2 + y^2 + z^2)^{-3/2}[x,\quad y,\quad z]$
5. $[\sin xy,\quad \sin xy,\quad z\cos xy]$
6. $[v_1(y, z),\quad v_2(z, x),\quad v_3(x, y)]$
7. $x^2y^2z^2[x,\quad y,\quad z]$
8. Let $\mathbf{v} = [x,\quad y,\quad v_3]$. Find a v_3 such that (a) $\operatorname{div}\mathbf{v} > 0$ everywhere, (b) $\operatorname{div}\mathbf{v} > 0$ if $|z| < 1$ and $\operatorname{div}\mathbf{v} < 0$ if $|z| > 1$.
9. **(Incompressible flow)** Show that the flow with velocity vector $\mathbf{v} = y\mathbf{i}$ is incompressible. Show that the particles that at time $t = 0$ are in the cube whose faces are portions of the planes $x = 0$, $x = 1$, $y = 0$, $y = 1$, $z = 0$, $z = 1$ occupy at $t = 1$ the volume 1.
10. **(Compressible flow)** Consider the flow with velocity vector $\mathbf{v} = x\mathbf{i}$. Show that the individual particles have the position vectors $\mathbf{r}(t) = c_1e^t\mathbf{i} + c_2\mathbf{j} + c_3\mathbf{k}$ with constant c_1, c_2, c_3. Show that the particles that at $t = 0$ are in the cube of Prob. 9 at $t = 1$ occupy the volume e.
11. **(Rotational flow)** The velocity vector $\mathbf{v}(x, y, z)$ of an incompressible fluid rotating in a cylindrical vessel is of the form $\mathbf{v} = \mathbf{w} \times \mathbf{r}$, where $\mathbf{w}$ is the (constant) rotation vector; see Example 5 in Sec. 9.3. Show that $\operatorname{div}\mathbf{v} = 0$. Is this plausible because of our present Example 2?
12. **CAS PROJECT. Visualizing the Divergence.** Graph the given velocity field $\mathbf{v}$ of a fluid flow in a square centered at the origin with sides parallel to the coordinate axes. Recall that the divergence measures outflow minus inflow. By looking at the flow near the sides of the square, can you see whether $\operatorname{div}\mathbf{v}$ must be positive or negative or may perhaps be zero? Then calculate $\operatorname{div}\mathbf{v}$. First do the given flows and then do some of your own. Enjoy it.
    (a) $\mathbf{v} = \mathbf{i}$
    (b) $\mathbf{v} = x\mathbf{i}$
    (c) $\mathbf{v} = x\mathbf{i} - y\mathbf{j}$
    (d) $\mathbf{v} = x\mathbf{i} + y\mathbf{j}$
    (e) $\mathbf{v} = -x\mathbf{i} - y\mathbf{j}$
    (f) $\mathbf{v} = (x^2 + y^2)^{-1}(-y\mathbf{i} + x\mathbf{j})$

13. PROJECT. Useful Formulas for the Divergence. Prove

(a) $\operatorname{div}(k\mathbf{v}) = k \operatorname{div} \mathbf{v}$ (k constant)

(b) $\operatorname{div}(f\mathbf{v}) = f \operatorname{div} \mathbf{v} + \mathbf{v} \cdot \nabla f$

(c) $\operatorname{div}(f\nabla g) = f\nabla^2 g + \nabla f \cdot \nabla g$

(d) $\operatorname{div}(f\nabla g) - \operatorname{div}(g\nabla f) = f\nabla^2 g - g\nabla^2 f$.

Verify (b) for $f = e^{xyz}$ and $\mathbf{v} = ax\mathbf{i} + by\mathbf{j} + cz\mathbf{k}$. Obtain the answer to Prob. 4 from (b). Verify (c) for $f = x^2 - y^2$ and $g = e^{x+y}$. Give examples of your own for which (a)–(d) are advantageous.

14–20 CALCULATION OF THE LAPLACIAN BY (3)

Find $\nabla^2 f$ by (3). Check by differentiation. Indicate when (3) is simpler. (Show the details of your work.)

14. $f = xy/z^2$

15. $f = (y + x)/(y - x)$

16. $f = z - 4\sqrt{x^2 + y^2}$

17. $f = e^{x^2 - y^2} \cos 2xy$

18. $f = \arctan(y/x)$

19. $f = e^{xyz}$

20. $f = \cos^2 x - \sin^2 y$

9.9 Curl of a Vector Field

Gradient (Sec. 9.7), divergence (Sec. 9.8), and curl are basic in connection with fields, and we now define and discuss the curl.

Let $\mathbf{v}(x, y, z) = [v_1, v_2, v_3] = v_1\mathbf{i} + v_2\mathbf{j} + v_3\mathbf{k}$ be a differentiable vector function of the Cartesian coordinates x, y, z. Then the **curl** *of the vector function* $\mathbf{v}$ or *of the vector field given by* $\mathbf{v}$ is defined by the "symbolic" determinant

$$\text{(1)} \qquad \operatorname{curl} \mathbf{v} = \nabla \times \mathbf{v} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ v_1 & v_2 & v_3 \end{vmatrix}$$

$$= \left(\frac{\partial v_3}{\partial y} - \frac{\partial v_2}{\partial z}\right)\mathbf{i} + \left(\frac{\partial v_1}{\partial z} - \frac{\partial v_3}{\partial x}\right)\mathbf{j} + \left(\frac{\partial v_2}{\partial x} - \frac{\partial v_1}{\partial y}\right)\mathbf{k}.$$

This is the formula when x, y, z are *right-handed.* If they are *left-handed,* the determinant has a minus sign in front (just as in (2**) in Sec. 9.3).

Instead of curl $\mathbf{v}$ one also uses the notation rot $\mathbf{v}$ (suggested by "rotation"; see Example 2).

EXAMPLE 1 Curl of a Vector Function

Let $\mathbf{v} = [yz, \ 3zx, \ z] = yz\mathbf{i} + 3zx\mathbf{j} + z\mathbf{k}$ with right-handed x, y, z. Then (1) gives

$$\operatorname{curl} \mathbf{v} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \partial/\partial x & \partial/\partial y & \partial/\partial z \\ yz & 3zx & z \end{vmatrix} = -3x\mathbf{i} + y\mathbf{j} + (3z - z)\mathbf{k} = -3x\mathbf{i} + y\mathbf{j} + 2z\mathbf{k}. \qquad \blacksquare$$

The curl plays an important role in many applications. Let us illustrate this with a typical basic example. More about the nature and significance of the curl will be said in Sec. 10.9.

EXAMPLE 2 Rotation of a Rigid Body. Relation to the Curl

We have seen in Example 5, Sec. 9.3, that a rotation of a rigid body B about a fixed axis in space can be described by a vector $\mathbf{w}$ of magnitude ω in the direction of the axis of rotation, where ω (> 0) is the angular speed of the rotation, and $\mathbf{w}$ is directed so that the rotation appears clockwise if we look in the direction of $\mathbf{w}$. According to (9), Sec. 9.3, the velocity field of the rotation can be represented in the form

$$\mathbf{v} = \mathbf{w} \times \mathbf{r}$$

where $\mathbf{r}$ is the position vector of a moving point with respect to a Cartesian coordinate system *having the origin on the axis of rotation.* Let us choose right-handed Cartesian coordinates such that the axis of rotation is the z-axis. Then (see Example 2 in Sec. 9.4)

$$\mathbf{w} = [0, \quad 0, \quad \omega] = \omega\mathbf{k}, \qquad \mathbf{v} = \mathbf{w} \times \mathbf{r} = [-\omega y, \quad \omega x, \quad 0] = -\omega y\mathbf{i} + \omega x\mathbf{j}.$$

Hence

$$\operatorname{curl} \mathbf{v} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ -\omega y & \omega x & 0 \end{vmatrix} = [0, \quad 0, \quad 2\omega] = 2\omega\mathbf{k} = 2\mathbf{w}.$$

This proves the following theorem. ∎

THEOREM 1

Rotating Body and Curl

The curl of the velocity field of a rotating rigid body has the direction of the axis of the rotation, and its magnitude equals twice the angular speed of the rotation.

The following two relations among grad, div, and curl are basic and shed further light on the nature of the curl.

THEOREM 2

Grad, Div, Curl

***Gradient fields** are **irrotational.** That is, if a continuously differentiable vector function is the gradient of a scalar function f, then its curl is the zero vector,*

$$\operatorname{curl} (\operatorname{grad} f) = \mathbf{0}. \tag{2}$$

Furthermore, the divergence of the curl of a twice continuously differentiable vector function $\mathbf{v}$ is zero,

$$\operatorname{div} (\operatorname{curl} \mathbf{v}) = 0. \tag{3}$$

PROOF Both (2) and (3) follow directly from the definitions by straightforward calculation. In the proof of (3) the six terms cancel in pairs. ∎

EXAMPLE 3 **Rotational and Irrotational Fields**

The field in Example 2 is not irrotational. A similar velocity field is obtained by stirring tea or coffee in a cup. The gravitational field in Theorem 3 of Sec. 9.7 has curl $\mathbf{p} = \mathbf{0}$. It is an irrotational gradient field. ∎

The term "irrotational" for curl $\mathbf{v} = \mathbf{0}$ is suggested by the use of the curl for characterizing the rotation in a field. If a gradient field occurs elsewhere, not as a velocity field, it is usually called **conservative** (see Sec. 9.7). Relation (3) is plausible because of the interpretation of the curl as a rotation and of the divergence as a flux (see Example 2 in Sec. 9.8).

Finally, since the curl is defined in terms of coordinates, we should do what we did for the gradient in Sec. 9.7, namely, to find out whether the curl is a vector. This is true, as follows.

THEOREM 3

Invariance of the Curl

curl **v** *is a vector. That is, it has a length and direction that are independent of the particular choice of a Cartesian coordinate system in space.* (Proof in App. 4.)

PROBLEM SET 9.9

1–6 CALCULATION OF CURL

Find curl **v** for **v** given with respect to right-handed Cartesian coordinates. Show the details of your work.

1. $[y, \quad 2x^2, \quad 0]$
2. $[y^n, \quad z^n, \quad x^n]$ $\quad (n > 0, \text{ integer})$
3. $[e^x \cos y, \quad e^x \sin y, \quad 0]$
4. $(x^2 + y^2 + z^2)^{-3/2}[x, \quad y, \quad z]$
5. $[\ln (x^2 + y^2), \ 2 \arctan (y/x), \quad 0]$
6. $[\sin y, \quad \cos z, \quad -\tan x]$
7. What direction does curl **v** have if **v** is a vector parallel to the xz-plane?
8. Prove Theorem 2. Give two examples for (2) and (3) each.

9–14 FLUID FLOW

Let **v** be the velocity vector of a steady fluid flow. Is the flow irrotational? Incompressible? Find the **streamlines** (the paths of the particles). *Hint.* See the answers to Probs. 9 and 11 for a determination of a path.

9. $\mathbf{v} = [0, \quad z^2, \quad 0]$
10. $\mathbf{v} = [-y^2, \quad 4, \quad 0]$
11. $\mathbf{v} = [y, \quad -x, \quad 0]$
12. $\mathbf{v} = [\csc x, \quad \sec x, \quad 0]$
13. $\mathbf{v} = [x, \quad -y, \quad z]$
14. $\mathbf{v} = [y^3, \quad -x^3, \quad 0]$
15. **WRITING PROJECT. Summary on Grad, Div, Curl.** List the definition and most important facts and formulas for grad, div, curl, and ∇^2. Use your list to write a corresponding essay of 3–4 pages. Include typical examples of your own.
16. **PROJECT. Useful Formulas for the Curl.** Assuming sufficient differentiability, show that
    - **(a)** curl $(\mathbf{u} + \mathbf{v}) =$ curl $\mathbf{u}$ + curl $\mathbf{v}$
    - **(b)** div (curl $\mathbf{v}$) $= 0$
    - **(c)** curl $(f\mathbf{v}) =$ (grad f) $\times \mathbf{v} + f$ curl $\mathbf{v}$
    - **(d)** curl (grad f) $= \mathbf{0}$
    - **(e)** div $(\mathbf{u} \times \mathbf{v}) = \mathbf{v} \cdot$ curl $\mathbf{u} - \mathbf{u} \cdot$ curl $\mathbf{v}$.

17–20 EXPRESSIONS INVOLVING THE CURL

With respect to right-handed coordinates, let $\mathbf{u} = [y^2, \quad z^2, \quad x^2]$, $\mathbf{v} = [yz, \quad zx, \quad xy]$, $f = xyz$, and $g = x + y + z$. Find the following expressions. If one of the formulas in Project 16 applies, use it to check your result. (Show the details of your work.)

17. curl **v**, curl $(f\mathbf{v})$, curl $(g\mathbf{v})$
18. curl $(f\mathbf{u})$, curl $(g\mathbf{u})$
19. $\mathbf{u} \times$ curl **v**, $\mathbf{v} \times$ curl **v**, $\mathbf{u} \cdot$ curl **v**, $\mathbf{v} \cdot$ curl **u**
20. curl $(\mathbf{u} \times \mathbf{v})$, curl $(\mathbf{v} \times \mathbf{u})$

CHAPTER 9 REVIEW QUESTIONS AND PROBLEMS

1. Why did we discuss vectors in R^2 and R^3 in a separate chapter, in addition to Chap. 7 on R^n?
2. What are applications that motivate inner products, vector products, scalar triple products?
3. What is wrong with the expression $\mathbf{a} \times \mathbf{b} \times \mathbf{c}$? With $\mathbf{a} \cdot \mathbf{b} \cdot \mathbf{c}$? With $(\mathbf{a} \cdot \mathbf{b}) \times \mathbf{c}$?
4. What are scalar fields? Vector fields? Potentials? Give examples.
5. What is the gradient? How is it related to directional derivatives?
6. Explain "right-handed coordinates," "orthonormal basis," "tangential acceleration."
7. What is the definition of the divergence? Its physical meaning? Its relation to the Laplacian?
8. Granted sufficient differentiability of a scalar function f and a vector function **v**, which of the following make sense? grad f, f grad f, **v** grad f, $\mathbf{v} \cdot$ grad f, div f, div **v**, div $(f\mathbf{v})$, curl $(f\mathbf{v})$, curl f, f curl **v**, **v** curl f.
9. If $\mathbf{r}(t)$ represents a motion, what is $\mathbf{r}'(t)$, $|\mathbf{r}'(t)|$, $\mathbf{r}''(t)$, $|\mathbf{r}''(t)|$?

10. How do you express the resultant of forces, the moment of a force, and the work done by a force in terms of vectors?

11–20 VECTOR ADDITION, SCALAR MULTIPLICATION, PRODUCTS

In right-handed coordinates let $\mathbf{a} = [3, 2, 7]$, $\mathbf{b} = [6, 5, -4]$, $\mathbf{c} = [1, 8, 0]$, $\mathbf{d} = [9, -2, 0]$. Find

11. $4\mathbf{a} + \mathbf{b} - \mathbf{c} - 2\mathbf{d}$
12. $\mathbf{a} \cdot \mathbf{b}$, $\mathbf{a} \cdot \mathbf{c}$, $\mathbf{a} \times \mathbf{c}$
13. $\mathbf{b} \times \mathbf{b}$, $\mathbf{a} \times \mathbf{b}$, $\mathbf{b} \times \mathbf{a}$
14. $3\mathbf{a} \cdot 4\mathbf{a}$, $12\mathbf{a} \cdot \mathbf{a}$, $12|\mathbf{a}|^2$, $|\mathbf{b}|^2$
15. $2\mathbf{c} \times 5\mathbf{d}$, $10\mathbf{c} \times \mathbf{d}$
16. $(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c}$, $\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})$, $(\mathbf{a} \quad \mathbf{b} \quad \mathbf{c})$
17. $(\mathbf{a} \times \mathbf{b}) \times \mathbf{c}$, $\mathbf{a} \times (\mathbf{b} \times \mathbf{c})$
18. $(1/|\mathbf{a}|)\mathbf{a}$, $(1/|\mathbf{c}|)\mathbf{c}$
19. $(\mathbf{a} \quad \mathbf{b} \quad \mathbf{d})$, $(\mathbf{d} \quad \mathbf{a} \quad \mathbf{b})$
20. $\big||\mathbf{a}| - |\mathbf{b}|\big|$, $|\mathbf{a} + \mathbf{b}|$, $|\mathbf{a}| + |\mathbf{b}|$
21. **(Angle)** Find the angle between $\mathbf{a}$ and $\mathbf{b}$. Between $\mathbf{c}$ and $\mathbf{d}$. Sketch $\mathbf{c}$ and $\mathbf{d}$.
22. **(Angle)** Find the angle between the planes $4x + 3y - z = 2$ and $x + y + z = 1$.
23. In what case is $\mathbf{u} \times \mathbf{v} = \mathbf{v} \times \mathbf{u}$? $\mathbf{u} \cdot \mathbf{v} = \mathbf{v} \cdot \mathbf{u}$?
24. **(Resultant)** Find $\mathbf{u}$ such that $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}$ above, and $\mathbf{u}$ are in equilibrium.
25. **(Resultant)** Find the most general $\mathbf{v}$ such that the resultant of $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}$ above, and $\mathbf{v}$ is parallel to the xy-plane.
26. **(Work)** Find the work done by $\mathbf{q} = [5, 1, 0]$ in the displacement from (4, 4, 0) to (6, −1, 0).
27. **(Component)** Find the component of $\mathbf{u} = [-1, 5, 0]$ in the direction of $\mathbf{v} = [3, 4, 0]$.
28. **(Component)** In what cases is the component of $\mathbf{a}$ in the direction of $\mathbf{b}$ equal to the component of $\mathbf{b}$ in the direction of $\mathbf{a}$?
29. **(Component)** When is the component of $\mathbf{a}$ in the direction of $\mathbf{b}$ negative? Zero?
30. **(Moment)** Find the moment vector $\mathbf{m}$ of $\mathbf{p} = [4, 2, 0]$ about P: (5, 1, 0) if $\mathbf{p}$ acts on a line through (1, 4, 0). Make a sketch.
31. **(Moment)** In what cases is the moment of a force $\mathbf{p} \neq \mathbf{0}$ zero?
32. **(Velocity, acceleration)** Find the velocity, speed, and acceleration of the motion given by

$$\mathbf{r}(t) = [5 \cos t, \quad \sin t, \quad 2t]$$

at the point P: $[5/\sqrt{2}, \quad 1/\sqrt{2}, \quad \pi/2]$. What kind of curve is the path?

33. **(Tetrahedron)** Find the volume of the tetrahedron with vertices (0, 0, 0), (1, 2, 0), (3, −3, 0), (1, 1, 5).
34. **(Plane)** Find an equation of the plane through (1, 0, 2), (2, 3, 5), (3, 5, 7).
35. **(Linear dependence)** Are [2, −1, 3], [4, 2, −5], [−1, 6, 0] linearly dependent? (Give reason.)

36–45 GRAD, DIV, CURL, ∇^2, DIRECTIONAL DERIVATIVE

Let $f = zy + yx$, $\mathbf{v} = [y, z, 4z - x]$, $\mathbf{w} = [y^2, z^2, x^2]$. Find

36. grad f and f grad f at (3, 4, 0)
37. (grad f) × grad f, (grad f) · grad f
38. div $\mathbf{v}$, div $\mathbf{w}$
39. curl $\mathbf{v}$, curl $\mathbf{w}$
40. curl (grad f), div (grad f), div $\mathbf{v}$
41. $\nabla^2(f)$, $\nabla^2(f^2)$
42. $D_{\mathbf{w}}f$ at (1, 2, 0)
43. $D_{\mathbf{v}}f$ at (3, 7, 5)
44. div ($\mathbf{v} \times \mathbf{w}$)
45. curl ($\mathbf{v} \times \mathbf{w}$) + curl ($\mathbf{w} \times \mathbf{v}$)

SUMMARY OF CHAPTER 9
Vector Differential Calculus. Grad, Div, Curl

All vectors of the form $\mathbf{a} = [a_1, a_2, a_3] = a_1\mathbf{i} + a_2\mathbf{j} + a_3\mathbf{k}$ constitute the **real vector space** R^3 with componentwise vector addition

(1) $$[a_1, a_2, a_3] + [b_1, b_2, b_3] = [a_1 + b_1, a_2 + b_2, a_3 + b_3]$$

and componentwise scalar multiplication (c a scalar, a real number)

(2) $$c[a_1, a_2, a_3] = [ca_1, ca_2, ca_3]$$ (Sec. 9.1).

For instance, the *resultant* of forces **a** and **b** is the sum **a** + **b.**

The **inner product** or **dot product** of two vectors is defined by

$$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}||\mathbf{b}| \cos \gamma = a_1 b_1 + a_2 b_2 + a_3 b_3 \tag{3}$$ (Sec. 9.2)

where γ is the angle between **a** and **b.** This gives for the **norm** or **length** $|\mathbf{a}|$ of **a**

$$|\mathbf{a}| = \sqrt{\mathbf{a} \cdot \mathbf{a}} = \sqrt{a_1^2 + a_2^2 + a_3^2} \tag{4}$$

as well as a formula for γ. If $\mathbf{a} \cdot \mathbf{b} = 0$, we call **a** and **b** **orthogonal.** The dot product is suggested by the *work* $W = \mathbf{p} \cdot \mathbf{d}$ done by a force **p** in a displacement **d.**

The **vector product** or **cross product** $\mathbf{v} = \mathbf{a} \times \mathbf{b}$ is a vector of length

$$|\mathbf{a} \times \mathbf{b}| = |\mathbf{a}|\,|\mathbf{b}| \sin \gamma \tag{5}$$ (Sec. 9.3)

and perpendicular to both **a** and **b** such that **a, b, v** form a *right-handed* triple. In terms of components with respect to right-handed coordinates,

$$\mathbf{a} \times \mathbf{b} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix} \tag{6}$$ (Sec. 9.3).

The vector product is suggested, for instance, by moments of forces or by rotations. CAUTION! This multiplication is *anti*commutative, $\mathbf{a} \times \mathbf{b} = -\mathbf{b} \times \mathbf{a}$, and is *not* associative.

An (oblique) box with edges **a, b, c** has volume equal to the absolute value of the **scalar triple product**

$$(\mathbf{a} \quad \mathbf{b} \quad \mathbf{c}) = \mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = (\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c}. \tag{7}$$

Sections 9.4–9.9 extend differential calculus to vector functions

$$\mathbf{v}(t) = [v_1(t), v_2(t), v_3(t)] = v_1(t)\mathbf{i} + v_2(t)\mathbf{j} + v_3(t)\mathbf{k}$$

and to vector functions of more than one variable (see below). The derivative of $\mathbf{v}(t)$ is

$$\mathbf{v}' = \frac{d\mathbf{v}}{dt} = \lim_{\Delta t \to 0} \frac{\mathbf{v}(t + \Delta t) - \mathbf{v}(t)}{\Delta t} = [v_1', v_2', v_3'] = v_1'\mathbf{i} + v_2'\mathbf{j} + v_3'\mathbf{k}. \tag{8}$$

Differentiation rules are as in calculus. They imply (Sec. 9.4)

$$(\mathbf{u} \cdot \mathbf{v})' = \mathbf{u}' \cdot \mathbf{v} + \mathbf{u} \cdot \mathbf{v}', \qquad (\mathbf{u} \times \mathbf{v})' = \mathbf{u}' \times \mathbf{v} + \mathbf{u} \times \mathbf{v}'.$$

Curves C in space represented by the position vector $\mathbf{r}(t)$ have $\mathbf{r}'(t)$ as a **tangent vector** (the **velocity** in mechanics when t is time), $\mathbf{r}'(s)$ (s arc length, Sec. 9.5) as the *unit tangent vector,* and $|\mathbf{r}''(s)| = \kappa$ as the *curvature* (the *acceleration* in mechanics).

Vector functions $\mathbf{v}(x, y, z) = [v_1(x, y, z), v_2(x, y, z), v_3(x, y, z)]$ represent vector fields in space. Partial derivatives with respect to the Cartesian coordinates x, y, z are obtained componentwise, for instance,

$$\frac{\partial \mathbf{v}}{\partial x} = \left[\frac{\partial v_1}{\partial x}, \frac{\partial v_2}{\partial x}, \frac{\partial v_3}{\partial x}\right] = \frac{\partial v_1}{\partial x}\mathbf{i} + \frac{\partial v_2}{\partial x}\mathbf{j} + \frac{\partial v_3}{\partial x}\mathbf{k} \quad \text{(Sec. 9.6).}$$

The **gradient** of a scalar function f is

$$\text{grad } f = \nabla f = \left[\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}\right] \tag{9}$$ (Sec. 9.7).

The **directional derivative** of f in the direction of a vector $\mathbf{a}$ is

$$D_{\mathbf{a}} f = \frac{df}{ds} = \frac{1}{|\mathbf{a}|}\,\mathbf{a} \bullet \nabla f \tag{10}$$ (Sec. 9.7).

The **divergence** of a vector function $\mathbf{v}$ is

$$\text{div } \mathbf{v} = \nabla \bullet \mathbf{v} = \frac{\partial v_1}{\partial x} + \frac{\partial v_2}{\partial y} + \frac{\partial v_3}{\partial z}, \tag{11}$$ (Sec. 9.8).

The **curl** of $\mathbf{v}$ is

$$\text{curl } \mathbf{v} = \nabla \times \mathbf{v} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ v_1 & v_2 & v_3 \end{vmatrix} \tag{12}$$ (Sec. 9.9)

or minus the determinant if the coordinates are left-handed.

Some basic formulas for grad, div, curl are (Secs. 9.7–9.9)

$$\nabla(fg) = f\nabla g + g\nabla f$$
$$\nabla(f/g) = (1/g^2)(g\nabla f - f\nabla g) \tag{13}$$

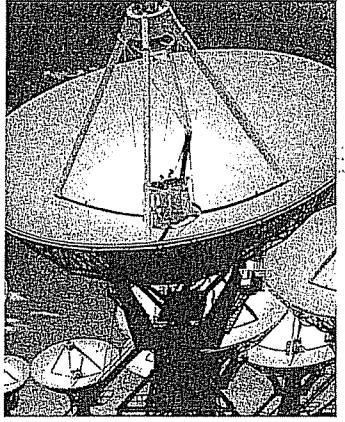
$$\text{div } (f\mathbf{v}) = f \text{ div } \mathbf{v} + \mathbf{v} \bullet \nabla f$$
$$\text{div } (f\nabla g) = f\nabla^2 g + \nabla f \bullet \nabla g \tag{14}$$

$$\nabla^2 f = \text{div } (\nabla f)$$
$$\nabla^2(fg) = g\nabla^2 f + 2\nabla f \bullet \nabla g + f\nabla^2 g \tag{15}$$

$$\text{curl } (f\mathbf{v}) = \nabla f \times \mathbf{v} + f \text{ curl } \mathbf{v}$$
$$\text{div } (\mathbf{u} \times \mathbf{v}) = \mathbf{v} \bullet \text{curl } \mathbf{u} - \mathbf{u} \bullet \text{curl } \mathbf{v} \tag{16}$$

$$\text{curl } (\nabla f) = \mathbf{0}$$
$$\text{div } (\text{curl } \mathbf{v}) = 0. \tag{17}$$

For grad, div, curl, and ∇^2 in **curvilinear coordinates** see App. A3.4.



CHAPTER 10

Vector Integral Calculus. Integral Theorems

This chapter is the companion to Chap. 9. Whereas the previous chapter dealt with differentiation in vector calculus, this chapter concerns integration. This vector integral calculus extends integrals as known from calculus to integrals over curves ("line integrals"), surfaces ("surface integrals"), and solids. We shall see that these integrals have basic engineering applications in solid mechanics, in fluid flow, and in heat problems.

These different kinds of integrals can be transformed into one another. This is done to simplify evaluations or to gain useful general formulas, for instance, in potential theory (see Sec. 10.8). Such transformations are done by the powerful formulas of Green (line integrals into double integrals or conversely, Sec. 10.4), Gauss (surface integrals into triple integrals or conversely, Sec. 10.7), and Stokes (line integrals into surface integrals or conversely, Sec. 10.9).

The root of these transformations was largely physical intuition. The corresponding formulas involve the divergence and the curl and will thus lead to a deeper physical understanding of these two operations.

Prerequisite: Elementary integral calculus, Secs. 9.7–9.9
Sections that may be omitted in a shorter course: 10.3, 10.5, 10.8
References and Answers to Problems: App. 1 Part B, App. 2

10.1 Line Integrals

The concept of a line integral is a simple and natural generalization of a definite integral

$$\int_a^b f(x)\,dx \tag{1}$$

known from calculus. In (1) we integrate the *integrand* $f(x)$ from $x = a$ along the x-axis to $x = b$. In a line integral we shall integrate a given function, also called the **integrand,** along a curve C in space (or in the plane). Hence *curve integral* would be a better name, but line integral is standard.

We represent the curve C by a parametric representation (as in Sec. 9.5)

$$\mathbf{r}(t) = [x(t), y(t), z(t)] = x(t)\mathbf{i} + y(t)\mathbf{j} + z(t)\mathbf{k} \qquad (a \leqq t \leqq b). \tag{2}$$

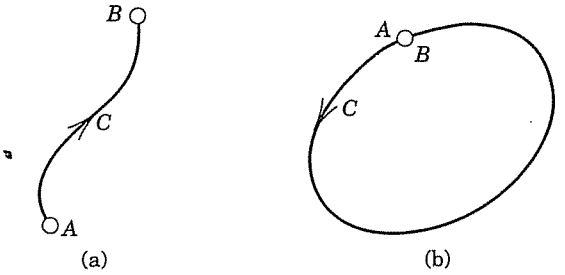


Fig. 217. Oriented curve

The curve C is called the **path of integration,** A: $\mathbf{r}(a)$ its *initial point,* and B: $\mathbf{r}(b)$ its terminal point. C is now *oriented.* The direction from A to B, in which t increases, is called the *positive direction* on C and can be marked by an arrow (as in Fig. 217a). The points A and B may coincide (as in Fig. 217b). Then C is called a *closed path.*

C is called a **smooth curve** if it has at each point a unique tangent whose direction varies continuously as we move along C. Technically: $\mathbf{r}(t)$ in (2) is differentiable and the derivative $\mathbf{r}'(t) = d\mathbf{r}/dt$ is continuous and different from the zero vector at every point of C.

General Assumption

In this book, every path of integration of a line integral is assumed to be **piecewise smooth;** *that is, it consists of* ***finitely many*** *smooth curves.*

For example, the boundary curve of a square is piecewise smooth, consisting of four smooth curves (segments, the four sides).

Definition and Evaluation of Line Integrals

A **line integral** of a vector function $\mathbf{F}(\mathbf{r})$ over a curve C: $\mathbf{r}(t)$ [as in (2)] is defined by

$$\int_C \mathbf{F}(\mathbf{r}) \cdot d\mathbf{r} = \int_a^b \mathbf{F}(\mathbf{r}(t)) \cdot \mathbf{r}'(t)\, dt \qquad \mathbf{r}' = \frac{d\mathbf{r}}{dt} \tag{3}$$

(see Sec. 9.2 for the dot product). In terms of components, with $d\mathbf{r} = [dx, \quad dy, \quad dz]$ as in Sec. 9.5 and $' = d/dt$, formula (3) becomes

$$\begin{aligned} \int_C \mathbf{F}(\mathbf{r}) \cdot d\mathbf{r} &= \int_C (F_1\, dx + F_2\, dy + F_3\, dz) \\ &= \int_a^b (F_1 x' + F_2 y' + F_3 z')\, dt. \end{aligned} \tag{3'}$$

If the path of integration C in (3) is a *closed* curve, then instead of

$$\int_C \qquad \text{we also write} \qquad \oint_C.$$

Note that the integrand in (3) is a scalar, not a vector, because we take the dot product. Indeed, $\mathbf{F} \cdot \mathbf{r}'/|\mathbf{r}'|$ is the tangential component of $\mathbf{F}$. (For "component" see (11) in Sec. 9.2.)

We see that the integral in (3) on the right is a definite integral of a function of t taken over the interval $a \leqq t \leqq b$ on the t-axis in the ***positive*** direction (the direction of increasing t). This definite integral exists for continuous $\mathbf{F}$ and piecewise smooth C, because this makes $\mathbf{F} \cdot \mathbf{r}'$ piecewise continuous.

Line integrals (3) arise naturally in mechanics, where they give the work done by a force $\mathbf{F}$ in a displacement along C (details and examples below). We may thus call the line integral (3) the **work integral.** Other forms of the line integral will be discussed later in this section.

EXAMPLE 1 **Evaluation of a Line Integral in the Plane**

Find the value of the line integral (3) when $\mathbf{F}(\mathbf{r}) = [-y, \quad -xy] = -y\mathbf{i} - xy\mathbf{j}$ and C is the circular arc in Fig. 218 from A to B.

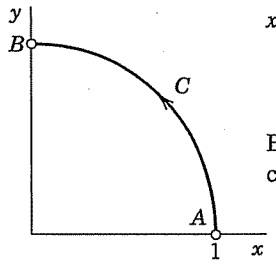


Fig. 218. Example 1

Solution. We may represent C by $\mathbf{r}(t) = [\cos t, \sin t] = \cos t\ \mathbf{i} + \sin t\ \mathbf{j}$, where $0 \leqq t \leqq \pi/2$. Then $x(t) = \cos t, \quad y(t) = \sin t, \quad$ and

$$\mathbf{F}(\mathbf{r}(t)) = -y(t)\mathbf{i} - x(t)y(t)\mathbf{j} = [-\sin t, \quad -\cos t \sin t] = -\sin t\,\mathbf{i} - \cos t \sin t\,\mathbf{j}.$$

By differentiation, $\mathbf{r}'(t) = [-\sin t, \quad \cos t] = -\sin t\ \mathbf{i} + \cos t\ \mathbf{j}$, so that by (3) [use (10) in App. 3.1; set $\cos t = u$ in the second term]

$$\int_C \mathbf{F}(\mathbf{r}) \cdot d\mathbf{r} = \int_0^{\pi/2} [-\sin t, -\cos t \sin t] \cdot [-\sin t, \cos t]\, dt = \int_0^{\pi/2} (\sin^2 t - \cos^2 t \sin t)\, dt$$

$$= \int_0^{\pi/2} \frac{1}{2}(1 - \cos 2t)\, dt - \int_1^0 u^2(-du) = \frac{\pi}{4} - 0 - \frac{1}{3} \approx 0.4521. \qquad \blacksquare$$

EXAMPLE 2 **Line Integral in Space**

The evaluation of line integrals in space is practically the same as it is in the plane. To see this, find the value of (3) when $\mathbf{F}(\mathbf{r}) = [z, x, y] = z\mathbf{i} + x\mathbf{j} + y\mathbf{k}$ and C is the helix (Fig. 219)

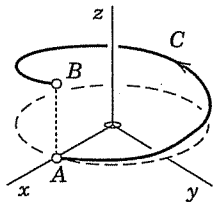


Fig. 219. Example 2

$$\text{(4)} \qquad \mathbf{r}(t) = [\cos t, \sin t, 3t] = \cos t\,\mathbf{i} + \sin t\,\mathbf{j} + 3t\mathbf{k} \qquad (0 \leqq t \leqq 2\pi).$$

Solution. From (4) we have $x(t) = \cos t$, $y(t) = \sin t$, $z(t) = 3t$. Thus

$$\mathbf{F}(\mathbf{r}(t)) \cdot \mathbf{r}'(t) = (3t\mathbf{i} + \cos t\,\mathbf{j} + \sin t\,\mathbf{k}) \cdot (-\sin t\,\mathbf{i} + \cos t\,\mathbf{j} + 3\mathbf{k}).$$

The dot product is $3t(-\sin t) + \cos^2 t + 3 \sin t$. Hence (3) gives

$$\int_C \mathbf{F}(\mathbf{r}) \cdot d\mathbf{r} = \int_0^{2\pi} (-3t \sin t + \cos^2 t + 3 \sin t)\, dt = 6\pi + \pi + 0 = 7\pi \approx 21.99. \qquad \blacksquare$$

Simple general properties of the line integral (3) follow directly from corresponding properties of the definite integral in calculus, namely,

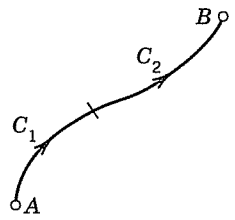


Fig. 220. Formula (5c)

$$\text{(5a)} \qquad \int_C k\mathbf{F} \cdot d\mathbf{r} = k \int_C \mathbf{F} \cdot d\mathbf{r} \qquad (k \text{ constant})$$

$$\text{(5b)} \qquad \int_C (\mathbf{F} + \mathbf{G}) \cdot d\mathbf{r} = \int_C \mathbf{F} \cdot d\mathbf{r} + \int_C \mathbf{G} \cdot d\mathbf{r}$$

$$\text{(5c)} \qquad \int_C \mathbf{F} \cdot d\mathbf{r} = \int_{C_1} \mathbf{F} \cdot d\mathbf{r} + \int_{C_2} \mathbf{F} \cdot d\mathbf{r} \qquad \text{(Fig. 220)}$$

where in (5c) the path C is subdivided into two arcs C_1 and C_2 that have the same orientation as C (Fig. 220). In (5b) the orientation of C is the same in all three integrals. If the sense of integration along C is reversed, the value of the integral is multiplied by -1. However, we note the following independence if the sense is preserved.

THEOREM 1

Direction-Preserving Parametric Transformations

Any representations of C that give the same positive direction on C also yield the same value of the line integral (3).

PROOF A proof follows by the chain rule, where $\mathbf{r}(t)$ is the given representation, $t = \phi(t^*)$ with a positive derivative dt/dt^* is the transformation, with $a^* \leqq t^* \leqq b^*$ corresponding to $a \leqq t \leqq b$ in (3), and we write $\mathbf{r}(t) = \mathbf{r}(\phi(t^*)) = \mathbf{r}^*(t^*)$. Then $dt = (dt/dt^*)\, dt^*$ and

$$\int_C \mathbf{F}(\mathbf{r}^*) \cdot d\mathbf{r}^* = \int_{a^*}^{b^*} \mathbf{F}(\mathbf{r}(\phi(t^*))) \cdot \frac{d\mathbf{r}}{dt} \frac{dt}{dt^*}\, dt^*$$

$$= \int_a^b \mathbf{F}(\mathbf{r}(t)) \cdot \frac{d\mathbf{r}}{dt}\, dt = \int_C \mathbf{F}(\mathbf{r}) \cdot d\mathbf{r}. \qquad \blacksquare$$

Motivation of the Line Integral (3): Work Done by a Force

The work W done by a *constant* force $\mathbf{F}$ in the displacement along a *straight* segment $\mathbf{d}$ is $W = \mathbf{F} \cdot \mathbf{d}$; see Example 2 in Sec. 9.2. This suggests that we define the work W done by a *variable* force $\mathbf{F}$ in the displacement along a curve C: $\mathbf{r}(t)$ as the limit of sums of works done in displacements along small chords of C. We show that this definition amounts to defining W by the line integral (3).

For this we choose points $t_0\ (= a) < t_1 < \cdots < t_n\ (= b)$. Then the work ΔW_m done by $\mathbf{F}(\mathbf{r}(t_m))$ in the straight displacement from $\mathbf{r}(t_m)$ to $\mathbf{r}(t_{m+1})$ is

$$\Delta W_m = \mathbf{F}(\mathbf{r}(t_m)) \cdot [\mathbf{r}(t_{m+1}) - \mathbf{r}(t_m)] \approx \mathbf{F}(\mathbf{r}(t_m)) \cdot \mathbf{r}'(t_m) \Delta t_m \qquad (\Delta t_m = t_{m+1} - t_m).$$

The sum of these n works is $W_n = \Delta W_0 + \cdots + \Delta W_{n-1}$. If we choose points and consider W_n for every n arbitrarily but so that the greatest Δt_m approaches zero as $n \to \infty$, then the limit of W_n as $n \to \infty$ is the line integral (3). This integral exists because of our general assumption that $\mathbf{F}$ is continuous and C is piecewise smooth; this makes $\mathbf{r}'(t)$ continuous, except at finitely many points where C may have corners or cusps. $\blacksquare$

EXAMPLE 3 **Work Done by a Variable Force**

If $\mathbf{F}$ in Example 1 is a force, the work done by $\mathbf{F}$ in the displacement along the quarter-circle is 0.4521, measured in suitable units, say, newton-meters (nt·m, also called joules, abbreviation J; see also front cover). Similarly in Example 2. $\blacksquare$

EXAMPLE 4 **Work Done Equals the Gain in Kinetic Energy**

Let $\mathbf{F}$ be a force, so that (3) is work. Let t be time, so that $d\mathbf{r}/dt = \mathbf{v}$, velocity. Then we can write (3) as

$$W = \int_C \mathbf{F} \cdot d\mathbf{r} = \int_a^b \mathbf{F}(\mathbf{r}(t)) \cdot \mathbf{v}(t)\, dt. \tag{6}$$

Now by Newton's second law (force = mass × acceleration),

$$\mathbf{F} = m\mathbf{r}''(t) = m\mathbf{v}'(t),$$

where m is the mass of the body displaced. Substitution into (5) gives [see (11), Sec. 9.4]

$$W = \int_a^b m\mathbf{v}' \cdot \mathbf{v}\, dt = \int_a^b m \left(\frac{\mathbf{v} \cdot \mathbf{v}}{2} \right)' dt = \frac{m}{2} |\mathbf{v}|^2 \Big|_{t=a}^{t=b}.$$

On the right, $m|\mathbf{v}|^2/2$ is the kinetic energy. Hence *the work done equals the gain in kinetic energy.* This is a basic law in mechanics. ■

Other Forms of Line Integrals

The line integrals

$$\int_C F_1\, dx, \qquad \int_C F_2\, dy, \qquad \int_C F_3\, dz \tag{7}$$

are special cases of (3) when $\mathbf{F} = F_1\mathbf{i}$ or $F_2\mathbf{j}$ or $F_3\mathbf{k}$, respectively.

Furthermore, without taking a dot product as in (3) we can obtain a line integral whose value is a vector rather than a scalar, namely,

$$\int_C \mathbf{F}(\mathbf{r})\, dt = \int_a^b \mathbf{F}(\mathbf{r}(t))\, dt = \int_a^b [F_1(\mathbf{r}(t)),\ F_2(\mathbf{r}(t)),\ F_3(\mathbf{r}(t))]\, dt. \tag{8}$$

Obviously, a special case of (7) is obtained by taking $F_1 = f$, $F_2 = F_3 = 0$. Then

$$\int_C f(\mathbf{r})\, dt = \int_a^b f(\mathbf{r}(t))\, dt \tag{8*}$$

with C as in (2). The evaluation is similar to that before.

EXAMPLE 5 **A Line Integral of the Form (8)**

Integrate $\mathbf{F}(\mathbf{r}) = [xy, yz, z]$ along the helix in Example 2.

Solution. $\mathbf{F}(\mathbf{r}(t)) = [\cos t \sin t, 3t \sin t, 3t]$ integrated with respect to t from 0 to 2π gives

$$\int_0^{2\pi} \mathbf{F}(\mathbf{r}(t))\, dt = \left[-\frac{1}{2}\cos^2 t,\ \ 3\sin t - 3t\cos t,\ \ \frac{3}{2}t^2 \right] \Big|_0^{2\pi} = [0,\ \ -6\pi,\ \ 6\pi^2].$$

■

Path Dependence

Path dependence of line integrals is practically and theoretically so important that we formulate it as a theorem. And a whole section (Sec. 10.2) will be devoted to conditions under which path dependence does not occur.

THEOREM 2

Path Dependence

The line integral (3) *generally depends not only on* $\mathbf{F}$ *and on the endpoints A and B of the path, but also on the path itself along which the integral is taken.*

PROOF Almost any example will show this. Take, for instance, the straight segment C_1: $\mathbf{r}_1(t) = [t, t, 0]$ and the parabola C_2: $\mathbf{r}_2(t) = [t, t^2, 0]$ with $0 \leqq t \leqq 1$ (Fig. 221) and integrate $\mathbf{F} = [0, xy, 0]$. Then $\mathbf{F}(\mathbf{r}_1(t)) \bullet \mathbf{r}_1'(t) = t^2$, $\mathbf{F}(\mathbf{r}_2(t)) \bullet \mathbf{r}_2'(t) = 2t^4$, so that integration gives 1/3 and 2/5, respectively. ∎

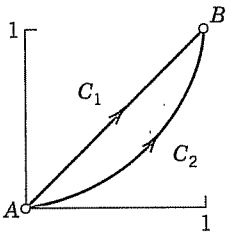


Fig. 221. Proof of Theorem 2

PROBLEM SET 10.1

1–12 LINE INTEGRAL. WORK DONE BY A FORCE

Calculate $\int_C \mathbf{F}(\mathbf{r}) \bullet d\mathbf{r}$ for the following data. If $\mathbf{F}$ is a force, this gives the work done in the displacement along C. (Show the details.)

1. $\mathbf{F} = [y^3, \quad x^3]$, C the parabola $y = 5x^2$ from A: (0, 0) to B: (2, 20)
2. $\mathbf{F}$ as in Prob. 1, C the shortest path from A to B. Is the integral smaller? Give reason.
3. $\mathbf{F}$ as in Prob. 1, C from A straight to (2, 0), then vertically up to B
4. $\mathbf{F} = [x^2, \quad y^2, \quad 0]$, C the semicircle from (2, 0) to $(-2, 0)$, $y \geqq 0$
5. $\mathbf{F} = [xy^2, \quad x^2y]$, C: $\mathbf{r} = [\cosh t, \quad \sinh t, \quad 0]$, $0 \leqq t \leqq 2$. Sketch C.
6. $\mathbf{F} = [e^x, \quad e^y]$ clockwise along the circle with center (0, 0) from (1, 0) to $(0, -1)$
7. $\mathbf{F} = [z, \quad x, \quad y]$, C: $\mathbf{r} = [\cos t, \quad \sin t, \quad t]$ from (1, 0, 0) to $(1, 0, 4\pi)$
8. $\mathbf{F} = [\cosh x, \quad \sinh y, \quad e^z]$, C: $\mathbf{r} = [t, \quad t^2, \quad t^3]$ from (0, 0, 0) to $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8})$
9. $\mathbf{F}$ as in Prob. 8, C the straight segment from (0, 0, 0) to $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8})$
10. $\mathbf{F} = [x, \quad -z, \quad 2y]$ from (0, 0, 0) straight to (1, 1, 0), then to (1, 1, 1), back to (0, 0, 0)
11. $\mathbf{F} = [e^x, \quad e^y, \quad e^z]$, $\mathbf{r} = [t, \quad t^2, \quad t^2]$ from (0, 0, 0) to (2, 4, 4). Sketch C.
12. $\mathbf{F} = [y^2, \quad x^2, \quad \cos^2 z]$, C as in Prob. 7. Sketch C.
13. **WRITING PROJECT. From Definite Integrals to Line Integrals.** Write a short report (1–2 pages) with examples on line integrals as generalizations of definite integrals. The latter give the area under a curve. Explain the corresponding geometric interpretation of a line integral.
14. **PROJECT. Independence of Representation.**

    Dependence on Path. Consider the integral $\int_C \mathbf{F}(\mathbf{r}) \bullet d\mathbf{r}$, where $\mathbf{F} = [xy, -y^2]$.

    (a) One path, several representations. Find the value of the integral when $\mathbf{r} = [\cos t, \sin t]$, $0 \leqq t \leqq \pi/2$. Show that the value remains the same if you set $t = -p$ or $t = p^2$ or apply two other parametric transformations of your own choice.

    (b) Several paths. Evaluate the integral when C: $y = x^n$, thus $\mathbf{r} = [t, t^n]$, $0 \leqq t \leqq 1$, where $n = 1, 2, 3, \cdots$. Note that these infinitely many paths have the same endpoints.

    (c) Limit. What is the limit in (b) as $n \to \infty$? Can you confirm your result by direct integration without referring to (b)?

    (d) Show path dependence with a simple example of your choice involving two paths.

15–18 INTEGRALS OF THE FORMS (8) AND (8*)

Evaluate (8) or (8*) with $\mathbf{F}$ or f and C as follows.

15. $f = x^2 + y^2$, C: $\mathbf{r} = [t, \quad 4t, \quad 0]$, $0 \leqq t \leqq 1$

16. $f = 1 - \sinh^2 x$, C the catenary $\mathbf{r} = [t, \quad \cosh t]$, $0 \leqq t \leqq 2$

17. $\mathbf{F} = [y^2, \quad z^2, \quad x^2]$, C the helix $[3\cos t, \quad 3\sin t, \quad 2t]$, $0 \leqq t \leqq 8\pi$

18. $\mathbf{F} = [(xy)^{1/3}, \quad (y/x)^{1/3}, \quad 0]$, C the hypocycloid $\mathbf{r} = [\cos^3 t, \quad \sin^3 t, \quad 0]$, $0 \leqq t \leqq \pi/4$

19. (***ML*-Inequality, Estimation of Line Integrals**) Let $\mathbf{F}$ be a vector function defined on a curve C. Let $|\mathbf{F}|$ be bounded, say, $|\mathbf{F}| \leqq M$ on C, where M is some positive number. Show that

$$\text{(9)} \qquad \left| \int_C \mathbf{F} \cdot d\mathbf{r} \right| \leqq ML \qquad (L = \text{Length of } C).$$

20. Using (9), find a bound for the absolute value of the work W done by the force $\mathbf{F} = [x^2, \quad y]$ in the displacement along the segment from (0, 0) to (3, 4).

10.2 Path Independence of Line Integrals

In this section we consider line integrals

$$\text{(1)} \qquad \int_C \mathbf{F}(\mathbf{r}) \cdot d\mathbf{r} = \int_C (F_1\,dx + F_2\,dy + F_3\,dz) \qquad (d\mathbf{r} = [dx, \quad dy, \quad dz])$$

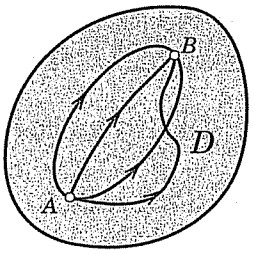


Fig. 222. Path independence

as before, and we shall now find conditions under which (1) is **path independent in a domain D in space.** By definition this means that for every pair of endpoints A, B in D the integral (1) has the same value for all paths in D that begin at A and end at B. (See Fig. 222. See Sec. 9.6 for "domain.")

Path independence is important. For instance, in mechanics it may mean that we have to do the same amount of work regardless of the path to the mountaintop, be it short and steep or long and gentle. Or it may mean that in releasing an elastic spring we get back the work done in expanding it. Not all forces are of this type—think of swimming in a big round pool in which the water is rotating as in a whirlpool.

We shall follow up three ideas that will give path independence of (1) in a domain D if and only if:

(*Theorem* 1) $\mathbf{F} = \text{grad } f$ (see Sec. 9.7 for the gradient).

(*Theorem* 2) Integration around closed curves C in D always gives 0.

(*Theorem* 3) curl $\mathbf{F} = \mathbf{0}$ (provided D is simply connected, as defined below).

Do you see that these theorems can help in understanding the examples and counterexample just mentioned?

Let us begin our discussion with the following very practical criterion for path independence.

THEOREM 1

Path Independence

A line integral (1) *with continuous* F_1, F_2, F_3 *in a domain* D *in space is path independent in* D *if and only if* $\mathbf{F} = [F_1, \quad F_2, \quad F_3]$ *is the gradient of some function* f *in* D,

$$\text{(2)} \qquad \mathbf{F} = \text{grad } f, \qquad \text{thus,} \qquad F_1 = \frac{\partial f}{\partial x}, \qquad F_2 = \frac{\partial f}{\partial y}, \qquad F_3 = \frac{\partial f}{\partial z}.$$

PROOF **(a)** We assume that (2) holds for some function f in D and show that this implies path independence. Let C be any path in D from any point A to any point B in D, given by $\mathbf{r}(t) = [x(t), \quad y(t), \quad z(t)]$, where $a \leqq t \leqq b$. Then from (2), the chain rule in Sec. 9.6, and (3′) in the last section we obtain

$$\begin{aligned}\int_C (F_1\,dx + F_2\,dy + F_3\,dz) &= \int_C \left(\frac{\partial f}{\partial x}\,dx + \frac{\partial f}{\partial y}\,dy + \frac{\partial f}{\partial z}\,dz\right) \\ &= \int_a^b \left(\frac{\partial f}{\partial x}\,\frac{dx}{dt} + \frac{\partial f}{\partial y}\,\frac{dy}{dt} + \frac{\partial f}{\partial z}\,\frac{dz}{dt}\right) dt \\ &= \int_a^b \frac{df}{dt}\,dt = f[x(t), y(t), z(t)]\,\Big|_{t=a}^{t=b} \\ &= f(x(b), y(b), z(b)) - f(x(a), y(a), z(a)) \\ &= f(B) - f(A).\end{aligned}$$

(b) The more complicated proof of the converse, that path independence implies (2) for some f, is given in App. 4. ▩

The last formula in part (a) of the proof,

$$\text{(3)} \qquad \int_A^B (F_1\,dx + F_2\,dy + F_3\,dz) = f(B) - f(A) \qquad [\mathbf{F} = \operatorname{grad} f]$$

is the analog of the usual formula for definite integrals in calculus,

$$\int_a^b g(x)\,dx = G(x)\,\Big|_a^b = G(b) - G(a) \qquad [G'(x) = g(x)].$$

Formula (3) should be applied whenever a line integral is independent of path.

Potential theory relates to our present discussion if we remember from Sec. 9.7 that f is called a *potential* of $\mathbf{F} = \operatorname{grad} f$. Thus the integral (1) is independent of path in D if and only if $\mathbf{F}$ is the gradient of a potential in D.

EXAMPLE 1 **Path Independence**

Show that the integral $\int_C \mathbf{F} \cdot d\mathbf{r} = \int_C (2x\,dx + 2y\,dy + 4z\,dz)$ is path independent in any domain in space and find its value in the integration from *A:* (0, 0, 0) to *B:* (2, 2, 2).

Solution. $\mathbf{F} = [2x, 2y, 4z] = \operatorname{grad} f$, where $f = x^2 + y^2 + 2z^2$ because $\partial f/\partial x = 2x = F_1$, $\partial f/\partial y = 2y = F_2$, $\partial f/\partial z = 4z = F_3$. Hence the integral is independent of path according to Theorem 1, and (3) gives $f(B) - f(A) = f(2, 2, 2) - f(0, 0, 0) = 4 + 4 + 8 = 16$.

If you want to check this, use the most convenient path *C:* $\mathbf{r}(t) = [t, \quad t, \quad t]$, $0 \leqq t \leqq 2$, on which $\mathbf{F}(\mathbf{r}(t)) = [2t, \quad 2t, \quad 4t]$, so that $\mathbf{F}(\mathbf{r}(t)) \cdot \mathbf{r}'(t) = 2t + 2t + 4t = 8t$, and integration from 0 to 2 gives $8 \cdot 2^2/2 = 16$.

If you did not see the potential by inspection, use the method in the next example. ▩

EXAMPLE 2 **Path Independence. Determination of a Potential**

Evaluate the integral $I = \int_C (3x^2\,dx + 2yz\,dy + y^2\,dz)$ from *A:* (0, 1, 2) to *B:* (1, −1, 7) by showing that $\mathbf{F}$ has a potential and applying (3).

Solution. If $\mathbf{F}$ has a potential f, we should have

$$f_x = F_1 = 3x^2, \qquad f_y = F_2 = 2yz, \qquad f_z = F_3 = y^2.$$

We show that we can satisfy these conditions. By integration of f_x and differentiation,

$$f = x^3 + g(y, z), \qquad f_y = g_y = 2yz, \qquad g = y^2z + h(z), \qquad f = x^3 + y^2z + h(z)$$
$$f_z = y^2 + h' = y^2, \qquad h' = 0 \qquad\qquad h = 0, \quad \text{say.}$$

This gives $f(x, y, z) = x^3 + y^2z$ and by (3),

$$I = f(1, -1, 7) - f(0, 1, 2) = 1 + 7 - (0 + 2) = 6. \qquad \blacksquare$$

Path Independence and Integration Around Closed Curves

The simple idea is that two paths with common endpoints (Fig. 223) make up a single closed curve. This gives almost immediately

THEOREM 2

Path Independence

The integral (1) *is path independent in a domain D if and only if its value around every closed path in D is zero.*

PROOF

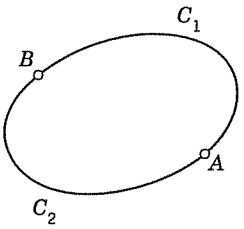


Fig. 223. Proof of Theorem 2

If we have path independence, then integration from A to B along C_1 and along C_2 in Fig. 223 gives the same value. Now C_1 and C_2 together make up a closed curve C, and if we integrate from A along C_1 to B as before, but then in the opposite sense along C_2 back to A (so that this second integral is multiplied by -1), the sum of the two integrals is zero, but this is the integral around the closed curve C.

Conversely, assume that the integral around any closed path C in D is zero. Given any points A and B and any two curves C_1 and C_2 from A to B in D, we see that C_1 with the orientation reversed and C_2 together form a closed path C. By assumption, the integral over C is zero. Hence the integrals over C_1 and C_2, both taken from A to B, must be equal. This proves the theorem. $\blacksquare$

Work. Conservative and Nonconservative (Dissipative) Physical Systems

Recall from the last section that in mechanics, the integral (1) gives the work done by a force $\mathbf{F}$ in the displacement of a body along the curve C. Then Theorem 2 states that work is path independent in D if and only if its value is zero for displacement around every closed path in D. Furthermore, Theorem 1 tells us that this happens if and only if $\mathbf{F}$ is the gradient of a potential in D. In this case, $\mathbf{F}$ and the vector field defined by $\mathbf{F}$ are called **conservative** in D because in this case mechanical energy is conserved; that is, no work is done in the displacement from a point A and back to A. Similarly for the displacement of an electrical charge (an electron, for instance) in a conservative electrostatic field.

Physically, the kinetic energy of a body can be interpreted as the ability of the body to do work by virtue of its motion, and if the body moves in a conservative field of force, after the completion of a round trip the body will return to its initial position with the same kinetic energy it had originally. For instance, the gravitational force is conservative; if we throw a ball vertically up, it will (if we assume air resistance to be negligible) return to our hand with the same kinetic energy it had when it left our hand.

Friction, air resistance, and water resistance always act against the direction of motion, tending to diminish the total mechanical energy of a system (usually converting it into heat or mechanical energy of the surrounding medium, or both), and if in the motion of a body these forces are so large that they can no longer be neglected, then the resultant $\mathbf{F}$ of the forces acting on the body is no longer conservative. Quite generally, a physical system is called **conservative** if all the forces acting in it are conservative; otherwise it is called **nonconservative** or **dissipative.**

Path Independence and Exactness of Differential Forms

Theorem 1 relates path independence of the line integral (1) to the gradient and Theorem 2 to integration around closed curves. A third idea (leading to Theorems 3* and 3, below) relates path independence to the exactness of the **differential form** (or *Pfaffian form*[1])

$$\mathbf{F} \cdot d\mathbf{r} = F_1\,dx + F_2\,dy + F_3\,dz \tag{4}$$

under the integral sign in (1). This form (4) is called **exact** in a domain D in space if it is the differential

$$df = \frac{\partial f}{\partial x}\,dx + \frac{\partial f}{\partial y}\,dy + \frac{\partial f}{\partial z}\,dz = (\text{grad } f) \cdot d\mathbf{r}$$

of a differentiable function $f(x, y, z)$ everywhere in D, that is, if we have

$$\mathbf{F} \cdot d\mathbf{r} = df.$$

Comparing these two formulas, we see that the form (4) is exact if and only if there is a differentiable function $f(x, y, z)$ in D such that everywhere in D,

$$\text{(5)} \qquad \mathbf{F} = \text{grad } f, \qquad \text{thus,} \qquad F_1 = \frac{\partial f}{\partial x}, \qquad F_2 = \frac{\partial f}{\partial y}, \qquad F_3 = \frac{\partial f}{\partial z}.$$

Hence Theorem 1 implies

THEOREM 3*

Path Independence

The integral (1) *is path independent in a domain D in space if and only if the differential form* (4) *has continuous coefficient functions F_1, F_2, F_3 and is exact in D.*

This theorem is practically important because there is a useful exactness criterion. To formulate the criterion, we need the following concept, which is of general interest.

A domain D is called **simply connected** if every closed curve in D can be continuously shrunk to any point in D without leaving D.

For example, the interior of a sphere or a cube, the interior of a sphere with finitely many points removed, and the domain between two concentric spheres are simply

[1]JOHANN FRIEDRICH PFAFF (1765–1825), German mathematician.

connected, while the interior of a torus (a doughnut; see Fig. 247 in Sec. 10.6) and the interior of a cube with one space diagonal removed are not simply connected.

The criterion for exactness (and path independence by Theorem 3*) is now as follows.

THEOREM 3

Criterion for Exactness and Path Independence

Let F_1, F_2, F_3 in the line integral (1),

$$\int_C \mathbf{F}(\mathbf{r}) \cdot d\mathbf{r} = \int_C (F_1\,dx + F_2\,dy + F_3\,dz),$$

be continuous and have continuous first partial derivatives in a domain D in space. Then:

(a) *If the differential form* (4) *is exact in D—and thus* (1) *is path independent by Theorem* 3*—, *then in D,*

$$\textbf{(6)} \qquad \operatorname{curl} \mathbf{F} = \mathbf{0};$$

in components (see Sec. 9.9)

$$\textbf{(6')} \qquad \frac{\partial F_3}{\partial y} = \frac{\partial F_2}{\partial z}, \qquad \frac{\partial F_1}{\partial z} = \frac{\partial F_3}{\partial x}, \qquad \frac{\partial F_2}{\partial x} = \frac{\partial F_1}{\partial y}.$$

(b) *If* (6) *holds in D and D is simply connected, then* (4) *is exact in D—and thus* (1) *is path independent by Theorem* 3*.

PROOF **(a)** If (4) is exact in D, then $\mathbf{F} = \operatorname{grad} f$ in D by Theorem 3*, and, furthermore, $\operatorname{curl} \mathbf{F} = \operatorname{curl} (\operatorname{grad} f) = \mathbf{0}$ by (2) in Sec. 9.9, so that (6) holds.

(b) The proof needs "Stokes's theorem" and will be given in Sec. 10.9. ∎

Line Integral in the Plane. For $\int_C \mathbf{F}(\mathbf{r}) \cdot d\mathbf{r} = \int_C (F_1\,dx + F_2\,dy)$ the curl has only one component (the z-component), so that (6′) reduces to the single relation

$$\textbf{(6'')} \qquad \frac{\partial F_2}{\partial x} = \frac{\partial F_1}{\partial y}$$

(which also occurs in (5) of Sec. 1.4 on exact ODEs).

EXAMPLE 3 **Exactness and Independence of Path. Determination of a Potential**

Using (6′), show that the differential form under the integral sign of

$$I = \int_C \left[2xyz^2\,dx + (x^2z^2 + z\cos yz)\,dy + (2x^2yz + y\cos yz)\,dz\right]$$

is exact, so that we have independence of path in any domain, and find the value of I from A: (0, 0, 1) to B: (1, π/4, 2).

Solution. Exactness follows from (6′), which gives

$$(F_3)_y = 2x^2z + \cos yz - yz \sin yz = (F_2)_z$$
$$(F_1)_z = 4xyz = (F_3)_x$$
$$(F_2)_x = 2xz^2 = (F_1)_y.$$

To find f, we integrate F_2 (which is "long," so that we save work) and then differentiate to compare with F_1 and F_3,

$$f = \int F_2\, dy = \int (x^2z^2 + z \cos yz)\, dy = x^2z^2y + \sin yz + g(x, z)$$
$$f_x = 2xz^2y + g_x = F_1 = 2xyz^2, \qquad g_x = 0, \qquad g = h(z)$$
$$f_z = 2x^2zy + y \cos yz + h' = F_3 = 2x^2zy + y \cos yz, \qquad h' = 0.$$

$h' = 0$ implies $h = const$ and we can take $h = 0$, so that $g = 0$ in the first line. This gives, by (3),

$$f(x, y, z) = x^2yz^2 + \sin yz, \qquad f(B) - f(A) = 1 \cdot \frac{\pi}{4} \cdot 4 + \sin \frac{\pi}{2} - 0 = \pi + 1.$$ ▩

The assumption in Theorem 3 that D is simply connected is essential and cannot be omitted. Perhaps the simplest example to see this is the following.

EXAMPLE 4 **On the Assumption of Simple Connectedness in Theorem 3**

Let

$$F_1 = -\frac{y}{x^2 + y^2}, \qquad F_2 = \frac{x}{x^2 + y^2}, \qquad F_3 = 0. \tag{7}$$

Differentiation shows that (6′) is satisfied in any domain of the xy-plane not containing the origin, for example, in the domain D: $\frac{1}{2} < \sqrt{x^2 + y^2} < \frac{3}{2}$ shown in Fig. 224. Indeed, F_1 and F_2 do not depend on z, and $F_3 = 0$, so that the first two relations in (6′) are trivially true, and the third is verified by differentiation:

$$\frac{\partial F_2}{\partial x} = \frac{x^2 + y^2 - x \cdot 2x}{(x^2 + y^2)^2} = \frac{y^2 - x^2}{(x^2 + y^2)^2},$$

$$\frac{\partial F_1}{\partial y} = -\frac{x^2 + y^2 - y \cdot 2y}{(x^2 + y^2)^2} = \frac{y^2 - x^2}{(x^2 + y^2)^2}.$$

Clearly, D in Fig. 224 is not simply connected. If the integral

$$I = \int_C (F_1\, dx + F_2\, dy) = \int_C \frac{-y\, dx + x\, dy}{x^2 + y^2}$$

were independent of path in D, then $I = 0$ on any closed curve in D, for example, on the circle $x^2 + y^2 = 1$. But setting $x = r \cos \theta$, $y = r \sin \theta$ and noting that the circle is represented by $r = 1$, we have

$$x = \cos \theta, \qquad dx = -\sin \theta\, d\theta, \qquad y = \sin \theta, \qquad dy = \cos \theta\, d\theta,$$

so that $-y\, dx + x\, dy = \sin^2 \theta\, d\theta + \cos^2 \theta\, d\theta = d\theta$ and counterclockwise integration gives

$$I = \int_0^{2\pi} \frac{d\theta}{1} = 2\pi.$$

Since D is not simply connected, we cannot apply Theorem 3 and cannot conclude that I is independent of path in D.

Although $\mathbf{F} = \operatorname{grad} f$, where $f = \arctan (y/x)$ (verify!), we cannot apply Theorem 1 either because the polar angle $f = \theta = \arctan (y/x)$ is not single-valued, as it is required for a function in calculus. ▩

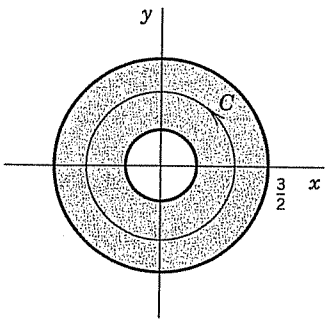


Fig. 224. Example 4

PROBLEM SET 10.2

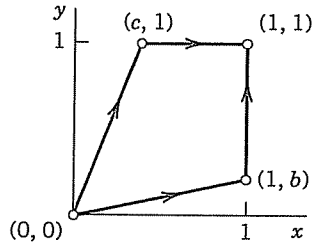
1–8 PATH-INDEPENDENT INTEGRALS

Show that the form under the integral sign is exact in the plane (Probs. 1–4) or in space (Probs. 5–8) and evaluate the integral. (Show the details of your work.)

1. $\int_{(0,0)}^{(4,\pi/8)} (y \cos xy\, dx + x \cos xy\, dy)$
2. $\int_{(5,0)}^{(0,5)} (y^2 e^{2x}\, dx + y e^{2x}\, dy)$
3. $\int_{(-1,-1)}^{(1,1)} e^{-x^2-y^2}(x\, dx + y\, dy)$
4. $\int_{(2,0)}^{(6,\pi)} (\cos^2 y\, dx - 2x \cos y \sin y\, dy)$
5. $\int_{(2,3,0)}^{(0,1,2)} (z\, e^{xz}\, dx + dy + x e^{xz}\, dz)$
6. $\int_{(0,0,0)}^{(1,1,0)} e^{x^2+y^2-2z}\, (x\, dx + y\, dy - dz)$
7. $\int_{(1,0,0)}^{(7,8,0)} (2xy\, dx + x^2\, dy + \sinh z\, dz)$
8. $\int_{(2,0,1)}^{(4,4,0)} [2x(y^3 - z^3)\, dx + 3x^2y^2\, dy - 3x^2z^2\, dz]$
9. Show that in Example 4 of the text we have $\mathbf{F} = \text{grad}\ (\arctan\ (y/x))$. Give examples of domains in which the integral is path independent.
10. **PROJECT. Path Dependence. (a)** Show that $I = \int_C (x^2y\, dx + 2xy^2\, dy)$ is path dependent in the xy-plane.

    (b) Integrate from (0, 0) along the straight-line segment to $(1, b)$, $0 \leqq b \leqq 1$, and then vertically up to (1, 1); see the figure. For which of these paths is I maximum? What is its maximum value?

    (c) Integrate from (0, 0) along the straight-line segment to $(c, 1)$, $0 \leqq c \leqq 1$, and then horizontally to (1, 1). For $c = 1$, do you get the same value as for $b = 1$ in (b)? For which c is I maximum? What is its maximum value?



Project 10. Path Dependence

11–19 CHECK FOR PATH INDEPENDENCE

and, if independent, integrate from (0, 0, 0) to (a, b, c).

11. $(\cosh xz)(z\, dx + x\, dz)$
12. $(3x^2e^{2y} + x)\, dx + 2x^3e^{2y}\, dy$
13. $3x^2y\, dx + x^3\, dy + y\, dz$
14. $2x \sin y\, dx + x^2 \cos y\, dy + y^2\, dz$
15. $(ze^x - e^y)\, dx - xe^y\, dy + e^x\, dz$
16. $e^x \cos 2y\, dx - 2e^x \sin 2y\, dy - xz\, dz$
17. $xy\, z^2\, dx + \frac{1}{2}x^2z^2\, dy + x^2yz\, dz$
18. $yz \cosh x\, dx + z \sinh x\, dy + y \sinh x\, dz$
19. $y\, dx + (x - 2y)\, dy + 4x\, dz$
20. **WRITING PROJECT. Ideas on Path Independence.** Make a list of the main ideas on path independence and dependence in this section. Then work this list into an essay, including explanations of all definitions and on the practical usefulness of the theorems, but no proofs. Include illustrating examples of your own. Explain what happens in Example 4 if you take the domain $0 < \sqrt{x^2 + y^2} < \frac{3}{2}$.

10.3 Calculus Review: Double Integrals. *Optional*

Students familiar with double integrals from calculus should go on to the next section, skipping the present review, which is included to make the book reasonably self-contained.

In a definite integral (1), Sec. 10.1, we integrate a function $f(x)$ over an interval (a segment) of the x-axis. In a double integral we integrate a function $f(x, y)$, called the *integrand,* over a closed bounded region[2] R in the xy-plane, whose boundary curve has a unique tangent at each point, but may perhaps have finitely many cusps (such as the vertices of a triangle or rectangle).

The definition of the double integral is quite similar to that of the definite integral. We subdivide the region R by drawing parallels to the x- and y-axes (Fig. 225). We number the rectangles that are entirely within R from 1 to n. In each such rectangle we choose a point, say, (x_k, y_k) in the kth rectangle, whose area we denote by ΔA_k. Then we form the sum

$$J_n = \sum_{k=1}^{n} f(x_k, y_k)\, \Delta A_k.$$

This we do for larger and larger positive integers n in a completely independent manner, but so that the length of the maximum diagonal of the rectangles approaches zero as n approaches infinity. In this fashion we obtain a sequence of real numbers $J_{n_1}, J_{n_2}, \cdots$. Assuming that $f(x, y)$ is continuous in R and R is bounded by finitely many smooth curves (see Sec. 10.1), one can show (see Ref. [GR4] in App. 1) that this sequence converges and its limit is independent of the choice of subdivisions and corresponding points (x_k, y_k). This limit is called the **double integral** *of* $f(x, y)$ *over the region* R, and is denoted by

$$\iint_R f(x, y)\, dx\, dy \qquad \text{or} \qquad \iint_R f(x, y)\, dA.$$

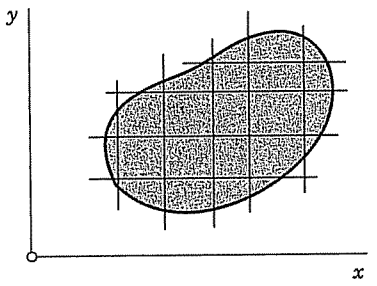


Fig. 225. Subdivision of a region R

[2]A **region** R is a domain (Sec. 9.6) plus, perhaps, some or all of its boundary points. R is **closed** if its **boundary** (all its boundary points) are regarded as belonging to R; and R is **bounded** if it can be enclosed in a circle of sufficiently large radius. A **boundary point** P of R is a point (of R or not) such that every disk with center P contains points of R and also points not of R.

Double integrals have properties quite similar to those of definite integrals. Indeed, for any functions f and g of (x, y), defined and continuous in a region R,

$$\iint_R kf\,dx\,dy = k\iint_R f\,dx\,dy \qquad (k \text{ constant})$$

(1)
$$\iint_R (f + g)\,dx\,dy = \iint_R f\,dx\,dy + \iint_R g\,dx\,dy$$

$$\iint_R f\,dx\,dy = \iint_{R_1} f\,dx\,dy + \iint_{R_2} f\,dx\,dy \qquad \text{(Fig. 226).}$$

Furthermore, if R is simply connected (see Sec. 10.2), then there exists at least one point (x_0, y_0) in R such that we have

(2)
$$\iint_R f(x, y)\,dx\,dy = f(x_0, y_0)A,$$

where A is the area of R. This is called the **mean value theorem** *for double integrals.*

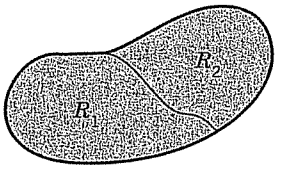


Fig. 226. Formula (1)

Evaluation of Double Integrals by Two Successive Integrations

Double integrals over a region R may be evaluated by two successive integrations. We may integrate first over y and then over x. Then the formula is

(3)
$$\iint_R f(x, y)\,dx\,dy = \int_a^b \left[\int_{g(x)}^{h(x)} f(x, y)\,dy\right] dx \qquad \text{(Fig. 227).}$$

Here $y = g(x)$ and $y = h(x)$ represent the boundary curve of R (see Fig. 227) and, keeping x constant, we integrate $f(x, y)$ over y from $g(x)$ to $h(x)$. The result is a function of x, and we integrate it from $x = a$ to $x = b$ (Fig. 227).

Similarly, for integrating first over x and then over y the formula is

(4)
$$\iint_R f(x, y)\,dx\,dy = \int_c^d \left[\int_{p(y)}^{q(y)} f(x, y)\,dx\right] dy \qquad \text{(Fig. 228).}$$

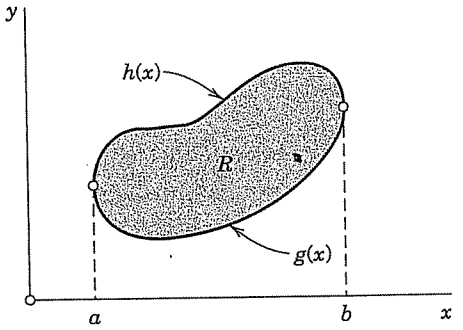


Fig. 227. Evaluation of a double integral

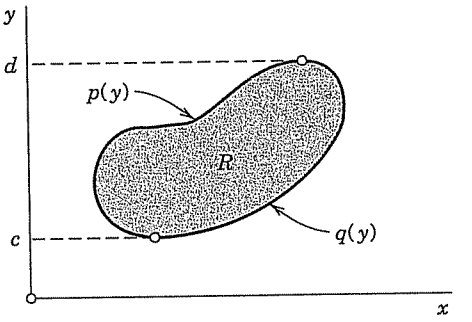


Fig. 228. Evaluation of a double integral

The boundary curve of R is now represented by $x = p(y)$ and $x = q(y)$. Treating y as a constant, we first integrate $f(x, y)$ over x from $p(y)$ to $q(y)$ (see Fig. 228) and then the resulting function of y from $y = c$ to $y = d$.

In (3) we assumed that R can be given by inequalities $a \leqq x \leqq b$ and $g(x) \leqq y \leqq h(x)$. Similarly in (4) by $c \leqq y \leqq d$ and $p(y) \leqq x \leqq q(y)$. If a region R has no such representation, then in any practical case it will at least be possible to subdivide R into finitely many portions each of which can be given by those inequalities. Then we integrate $f(x, y)$ over each portion and take the sum of the results. This will give the value of the integral of $f(x, y)$ over the entire region R.

Applications of Double Integrals

Double integrals have various physical and geometric applications. For instance, the **area** A of a region R in the xy-plane is given by the double integral

$$A = \iint_R dx\, dy.$$

The **volume** V beneath the surface $z = f(x, y)$ (> 0) and above a region R in the xy-plane is (Fig. 229)

$$V = \iint_R f(x, y)\, dx\, dy$$

because the term $f(x_k, y_k)\, \Delta A_k$ in J_n at the beginning of this section represents the volume of a rectangular box with base of area ΔA_k and altitude $f(x_k, y_k)$.

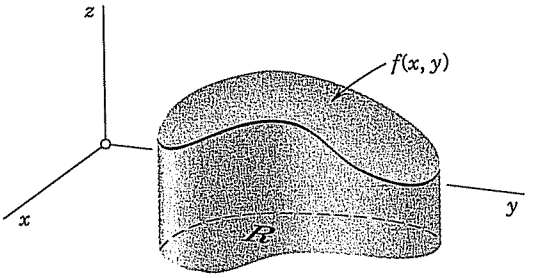


Fig. 229. Double integral as volume

As another application, let $f(x, y)$ be the density (= mass per unit area) of a distribution of mass in the xy-plane. Then the **total mass** M in R is

$$M = \iint_R f(x, y)\, dx\, dy;$$

the **center of gravity** of the mass in R has the coordinates $\bar{x}, \bar{y}$, where

$$\bar{x} = \frac{1}{M} \iint_R x f(x, y)\, dx\, dy \qquad \text{and} \qquad \bar{y} = \frac{1}{M} \iint_R y f(x, y)\, dx\, dy;$$

the **moments of inertia** I_x and I_y of the mass in R about the x- and y-axes, respectively, are

$$I_x = \iint_R y^2 f(x, y)\, dx\, dy, \qquad I_y = \iint_R x^2 f(x, y)\, dx\, dy;$$

and the **polar moment of inertia** I_0 about the origin of the mass in R is

$$I_0 = I_x + I_y = \iint_R (x^2 + y^2) f(x, y)\, dx\, dy.$$

An example is given below.

Change of Variables in Double Integrals. Jacobian

Practical problems often require a change of the variables of integration in double integrals. Recall from calculus that for a definite integral the formula for the change from x to u is

$$\int_a^b f(x)\, dx = \int_\alpha^\beta f(x(u)) \frac{dx}{du}\, du. \tag{5}$$

Here we assume that $x = x(u)$ is continuous and has a continuous derivative in some interval $\alpha \leqq u \leqq \beta$ such that $x(\alpha) = a$, $x(\beta) = b$ [or $x(\alpha) = b$, $x(\beta) = a$] and $x(u)$ varies between a and b when u varies between α and β.

The formula for a change of variables in double integrals from x, y to u, v is

$$\iint_R f(x, y)\, dx\, dy = \iint_{R^*} f(x(u, v), y(u, v)) \left| \frac{\partial(x, y)}{\partial(u, v)} \right| du\, dv; \tag{6}$$

that is, the integrand is expressed in terms of u and v, and $dx\, dy$ is replaced by $du\, dv$ times the absolute value of the **Jacobian**[3]

$$J = \frac{\partial(x, y)}{\partial(u, v)} = \begin{vmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{vmatrix} = \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u}. \tag{7}$$

[3]Named after the German mathematician CARL GUSTAV JACOB JACOBI (1804–1851), known for his contributions to elliptic functions, partial differential equations, and mechanics.

Here we assume the following. The functions

$$x = x(u, v), \qquad y = y(u, v)$$

effecting the change are continuous and have continuous partial derivatives in some region R^* in the uv-plane such that for every (u, v) in R^* the corresponding point (x, y) lies in R and, conversely, to every (x, y) in R there corresponds one and only one (u, v) in R^*; furthermore, the Jacobian J is either positive throughout R^* or negative throughout R^*. For a proof, see Ref. [GR4] in App. 1.

EXAMPLE 1 **Change of Variables in a Double Integral**

Evaluate the following double integral over the square R in Fig. 230.

$$\iint_R (x^2 + y^2)\, dx\, dy$$

Solution. The shape of R suggests the transformation $x + y = u$, $x - y = v$. Then $x = \frac{1}{2}(u + v)$, $y = \frac{1}{2}(u - v)$. The Jacobian is

$$J = \frac{\partial(x, y)}{\partial(u, v)} = \begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{vmatrix} = -\frac{1}{2}.$$

R corresponds to the square $0 \leqq u \leqq 2$, $0 \leqq v \leqq 2$. Therefore,

$$\iint_R (x^2 + y^2)\, dx\, dy = \int_0^2 \int_0^2 \frac{1}{2}(u^2 + v^2)\, \frac{1}{2}\, du\, dv = \frac{8}{3}.$$

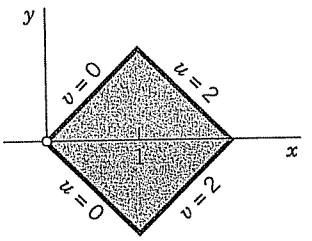


Fig. 230. Region R in Example 1

Of particular practical interest are **polar coordinates** r and θ, which can be introduced by setting $x = r\cos\theta$, $y = r\sin\theta$. Then

$$J = \frac{\partial(x, y)}{\partial(r, \theta)} = \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} = r$$

and

$$\iint_R f(x, y)\, dx\, dy = \iint_{R^*} f(r\cos\theta, r\sin\theta)\, r\, dr\, d\theta \tag{8}$$

where R^* is the region in the $r\theta$-plane corresponding to R in the xy-plane.

EXAMPLE 2 Double Integrals in Polar Coordinates. Center of Gravity. Moments of Inertia

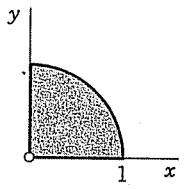


Fig. 231. Example 2

Let $f(x, y) \equiv 1$ be the mass density in the region in Fig. 231. Find the total mass, the center of gravity, and the moments of inertia I_x, I_y, I_0.

Solution. We use the polar coordinates just defined and formula (8). This gives the total mass

$$M = \int_R\int dx\,dy = \int_0^{\pi/2}\int_0^1 r\,dr\,d\theta = \int_0^{\pi/2}\frac{1}{2}\,d\theta = \frac{\pi}{4}.$$

The center of gravity has the coordinates

$$\bar{x} = \frac{4}{\pi}\int_0^{\pi/2}\int_0^1 r\cos\theta\, r\,dr\,d\theta = \frac{4}{\pi}\int_0^{\pi/2}\frac{1}{3}\cos\theta\,d\theta = \frac{4}{3\pi} = 0.4244$$

$$\bar{y} = \frac{4}{3\pi} \qquad \text{for reasons of symmetry.}$$

The moments of inertia are

$$I_x = \int_R\int y^2\,dx\,dy = \int_0^{\pi/2}\int_0^1 r^2\sin^2\theta\, r\,dr\,d\theta = \int_0^{\pi/2}\frac{1}{4}\sin^2\theta\,d\theta$$

$$= \int_0^{\pi/2}\frac{1}{8}(1-\cos 2\theta)\,d\theta = \frac{1}{8}\left(\frac{\pi}{2}-0\right) = \frac{\pi}{16} = 0.1963$$

$$I_y = \frac{\pi}{16} \qquad \text{for reasons of symmetry,} \qquad I_0 = I_x + I_y = \frac{\pi}{8} = 0.3927.$$

Why are $\bar{x}$ and $\bar{y}$ less than $\frac{1}{2}$?

This is the end of our review on double integrals. These integrals will be needed in this chapter, beginning in the next section.

PROBLEM SET 10.3

1. (Mean value theorem) Illustrate (2) with an example.

2–9 DOUBLE INTEGRALS

Describe the region of integration and evaluate. (Show the details.)

2. $\int_0^1\int_x^{2x}(x+y)^2\,dy\,dx$

3. $\int_0^1\int_{x^2}^{x}(1-2xy)\,dy\,dx$

4. As Prob. 3, order reversed

5. $\int_0^3\int_0^{y}\cosh(x+y)\,dx\,dy$

6. As Prob. 5, order reversed

7. $\int_0^4\int_{-x}^{x}e^{x+2y}\,dy\,dx$

8. $\int_0^1\int_{1-x}^{1-x^2}x^2y\,dy\,dx$

9. $\int_0^{\pi/2}\int_0^{\sin y}e^x\cos y\,dx\,dy$

10. Integrate $xye^{x^2-y^2}$ over the triangular region with vertices (0, 0), (1, 1), (1, 2).

11–13 VOLUME

Find the volume of the following regions in space.

11. The region beneath $z = x^2 + y^2$ and above the square with vertices (1, 1), (−1, 1), (−1, −1), (1, −1)

12. The tetrahedron cut from the first octant by the plane $\frac{3}{2}x + 2y + z = 6$. Check by vector methods.

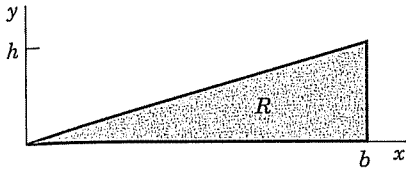
13. The first octant section cut from the region inside the cylinder $x^2 + z^2 = 1$ by the planes $y = 0$, $z = 0$, $x = y$.

14–16 CENTER OF GRAVITY

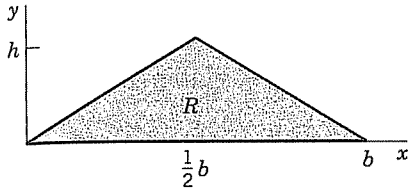
Find the center of gravity $(\bar{x}, \bar{y})$ of a mass of density $f(x, y) = 1$ in the given region R.

14. R the semidisk $x^2 + y^2 \leqq a^2$, $y \geqq 0$

15.



16.



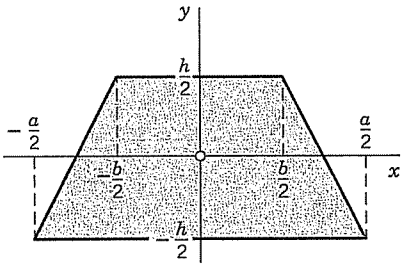
17–20 MOMENTS OF INERTIA

Find the moments of inertia I_x, I_y, I_0 of a mass of density $f(x, y) = 1$ in the region R shown in the figures (which the engineer is likely to need, along with other profiles listed in engineering handbooks).

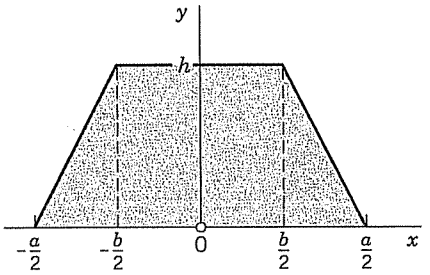
17. R as in Prob. 15.

18. R as in Prob. 16.

19.



20.



10.4 Green's Theorem in the Plane

Double integrals over a plane region may be transformed into line integrals over the boundary of the region and conversely. This is of practical interest because it may simplify the evaluation of an integral. It also helps in the theory whenever we want to switch from one kind of integral to the other. The transformation can be done by the following theorem.

THEOREM 1

Green's Theorem in the Plane[4]
(Transformation between Double Integrals and Line Integrals)

Let R be a closed bounded region (see Sec. 10.3) *in the xy-plane whose boundary C consists of finitely many smooth curves* (see Sec. 10.1). *Let $F_1(x, y)$ and $F_2(x, y)$ be functions that are continuous and have continuous partial derivatives $\partial F_1/\partial y$ and $\partial F_2/\partial x$ everywhere in some domain containing R. Then*

$$\text{(1)} \qquad \iint_R \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} \right) dx\, dy = \oint_C (F_1\, dx + F_2\, dy).$$

Here we integrate along the entire boundary C of R in such a sense that R is on the left as we advance in the direction of integration (see Fig. 232 on p. 440).

[4]GEORGE GREEN (1793–1841), English mathematician who was self-educated, started out as a baker, and at his death was fellow of Caius College, Cambridge. His work concerned potential theory in connection with electricity and magnetism, vibrations, waves, and elasticity theory. It remained almost unknown, even in England, until after his death.

A "domain containing R" in the theorem guarantees that the assumptions about F_1 and F_2 at boundary points of R are the same as at other points of R.

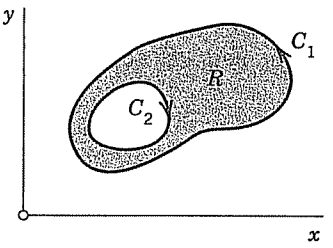


Fig. 232. Region R whose boundary C consists of two parts: C_1 is traversed counterclockwise, while C_2 is traversed clockwise in such a way that R is on the left for both curves

Setting $\mathbf{F} = [F_1, F_2] = F_1\mathbf{i} + F_2\mathbf{j}$ *and using* (1) *in* Sec. 9.9, we obtain (1) in vectorial form,

$$(1') \qquad \iint_R (\operatorname{curl} \mathbf{F}) \cdot \mathbf{k}\, dx\, dy = \oint_C \mathbf{F} \cdot d\mathbf{r}.$$

The proof follows after the first example. For $\oint$ see Sec. 10.1.

EXAMPLE 1 **Verification of Green's Theorem in the Plane**

Green's theorem in the plane will be quite important in our further work. Before proving it, let us get used to it by verifying it for $F_1 = y^2 - 7y$, $F_2 = 2xy + 2x$ and C the circle $x^2 + y^2 = 1$.

Solution. In (1) on the left we get

$$\iint_R \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} \right) dx\, dy = \iint_R [(2y + 2) - (2y - 7)]\, dx\, dy = 9 \iint_R dx\, dy = 9\pi$$

since the circular disk R has area π.

We now show that the line integral in (1) on the right gives the same value, 9π. We must orient C counterclockwise, say, $\mathbf{r}(t) = [\cos t, \sin t]$. Then $\mathbf{r}'(t) = [-\sin t, \cos t]$, and on C,

$$F_1 = y^2 - 7y = \sin^2 t - 7 \sin t, \qquad F_2 = 2xy + 2x = 2 \cos t \sin t + 2 \cos t.$$

Hence the line integral in (1) becomes, verifying Green's theorem,

$$\begin{aligned} \oint_C (F_1 x' + F_2 y')\, dt &= \int_0^{2\pi} \left[(\sin^2 t - 7 \sin t)(-\sin t) + 2(\cos t \sin t + \cos t)(\cos t) \right] dt \\ &= \int_0^{2\pi} (-\sin^3 t + 7 \sin^2 t + 2 \cos^2 t \sin t + 2 \cos^2 t)\, dt \\ &= 0 + 7\pi - 0 + 2\pi = 9\pi. \end{aligned}$$

PROOF We prove Green's theorem in the plane, first for a *special region* R that can be represented in both forms

$$a \leqq x \leqq b, \qquad u(x) \leqq y \leqq v(x) \qquad \text{(Fig. 233)}$$

and

$$c \leqq y \leqq d, \qquad p(y) \leqq x \leqq q(y) \qquad \text{(Fig. 234).}$$

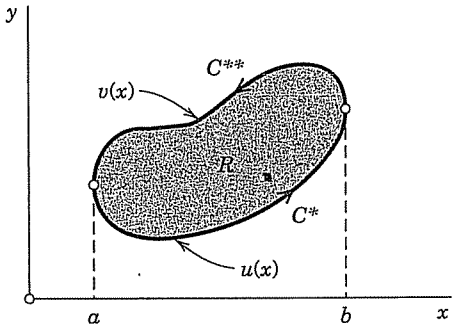


Fig. 233. Example of a special region

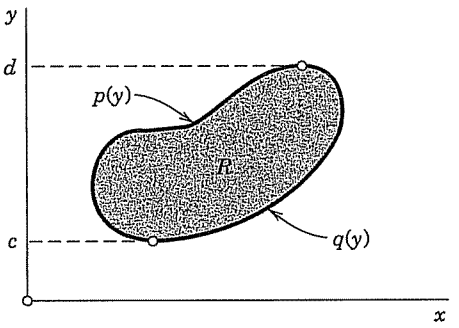


Fig. 234. Example of a special region

Using (3) in the last section, we obtain for the second term on the left side of (1) taken without the minus sign

$$\text{(2)} \qquad \iint_R \frac{\partial F_1}{\partial y}\,dx\,dy = \int_a^b \left[\int_{u(x)}^{v(x)} \frac{\partial F_1}{\partial y}\,dy \right] dx \qquad \text{(see Fig. 233).}$$

(The first term will be considered later.) We integrate the inner integral:

$$\int_{u(x)}^{v(x)} \frac{\partial F_1}{\partial y}\,dy = F_1(x, y) \Big|_{y=u(x)}^{y=v(x)} = F_1[x, v(x)] - F_1[x, u(x)].$$

By inserting this into (2) we find (changing a direction of integration)

$$\begin{aligned} \iint_R \frac{\partial F_1}{\partial y}\,dx\,dy &= \int_a^b F_1[x, v(x)]\,dx - \int_a^b F_1[x, u(x)]\,dx \\ &= -\int_b^a F_1[x, v(x)]\,dx - \int_a^b F_1[x, u(x)]\,dx. \end{aligned}$$

Since $y = v(x)$ represents the curve C^{**} (Fig. 233) and $y = u(x)$ represents C^*, the last two integrals may be written as line integrals over C^{**} and C^* (oriented as in Fig. 233); therefore,

$$\text{(3)} \qquad \begin{aligned} \iint_R \frac{\partial F_1}{\partial y}\,dx\,dy &= -\int_{C^{**}} F_1(x, y)\,dx - \int_{C^*} F_1(x, y)\,dx \\ &= -\oint_C F_1(x, y)\,dx. \end{aligned}$$

This proves (1) in Green's theorem if $F_2 = 0$.

The result remains valid if C has portions parallel to the y-axis (such as $\tilde{C}$ and $\tilde{\tilde{C}}$ in Fig. 235). Indeed, the integrals over these portions are zero because in (3) on the right we integrate with respect to x. Hence we may add these integrals to the integrals over C^* and C^{**} to obtain the integral over the whole boundary C in (3).

We now treat the first term in (1) on the left in the same way. Instead of (3) in the last section we use (4), and the second representation of the special region (see Fig. 234). Then (again changing a direction of integration)

$$\iint_R \frac{\partial F_2}{\partial x}\,dx\,dy = \int_c^d \left[\int_{p(y)}^{q(y)} \frac{\partial F_2}{\partial x}\,dx\right] dy$$

$$= \int_c^d F_2(q(y), y)\,dy + \int_d^c F_2(p(y), y)\,dy$$

$$= \oint_C F_2(x, y)\,dy.$$

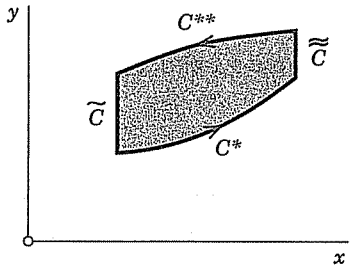


Fig. 235. Proof of Green's theorem

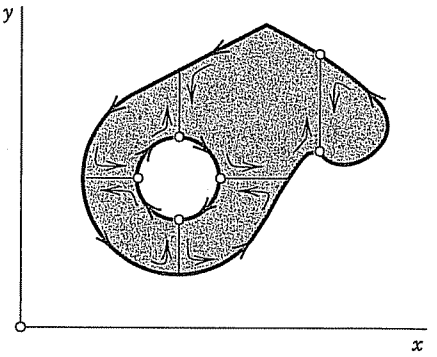


Fig. 236. Proof of Green's theorem

Together with (3) this gives (1) and proves Green's theorem for special regions.

We now prove the theorem for a region R that itself is not a special region but can be subdivided into finitely many special regions (Fig. 236). In this case we apply the theorem to each subregion and then add the results; the left-hand members add up to the integral over R while the right-hand members add up to the line integral over C plus integrals over the curves introduced for subdividing R. The simple ***key observation*** now is that each of the latter integrals occurs twice, taken once in each direction. Hence they cancel each other, leaving us with the line integral over C.

The proof thus far covers all regions that are of interest in practical problems. To prove the theorem for a most general region R satisfying the conditions in the theorem, we must approximate R by a region of the type just considered and then use a limiting process. For details of this see Ref. [GR4] in App. 1. ▩

Some Applications of Green's Theorem

EXAMPLE 2 **Area of a Plane Region as a Line Integral Over the Boundary**

In (1) we first choose $F_1 = 0$, $F_2 = x$ and then $F_1 = -y$, $F_2 = 0$. This gives

$$\iint_R dx\,dy = \oint_C x\,dy \qquad \text{and} \qquad \iint_R dx\,dy = -\oint_C y\,dx$$

respectively. The double integral is the area A of R. By addition we have

(4) $$A = \frac{1}{2}\oint_C (x\,dy - y\,dx)$$

where we integrate as indicated in Green's theorem. This interesting formula expresses the area of R in terms of a line integral over the boundary. It is used, for instance, in the theory of certain **planimeters** (mechanical instruments for measuring area). See also Prob. 17.

For an **ellipse** $x^2/a^2 + y^2/b^2 = 1$ or $x = a\cos t$, $y = b\sin t$ we get $x' = -a\sin t$, $y' = b\cos t$; thus from (4) we obtain the familiar formula for the area of the region bounded by an ellipse,

$$A = \frac{1}{2}\int_0^{2\pi} (xy' - yx')\,dt = \frac{1}{2}\int_0^{2\pi} \left[ab\cos^2 t - (-ab\sin^2 t)\right] dt = \pi ab. \qquad \blacksquare$$

EXAMPLE 3 **Area of a Plane Region in Polar Coordinates**

Let r and θ be polar coordinates defined by $x = r\cos\theta$, $y = r\sin\theta$. Then

$$dx = \cos\theta\,dr - r\sin\theta\,d\theta, \qquad dy = \sin\theta\,dr + r\cos\theta\,d\theta,$$

and (4) becomes a formula that is well known from calculus, namely,

$$A = \frac{1}{2}\oint_C r^2\,d\theta. \tag{5}$$

As an application of (5), we consider the **cardioid** $r = a(1 - \cos\theta)$, where $0 \leqq \theta \leqq 2\pi$ (Fig. 237). We find

$$A = \frac{a^2}{2}\int_0^{2\pi} (1 - \cos\theta)^2\,d\theta = \frac{3\pi}{2}a^2. \qquad \blacksquare$$

EXAMPLE 4 **Transformation of a Double Integral of the Laplacian of a Function into a Line Integral of Its Normal Derivative**

The Laplacian plays an important role in physics and engineering. A first impression of this was obtained in Sec. 9.7, and we shall discuss this further in Chap. 12. At present, let us use Green's theorem for deriving a basic integral formula involving the Laplacian.

We take a function $w(x, y)$ that is continuous and has continuous first and second partial derivatives in a domain of the xy-plane containing a region R of the type indicated in Green's theorem. We set $F_1 = -\partial w/\partial y$ and $F_2 = \partial w/\partial x$. Then $\partial F_1/\partial y$ and $\partial F_2/\partial x$ are continuous in R, and in (1) on the left we obtain

$$\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} = \frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2} = \nabla^2 w, \tag{6}$$

the Laplacian of w (see Sec. 9.7). Furthermore, using those expressions for F_1 and F_2, we get in (1) on the right

$$\oint_C (F_1\,dx + F_2\,dy) = \oint_C \left(F_1\frac{dx}{ds} + F_2\frac{dy}{ds}\right) ds = \oint_C \left(-\frac{\partial w}{\partial y}\frac{dx}{ds} + \frac{\partial w}{\partial x}\frac{dy}{ds}\right) ds \tag{7}$$

where s is the arc length of C, and C is oriented as shown in Fig. 238. The integrand of the last integral may be written as the dot product

$$(\text{grad } w)\cdot\mathbf{n} = \left[\frac{\partial w}{\partial x}, \frac{\partial w}{\partial y}\right]\cdot\left[\frac{dy}{ds}, -\frac{dx}{ds}\right] = \frac{\partial w}{\partial x}\frac{dy}{ds} - \frac{\partial w}{\partial y}\frac{dx}{ds}. \tag{8}$$

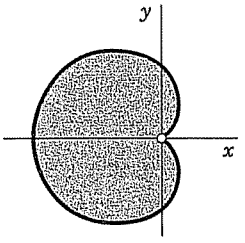


Fig. 237. Cardioid

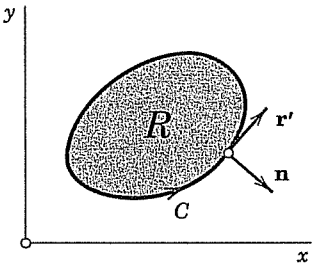


Fig. 238. Example 4

The vector $\mathbf{n}$ is a unit normal vector to C, because the vector $\mathbf{r}'(s) = d\mathbf{r}/ds = [dx/ds, \quad dy/ds]$ is the unit tangent vector of C, and $\mathbf{r}' \cdot \mathbf{n} = 0$, so that $\mathbf{n}$ is perpendicular to $\mathbf{r}'$. Also, $\mathbf{n}$ is directed to the *exterior* of C because in Fig. 238 the positive x-component dx/ds of $\mathbf{r}'$ is the negative y-component of $\mathbf{n}$, and similarly at other points. From this and (4) in Sec. 9.7 we see that the left side of (8) is the derivative of w in the direction of the outward normal of C. This derivative is called the **normal derivative** of w and is denoted by $\partial w/\partial n$; that is, $\partial w/\partial n = (\text{grad } w) \cdot \mathbf{n}$. Because of (6), (7), and (8), Green's theorem gives the desired formula relating the Laplacian to the normal derivative,

$$\textbf{(9)} \qquad \iint_R \nabla^2 w \, dx \, dy = \oint_C \frac{\partial w}{\partial n} \, ds.$$

For instance, $w = x^2 - y^2$ satisfies Laplace's equation $\nabla^2 w = 0$. Hence its normal derivative integrated over a closed curve must give 0. Can you verify this directly by integration, say, for the square $0 \leqq x \leqq 1$, $0 \leqq y \leqq 1$? ■

Green's theorem in the plane may facilitate the evaluation of integrals and can be used in both directions, depending on the kind of integral that is simpler in a concrete case. This is illustrated further in the problem set. Moreover, and perhaps more fundamentally, Green's theorem will be the essential tool in the proof of a very important integral theorem, namely, Stokes's theorem in Sec. 10.9.

PROBLEM SET 10.4

1–12 EVALUATION OF LINE INTEGRALS BY GREEN'S THEOREM

Using Green's theorem, evaluate $\int_C \mathbf{F}(\mathbf{r}) \cdot d\mathbf{r}$ counterclockwise around the boundary curve C of the region R, where

1. $\mathbf{F} = [\frac{1}{2}xy^4, \quad \frac{1}{2}x^4y]$, R the rectangle with vertices (0, 0), (3, 0), (3, 2), (0, 2)
2. $\mathbf{F} = [y \sin x, \quad 2x \cos y]$, R the square with vertices (0, 0), $(\frac{1}{2}\pi, 0)$, $\frac{1}{2}\pi, \frac{1}{2}\pi)$, $(0, \frac{1}{2}\pi)$
3. $\mathbf{F} = [-y^3, \quad x^3]$, C the circle $x^2 + y^2 = 25$
4. $\mathbf{F} = [-e^y, \quad e^x]$, R the triangle with vertices (0, 0), (2, 0), (2, 1)
5. $\mathbf{F} = [e^{x+y}, \quad e^{x-y}]$, R the triangle with vertices (0, 0), (1, 1), (1, 2)
6. $\mathbf{F} = [x \cosh y, \quad x^2 \sinh y]$, R: $x^2 \leqq y \leqq x$. Sketch R.
7. $\mathbf{F} = [x^2 + y^2, \quad x^2 - y^2]$, R: $1 \leqq y \leqq 2 - x^2$. Sketch R.
8. $\mathbf{F} = [e^x \cos y, \quad -e^x \sin y]$, R the semidisk $x^2 + y^2 \leqq a^2$, $x \geqq 0$
9. $\mathbf{F} = \text{grad } (x^3 \cos^2 (xy))$, R the region in Prob. 7
10. $\mathbf{F} = [x \ln y, \quad ye^x]$, R the rectangle with vertices (0, 1), (3, 1), (3, 2), (0, 2)
11. $\mathbf{F} = [2x - 3y, \quad x + 5y]$, R: $16x^2 + 25y^2 \leqq 400$, $y \geqq 0$
12. $\mathbf{F} = [x^2y^2, \quad -x/y^2]$, R: $1 \leqq x^2 + y^2 \leqq 4$, $x \geqq 0$, $y \geqq x$. Sketch R.

13–16 INTEGRAL OF THE NORMAL DERIVATIVE

Using (9), evaluate $\oint_C \frac{\partial w}{\partial n} \, ds$ counterclockwise over the boundary curve C of the region R.

13. $w = \sinh x$, R the triangle with vertices (0, 0), (2, 0), (2, 1)
14. $w = x^2 + y^2$, C: $x^2 + y^2 = 1$. Confirm the answer by direct integration.
15. $w = 2 \ln (x^2 + y^2) + xy^3$, R: $1 \leqq y \leqq 5 - x^2$, $x \geqq 0$
16. $w = x^6y + xy^6$, R: $x^2 + y^2 \leqq 4$, $y \geqq 0$

17. **CAS EXPERIMENT.** Apply (4) to figures of your choice whose area can also be obtained by another method and compare the results.
18. **(Laplace's equation)** Show that for a solution $w(x, y)$ of Laplace's equation $\nabla^2 w = 0$ in a region R with boundary curve C and outer unit normal vector $\mathbf{n}$,

$$\textbf{(10)} \qquad \iint_R \left[\left(\frac{\partial w}{\partial x} \right)^2 + \left(\frac{\partial w}{\partial y} \right)^2 \right] dx \, dy = \oint_C w \frac{\partial w}{\partial n} \, ds.$$

19. Show that $w = 2e^x \cos y$ satisfies Laplace's equation $\nabla^2 w = 0$ and, using (10), integrate $w(\partial w/\partial n)$ counterclockwise around the boundary curve C of the square $0 \leqq x \leqq 2$, $0 \leqq y \leqq 2$.

20. PROJECT. Other Forms of Green's Theorem in the Plane. Let R and C be as in Green's theorem, $\mathbf{r}'$ a unit tangent vector, and $\mathbf{n}$ the outer unit normal vector of C (Fig. 238 in Example 4). Show that (1) may be written

$$\text{(11)} \qquad \iint_R \operatorname{div} \mathbf{F}\, dx\, dy = \oint_C \mathbf{F} \cdot \mathbf{n}\, ds$$

or

$$\text{(12)} \qquad \iint_R (\operatorname{curl} \mathbf{F}) \cdot \mathbf{k}\, dx\, dy = \oint_C \mathbf{F} \cdot \mathbf{r}'\, ds$$

where $\mathbf{k}$ is a unit vector perpendicular to the xy-plane. Verify (11) and (12) for $\mathbf{F} = [7x, -3y]$ and C the circle $x^2 + y^2 = 4$ as well as for an example of your own choice.

10.5 Surfaces for Surface Integrals

Having introduced double integrals over regions in the plane, we turn next to surface integrals, in which we integrate over surfaces in space, such as a sphere or a portion of a cylinder. For this we must first see how to represent a surface. And we must discuss surface normals, since they are also needed in surface integrals. For simplicity we shall say "surface" also for a portion of a surface.

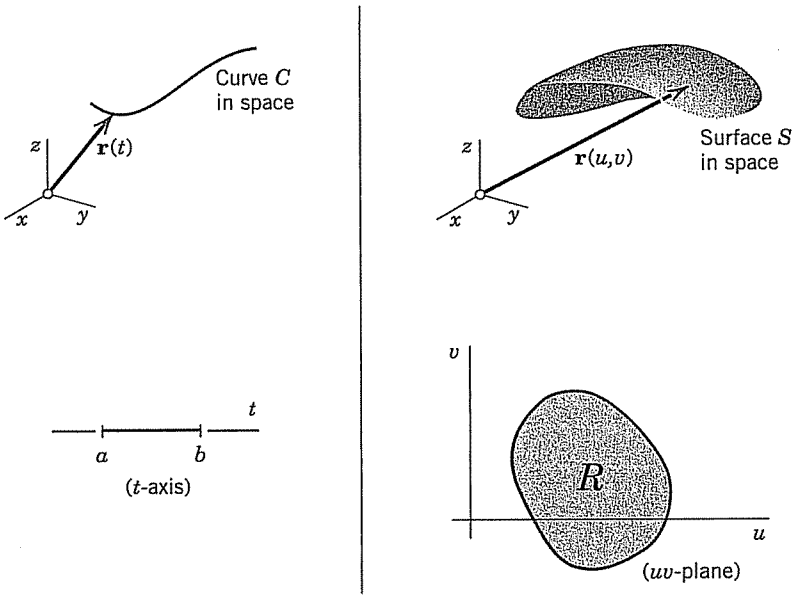
Representation of Surfaces

Representations of a surface S in xyz-space are

$$\text{(1)} \qquad z = f(x, y) \qquad \text{or} \qquad g(x, y, z) = 0.$$

For example, $z = +\sqrt{a^2 - x^2 - y^2}$ or $x^2 + y^2 + z^2 - a^2 = 0$ $(z \geqq 0)$ represents a hemisphere of radius a and center 0.

Now for *curves* C in line integrals, it was more practical and gave greater flexibility to use a *parametric* representation $\mathbf{r} = \mathbf{r}(t)$, where $a \leqq t \leqq b$. This is a mapping of the interval $a \leqq t \leqq b$, located on the t-axis, onto the curve C (actually a portion of it) in xyz-space. It maps every t in that interval onto the point of C with position vector $\mathbf{r}(t)$. See Fig. 239A.



(A) Curve (B) Surface

Fig. 239. Parametric representations of a curve and a surface

Similarly, for surfaces S in surface integrals, it will often be more practical to use a *parametric* representation. Surfaces are *two*-dimensional. Hence we need *two* parameters, which we call u and v. Thus a **parametric representation** of a surface S in space is of the form

$$\mathbf{r}(u, v) = [x(u, v), y(u, v), z(u, v)] = x(u, v)\mathbf{i} + y(u, v)\mathbf{j} + z(u, v)\mathbf{k} \tag{2}$$

where (u, v) varies in some region R of the uv-plane. This mapping (2) maps every point (u, v) in R onto the point of S with position vector $\mathbf{r}(u, v)$. See Fig. 239B.

EXAMPLE 1 **Parametric Representation of a Cylinder**

The circular cylinder $x^2 + y^2 = a^2$, $-1 \leqq z \leqq 1$, has radius a, height 2, and the z-axis as axis. A parametric representation is

$$\mathbf{r}(u, v) = [a \cos u, a \sin u, v] = a \cos u\, \mathbf{i} + a \sin u\, \mathbf{j} + v\mathbf{k} \qquad \text{(Fig. 240).}$$

The components of $\mathbf{r}$ are $x = a \cos u$, $y = a \sin u$, $z = v$. The parameters u, v vary in the rectangle R: $0 \leqq u \leqq 2\pi$, $-1 \leqq v \leqq 1$ in the uv-plane. The curves $u = const$ are vertical straight lines. The curves $v = const$ are parallel circles. The point P in Fig. 240 corresponds to $u = \pi/3 = 60°$, $v = 0.7$. ■

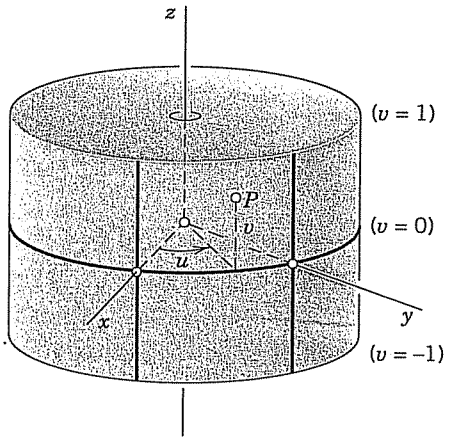


Fig. 240. Parametric representation of a cylinder

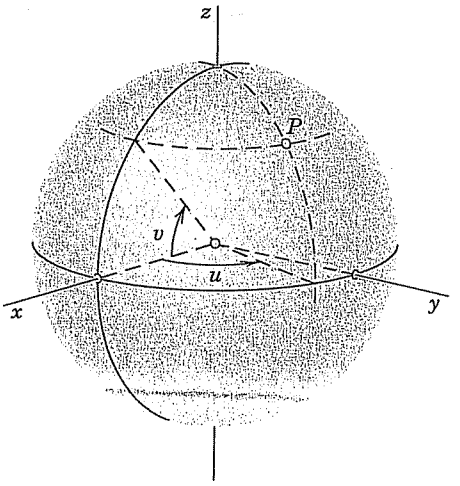


Fig. 241. Parametric representation of a sphere

EXAMPLE 2 **Parametric Representation of a Sphere**

A sphere $x^2 + y^2 + z^2 = a^2$ can be represented in the form

$$\mathbf{r}(u, v) = a \cos v \cos u\, \mathbf{i} + a \cos v \sin u\, \mathbf{j} + a \sin v\, \mathbf{k} \tag{3}$$

where the parameters u, v vary in the rectangle R in the uv-plane given by the inequalities $0 \leqq u \leqq 2\pi$, $-\pi/2 \leqq v \leqq \pi/2$. The components of $\mathbf{r}$ are

$$x = a \cos v \cos u, \qquad y = a \cos v \sin u, \qquad z = a \sin v.$$

The curves $u = const$ and $v = const$ are the "meridians" and "parallels" on S (see Fig. 241). *This representation is used in* ***geography*** *for measuring the latitude and longitude of points on the globe.*

Another parametric representation of the sphere also used in mathematics is

$$\mathbf{r}(u, v) = a \cos u \sin v\, \mathbf{i} + a \sin u \sin v\, \mathbf{j} + a \cos v\, \mathbf{k} \tag{3*}$$

where $0 \leqq u \leqq 2\pi$, $0 \leqq v \leqq \pi$. ■

EXAMPLE 3 Parametric Representation of a Cone

A circular cone $z = \sqrt{x^2 + y^2}$, $0 \leqq t \leqq H$ can be represented by

$$\mathbf{r}(u, v) = [u \cos v, u \sin v, u] = u \cos v\, \mathbf{i} + u \sin v\, \mathbf{j} + u\mathbf{k},$$

in components $x = u \cos v$, $y = u \sin v$, $z = u$. The parameters vary in the rectangle R: $0 \leqq u \leqq H$, $0 \leqq v \leqq 2\pi$. Check that $x^2 + y^2 = z^2$, as it should be. What are the curves $u = const$ and $v = const$? ▩

Tangent Plane and Surface Normal

Recall from Sec. 9.7 that the tangent vectors of all the curves on a surface S through a point P of S form a plane, called the **tangent plane** of S at P (Fig. 242). Exceptions are points where S has an edge or a cusp (like a cone), so that S cannot have a tangent plane at such a point. Furthermore, a vector perpendicular to the tangent plane is called a **normal vector** of S at P.

Now since S can be given by $\mathbf{r} = \mathbf{r}(u, v)$ in (2), the new idea is that we get a curve C on S by taking a pair of differentiable functions

$$u = u(t), \qquad v = v(t)$$

whose derivatives $u' = du/dt$ and $v' = dv/dt$ are continuous. Then C has the position vector $\tilde{\mathbf{r}}(t) = \mathbf{r}(u(t), v(t))$. By differentiation and the use of the chain rule (Sec. 9.6) we obtain a tangent vector of C on S

$$\tilde{\mathbf{r}}'(t) = \frac{d\tilde{\mathbf{r}}}{dt} = \frac{\partial \mathbf{r}}{\partial u} u' + \frac{\partial \mathbf{r}}{\partial v} v'.$$

Hence the partial derivatives $\mathbf{r}_u$ and $\mathbf{r}_v$ *at P are tangential to S at P.* We assume that they are linearly independent, which geometrically means that the curves $u = const$ and $v = const$ on S intersect at P at a nonzero angle. Then $\mathbf{r}_u$ and $\mathbf{r}_v$ span the tangent plane of S at P. Hence their cross product gives a **normal vector** $\mathbf{N}$ of S at P.

(4) $$\mathbf{N} = \mathbf{r}_u \times \mathbf{r}_v \neq \mathbf{0}.$$

The corresponding **unit normal vector** $\mathbf{n}$ of S at P is (Fig. 242)

(5) $$\mathbf{n} = \frac{1}{|\mathbf{N}|}\mathbf{N} = \frac{1}{|\mathbf{r}_u \times \mathbf{r}_v|}\mathbf{r}_u \times \mathbf{r}_v.$$

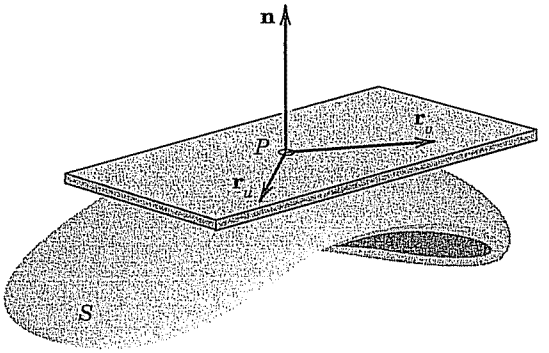


Fig. 242. Tangent plane and normal vector

Also, if S is represented by $g(x, y, z) = 0$, then, by Theorem 2 in Sec. 9.7,

$$\mathbf{n} = \frac{1}{|\text{grad } g|} \text{grad } g. \tag{5*}$$

A surface S is called a **smooth surface** if its surface normal depends continuously on the points of S.

S is called **piecewise smooth** if it consists of finitely many smooth portions.

For instance, a sphere is smooth, and the surface of a cube is piecewise smooth (explain!). We can now summarize our discussion as follows.

THEOREM 1

Tangent Plane and Surface Normal

If a surface S is given by (2) *with continuous* $\mathbf{r}_u = \partial\mathbf{r}/\partial u$ *and* $\mathbf{r}_v = \partial\mathbf{r}/\partial v$ *satisfying* (4) *at every point of S, then S has at every point P a unique tangent plane passing through P and spanned by* $\mathbf{r}_u$ *and* $\mathbf{r}_v$, *and a unique normal whose direction depends continuously on the points of S. A normal vector is given by* (4) *and the corresponding unit normal vector by* (5). (See Fig. 242.)

EXAMPLE 4 **Unit Normal Vector of a Sphere**

From (5*) we find that the sphere $g(x, y, z) = x^2 + y^2 + z^2 - a^2 = 0$ has the unit normal vector

$$\mathbf{n}(x, y, z) = \left[\frac{x}{a}, \frac{y}{a}, \frac{z}{a}\right] = \frac{x}{a}\mathbf{i} + \frac{y}{a}\mathbf{j} + \frac{z}{a}\mathbf{k}.$$

We see that $\mathbf{n}$ has the direction of the position vector $[x, y, z]$ of the corresponding point. Is it obvious that this must be the case? ▩

EXAMPLE 5 **Unit Normal Vector of a Cone**

At the apex of the cone $g(x, y, z) = -z + \sqrt{x^2 + y^2} = 0$ in Example 3, the unit normal vector $\mathbf{n}$ becomes undetermined because from (5*) we get

$$\mathbf{n} = \left[\frac{x}{\sqrt{2(x^2 + y^2)}}, \frac{y}{\sqrt{2(x^2 + y^2)}}, \frac{-1}{\sqrt{2}}\right] = \frac{1}{\sqrt{2}}\left(\frac{x}{\sqrt{x^2 + y^2}}\mathbf{i} + \frac{y}{\sqrt{x^2 + y^2}}\mathbf{j} - \mathbf{k}\right). \blacksquare$$

We are now ready to discuss surface integrals and their applications, beginning in the next section.

PROBLEM SET 10.5

1–10 PREPARATION FOR SURFACE INTEGRALS: PARAMETRIC REPRESENTATION, NORMAL

Familiarize yourself with parametric representations of important surfaces by deriving a representation (1), by finding the **parameter curves** (curves $u = const$ and $v = const$) of the surface and a normal vector $\mathbf{N} = \mathbf{r}_u \times \mathbf{r}_v$ of the surface. (Show the details of your work.)

1. xy-plane $\mathbf{r}(u, v) = [u, v]$ (thus $u\mathbf{i} + v\mathbf{j}$; similarly in Probs. 2–10)
2. xy-plane in polar coordinates $\mathbf{r}(u, v) = [u \cos v, u \sin v]$ (thus $u = r, v = \theta$)
3. Elliptic cylinder $\mathbf{r}(u, v) = [a \cos v, b \sin v, u]$
4. Paraboloid of revolution $\mathbf{r}(u, v) = [u \cos v, u \sin v, u^2]$
5. Cone $\mathbf{r}(u, v) = [au \cos v, au \sin v, cu]$
6. Hyperbolic paraboloid $\mathbf{r}(u, v) = [4u \cosh v, u \sinh v, u^2]$
7. Elliptic paraboloid $\mathbf{r}(u, v) = [3u \cos v, 4u \sin v, u^2]$

8. Helicoid $\mathbf{r}(u, v) = [u \cos v, \quad u \sin v, \quad v]$. Explain the name.

9. Ellipsoid
$\mathbf{r}(u, v) = [2 \cos v \cos u, \quad 3 \cos v \sin u, \quad 4 \sin v]$

10. Ellipsoid
$\mathbf{r}(u, v) = [a \cos v \cos u, \quad b \cos v \sin u, \quad c \sin v]$

11. **CAS EXPERIMENT. Graphing Surfaces, Dependence on *a, b, c.*** Graph the surfaces in Probs. 1–10. In Probs. 6–9 generalize the surfaces by introducing parameters a, b, c and then find out in Probs. 3–10 how the shape of the surfaces depends on a, b. c.

12–19 DERIVATION OF PARAMETRIC REPRESENTATIONS

Find a parametric representation and a normal vector. (The answer gives *one* of them. There are many.)

12. Plane $5x + y - 3z = 30$
13. Plane $4x - 2y + 10z = 16$
14. Sphere $(x - 1)^2 + (y + 2)^2 + z^2 = 25$
15. Sphere $(x + 2)^2 + y^2 + (z - 2)^2 = 1$
16. Elliptic paraboloid $z = 4x^2 + y^2$
17. Parabolic cylinder $z = 3y^2$
18. Hyperbolic cylinder $9x^2 - 4y^2 = 36$
19. Elliptic cone $z = \sqrt{9x^2 + y^2}$

20. **(Representation $z = f(x, y)$)** Show that $z = f(x, y)$ or $g = z - f(x, y) = 0$ can be written ($f_u = \partial f/\partial u$, etc.)

$$\text{(6)} \qquad \mathbf{r}(u, v) = [u, \quad v, \quad f(u, v)] \quad \text{and} \quad \mathbf{N} = \operatorname{grad} g = [\mp f_u, \quad -f_v, \quad 1].$$

21. **(Orthogonal parameters)** Show that the parameter curves $u = const$ and $v = const$ on a surface $\mathbf{r}(u, v)$ are orthogonal (intersect at right angles) if and only if $\mathbf{r}_u \bullet \mathbf{r}_v = 0$.

22. **(Condition (4))** Find the points in Probs. 2–7 at which (4) $\mathbf{N} \neq \mathbf{0}$ does not hold and state whether this is owing to the shape of the surface or to the choice of the representation.

23. **(Change of representation)** Represent the paraboloid in Prob. 4 so that $\mathbf{N}(0, 0) \neq \mathbf{0}$, and show $\mathbf{N}$.

24. **PROJECT. Tangent Planes $T(P)$** will be less important in our work, but you should know how to represent them.
(a) If S: $\mathbf{r}(u, v)$, then $T(P)$: $(\mathbf{r}^* - \mathbf{r} \quad \mathbf{r}_u \quad \mathbf{r}_v) = 0$ (a scalar triple product) or
$\mathbf{r}^*(p, q) = \mathbf{r}(P) + p\mathbf{r}_u(P) + q\mathbf{r}_v(P)$.
(b) If S: $g(x, y, z) = 0$, then $T(P)$: $(\mathbf{r}^* - \mathbf{r}(P)) \bullet \nabla g = 0$.
(c) If S: $z = f(x, y)$, then
$T(P)$: $z^* - z = (x^* - x)f_x(P) + (y^* - y)f_y(P)$.
Interpret (a)–(c) geometrically. Give two examples for (a), two for (b), and two for (c).

10.6 Surface Integrals

To define a surface integral, we take a surface S, given by a parametric representation as just discussed,

$$\text{(1)} \qquad \mathbf{r}(u, v) = [x(u, v), y(u, v), z(u, v)] = x(u, v)\mathbf{i} + y(u, v)\mathbf{j} + z(u, v)\mathbf{k}$$

where (u, v) varies over a region R in the uv-plane. We assume S to be piecewise smooth (Sec. 10.5), so that S has a normal vector

$$\text{(2)} \qquad \mathbf{N} = \mathbf{r}_u \times \mathbf{r}_v \qquad \text{and unit normal vector} \qquad \mathbf{n} = \frac{1}{|\mathbf{N}|}\mathbf{N}$$

at every point (except perhaps for some edges or cusps, as for a cube or cone). For a given vector function $\mathbf{F}$ we can now define the **surface integral** over S by

$$\text{(3)} \qquad \iint_S \mathbf{F} \bullet \mathbf{n}\, dA = \iint_R \mathbf{F}(\mathbf{r}(u, v)) \bullet \mathbf{N}(u, v)\, du\, dv.$$

Here $\mathbf{N} = |\mathbf{N}|\mathbf{n}$ by (2), and $|\mathbf{N}| = |\mathbf{r}_u \times \mathbf{r}_v|$ is the area of the parallelogram with sides $\mathbf{r}_u$ and $\mathbf{r}_v$, by the definition of cross product. Hence

$$\mathbf{n}\, dA = \mathbf{n}\,|\mathbf{N}|\, du\, dv = \mathbf{N}\, du\, dv. \tag{3*}$$

And we see that $dA = |\mathbf{N}|\, du\, dv$ is the element of area of S.

Also $\mathbf{F} \bullet \mathbf{n}$ is the normal component of $\mathbf{F}$. This integral arises naturally in flow problems, where it gives the **flux** across S (= mass of fluid crossing S per unit time; see Sec. 9.8) when $\mathbf{F} = \rho\mathbf{v}$. Here, ρ is the density of the fluid and $\mathbf{v}$ the velocity vector of the flow (example below). We may thus call the surface integral (3) the **flux integral.**

We can write (3) in components, using $\mathbf{F} = [F_1, \ F_2, \ F_3]$, $\mathbf{N} = [N_1, \ N_2, \ N_3]$, and $\mathbf{n} = [\cos\alpha, \cos\beta, \cos\gamma]$. Here, α, β, γ are the angles between $\mathbf{n}$ and the coordinate axes; indeed, for the angle between $\mathbf{n}$ and $\mathbf{i}$, formula (4) in Sec. 9.2 gives $\cos\alpha = \mathbf{n} \bullet \mathbf{i}/|\mathbf{n}||\mathbf{i}| = \mathbf{n} \bullet \mathbf{i}$, and so on. We thus obtain from (3)

$$\begin{aligned} \iint_S \mathbf{F} \bullet \mathbf{n}\, dA &= \iint_S (F_1 \cos\alpha + F_2 \cos\beta + F_3 \cos\gamma)\, dA \\ &= \iint_R (F_1 N_1 + F_2 N_2 + F_3 N_3)\, du\, dv. \end{aligned} \tag{4}$$

In (4) we can write $\cos\alpha\, dA = dy\, dz$, $\cos\beta\, dA = dz\, dx$, $\cos\gamma\, dA = dx\, dy$. Then (4) becomes the following integral for the flux:

$$\iint_S \mathbf{F} \bullet \mathbf{n}\, dA = \iint_S (F_1\, dy\, dz + F_2\, dz\, dx + F_3\, dx\, dy). \tag{5}$$

We can use this formula to evaluate surface integrals by converting them to double integrals over regions in the coordinate planes of the xyz-coordinate system. But we must carefully take into account the orientation of S (the choice of $\mathbf{n}$). We explain this for the integrals of the F_3-terms,

$$\iint_S F_3 \cos\gamma\, dA = \iint_S F_3\, dx\, dy. \tag{5'}$$

If the surface S is given by $z = h(x, y)$ with (x, y) varying in a region $\overline{R}$ in the xy-plane, and if S is oriented so that $\cos\gamma > 0$, then (5′) gives

$$\iint_S F_3 \cos\gamma\, dA = + \iint_{\overline{R}} F_3(x, y, h(x, y))\, dx\, dy. \tag{5''}$$

But if $\cos\gamma < 0$, the integral on the right of (5″) gets a minus sign in front. This follows if we note that the element of area $dx\, dy$ in the xy-plane is the projection $|\cos\gamma|\, dA$ of the element of area dA of S; and we have $\cos\gamma = +|\cos\gamma|$ when $\cos\gamma > 0$, but $\cos\gamma = -|\cos\gamma|$ when $\cos\gamma < 0$. Similarly for the other two terms in (5). At the same time, this justifies the notations in (5).

Other forms of surface integrals will be discussed later in this section.

EXAMPLE 1 **Flux Through a Surface**

Compute the flux of water through the parabolic cylinder S: $y = x^2$, $0 \leqq x \leqq 2$, $0 \leqq z \leqq 3$ (Fig. 243) if the velocity vector is $\mathbf{v} = \mathbf{F} = [3z^2, \quad 6, \quad 6xz]$, speed being measured in meters/sec. (Generally, $\mathbf{F} = \rho\mathbf{v}$, but water has the density $\rho = 1$ gm/cm^3 $= 1$ ton/m^3.)

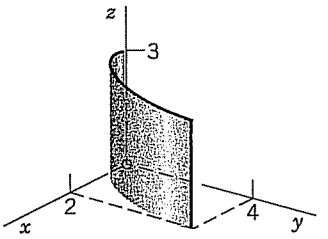


Fig. 243. Surface S in Example 1

Solution. Writing $x = u$ and $z = v$, we have $y = x^2 = u^2$. Hence a representation of S is

$$S: \qquad \mathbf{r} = [u, u^2, v] \qquad (0 \leqq u \leqq 2, \ 0 \leqq v \leqq 3).$$

By differentiation and by the definition of the cross product,

$$\mathbf{N} = \mathbf{r}_u \times \mathbf{r}_v = [1, \quad 2u, \quad 0] \times [0, \quad 0, \quad 1] = [2u, \quad -1, \quad 0].$$

On S, writing simply $\mathbf{F}(S)$ for $\mathbf{F}[\mathbf{r}(u, v)]$, we have $\mathbf{F}(S) = [3v^2, 6, 6uv]$. Hence $\mathbf{F}(S) \cdot \mathbf{N} = 6uv^2 - 6$. By integration we thus get from (3) the flux

$$\iint_S \mathbf{F} \cdot \mathbf{n} \, dA = \int_0^3 \int_0^2 (6uv^2 - 6) \, du \, dv = \int_0^3 (3u^2v^2 - 6u) \Big|_{u=0}^{2} dv$$

$$= \int_0^3 (12v^2 - 12) \, dv = (4v^3 - 12v) \Big|_{v=0}^{3} = 108 - 36 = 72 \ [\text{m}^3/\text{sec}]$$

or 72 000 liters/sec. Note that the y-component of $\mathbf{F}$ is positive (equal to 6), so that in Fig. 243 the flow goes from left to right.

Let us confirm this result by (5). Since

$$\mathbf{N} = |\mathbf{N}|\mathbf{n} = |\mathbf{N}|[\cos \alpha, \quad \cos \beta, \quad \cos \gamma] = [2u, \quad -1, \quad 0] = [2x, \quad -1, \quad 0]$$

we see that $\cos \alpha > 0$, $\cos \beta < 0$, and $\cos \gamma = 0$. Hence the second term of (5) on the right gets a minus sign, and the last term is absent. This gives, in agreement with the previous result,

$$\iint_S \mathbf{F} \cdot \mathbf{n} \, dA = \int_0^3 \int_0^4 3z^2 \, dy \, dz - \int_0^2 \int_0^3 6 \, dz \, dx = \int_0^3 4(3z^2) \, dz - \int_0^2 6 \cdot 3 \, dx = 4 \cdot 3^3 - 6 \cdot 3 \cdot 2 = 72. \ \blacksquare$$

EXAMPLE 2 **Surface Integral**

Evaluate (3) when $\mathbf{F} = [x^2, 0, 3y^2]$ and S is the portion of the plane $x + y + z = 1$ in the first octant (Fig. 244).

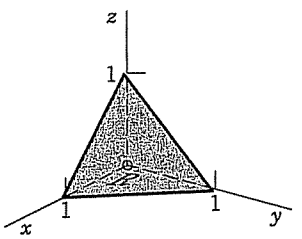


Fig. 244. Portion of a plane in Example 2

Solution. Writing $x = u$ and $y = v$, we have $z = 1 - x - y = 1 - u - v$. Hence we can represent the plane $x + y + z = 1$ in the form $\mathbf{r}(u, v) = [u, v, 1 - u - v]$. We obtain the first-octant portion S of this plane by restricting $x = u$ and $y = v$ to the projection R of S in the xy-plane. R is the triangle bounded by the two coordinate axes and the straight line $x + y = 1$, obtained from $x + y + z = 1$ by setting $z = 0$. Thus $0 \leqq x \leqq 1 - y$, $0 \leqq y \leqq 1$.

By inspection or by differentiation,

$$\mathbf{N} = \mathbf{r}_u \times \mathbf{r}_v = [1, 0, -1] \times [0, 1, -1] = [1, 1, 1].$$

Hence $\mathbf{F}(S) \cdot \mathbf{N} = [u^2, 0, 3v^2] \cdot [1, 1, 1] = u^2 + 3v^2$. By (3),

$$\iint_S \mathbf{F} \cdot \mathbf{n}\, dA = \iint_R (u^2 + 3v^2)\, du\, dv = \int_0^1 \int_0^{1-v} (u^2 + 3v^2)\, du\, dv$$

$$= \int_0^1 \left[\frac{1}{3}(1 - v)^3 + 3v^2(1 - v) \right] dv = \frac{1}{3}. \quad \blacksquare$$

Orientation of Surfaces

From (3) or (4) we see that the value of the integral depends on the choice of the unit normal vector $\mathbf{n}$. (Instead of $\mathbf{n}$ we could choose $-\mathbf{n}$.) We express this by saying that such an integral is an *integral over an* **oriented surface** S, that is, over a surface S on which we have chosen one of the two possible unit normal vectors in a continuous fashion. (For a piecewise smooth surface, this needs some further discussion, which we give below.) If we change the orientation of S, this means that we replace $\mathbf{n}$ with $-\mathbf{n}$. Then each component of $\mathbf{n}$ in (4) is multiplied by -1, so that we have

THEOREM 1

Change of Orientation in a Surface Integral

The replacement of $\mathbf{n}$ *by* $-\mathbf{n}$ (*hence of* $\mathbf{N}$ *by* $-\mathbf{N}$) *corresponds to the multiplication of the integral in* (3) *or* (4) *by* -1.

How do we effect such a change of $\mathbf{N}$ in practice if S is given in the form (1)? The simplest way is to interchange u and v, because then $\mathbf{r}_u$ becomes $\mathbf{r}_v$ and conversely, so that $\mathbf{N} = \mathbf{r}_u \times \mathbf{r}_v$ becomes $\mathbf{r}_v \times \mathbf{r}_u = -\mathbf{r}_u \times \mathbf{r}_v = -\mathbf{N}$, as wanted. Let us illustrate this.

EXAMPLE 3 **Change of Orientation in a Surface Integral**

In Example 1 we now represent S by $\tilde{\mathbf{r}} = [v, v^2, u]$, $0 \leqq v \leqq 2$, $0 \leqq u \leqq 3$. Then

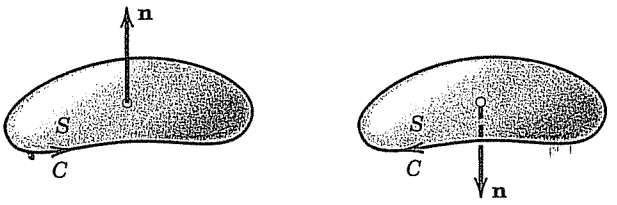
$$\tilde{\mathbf{N}} = \tilde{\mathbf{r}}_u \times \tilde{\mathbf{r}}_v = [0, 0, 1] \times [1, 2v, 0] = [-2v, 1, 0].$$

For $\mathbf{F} = [3z^2, 6, 6xz]$ we now get $\tilde{\mathbf{F}}(S) = [3u^2, 6, 6uv]$. Hence $\tilde{\mathbf{F}}(S) \cdot \tilde{\mathbf{N}} = -6u^2v + 6$ and integration gives the old result times -1,

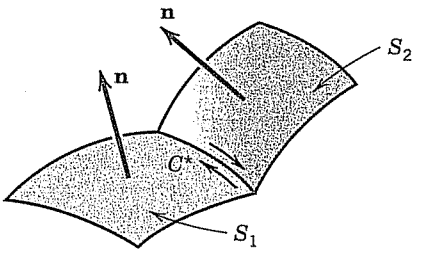
$$\iint_R \tilde{\mathbf{F}}(S) \cdot \tilde{\mathbf{N}}\, dv\, du = \int_0^3 \int_0^2 (-6u^2v + 6)\, dv\, du = \int_0^3 (-12u^2 + 12)\, du = -72. \quad \blacksquare$$

Orientation of Smooth Surfaces

A smooth surface S (see Sec. 10.5) is called **orientable** if the positive normal direction, when given at an arbitrary point P_0 of S, can be continued in a unique and continuous way to the entire surface. For smooth surfaces occurring in applications this is always true.



(a) Smooth surface



(b) Piecewise smooth surface

Fig. 245. Orientation of a surface

Orientation of Piecewise Smooth Surfaces

Here the following idea will do it. For a *smooth* orientable surface S with boundary curve C we may associate with each of the two possible orientations of S an orientation of C, as shown in Fig. 245a. Then a *piecewise smooth* surface is called **orientable** if we can orient each smooth piece of S so that along each curve C^* which is a common boundary of two pieces S_1 and S_2 the positive direction of C^* relative to S_1 is opposite to the direction of C^* relative to S_2. See Fig. 245b for two adjacent pieces; note the arrows along C^*.

Theory: Nonorientable Surfaces

A sufficiently small piece of a smooth surface is always orientable. This may not hold for entire surfaces. A well-known example is the **Möbius strip**[5], shown in Fig. 246. To make a model, take the rectangular paper in Fig. 246, make a half-twist, and join the short sides together so that A goes onto A, and B onto B. At P_0 take a normal vector pointing, say, to the *left.* Displace it along C to the right (in the lower part of the figure) around the strip until you return to P_0 and see that you get a normal vector pointing to the *right,* opposite to the given one. See also Prob. 21.

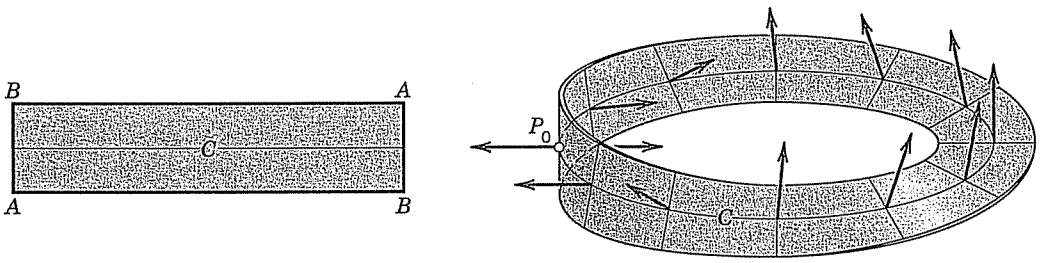


Fig. 246. Möbius strip

[5]AUGUST FERDINAND MÖBIUS (1790–1868), German mathematician, student of Gauss, known for his work in surface theory, geometry, and complex analysis (see Sec. 17.2).

Surface Integrals Without Regard to Orientation

Another type of surface integral is

$$\iint_S G(\mathbf{r})\, dA = \iint_R G(\mathbf{r}(u, v))|\mathbf{N}(u, v)|\, du\, dv. \tag{6}$$

Here $dA = |\mathbf{N}|\, du\, dv = |\mathbf{r}_u \times \mathbf{r}_v|\, du\, dv$ is the element of area of the surface S represented by (1) and we disregard the orientation.

We shall need later (in Sec. 10.9) the **mean value theorem** for surface integrals, which states that if R in (6) is simply connected (see Sec. 10.2) and $G(\mathbf{r})$ is continuous in a domain containing R, then there is a point (u_0, v_0) in R such that

$$\iint_S G(\mathbf{r})\, dA = G(\mathbf{r}(u_0, v_0))A \qquad (A = \text{Area of } S). \tag{7}$$

As for applications, if $G(\mathbf{r})$ is the mass density of S, then (6) is the total mass of S. If $G = 1$, then (6) gives the **area** $A(S)$ of S,

$$A(S) = \iint_S dA = \iint_R |\mathbf{r}_u \times \mathbf{r}_v|\, du\, dv. \tag{8}$$

Examples 4 and 5 show how to apply (8) to a sphere and a torus. The final example, Example 6, explains how to calculate moments of inertia for a surface.

EXAMPLE 4 **Area of a Sphere**

For a sphere $\mathbf{r}(u, v) = [a\cos v\cos u,\quad a\cos v\sin u,\quad a\sin v]$, $0 \leqq u \leqq 2\pi$, $-\pi/2 \leqq v \leqq \pi/2$, [see (3) in Sec. 10.5] we obtain by direct calculation (verify!)

$$\mathbf{r}_u \times \mathbf{r}_v = [a^2\cos^2 v\cos u,\qquad a^2\cos^2 v\sin u,\qquad a^2\cos v\sin v].$$

Using $\cos^2 u + \sin^2 u = 1$ and then $\cos^2 v + \sin^2 v = 1$, we obtain

$$|\mathbf{r}_u \times \mathbf{r}_v| = a^2(\cos^4 v\cos^2 u + \cos^4 v\sin^2 u + \cos^2 v\sin^2 v)^{1/2} = a^2|\cos v|.$$

With this, (8) gives the familiar formula (note that $|\cos v| = \cos v$ when $-\pi/2 \leqq v \leqq \pi/2$)

$$A(S) = a^2\int_{-\pi/2}^{\pi/2}\int_0^{2\pi} |\cos v|\, du\, dv = 2\pi a^2\int_{-\pi/2}^{\pi/2}\cos v\, dv = 4\pi a^2. \qquad \blacksquare$$

EXAMPLE 5 **Torus Surface (Doughnut Surface): Representation and Area**

A *torus surface* S is obtained by rotating a circle C about a straight line L in space so that C does not intersect or touch L but its plane always passes through L. If L is the z-axis and C has radius b and its center has distance a ($> b$) from L, as in Fig. 247, then S can be represented by

$$\mathbf{r}(u, v) = (a + b\cos v)\cos u\,\mathbf{i} + (a + b\cos v)\sin u\,\mathbf{j} + b\sin v\,\mathbf{k}$$

where $0 \leqq u \leqq 2\pi$, $0 \leqq v \leqq 2\pi$. Thus

$$\mathbf{r}_u = -(a + b\cos v)\sin u\,\mathbf{i} + (a + b\cos v)\cos u\,\mathbf{j}$$

$$\mathbf{r}_v = -b\sin v\cos u\,\mathbf{i} - b\sin v\sin u\,\mathbf{j} + b\cos v\,\mathbf{k}$$

$$\mathbf{r}_u \times \mathbf{r}_v = b(a + b\cos v)(\cos u\cos v\,\mathbf{i} + \sin u\cos v\,\mathbf{j} + \sin v\,\mathbf{k}).$$

Hence $|\mathbf{r}_u \times \mathbf{r}_v| = b(a + b\cos v)$, and (8) gives the total area of the torus,

$$A(S) = \int_0^{2\pi}\int_0^{2\pi} b(a + b\cos v)\,du\,dv = 4\pi^2 ab. \tag{9}$$

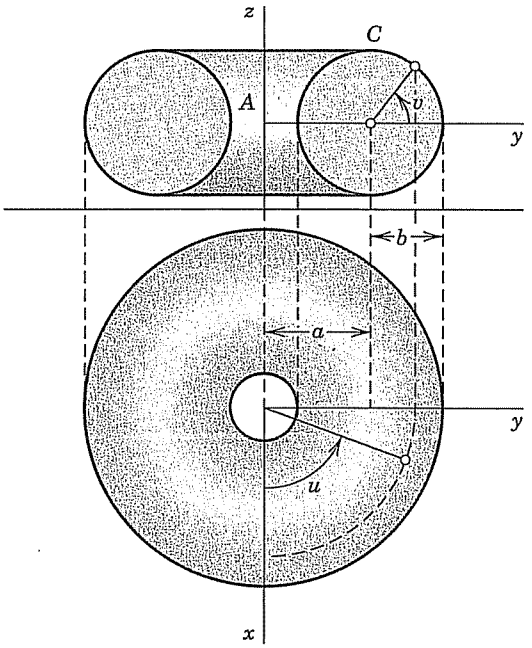


Fig. 247. Torus in Example 5

EXAMPLE 6 Moment of Inertia of a Surface

Find the moment of inertia I of a spherical lamina S: $x^2 + y^2 + z^2 = a^2$ of constant mass density and total mass M about the z-axis.

Solution. If a mass is distributed over a surface S and $\mu(x, y, z)$ is the density of the mass (= mass per unit area), then the moment of inertia I of the mass with respect to a given axis L is defined by the surface integral

$$I = \iint_S \mu D^2\,dA \tag{10}$$

where $D(x, y, z)$ is the distance of the point (x, y, z) from L. Since, in the present example, μ is constant and S has the area $A = 4\pi a^2$, we have $\mu = M/A = M/(4\pi a^2)$.

For S we use the same representation as in Example 4. Then $D^2 = x^2 + y^2 = a^2\cos^2 v$. Also, as in that example, $dA = a^2\cos v\,du\,dv$. This gives the following result. [In the integration, use $\cos^3 v = \cos v\,(1 - \sin^2 v)$.]

$$I = \iint_S \mu D^2\,dA = \frac{M}{4\pi a^2}\int_{-\pi/2}^{\pi/2}\int_0^{2\pi} a^4\cos^3 v\,du\,dv = \frac{Ma^2}{2}\int_{-\pi/2}^{\pi/2}\cos^3 v\,dv = \frac{2Ma^2}{3}.$$

Representations $z = f(x, y)$. If a surface S is given by $z = f(x, y)$, then setting $u = x$, $v = y$, $\mathbf{r} = [u, v, f]$ gives

$$|\mathbf{N}| = |\mathbf{r}_u \times \mathbf{r}_v| = |[1, 0, f_u] \times [0, 1, f_v]| = |[-f_u, -f_v, 1]| = \sqrt{1 + f_u^2 + f_v^2}$$

and, since $f_u = f_x$, $f_v = f_y$, formula (6) becomes

$$\iint_S G(\mathbf{r})\,dA = \iint_{R^*} G(x, y, f(x, y))\sqrt{1 + \left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2}\,dx\,dy. \tag{11}$$

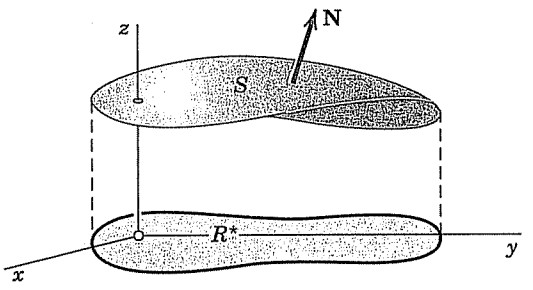


Fig. 248. Formula (11)

Here R^* is the projection of S into the xy-plane (Fig. 248) and the normal vector $\mathbf{N}$ on S points *up*. If it points *down*, the integral on the right is preceded by a minus sign.

From (11) with $G = 1$ we obtain for the **area** $A(S)$ of S: $z = f(x, y)$ the formula

$$\textbf{(12)} \qquad A(S) = \iint_{R^*} \sqrt{1 + \left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2}\, dx\, dy$$

where R^* is the projection of S into the xy-plane, as before.

PROBLEM SET 10.6

1–12 FLUX INTEGRALS (3) $\int_S \mathbf{F} \cdot \mathbf{n}\, dA$

Evaluate these integrals for the following data. Indicate the kind of surface. (Show the details of your work.)

1. $\mathbf{F} = [2x, \quad 5y, \quad 0]$, S: $\mathbf{r} = [u, \quad v, \quad 4u + 3v]$, $0 \leqq u \leqq 1$, $-8 \leqq v \leqq 8$
2. $\mathbf{F} = [x^2, \quad y^2, \quad z^2]$, S: $x + y + z = 4$, $x \geqq 0$, $y \geqq 0$, $z \geqq 0$
3. $\mathbf{F} = [x - z, \quad y - x, \quad z - y]$, S: $\mathbf{r} = [u \cos v, \quad u \sin v, \quad u]$, $0 \leqq u \leqq 3$, $0 \leqq v \leqq \pi$
4. $\mathbf{F} = [e^y, \quad -e^z, \quad e^x]$, S: $x^2 + y^2 = 9$, $x \geqq 0$, $y \geqq 0$, $0 \leqq z \leqq 2$
5. $\mathbf{F} = [x, \quad y, \quad z]$, S: $\mathbf{r} = [u \cos v, \quad u \sin v, \quad u^2]$, $0 \leqq u \leqq 4$, $-\pi \leqq v \leqq \pi$
6. $\mathbf{F} = [\cosh yz, \quad 0, \quad y^4]$, S: $y^2 + z^2 = 1$, $0 \leqq x \leqq 20$, $z \geqq 0$
7. $\mathbf{F} = [1, \quad 1, \quad 1]$, S the sphere of radius 1 and center 0
8. $\mathbf{F} = [\tan xy, \quad x^2y, \quad -z]$, S: $y^2 + \frac{1}{9}z^2 = 1$, $1 \leqq x \leqq 4$
9. $\mathbf{F} = [0, \quad x, \quad 0]$, S: $x^2 + y^2 + z^2 = a^2$, $x \geqq 0$, $y \geqq 0$, $z \geqq 0$
10. $\mathbf{F} = [y^2, \quad x^2, \quad z^4]$, S: $z = 4\sqrt{x^2 + y^2}$, $0 \leqq z \leqq 8$, $y \geqq 0$
11. $\mathbf{F} = [y^3, \quad x^3, \quad z^3]$, S: $x^2 + 4y^2 = 4$, $x \geqq 0$, $y \geqq 0$, $0 \leqq z \leqq h$
12. $\mathbf{F} = [\cosh y, \quad 0, \quad \sinh x]$, S: $z = x + y^2$, $0 \leqq y \leqq x$, $0 \leqq x \leqq 1$
13. **CAS EXPERIMENT.** Write a program for evaluating surface integrals (3) that prints intermediate results ($\mathbf{F}$, $\mathbf{F} \cdot \mathbf{N}$, the integral over one of the two variables). Can you experimentally obtain rules on functions and surfaces giving integrals that can be evaluated by the usual methods of calculus? Make a list of positive and negative results.

14–20 SURFACE INTEGRALS (6) $\iint_S G(\mathbf{r})\, dA$

Evaluate these integrals for the following data. Indicate the kind of surface. (Show the details.)

14. $G = \cos y + \sin x$, S: $x + y + z = 2$, $x \geqq 0$, $y \geqq 0$, $z \geqq 0$
15. $G = 5(x + y + z)$, S: $z = x + 2y$, $0 \leqq y \leqq x$, $0 \leqq x \leqq 2$
16. $G = ye^x + xe^y + e^z$, S: $x^2 + y^2 = 16$, $y \geqq 0$, $0 \leqq z \leqq 4$
17. $G = (x^2 + y^2 + z^2)^2$, S: $z = \sqrt{x^2 + y^2}$, $y \geqq 0$, $0 \leqq z \leqq 2$
18. $G = ax + by + cz$, S: $x^2 + y^2 + z^2 = 1$, $y \geqq 0$, $z \geqq 0$
19. $G = \arctan(y/x)$, S: $z = x^2 + y^2$, $1 \leqq z \leqq 9$, $x \geqq 0$, $y \geqq 0$
20. $G = 3xy$, S: $z = xy$, $0 \leqq x \leqq 1$, $0 \leqq y \leqq 1$
21. **(Fun with Möbius)** Make Möbius strips from long slim rectangles R of grid paper (graph paper) by pasting the short sides together after giving the paper a half-twist. In each case count the number of parts obtained

by cutting along lines parallel to the edge. (a) Make R three squares wide and cut until you reach the beginning. (b) Make R four squares wide. Begin cutting one square away from the edge until you reach the beginning. Then cut the portion that is still two squares wide. (c) Make R five squares wide and cut similarly. (d) Make R six squares wide and cut. Formulate a conjecture about the number of parts obtained.

APPLICATIONS

22. (Center of gravity) Justify the following formulas for the mass M and the center of gravity $(\bar{x}, \bar{y}, \bar{z})$ of a lamina S of density (mass per unit area) $\sigma(x, y, z)$ in space:

$$M = \iint_S \sigma \, dA, \quad \bar{x} = \frac{1}{M} \iint_S x\sigma \, dA,$$

$$\bar{y} = \frac{1}{M} \iint_S y\sigma \, dA, \quad \bar{z} = \frac{1}{M} \iint_S z\sigma \, dA.$$

23. (Moments of inertia) Justify the following formulas for the moments of inertia of the lamina in Prob. 22 about the x-, y-, and z-axes, respectively:

$$I_x = \iint_S (y^2 + z^2)\sigma \, dA, \quad I_y = \iint_S (x^2 + z^2)\sigma \, dA,$$

$$I_z = \iint_S (x^2 + y^2)\sigma \, dA.$$

24. Find a formula for the moment of inertia of the lamina in Prob. 22 about the line $y = x$, $z = 0$.

Find the moment of inertia of a lamina S of density 1 about an axis A, where

25. S: $x^2 + y^2 = 1$, $0 \leqq z \leqq h$, A: the z-axis

26. S as in Prob. 25, A: the line $z = h/2$ in the xz-plane

27. S: $x^2 + y^2 = z^2$, $0 \leqq z \leqq h$, A: the z-axis

28. (Steiner's theorem[6]) If I_A is the moment of inertia of a mass distribution of total mass M with respect to an axis A through the center of gravity, show that its moment of inertia I_B with respect to an axis B, which is parallel to A and has the distance k from it, is

$$I_B = I_A + k^2 M.$$

29. Using Steiner's theorem, find the moment of inertia of S in Prob. 26 about the x-axis.

30. TEAM PROJECT. First Fundamental Form of a Surface. Given a surface S: $\mathbf{r}(u, v)$, the corresponding quadratic differential form

$$(13) \quad ds^2 = E\,du^2 + 2F\,du\,dv + G\,dv^2$$

with coefficients

$$(14) \quad E = \mathbf{r}_u \cdot \mathbf{r}_u, \quad F = \mathbf{r}_u \cdot \mathbf{r}_v, \quad G = \mathbf{r}_v \cdot \mathbf{r}_v$$

is called the **first fundamental form** of S. (E, F, G are standard notations that have nothing to do with F and G that occur at some other places in this chapter.) The first fundamental form is basic in the theory of surfaces, since with its help we can determine lengths, angles, and areas on S. To show this, prove the following.

(a) For a curve C: $u = u(t)$, $v = v(t)$, $a \leqq t \leqq b$, on S, formulas (10), Sec. 9.5, and (14) give the length

$$(15) \quad l = \int_a^b \sqrt{\mathbf{r}'(t) \cdot \mathbf{r}'(t)}\, dt = \int_a^b \sqrt{Eu'^2 + 2Fu'v' + Gv'^2}\, dt.$$

(b) The angle γ between two intersecting curves C_1: $u = g(t)$, $v = h(t)$ and C_2: $u = p(t)$, $v = q(t)$ on S: $\mathbf{r}(u, v)$ is obtained from

$$(16) \quad \cos\gamma = \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}|\,|\mathbf{b}|}$$

where $\mathbf{a} = \mathbf{r}_u g' + \mathbf{r}_v h'$ and $\mathbf{b} = \mathbf{r}_u p' + \mathbf{r}_v q'$ are tangent vectors of C_1 and C_2.

(c) The square of the length of the normal vector $\mathbf{N}$ can be written

$$(17) \quad |\mathbf{N}|^2 = |\mathbf{r}_u \times \mathbf{r}_v|^2 = EG - F^2,$$

so that formula (8) for the area $A(S)$ of S becomes

$$(18) \quad A(S) = \iint_S dA = \iint_R |\mathbf{N}|\, du\, dv = \iint_R \sqrt{EG - F^2}\, du\, dv.$$

(d) For polar coordinates u ($= r$) and v ($= \theta$) defined by $x = u\cos v$, $y = u \sin v$ we have $E = 1$, $F = 0$, $G = u^2$, so that

$$ds^2 = du^2 + u^2\, dv^2 = dr^2 + r^2\, d\theta^2.$$

Calculate from this and (18) the area of a disk of radius a.

(e) Find the first fundamental form of the torus in Example 5. Use it to calculate the area A of the torus. Show that A can also be obtained by the **theorem of**

[6]JACOB STEINER (1796–1863), Swiss geometer, born in a small village, learned to write only at age 14, became a pupil of Pestalozzi at 18, later studied at Heidelberg and Berlin and, finally, because of his outstanding research, was appointed professor at Berlin University.

Pappus,[7] which states that the area of a surface of revolution equals the product of the length of a meridian C and the length of the path of the center of gravity of C when C is rotated through the angle 2π.

(f) Calculate the first fundamental form for the usual representations of important surfaces of your own choice (cylinder, cone, etc.) and apply them to the calculation of lengths and areas on these surfaces.

10.7 Triple Integrals. Divergence Theorem of Gauss

In this section we discuss another "big" integral theorem, the divergence theorem, which transforms surface integrals into triple integrals. So let us begin with a review of the latter.

A **triple integral** is an integral of a function $f(x, y, z)$ taken over a closed bounded (three-dimensional) region T in space (where "closed" and "bounded" are defined as in footnote 2 of Sec. 10.3, with "sphere" substituted for "circle"). We subdivide T by planes parallel to the coordinate planes. Then we consider those boxes of the subdivision (rectangular parallelepipeds) that lie entirely inside T, and number them from 1 to n. In each such box we choose an arbitrary point, say, (x_k, y_k, z_k) in box k. The volume of box k we denote by ΔV_k. We now form the sum

$$J_n = \sum_{k=1}^{n} f(x_k, y_k, z_k)\, \Delta V_k.$$

This we do for larger and larger positive integers n arbitrarily but so that the maximum length of all the edges of those n boxes approaches zero as n approaches infinity. This gives a sequence of real numbers $J_{n_1}, J_{n_2}, \cdots$. We assume that $f(x, y, z)$ is continuous in a domain containing T, and T is bounded by finitely many *smooth surfaces* (see Sec. 10.5). Then it can be shown (see Ref. [GR4] in App. 1) that the sequence converges to a limit that is independent of the choice of subdivisions and corresponding points (x_k, y_k, z_k). This limit is called the **triple integral** *of* $f(x, y, z)$ *over the region* T and is denoted by

$$\iiint_T f(x, y, z)\, dx\, dy\, dz \qquad \text{or by} \qquad \iiint_T f(x, y, z)\, dV.$$

Triple integrals can be evaluated by three successive integrations. This is similar to the evaluation of double integrals by two successive integrations, as discussed in Sec. 10.3. An example is shown below (Example 1).

Divergence Theorem of Gauss

Triple integrals can be transformed into surface integrals over the boundary surface of a region in space and conversely. Such a transformation is of practical interest because one of the two kinds of integral is often simpler than the other. It also helps in establishing fundamental equations in fluid flow, heat conduction, etc., as we shall see. The transformation is done by the *divergence theorem,* which involves the **divergence** of a vector function $\mathbf{F} = [F_1, F_2, F_3] = F_1\mathbf{i} + F_2\mathbf{j} + F_3\mathbf{k}$, namely,

[7] PAPPUS OF ALEXANDRIA (about A.D. 300), Greek mathematician. The theorem is also called Guldin's theorem. HABAKUK GULDIN (1577–1643) was born in St. Gallen, Switzerland, and later became professor in Graz and Vienna.

(1) $$\operatorname{div} \mathbf{F} = \frac{\partial F_1}{\partial x} + \frac{\partial F_2}{\partial y} + \frac{\partial F_3}{\partial z}$$ (Sec. 9.8).

THEOREM 1

Divergence Theorem of Gauss
(Transformation Between Triple and Surface Integrals)

Let T be a closed bounded region in space whose boundary is a piecewise smooth orientable surface S. Let $\mathbf{F}(x, y, z)$ be a vector function that is continuous and has continuous first partial derivatives in some domain containing T. Then

(2) $$\iiint_T \operatorname{div} \mathbf{F}\, dV = \iint_S \mathbf{F} \cdot \mathbf{n}\, dA.$$

***In components** of $\mathbf{F} = [F_1, \quad F_2, \quad F_3]$ and of the outer unit normal vector $\mathbf{n} = [\cos \alpha, \quad \cos \beta, \quad \cos \gamma]$ of S (as in Fig. 250), formula (2) becomes*

(2*) $$\begin{aligned} &\iiint_T \left(\frac{\partial F_1}{\partial x} + \frac{\partial F_2}{\partial y} + \frac{\partial F_3}{\partial z} \right) dx\, dy\, dz \\ &= \iint_S (F_1 \cos \alpha + F_2 \cos \beta + F_3 \cos \gamma)\, dA \\ &= \iint_S (F_1\, dy\, dz + F_2\, dz\, dx + F_3\, dx\, dy). \end{aligned}$$

The proof follows after Example 1. "Closed bounded region" is explained above, "piecewise smooth orientable" in Sec. 10.5, and "domain containing T" in footnote 4, Sec. 10.4, for the two-dimensional case.

EXAMPLE 1 **Evaluation of a Surface Integral by the Divergence Theorem**

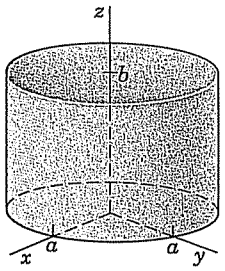


Fig. 249. Surface S in Example 1

Before we prove the theorem, let us show a typical application. Evaluate

$$I = \iint_S (x^3\, dy\, dz + x^2 y\, dz\, dx + x^2 z\, dx\, dy)$$

where S is the closed surface in Fig. 249 consisting of the cylinder $x^2 + y^2 = a^2$ $(0 \leqq z \leqq b)$ and the circular disks $z = 0$ and $z = b$ $(x^2 + y^2 \leqq a^2)$.

Solution. $F_1 = x^3$, $F_2 = x^2 y$, $F_3 = x^2 z$. Hence $\operatorname{div} \mathbf{F} = 3x^2 + x^2 + x^2 = 5x^2$. The form of the surface suggests that we introduce polar coordinates r, θ defined by $x = r \cos \theta$, $y = r \sin \theta$ (thus cylindrical coordinates r, θ, z). Then the volume element is $dx\, dy\, dz = r\, dr\, d\theta\, dz$, and we obtain

$$I = \iiint_T 5x^2\, dx\, dy\, dz = \int_{z=0}^{b} \int_{\theta=0}^{2\pi} \int_{r=0}^{a} (5r^2 \cos^2 \theta) r\, dr\, d\theta\, dz$$

$$= 5 \int_{z=0}^{b} \int_{\theta=0}^{2\pi} \frac{a^4}{4} \cos^2 \theta\, d\theta\, dz = 5 \int_{z=0}^{b} \frac{a^4 \pi}{4}\, dz = \frac{5\pi}{4} a^4 b. \qquad \blacksquare$$

PROOF We prove the divergence theorem, beginning with the first equation in (2*). This equation is true if and only if the integrals of each component on both sides are equal; that is,

$$\iiint_T \frac{\partial F_1}{\partial x}\,dx\,dy\,dz = \iint_S F_1 \cos\alpha\,dA, \tag{3}$$

$$\iiint_T \frac{\partial F_2}{\partial y}\,dx\,dy\,dz = \iint_S F_2 \cos\beta\,dA, \tag{4}$$

$$\iiint_T \frac{\partial F_3}{\partial z}\,dx\,dy\,dz = \iint_S F_3 \cos\gamma\,dA. \tag{5}$$

We first prove (5) for a ***special region*** T that is bounded by a piecewise smooth orientable surface S and has the property that any straight line parallel to any one of the coordinate axes and intersecting T has at most *one* segment (or a single point) in common with T. This implies that T can be represented in the form

$$g(x, y) \leqq z \leqq h(x, y) \tag{6}$$

where (x, y) varies in the orthogonal projection $\bar{R}$ of T in the xy-plane. Clearly, $z = g(x, y)$ represents the "bottom" S_2 of S (Fig. 250), whereas $z = h(x, y)$ represents the "top" S_1 of S, and there may be a remaining vertical portion S_3 of S. (The portion S_3 may degenerate into a curve, as for a sphere.)

To prove (5), we use (6). Since $\mathbf{F}$ is continuously differentiable in some domain containing T, we have

$$\iiint_T \frac{\partial F_3}{\partial z}\,dx\,dy\,dz = \iint_{\bar{R}} \left[\int_{g(x,\,y)}^{h(x,\,y)} \frac{\partial F_3}{\partial z}\,dz\right] dx\,dy. \tag{7}$$

Integration of the inner integral $[\cdots]$ gives $F_3[x, y, h(x, y)] - F_3[x, y, g(x, y)]$. Hence the triple integral in (7) equals

$$\iint_{\bar{R}} F_3[x, y, h(x, y)]\,dx\,dy - \iint_{\bar{R}} F_3[x, y, g(x, y)]\,dx\,dy. \tag{8}$$

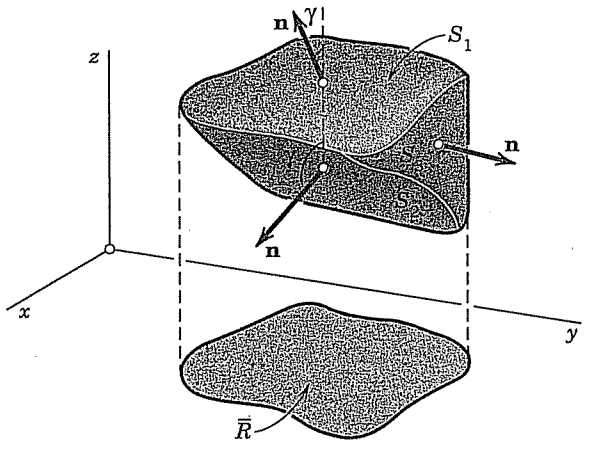


Fig. 250. Example of a special region

But the same result is also obtained by evaluating the right side of (5); that is [see also the last line of (2*)],

$$\iint_S F_3 \cos\gamma \, dA = \iint_S F_3 \, dx\, dy$$

$$= + \iint_{\bar{R}} F_3[x, y, h(x, y)] \, dx\, dy - \iint_{\bar{R}} F_3[x, y, g(x, y)] \, dx\, dy,$$

where the first integral over $\bar{R}$ gets a plus sign because $\cos\gamma > 0$ on S_1 in Fig. 250 [as in (5″), Sec. 10.6], and the second integral gets a minus sign because $\cos\gamma < 0$ on S_2. This proves (5).

The relations (3) and (4) now follow by merely relabeling the variables and using the fact that, by assumption, T has representations similar to (6), namely,

$$\tilde{g}(y, z) \leqq x \leqq \tilde{h}(y, z) \qquad \text{and} \qquad \tilde{\tilde{g}}(z, x) \leqq y \leqq \tilde{\tilde{h}}(z, x).$$

This proves the first equation in (2*) for special regions. It implies (2) because the left side of (2*) is just the definition of the divergence, and the right sides of (2) and of the first equation in (2*) are equal, as was shown in the first line of (4) in the last section. Finally, equality of the right sides of (2) and (2*), last line, is seen from (5) in the last section.

This establishes the divergence theorem for special regions.

For any region T that can be subdivided into *finitely many* special regions by means of auxiliary surfaces, the theorem follows by adding the result for each part separately; this procedure is analogous to that in the proof of Green's theorem in Sec. 10.4. The surface integrals over the auxiliary surfaces cancel in pairs, and the sum of the remaining surface integrals is the surface integral over the whole boundary surface S of T; the triple integrals over the parts of T add up to the triple integral over T.

The divergence theorem is now proved for any bounded region that is of interest in practical problems. The extension to a most general region T of the type indicated in the theorem would require a certain limit process; this is similar to the situation in the case of Green's theorem in Sec. 10.4. ∎

EXAMPLE 2 Verification of the Divergence Theorem

Evaluate $\iint_S (7x\mathbf{i} - z\mathbf{k})\cdot\mathbf{n}\, dA$ over the sphere S: $x^2 + y^2 + z^2 = 4$ (a) by (2), (b) directly.

Solution. (a) $\operatorname{div}\mathbf{F} = \operatorname{div}[7x, 0, -z] = \operatorname{div}[7x\mathbf{i} - z\mathbf{k}] = 7 - 1 = 6$. *Answer:* $6\cdot(4/3)\pi\cdot 2^3 = 64\pi$.

(b) We can represent S by (3), Sec. 10.5 (with $a = 2$), and we shall use $\mathbf{n}\, dA = \mathbf{N}\, du\, dv$ [see (3*), Sec. 10.6]. Accordingly,

$$S: \quad \mathbf{r} = [2\cos v\cos u, \quad 2\cos v\sin u, \quad 2\sin v].$$

Then

$$\mathbf{r}_u = [-2\cos v\sin u, \quad 2\cos v\cos u, \quad 0]$$

$$\mathbf{r}_v = [-2\sin v\cos u, \quad -2\sin v\sin u, \quad 2\cos v]$$

$$\mathbf{N} = \mathbf{r}_u \times \mathbf{r}_v = [4\cos^2 v\cos u, \quad 4\cos^2 v\sin u, \quad 4\cos v\sin v].$$

Now on S we have $x = 2\cos v\cos u$, $z = 2\sin v$, so that $\mathbf{F} = [7x, 0, -z]$ becomes on S

$$\mathbf{F}(S) = [14\cos v\cos u, \quad 0, \quad -2\sin v]$$

and

$$\mathbf{F}(S)\cdot\mathbf{N} = (14\cos v\cos u)\cdot 4\cos^2 v\cos u + (-2\sin v)\cdot 4\cos v\sin v$$

$$= 56\cos^3 v\cos^2 u - 8\cos v\sin^2 v.$$

On S we have to integrate over u from 0 to 2π. This gives

$$\pi \cdot 56 \cos^3 v - 2\pi \cdot 8 \cos v \sin^2 v.$$

The integral of $\cos v \sin^2 v$ equals $(\sin^3 v)/3$, and that of $\cos^3 v = \cos v\,(1 - \sin^2 v)$ equals $\sin v - (\sin^3 v)/3$. On S we have $-\pi/2 \leqq v \leqq \pi/2$, so that by substituting these limits we get

$$56\pi(2 - 2/3) - 16\pi \cdot 2/3 = 64\pi$$

as hoped for. To see the point of Gauss's theorem, compare the amounts of work. ∎

Coordinate Invariance of the Divergence. The divergence (1) is defined in terms of coordinates, but we can use the divergence theorem to show that div $\mathbf{F}$ has a meaning independent of coordinates.

For this purpose we first note that triple intgrals have properties quite similar to those of double integrals in Sec. 10.3. In particular, the **mean value theorem for triple integrals** asserts that for any continuous function $f(x, y, z)$ in a bounded and simply connected region T there is a point Q: (x_0, y_0, z_0) in T such that

$$\textbf{(9)} \qquad \iiint_T f(x, y, z)\, dV = f(x_0, y_0, z_0)V(T) \qquad (V(T) = \text{volume of } T).$$

In this formula we interchange the two sides, divide by $V(T)$, and set $f = \operatorname{div} \mathbf{F}$. Then by the divergence theorem we obtain for the divergence an integral over the boundary surface $S(T)$ of T,

$$\textbf{(10)} \qquad \operatorname{div} \mathbf{F}(x_0, y_0, z_0) = \frac{1}{V(T)} \iiint_T \operatorname{div} \mathbf{F}\, dV = \frac{1}{V(T)} \iint_{S(T)} \mathbf{F} \cdot \mathbf{n}\, dA.$$

We now choose a point P: (x_1, y_1, z_1) in T and let T shrink down onto P so that the maximum distance $d(T)$ of the points of T from P goes to zero. Then Q: (x_0, y_0, z_0) must approach P. Hence (10) becomes

$$\textbf{(11)} \qquad \operatorname{div} \mathbf{F}(P) = \lim_{d(T) \to 0} \frac{1}{V(T)} \iint_{S(T)} \mathbf{F} \cdot \mathbf{n}\, dA.$$

This proves

THEOREM 2

Invariance of the Divergence

The divergence of a vector function $\mathbf{F}$ *with continuous first partial derivatives in a region* T *is independent of the particular choice of Cartesian coordinates. For any* P *in* T *it is given by* (11).

Equation (11) is sometimes used as a *definition* of the divergence. Then the representation (1) in Cartesian coordinates can be derived from (11).

Further applications of the divergence theorem follow in the problem set and in the next section. The examples in the next section will also shed further light on the nature of the divergence.

PROBLEM SET 10.7

1–8 APPLICATION OF TRIPLE INTEGRALS: MASS DISTRIBUTION

Find the total mass of a mass distribution of density σ in a region T in space. (Show the details of your work.)

1. $\sigma = x^2y^2z^2$, T the box $|x| \leqq a$, $|y| \leqq b$, $|z| \leqq c$
2. $\sigma = x^2 + y^2 + z^2$, T the box $0 \leqq x \leqq 4$, $0 \leqq y \leqq 9$, $0 \leqq z \leqq 1$
3. $\sigma = \sin x \cos y$, T: $0 \leqq x \leqq \frac{1}{2}\pi$, $\frac{1}{2}\pi - x \leqq y \leqq \frac{1}{2}\pi$, $0 \leqq z \leqq 12$
4. $\sigma = e^{-x-y-z}$, T the tetrahedron with vertices (0, 0, 0), (2, 0, 0), (0, 2, 0), (0, 0, 2)
5. $\sigma = \frac{1}{2}(x^2 + y^2)^2$, T the cylinder $x^2 + y^2 \leqq 4$, $|z| \leqq 2$
6. $\sigma = 30z$, T the region in the first octant bounded by $y = 1 - x^2$ and $z = x$. Sketch it.
7. $\sigma = 1 + y + z^2$, T the cylinder $y^2 + z^2 \leqq 9$, $1 \leqq x \leqq 9$
8. $\sigma = x^2 + y^2$, T the ball $x^2 + y^2 + z^2 \leqq a^2$

9–14 APPLICATION OF TRIPLE INTEGRALS: MOMENT OF INERTIA

$I_x = \iiint_T (y^2 + z^2)\, dx\, dy\, dz$ of a mass of density 1 in a region T about the x-axis. Find I_x when T is as follows.

9. The cube $0 \leqq x \leqq a$, $0 \leqq y \leqq a$, $0 \leqq z \leqq a$
10. The box $0 \leqq x \leqq a$, $-b/2 \leqq y \leqq b/2$, $-c/2 \leqq z \leqq c/2$
11. The cylinder $y^2 + z^2 \leqq a^2$, $0 \leqq x \leqq h$
12. The ball $x^2 + y^2 + z^2 \leqq a^2$
13. The cone $y^2 + z^2 \leqq x^2$, $0 \leqq x \leqq h$
14. The paraboloid $y^2 + z^2 \leqq x$, $0 \leqq x \leqq h$
15. Show that for a solid of revolution, $I_x = \frac{1}{2}\pi \int_0^h r^4(x)\, dx$. Use this to solve Probs. 11–14.
16. Why is I_x in Prob. 13 for large h larger than I_x in Prob. 14? Why is it smaller for $h = 1$? Give physical reason.

17–25 APPLICATION OF THE DIVERGENCE THEOREM: SURFACE INTEGRALS $\iint_S \mathbf{F} \cdot \mathbf{n}\, dA$

Evaluate this integral by the divergence theorem. (Show the details.)

17. $\mathbf{F} = [x, y, z]$, S the sphere $x^2 + y^2 + z^2 = 9$
18. $\mathbf{F} = [4x, 3z, 5y]$, S the surface of the cone $x^2 + y^2 \leqq z^2$, $0 \leqq z \leqq 2$
19. $\mathbf{F} = [z - y \quad y^3, \quad 2z^3]$, S the surface of $y^2 + z^2 \leqq 4$, $-3 \leqq x \leqq 3$
20. $\mathbf{F} = [3xy^2, \quad yx^2 - y^3, \quad 3zx^2]$, S the surface of $x^2 + y^2 \leqq 25$, $0 \leqq z \leqq 2$
21. $\mathbf{F} = [\sin y, \quad \cos x, \quad \cos z]$, S the surface of $x^2 + y^2 \leqq 4$, $|z| \leqq 2$
22. $\mathbf{F} = [x^3 - y^3, \quad y^3 - z^3, \quad z^3 - x^3]$, S the surface of $x^2 + y^2 + z^2 \leqq 25$, $z \geqq 0$
23. $\mathbf{F} = [4x^2, \quad 2x + y^2, \quad x^2 + z^2]$, S the surface of the tetrahedron in Prob. 4
24. $\mathbf{F} = [4x^2, \quad y^2, \quad -2 \cos \pi z]$, S the surface of the tetrahedron with vertices (0, 0, 0), (1, 0, 0), (0, 1, 0), (0, 0, 1)
25. $\mathbf{F} = [5x^3, \quad 5y^3, \quad 5z^3]$, S: $x^2 + y^2 + z^2 = 4$

10.8 Further Applications of the Divergence Theorem

We show in this section that the divergence theorem has basic applications in *fluid flow,* where it helps characterize *sources* and *sinks* of fluid, in *heat flow,* where it leads to the basic *heat equation,* and in *potential theory,* where it gives basic properties of the solutions of *Laplace's equation.* Here the region T and its boundary surface S are assumed to be such that the divergence theorem applies.

EXAMPLE 1 Fluid Flow. Physical Interpretation of the Divergence

From the divergence theorem we may obtain an intuitive interpretation of the divergence of a vector. For this purpose we consider the flow of an incompressible fluid (see Sec. 9.8) of constant density $\rho = 1$ which is **steady,** that is, does not vary with time. Such a flow is determined by the field of its velocity vector $\mathbf{v}(P)$ at any point P.

Let S be the boundary surface of a region T in space, and let $\mathbf{n}$ be the outer unit normal vector of S. Then $\mathbf{v} \cdot \mathbf{n}$ is the normal component of $\mathbf{v}$ in the direction of $\mathbf{n}$, and $|\mathbf{v} \cdot \mathbf{n}\, dA|$ is the mass of fluid *leaving* T (if $\mathbf{v} \cdot \mathbf{n} > 0$ at some P) or *entering* T (if $\mathbf{v} \cdot \mathbf{n} < 0$ at P) per unit time at some point P of S through a small portion ΔS of S of area ΔA. Hence the total mass of fluid that flows across S from T to the outside per unit time is given by the surface integral

$$\iint_S \mathbf{v} \cdot \mathbf{n}\, dA.$$

Division by the volume V of T gives the **average flow** out of T:

$$\frac{1}{V} \iint_S \mathbf{v} \cdot \mathbf{n}\, dA. \tag{1}$$

Since the flow is steady and the fluid is incompressible, the amount of fluid flowing outward must be continuously supplied. Hence, if the value of the integral (1) is different from zero, there must be **sources** (*positive sources and negative sources, called* **sinks**) in T, that is, points where fluid is produced or disappears.

If we let T shrink down to a fixed point P in T, we obtain from (1) the **source intensity** at P given by the right side of (11) in the last section with $\mathbf{F} \cdot \mathbf{n}$ replaced by $\mathbf{v} \cdot \mathbf{n}$, that is,

$$\operatorname{div} \mathbf{v}(P) = \lim_{d(T) \to 0} \frac{1}{V(T)} \iint_{S(T)} \mathbf{v} \cdot \mathbf{n}\, dA. \tag{2}$$

Hence *the divergence of the velocity vector* $\mathbf{v}$ *of a steady incompressible flow is the source intensity of the flow at the corresponding point.*

There are no sources in T if and only if div $\mathbf{v}$ is zero everywhere in T. Then for any closed surface S in T we have

$$\iint_S \mathbf{v} \cdot \mathbf{n}\, dA = 0.$$

EXAMPLE 2 **Modeling of Heat Flow. Heat or Diffusion Equation**

Physical experiments show that in a body, heat flows in the direction of decreasing temperature, and the rate of flow is proportional to the gradient of the temperature. This means that the velocity $\mathbf{v}$ of the heat flow in a body is of the form

$$\mathbf{v} = -K \operatorname{grad} U \tag{3}$$

where $U(x, y, z, t)$ is temperature, t is time, and K is called the *thermal conductivity* of the body; in ordinary physical circumstances K is a constant. Using this information, set up the mathematical model of heat flow, the so-called **heat equation** or **diffusion equation.**

Solution. Let T be a region in the body bounded by a surface S with outer unit normal vector $\mathbf{n}$ such that the divergence theorem applies. Then $\mathbf{v} \cdot \mathbf{n}$ is the component of $\mathbf{v}$ in the direction of $\mathbf{n}$, and the amount of heat leaving T per unit time is

$$\iint_S \mathbf{v} \cdot \mathbf{n}\, dA.$$

This expression is obtained similarly to the corresponding surface integral in the last example. Using

$$\operatorname{div} (\operatorname{grad} U) = \nabla^2 U = U_{xx} + U_{yy} + U_{zz}$$

(the Laplacian; see (3) in Sec. 9.8), we have by the divergence theorem and (3)

$$\begin{aligned} \iint_S \mathbf{v} \cdot \mathbf{n}\, dA &= -K \iiint_T \operatorname{div} (\operatorname{grad} U)\, dx\, dy\, dz \\ &= -K \iiint_T \nabla^2 U\, dx\, dy\, dz. \end{aligned} \tag{4}$$

On the other hand, the total amount of heat H in T is

$$H = \iiint_T \sigma\rho U \, dx\, dy\, dz$$

where the constant σ is the specific heat of the material of the body and ρ is the density (= mass per unit volume) of the material. Hence the time rate of decrease of H is

$$-\frac{\partial H}{\partial t} = -\iiint_T \sigma\rho \frac{\partial U}{\partial t} \, dx\, dy\, dz$$

and this must be equal to the above amount of heat leaving T. From (4) we thus have

$$-\iiint_T \sigma\rho \frac{\partial U}{\partial t} \, dx\, dy\, dz = -K \iiint_T \nabla^2 U \, dx\, dy\, dz$$

or

$$\iiint_T \left(\sigma\rho \frac{\partial U}{\partial t} - K\nabla^2 U\right) dx\, dy\, dz = 0.$$

Since this holds for any region T in the body, the integrand (if continuous) must be zero everywhere; that is,

(5)
$$\frac{\partial U}{\partial t} = c^2 \nabla^2 U \qquad c^2 = \frac{K}{\sigma\rho}$$

where c^2 is called the *thermal diffusivity* of the material. This partial differential equation is called the **heat equation.** It is the fundamental equation for heat conduction. And our derivation is another impressive demonstration of the great importance of the divergence theorem. Methods for solving heat problems will be shown in Chap. 12.

The heat equation is also called the **diffusion equation** because it also models diffusion processes of motions of molecules tending to level off differences in density or pressure in gases or liquids.

If heat flow does not depend on time, it is called **steady-state heat flow.** Then $\partial U/\partial t = 0$, so that (5) reduces to *Laplace's equation* $\nabla^2 U = 0$. We met this equation in Secs. 9.7 and 9.8, and we shall now see that the divergence theorem adds basic insights into the nature of solutions of this equation. ■

Potential Theory. Harmonic Functions

The theory of solutions of **Laplace's equation**

(6)
$$\nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2} = 0$$

is called **potential theory.** A solution of (6) with ***continuous*** second-order partial derivatives is called a **harmonic function.** That continuity is needed for application of the divergence theorem in potential theory, where the theorem plays a key role that we want to explore. Further details of potential theory follow in Chaps. 12 and 18.

EXAMPLE 3 **A Basic Property of Solutions of Laplace's Equation**

The integrands in the divergence theorem are div $\mathbf{F}$ and $\mathbf{F} \cdot \mathbf{n}$ (Sec. 10.7). If $\mathbf{F}$ is the gradient of a scalar function, say, $\mathbf{F} = \text{grad } f$, then div $\mathbf{F}$ = div (grad f) = $\nabla^2 f$; see (3), Sec. 9.8. Also, $\mathbf{F} \cdot \mathbf{n} = \mathbf{n} \cdot \mathbf{F} = \mathbf{n} \cdot \text{grad } f$. This is the directional derivative of f in the outer normal direction of S, the boundary surface of the region T in the theorem. This derivative is called the (outer) **normal derivative** of f and is denoted by $\partial f/\partial n$. Thus the formula in the divergence theorem becomes

(7) $$\iiint_T \nabla^2 f \, dV = \iint_S \frac{\partial f}{\partial n} \, dA.$$

This is the three-dimensional analog of (9) in Sec. 10.4. Because of the assumptions in the divergence theorem this gives the following result.

THEOREM 1

A Basic Property of Harmonic Functions

Let $f(x, y, z)$ be a harmonic function in some domain D is space. Let S be any piecewise smooth closed orientable surface in D whose entire region it encloses belongs to D. Then the integral of the normal derivative of f taken over S is zero. (For "piecewise smooth" see Sec. 10.5.)

EXAMPLE 4 **Green's Theorems**

Let f and g be scalar functions such that $F = f \operatorname{grad} g$ satisfies the assumptions of the divergence theorem in some region T. Then

$$\begin{aligned} \operatorname{div} \mathbf{F} &= \operatorname{div} (f \operatorname{grad} g) \\ &= \operatorname{div} \left(\left[f \frac{\partial g}{\partial x}, f \frac{\partial g}{\partial y}, f \frac{\partial g}{\partial z} \right] \right) \\ &= \left(\frac{\partial f}{\partial x} \frac{\partial g}{\partial x} + f \frac{\partial^2 g}{\partial x^2} \right) + \left(\frac{\partial f}{\partial y} \frac{\partial g}{\partial y} + f \frac{\partial^2 g}{\partial y^2} \right) + \left(\frac{\partial f}{\partial z} \frac{\partial g}{\partial z} + f \frac{\partial^2 g}{\partial z^2} \right) \\ &= f \nabla^2 g + \operatorname{grad} f \cdot \operatorname{grad} g. \end{aligned}$$

Also, since f is a scalar function,

$$\begin{aligned} \mathbf{F} \cdot \mathbf{n} &= \mathbf{n} \cdot \mathbf{F} \\ &= \mathbf{n} \cdot (f \operatorname{grad} g) \\ &= (\mathbf{n} \cdot \operatorname{grad} g) f. \end{aligned}$$

Now $\mathbf{n} \cdot \operatorname{grad} g$ is the directional derivative $\partial g / \partial n$ of g in the outer normal direction of S. Hence the formula in the divergence theorem becomes **"Green's first formula"**

(8) $$\iiint_T (f \nabla^2 g + \operatorname{grad} f \cdot \operatorname{grad} g) \, dV = \iint_S f \frac{\partial g}{\partial n} \, dA.$$

Formula (8) together with the assumptions is known as the *first form of Green's theorem.*

Interchanging f and g we obtain a similar formula. Subtracting this formula from (8) we find

(9) $$\iiint_T (f \nabla^2 g - g \nabla^2 f) \, dV = \iint_S \left(f \frac{\partial g}{\partial n} - g \frac{\partial f}{\partial n} \right) dA.$$

This formula is called **Green's second formula** or (together with the assumptions) the *second form of Green's theorem.*

EXAMPLE 5 **Uniqueness of Solutions of Laplace's Equation**

Let f be harmonic in a domain D and let f be zero everywhere on a piecewise smooth closed orientable surface S in D whose entire region T it encloses belongs to D. Then $\nabla^2 g$ is zero in T, and the surface integral in (8) is zero, so that (8) with $g = f$ gives

$$\iiint_T \text{grad } f \cdot \text{grad } f \, dV = \iiint_T |\text{grad } f|^2 \, dV = 0.$$

Since f is harmonic, grad f and thus $|\text{grad } f|$ are continuous in T and on S, and since $|\text{grad } f|$ is nonnegative, to make the integral over T zero, grad f must be the zero vector everywhere in T. Hence $f_x = f_y = f_z = 0$, and f is constant in T and, because of continuity, it is equal to its value 0 on S. This proves the following theorem.

THEOREM 2

Harmonic Functions

Let $f(x, y, z)$ be harmonic in some domain D and zero at every point of a piecewise smooth closed orientable surface S in D whose entire region T it encloses belongs to D. Then f is identically zero in T.

This theorem has an important consequence. Let f_1 and f_2 be functions that satisfy the assumptions of Theorem 1 and take on the same values on S. Then their difference $f_1 - f_2$ satisfies those assumptions and has the value 0 everywhere on S. Hence, Theorem 2 implies that

$$f_1 - f_2 = 0 \qquad \text{throughout} \qquad T,$$

and we have the following fundamental result.

THEOREM 3

Uniqueness Theorem for Laplace's Equation

Let T be a region that satisfies the assumptions of the divergence theorem, and let $f(x, y, z)$ be a harmonic function in a domain D that contains T and its boundary surface S. Then f is uniquely determined in T by its values on S.

The problem of determining a solution u of a partial differential equation in a region T such that u assumes given values on the boundary surface S of T is called the **Dirichlet problem.**[8] We may thus reformulate Theorem 3 as follows.

THEOREM 3*

Uniqueness Theorem for the Dirichlet Problem

If the assumptions in Theorem 3 are satisfied and the Dirichlet problem for the Laplace equation has a solution in T, then this solution is unique.

These theorems demonstrate the extreme importance of the divergence theorem in potential theory. ■

[8]PETER GUSTAV LEJEUNE DIRICHLET (1805–1859), German mathematician, studied in Paris under Cauchy and others and succeeded Gauss at Göttingen in 1855. He became known by his important research on Fourier series (he knew Fourier personally) and in number theory.

PROBLEM SET 10.8

1. **(Harmonic functions)** Verify Theorem 1 for $f = 2x^2 + 2y^2 - 4z^2$ and S the surface of the cube $0 \leqq x \leqq 1, 0 \leqq y \leqq 1, 0 \leqq z \leqq 1$.
2. **(Harmonic functions)** Verify Theorem 1 for $f = y^2 - x^2$ and the surface of the cylinder $x^2 + y^2 \leqq 1, 0 \leqq z \leqq 5$.
3. **(Green's first formula)** Verify (8) for $f = 3y^2$, $g = x^2$, S the surface of the cube in Prob. 1.
4. **(Green's first formula)** Verify (8) for $f = x$, $g = y^2 + z^2$, S the surface of the box $0 \leqq x \leqq 1$, $0 \leqq y \leqq 2, 0 \leqq z \leqq 3$.
5. **(Green's second formula)** Verify (9) for the data in Prob. 3.
6. **(Green's second formula)** Verify (9) for $f = x^4$, $g = y^2$ and the cube in Prob. 1.
7. **(Volume as a surface integral)** Show that a region T with boundary surface S has the volume

$$V = \frac{1}{3}\iint_S r \cos\phi \, dA$$

where r is the distance of a variable point P: (x, y, z) on S from the origin O and ϕ is the angle between the directed line OP and the outer normal of S at P. (Make a sketch.)

8. Find the volume of a ball of radius a by means of the formula in Prob. 7.
9. Show that a region T with boundary surface S has the volume

$$V = \iint_S x \, dy \, dz$$

$$= \iint_S y \, dz \, dx$$

$$= \iint_S z \, dx \, dy$$

$$= \frac{1}{3}\iint_S (x \, dy \, dz + y \, dz \, dx + z \, dx \, dy).$$

10. **TEAM PROJECT. Divergence Theorem and Potential Theory.** The importance of the divergence theorem in potential theory is obvious from (7)–(9) and Theorems 1–3. To emphasize it further, consider functions f and g that are harmonic in some domain D containing a region T with boundary surface S such that T satisfies the assumptions in the divergence theorem. Prove and illustrate by examples that then:

(a) $$\iint_S g \frac{\partial g}{\partial n} \, dA = \iiint_T |\text{grad } g|^2 \, dV.$$

(b) If $\partial g/\partial n = 0$ on S, then g is constant in T.

(c) $$\iint_S \left(f \frac{\partial g}{\partial n} - g \frac{\partial f}{\partial n} \right) dA = 0.$$

(d) If $\partial f/\partial n = \partial g/\partial n$ on S, then $f = g + c$ in T, where c is a constant.

(e) The **Laplacian** can be represented ***independently of coordinate systems*** in the form

$$\nabla^2 f = \lim_{d(T)\to 0} \frac{1}{V(T)} \iint_{S(T)} \frac{\partial f}{\partial n} \, dA$$

where $d(T)$ is the maximum distance of the points of a region T bounded by $S(T)$ from the point at which the Laplacian is evaluated and $V(T)$ is the volume of T.

10.9 Stokes's Theorem

Having seen the great usefulness of Gauss's divergence theorem, we now turn to the second "big" theorem in this chapter, Stokes's theorem. This theorem transforms line integrals into surface integrals and conversely. Hence it generalizes Green's theorem of Sec. 10.4. Stokes's theorem involves the curl

(1) $$\text{curl } \mathbf{F} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \partial/\partial x & \partial/\partial y & \partial/\partial z \\ F_1 & F_2 & F_3 \end{vmatrix}$$ (see Sec. 9.9).

THEOREM 1

Stokes's Theorem[9]
(Transformation Between Surface and Line Integrals)

Let S be a piecewise smooth[9] oriented surface in space and let the boundary of S be a piecewise smooth simple closed curve C. Let $\mathbf{F}(x, y, z)$ be a continuous vector function that has continuous first partial derivatives in a domain in space containing S. Then

$$\iint_S (\operatorname{curl} \mathbf{F}) \cdot \mathbf{n}\, dA = \oint_C \mathbf{F} \cdot \mathbf{r}'(s)\, ds. \tag{2}$$

Here $\mathbf{n}$ is a unit normal vector of S and, depending on $\mathbf{n}$, the integration around C is taken in the sense shown in Fig. 251. Furthermore, $\mathbf{r}' = d\mathbf{r}/ds$ is the unit tangent vector and s the arc length of C.

In components, formula (2) becomes

$$\iint_R \left[\left(\frac{\partial F_3}{\partial y} - \frac{\partial F_2}{\partial z} \right) N_1 + \left(\frac{\partial F_1}{\partial z} - \frac{\partial F_3}{\partial x} \right) N_2 + \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} \right) N_3 \right] du\, dv$$
$$= \oint_{\bar{C}} (F_1\, dx + F_2\, dy + F_3\, dz). \tag{2*}$$

Here, $\mathbf{F} = [F_1, F_2, F_3]$, $\mathbf{N} = [N_1, N_2, N_3]$, $\mathbf{n}\, dA = \mathbf{N}\, du\, dv$, $\mathbf{r}'\, ds = [dx, dy, dz]$, and R is the region with boundary curve $\bar{C}$ in the uv-plane corresponding to S represented by $\mathbf{r}(u, v)$.

The proof follows after Example 1.

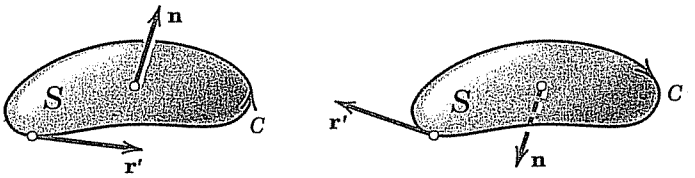


Fig. 251. Stokes's theorem

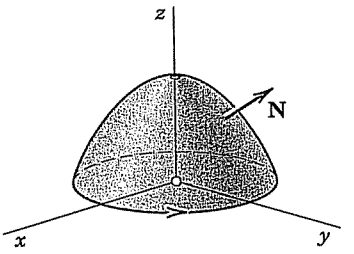


Fig. 252. Surface S in Example 1

EXAMPLE 1 **Verification of Stokes's Theorem**

Before we prove Stokes's theorem, let us first get used to it by verifying it for $\mathbf{F} = [y, z, x]$ and S the paraboloid (Fig. 252)

$$z = f(x, y) = 1 - (x^2 + y^2), \qquad z \geqq 0.$$

Solution. The curve C, oriented as in Fig. 252, is the circle $\mathbf{r}(s) = [\cos s, \sin s, 0]$. Its unit tangent vector is $\mathbf{r}'(s) = [-\sin s, \cos s, 0]$. The function $\mathbf{F} = [y, z, x]$ on C is $\mathbf{F}(\mathbf{r}(s)) = [\sin s, \ 0, \ \cos s]$. Hence

$$\oint_C \mathbf{F} \cdot d\mathbf{r} = \int_0^{2\pi} \mathbf{F}(\mathbf{r}(s)) \cdot \mathbf{r}'(s)\, ds = \int_0^{2\pi} [(\sin s)(-\sin s) + 0 + 0]\, ds = -\pi.$$

---

[9]Sir GEORGE GABRIEL STOKES (1819–1903), Irish mathematician and physicist who became a professor in Cambridge in 1849. He is also known for his important contribution to the theory of infinite series and to viscous flow (Navier–Stokes equations), geodesy, and optics.

"Piecewise smooth" curves and surfaces are defined in Secs. 10.1 and 10.5.

We now consider the surface integral. We have $F_1 = y$, $F_2 = z$, $F_3 = x$, so that in (2*) we obtain

$$\text{curl } \mathbf{F} = \text{curl } [F_1, \quad F_2, \quad F_3] = \text{curl } [y, \quad z, \quad x] = [-1, \quad -1, \quad -1].$$

A normal vector of S is $\mathbf{N} = \text{grad } (z - f(x, y)) = [2x, 2y, 1]$. Hence $(\text{curl } \mathbf{F}) \cdot \mathbf{N} = -2x - 2y - 1$. Now $\mathbf{n}\, dA = \mathbf{N}\, dx\, dy$ (see (3*) in Sec. 10.6 with x, y instead of u, v). Using polar coordinates r, θ defined by $x = r \cos \theta$, $y = r \sin \theta$ and denoting the projection of S into the xy-plane by R, we thus obtain

$$\iint_S (\text{curl } \mathbf{F}) \cdot \mathbf{n}\, dA = \iint_R (\text{curl } \mathbf{F}) \cdot \mathbf{N}\, dx\, dy = \iint_R (-2x - 2y - 1)\, dx\, dy$$

$$= \int_{\theta=0}^{2\pi} \int_{r=0}^{1} (-2r(\cos \theta + \sin \theta) - 1) r\, dr\, d\theta$$

$$= \int_{\theta=0}^{2\pi} \left(-\frac{2}{3} (\cos \theta + \sin \theta) - \frac{1}{2} \right) d\theta = 0 + 0 - \frac{1}{2} (2\pi) = -\pi. \qquad \blacksquare$$

PROOF We prove Stokes's theorem. Obviously, (2) holds if the integrals of each component on both sides of (2*) are equal; that is,

(3) $$\iint_R \left(\frac{\partial F_1}{\partial z} N_2 - \frac{\partial F_1}{\partial y} N_3 \right) du\, dv = \oint_{\bar{C}} F_1\, dx$$

(4) $$\iint_R \left(-\frac{\partial F_2}{\partial z} N_1 + \frac{\partial F_2}{\partial x} N_3 \right) du\, dv = \oint_{\bar{C}} F_2\, dy$$

(5) $$\iint_R \left(\frac{\partial F_3}{\partial y} N_1 - \frac{\partial F_3}{\partial x} N_2 \right) du\, dv = \oint_{\bar{C}} F_3\, dz.$$

We prove this first for a surface S that can be represented simultaneously in the forms

(6) (a) $z = f(x, y)$, (b) $y = g(x, z)$, (c) $x = h(y, z)$.

We prove (3), using (6a). Setting $u = x$, $v = y$, we have from (6a)

$$\mathbf{r}(u, v) = \mathbf{r}(x, y) = [x, y, f(x, y)] = x\mathbf{i} + y\mathbf{j} + f\mathbf{k}$$

and in (2), Sec. 10.6, by direct calculation

$$\mathbf{N} = \mathbf{r}_u \times \mathbf{r}_v = \mathbf{r}_x \times \mathbf{r}_y = [-f_x, \quad -f_y, \quad 1] = -f_x\mathbf{i} - f_y\mathbf{j} + \mathbf{k}.$$

Note that $\mathbf{N}$ is an *upper* normal vector of S, since it has a *positive* z-component. Also, $R = S^*$, the projection of S into the xy-plane, with boundary curve $\bar{C} = C^*$ (Fig. 253). Hence the left side of (3) is

(7) $$\iint_{S^*} \left[\frac{\partial F_1}{\partial z} (-f_y) - \frac{\partial F_1}{\partial y} \right] dx\, dy.$$

We now consider the right side of (3). We transform this line integral over $\bar{C} = C^*$ into a double integral over S^* by applying Green's theorem [formula (1) in Sec. 10.4 with $F_2 = 0$]. This gives

$$\oint_{C^*} F_1\, dx = \iint_{S^*} -\frac{\partial F_1}{\partial y}\, dx\, dy.$$

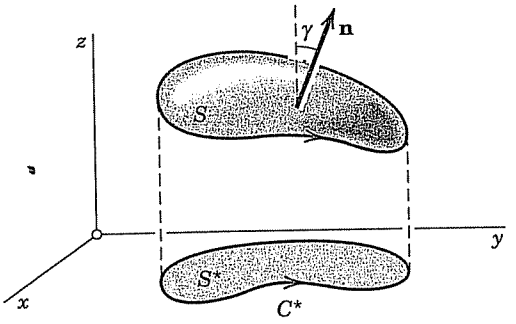


Fig. 253. Proof of Stokes's theorem

Here, $F_1 = F_1(x, y, f(x, y))$. Hence by the chain rule (see also Prob. 10 in Problem Set 9.6),

$$-\frac{\partial F_1(x, y, f(x, y))}{\partial y} = -\frac{\partial F_1(x, y, z)}{\partial y} - \frac{\partial F_1(x, y, z)}{\partial z}\,\frac{\partial f}{\partial y} \qquad [z = f(x, y)].$$

We see that the right side of this equals the integrand in (7). This proves (3). Relations (4) and (5) follow in the same way if we use (6b) and (6c), respectively. By addition we obtain (2*). This proves Stokes's theorem for a surface S that can be represented simultaneously in the forms (6a), (6b), (6c).

As in the proof of the divergence theorem, our result may be immediately extended to a surface S that can be decomposed into finitely many pieces, each of which is of the kind just considered. This covers most of the cases of practical interest. The proof in the case of a most general surface S satisfying the assumptions of the theorem would require a limit process; this is similar to the situation in the case of Green's theorem in Sec. 10.4. ▩

EXAMPLE 2 **Green's Theorem in the Plane as a Special Case of Stokes's Theorem**

Let $\mathbf{F} = [F_1, F_2] = F_1\mathbf{i} + F_2\mathbf{j}$ be a vector function that is continuously differentiable in a domain in the xy-plane containing a simply connected bounded closed region S whose boundary C is a piecewise smooth simple closed curve. Then, according to (1),

$$(\operatorname{curl}\mathbf{F})\cdot\mathbf{n} = (\operatorname{curl}\mathbf{F})\cdot\mathbf{k} = \frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y}.$$

Hence the formula in Stokes's theorem now takes the form

$$\iint_S \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y}\right) dA = \oint_C (F_1\,dx + F_2\,dy).$$

This shows that Green's theorem in the plane (Sec. 10.4) is a special case of Stokes's theorem (which we needed in the proof of the latter!). ▩

EXAMPLE 3 **Evaluation of a Line Integral by Stokes's Theorem**

Evaluate $\int_C \mathbf{F}\cdot\mathbf{r}'\,ds$, where C is the circle $x^2 + y^2 = 4$, $z = -3$, oriented counterclockwise as seen by a person standing at the origin, and, with respect to right-handed Cartesian coordinates,

$$\mathbf{F} = [y,\quad xz^3,\quad -zy^3] = y\mathbf{i} + xz^3\mathbf{j} - zy^3\mathbf{k}.$$

Solution. As a surface S bounded by C we can take the plane circular disk $x^2 + y^2 \leqq 4$ in the plane $z = -3$. Then $\mathbf{n}$ in Stokes's theorem points in the positive z-direction; thus $\mathbf{n} = \mathbf{k}$. Hence $(\operatorname{curl}\mathbf{F})\cdot\mathbf{n}$ is simply the component of curl $\mathbf{F}$ in the positive z-direction. Since $\mathbf{F}$ with $z = -3$ has the components $F_1 = y$, $F_2 = -27x$, $F_3 = 3y^3$, we thus obtain

$$(\operatorname{curl}\mathbf{F})\cdot\mathbf{n} = \frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} = -27 - 1 = -28.$$

Hence the integral over S in Stokes's theorem equals -28 times the area 4π of the disk S. This yields the answer $-28 \cdot 4\pi = -112\pi \approx -352$. Confirm this by direct calculation, which involves somewhat more work.

EXAMPLE 4 **Physical Meaning of the Curl in Fluid Motion. Circulation**

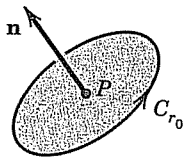


Fig. 254. Example 4

Let S_{r_0} be a circular disk of radius r_0 and center P bounded by the circle C_{r_0} (Fig. 254), and let $\mathbf{F}(Q) \equiv \mathbf{F}(x, y, z)$ be a continuously differentiable vector function in a domain containing S_{r_0}. Then by Stokes's theorem and the mean value theorem for surface integrals (see Sec. 10.6),

$$\oint_{C_{r_0}} \mathbf{F} \cdot \mathbf{r}' \, ds = \iint_{S_{r_0}} (\operatorname{curl} \mathbf{F}) \cdot \mathbf{n} \, dA = (\operatorname{curl} \mathbf{F}) \cdot \mathbf{n}(P^*) A_{r_0}$$

where A_{r_0} is the area of S_{r_0} and P^* is a suitable point of S_{r_0}. This may be written in the form

$$(\operatorname{curl} \mathbf{F}) \cdot \mathbf{n}(P^*) = \frac{1}{A_{r_0}} \oint_{C_{r_0}} \mathbf{F} \cdot \mathbf{r}' \, ds.$$

In the case of a fluid motion with velocity vector $\mathbf{F} = \mathbf{v}$, the integral

$$\oint_{C_{r_0}} \mathbf{v} \cdot \mathbf{r}' \, ds$$

is called the **circulation** of the flow around C_{r_0}. It measures the extent to which the corresponding fluid motion is a rotation around the circle C_{r_0}. If we now let r_0 approach zero, we find

$$(8) \qquad (\operatorname{curl} \mathbf{v}) \cdot \mathbf{n}(P) = \lim_{r_0 \to 0} \frac{1}{A_{r_0}} \oint_{C_{r_0}} \mathbf{v} \cdot \mathbf{r}' \, ds;$$

that is, the component of the curl in the positive normal direction can be regarded as the **specific circulation** (circulation per unit area) of the flow in the surface at the corresponding point.

EXAMPLE 5 **Work Done in the Displacement around a Closed Curve**

Find the work done by the force $\mathbf{F} = 2xy^3 \sin z \, \mathbf{i} + 3x^2y^2 \sin z \, \mathbf{j} + x^2y^3 \cos z \, \mathbf{k}$ in the displacement around the curve of intersection of the paraboloid $z = x^2 + y^2$ and the cylinder $(x - 1)^2 + y^2 = 1$.

Solution. This work is given by the line integral in Stokes's theorem. Now $\mathbf{F} = \operatorname{grad} f$, where $f = x^2y^3 \sin z$ and curl (grad f) $= \mathbf{0}$ (see (2) in Sec. 9.9), so that (curl $\mathbf{F}$)$\cdot \mathbf{n} = 0$ and the work is 0 by Stokes's theorem. This agrees with the fact that the present field is conservative (definition in Sec. 7.7).

Stokes's Theorem Applied to Path Independence

We emphasized in Sec. 10.2 that the value of a line integral generally depends not only on the function to be integrated and on the two endpoints A and B of the path of integration C, but also on the particular choice of a path from A to B. In Theorem 3 of Sec. 10.2 we proved that if a line integral

$$(9) \qquad \int_C \mathbf{F}(\mathbf{r}) \cdot d\mathbf{r} = \int_C (F_1 \, dx + F_2 \, dy + F_3 \, dz)$$

(involving continuous F_1, F_2, F_3 that have continuous first partial derivatives) is path independent in a domain D, then curl $\mathbf{F} = \mathbf{0}$ in D. And we claimed in Sec. 10.2 that, conversely, curl $\mathbf{F} = \mathbf{0}$ everywhere in D implies path independence of (9) in D provided D is *simply connected*. A proof of this needs Stokes's theorem and can now be given as follows.

Let C be any closed path in D. Since D is simply connected, we can find a surface S in D bounded by C. Stokes's theorem applies and gives

$$\oint_C (F_1 \, dx + F_2 \, dy + F_3 \, dz) = \oint_C \mathbf{F} \cdot \mathbf{r}' \, ds = \iint_S (\operatorname{curl} \mathbf{F}) \cdot \mathbf{n} \, dA$$

for proper direction on C and normal vector $\mathbf{n}$ on S. Since curl $\mathbf{F} = \mathbf{0}$ in D, the surface integral and hence the line integral are zero. This and Theorem 2 of Sec. 10.2 imply that the integral (9) is path independent in D. This completes the proof. ∎

PROBLEM SET 10.9

1–8 DIRECT INTEGRATION OF THE SURFACE INTEGRALS

Evaluate the integral $\iint_S (\text{curl } \mathbf{F}) \cdot \mathbf{n}\, dA$ directly for the given $\mathbf{F}$ and S.

1. $\mathbf{F} = [4z^2, \quad 16x, \quad 0]$, S: $z = y$ $(0 \leqq x \leqq 1, 0 \leqq y \leqq 1)$
2. $\mathbf{F} = [0, \quad 0, \quad 5x \cos z]$, S: $x^2 + y^2 = 4$, $y \geqq 0$, $0 \leqq z \leqq \frac{1}{2}\pi$
3. $\mathbf{F} = [-e^y, \quad e^z, \quad e^x]$, S: $z = x + y$ $(0 \leqq x \leqq 1, 0 \leqq y \leqq 1)$
4. $\mathbf{F} = [3 \cos y, \quad \cosh z, \quad x]$, S the square $0 \leqq x \leqq 2$, $0 \leqq y \leqq 2$, $z = 4$
5. $\mathbf{F} = [e^{2z}, \quad e^z \sin y, \quad e^z \cos y]$, S: $z = y^2$ $(0 \leqq x \leqq 4, 0 \leqq y \leqq 1)$
6. $\mathbf{F} = [z^2, \quad x^2, \quad y^2]$, S: $z^2 = x^2 + y^2$, $y \geqq 0$, $0 \leqq z \leqq 2$
7. $\mathbf{F} = [z^2, \quad \frac{3}{2}x, \quad 0]$, S the square $0 \leqq x \leqq a$, $0 \leqq y \leqq a$, $z = 1$
8. $\mathbf{F} = [y^3, \quad -x^3, \quad 0]$, S: $x^2 + y^2 \leqq 1$, $z = 0$
9. Verify Stokes's theorem for $\mathbf{F}$ and S in Prob. 7.
10. Verify Stokes's theorem for $\mathbf{F}$ and S in Prob. 8.

11–18 EVALUATION OF $\oint_C \mathbf{F} \cdot \mathbf{r}'\, ds$

Calculate this line integral by Stokes's theorem, clockwise as seen by a person standing at the origin, for the following $\mathbf{F}$ and C. Assume the Cartesian coordinates to be right-handed. (Show the details.)

11. $\mathbf{F} = [-3y, \quad 3x, \quad z]$, C the circle $x^2 + y^2 = 4$, $z = 1$
12. $\mathbf{F} = [4z, \quad -2x, \quad 2x]$, C the intersection of $x^2 + y^2 = 1$ and $z = y + 1$
13. $\mathbf{F} = [y^2, \quad x^2, \quad -x + z]$, around the triangle with vertices (0, 0, 1), (1, 0, 1), (1, 1, 1)
14. $\mathbf{F} = [y, \quad xy^3, \quad -zy^3]$, C the circle $x^2 + y^2 = a^2$, $z = b$ (> 0)
15. $\mathbf{F} = [y, \quad z^2, \quad x^3]$, C as in Prob. 12
16. $\mathbf{F} = [x^2, \quad y^2, \quad z^2]$, C the intersection of $x^2 + y^2 + z^2 = 4$ and $z = y^2$
17. $\mathbf{F} = [\cos \pi y, \quad \sin \pi x, \quad 0]$, around the rectangle with vertices (0, 1, 0), (0, 0, 1), (1, 0, 1), (1, 1, 0)
18. $\mathbf{F} = [z, \quad x, \quad y]$, C as in Prob. 13
19. **(Stokes's theorem not applicable)** Evaluate $\oint_C \mathbf{F} \cdot \mathbf{r}'\, ds$, $\mathbf{F} = (x^2 + y^2)^{-1}[-y, x]$, C: $x^2 + y^2 = 1$, $z = 0$, oriented clockwise. Why can Stokes's theorem not be applied? What (false) result would it give?
20. **WRITING PROJECT. Grad, Div, Curl in Connection with Integrals.** Make a list of ideas and results on this topic in this chapter. See whether you can rearrange or combine parts of your material. Then subdivide the material into 3–5 portions and work out the details of each portion. Include no proofs but simple typical examples of your own that lead to a better understanding of the material.

CHAPTER 10 REVIEW QUESTIONS AND PROBLEMS

1. List the kinds of integrals in this chapter and how the integral theorems relate some of them.
2. How can work of a variable force be expressed by an integral?
3. State from memory how you can evaluate a line integral. A double integral.
4. What do you remember about path independence? Why is it important?
5. How did we use Stokes's theorem in connection with path independence?
6. State the definition of curl. Why is it important in this chapter?
7. How can you transform a double integral or a surface integral into a line integral?
8. What is orientation of a surface? What is its role in connection with surface integrals?
9. State the divergence theorem and its applications from memory.
10. State Laplace's equation. Where in physics is it important? What properties of its solutions did we discuss?

11–20 LINE INTEGRALS $\int_C \mathbf{F}(\mathbf{r}) \cdot d\mathbf{r}$ (WORK INTEGRALS)

Evaluate, with $\mathbf{F}$ and C as given, by the method that seems most suitable. Recall that if $\mathbf{F}$ is a force, the integral gives the work done in a displacement along C. (Show the details.)

11. $\mathbf{F} = [x^2, \quad y^2, \quad z^2]$, C the straight-line segment from (4, 1, 8) to (0, 2, 3)

12. $\mathbf{F} = [\cos z, \ -\sin z, \ -x \sin z - y \cos z]$, C the straight-line segment from $(-2, 0, \frac{1}{2}\pi)$ to $(4, 3, 0)$
13. $\mathbf{F} = [x - y, \ 0, \ e^z]$, C: $y = 3x^2$, $z = 2x$ for x from 0 to 2
14. $\mathbf{F} = [yz, \ 2zx, \ xy]$, C the circle $x^2 + y^2 = 9$, $z = 1$, counterclockwise
15. $\mathbf{F} = [-3y^3, \ 3x^3 + \cos y, \ 0]$, C the circle $x^2 + y^2 = 16$, $z = 0$, counterclockwise
16. $\mathbf{F} = [\sin \pi y, \ \cos \pi x, \ \sin \pi x]$, C the boundary of $0 \leqq x \leqq 1/2$, $0 \leqq y \leqq 2$, $z = 2x$
17. $\mathbf{F} = [9z, \ 5x, \ 3y]$, C the ellipse $x^2 + y^2 = 9$, $z = x + 2$
18. $\mathbf{F} = [\cosh x, \ e^{4y}, \ \tan z]$, C: $x^2 + y^2 = 4$, $z = x^2$. (Sketch C.)
19. $\mathbf{F} = [z^2, \ x^3, \ y^2]$, C: $x^2 + y^2 = 4$, $x + y + z = 0$
20. $\mathbf{F} = [x^2, \ y^2, \ y^2x]$, C the helix $\mathbf{r} = [2 \cos t, \ 2 \sin t, \ 6t]$ from $(2, 0, 0)$ to $(0, 2, 3\pi)$

21–25 DOUBLE INTEGRALS, CENTER OF GRAVITY

Find the coordinates $\bar{x}$, $\bar{y}$ of the center of gravity of a mass of density $f(x, y)$ in the region R. (Sketch R. Show the details.)

21. $f = 2xy$, R the triangle with vertices $(0, 0)$, $(1, 0)$, $(1, 1)$
22. $f = 1$, R: $0 \leqq y \leqq 1 - x^2$
23. $f = 1$, R: $x^2 + y^2 \leqq a^2$, $y \geqq 0$
24. $f = x^2 + y^2$, R: $x^2 + y^2 \leqq 1$, $x \geqq 0$, $y \geqq 0$
25. $f = 2x^2$, R the region below $y = x + 2$ and above $y = x^2$

26–35 SURFACE INTEGRALS $\iint_S \mathbf{F} \cdot \mathbf{n} \, dA$

Evaluate this integral directly or, if possible, by the divergence theorem. (Show the details.)

26. $\mathbf{F} = [2x^2, \ 4y, \ 0]$, S: $x + y + z = 1$, $x \geqq 0$, $y \geqq 0$, $z \geqq 0$
27. $\mathbf{F} = [y, \ -x, \ 0]$, S: $3x + 2y + z = 6$, $x \geqq 0$, $y \geqq 0$, $z \geqq 0$
28. $\mathbf{F} = [x - y, \ y - z, \ z - x]$, S the sphere of radius 5 and center 0
29. $\mathbf{F} = [y^2, \ x^2, \ z^2]$, S the surface of $x^2 + y^2 \leqq 4$, $0 \leqq z \leqq 5$
30. $\mathbf{F} = [y^3, \ x^3, \ 3z^2]$, S the portion of the paraboloid $z = x^2 + y^2$, $z \leqq 4$
31. $\mathbf{F} = [\sin^2 x, \ -y \sin 2x, \ 5z]$, S the surface of the box $|x| \leqq a$, $|y| \leqq b$, $|z| \leqq c$
32. $\mathbf{F} = [1, \ 1, \ a]$, S: $x^2 + y^2 + 4z^2 = 4$, $z \geqq 0$
33. $\mathbf{F} = [x, \ xy, \ z]$, S: $x^2 + y^2 = 1$, $0 \leqq z \leqq h$
34. $\mathbf{F}$ as in Prob. 33, S the complete boundary of $x^2 + y^2 \leqq 1$, $0 \leqq z \leqq h$
35. $\mathbf{F} = [e^y, \ 0, \ ze^x]$, S the rectangle with vertices $(0, 0, 0)$, $(1, 2, 0)$, $(0, 0, 5)$, $(1, 2, 5)$

SUMMARY OF CHAPTER 10
Vector Integral Calculus. Integral Theorems

Chapter 9 extended *differential* calculus to vectors, that is, to vector functions $\mathbf{v}(x, y, z)$ or $\mathbf{v}(t)$. Similarly, Chapter 10 extends *integral* calculus to vector functions. This involves *line integrals* (Sec. 10.1), *double integrals* (Sec. 10.3), *surface integrals* (Sec. 10.6), and *triple integrals* (Sec. 10.7) and the three "big" theorems for transforming these integrals into one another, the theorems of Green (Sec. 10.4), Gauss (Sec. 10.7), and Stokes (Sec. 10.9).

The analog of the definite integral of calculus is the **line integral** (Sec. 10.1)

$$(1) \qquad \int_C \mathbf{F}(\mathbf{r}) \cdot d\mathbf{r} = \int_C (F_1 \, dx + F_2 \, dy + F_3 \, dz) = \int_a^b \mathbf{F}(\mathbf{r}(t)) \cdot \frac{d\mathbf{r}}{dt} \, dt$$

where C: $\mathbf{r}(t) = [x(t), y(t), z(t)] = x(t)\mathbf{i} + y(t)\mathbf{j} + z(t)\mathbf{k}$ $(a \leqq t \leqq b)$ is a curve in space (or in the plane). Physically, (1) may represent the work done by a (variable) force in a displacement. Other kinds of line integrals and their applications are also discussed in Sec. 10.1.

Independence of path of a line integral in a domain D means that the integral of a given function over any path C with endpoints P and Q has the same value for all paths from P to Q that lie in D; here P and Q are fixed. An integral (1) is independent of path in D if and only if the differential form $F_1\,dx + F_2\,dy + F_3\,dz$ with continuous F_1, F_2, F_3 is **exact** in D (Sec. 10.2). Also, if curl $\mathbf{F} = \mathbf{0}$, where $\mathbf{F} = [F_1, F_2, F_3]$, has continuous first partial derivatives in a *simply connected* domain D, then the integral (1) is independent of path in D (Sec. 10.2).

Integral Theorems. The formula of **Green's theorem in the plane** (Sec. 10.4)

$$\text{(2)} \qquad \iint_R \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} \right) dx\,dy = \oint_C (F_1\,dx + F_2\,dy)$$

transforms **double integrals** over a region R in the xy-plane into line integrals over the boundary curve C of R and conversely. For other forms of (2) see Sec. 10.4.

Similarly, the formula of the **divergence theorem of Gauss** (Sec. 10.7)

$$\text{(3)} \qquad \iiint_T \operatorname{div} \mathbf{F}\,dV = \iint_S \mathbf{F} \cdot \mathbf{n}\,dA$$

transforms **triple integrals** over a region T in space into surface integrals over the boundary surface S of T, and conversely. Formula (3) implies **Green's formulas**

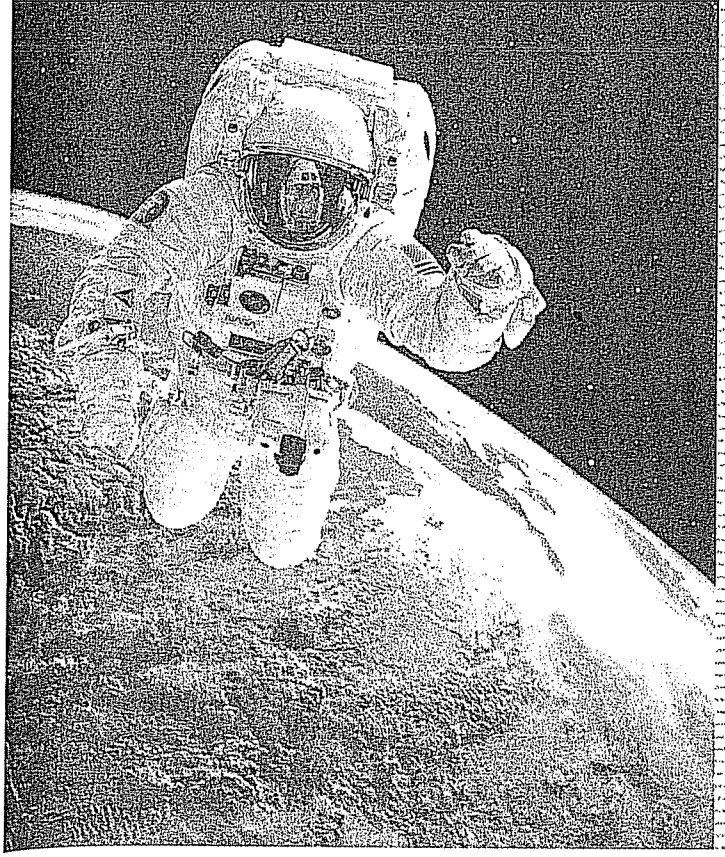
$$\text{(4)} \qquad \iiint_T (f \nabla^2 g + \nabla f \cdot \nabla g)\,dV = \iint_S f \frac{\partial g}{\partial n}\,dA,$$

$$\text{(5)} \qquad \iiint_T (f \nabla^2 g - g \nabla^2 f)\,dV = \iint_S \left(f \frac{\partial g}{\partial n} - g \frac{\partial f}{\partial n} \right) dA.$$

Finally, the formula of **Stokes's theorem** (Sec. 10.9)

$$\text{(6)} \qquad \iint_S (\operatorname{curl} \mathbf{F}) \cdot \mathbf{n}\,dA = \oint_C \mathbf{F} \cdot \mathbf{r}'(s)\,ds$$

transforms **surface integrals** over a surface S into line integrals over the boundary curve C of S and conversely.



PART C Fourier Analysis. Partial Differential Equations (PDEs)

CHAPTER 11 **Fourier Series, Integrals, and Transforms**

CHAPTER 12 **Partial Differential Equations (PDEs)**

Fourier analysis concerns **periodic phenomena,** as they occur quite frequently in engineering and elsewhere—think of rotating parts of machines, alternating electric currents, or the motion of planets. Related periodic functions may be complicated. This situation poses the important practical task of representing these complicated functions in terms of simple periodic functions, namely, cosines and sines. These representations will be infinite series, called **Fourier series.**[1]

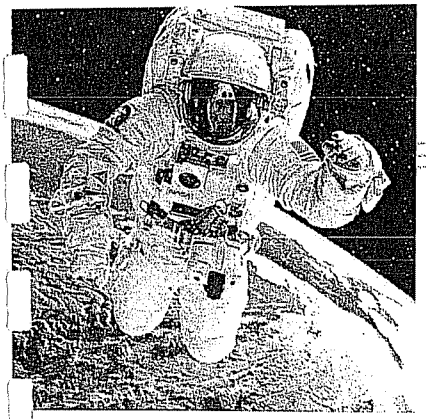
The creation of these series was one of the most path-breaking events in applied mathematics, and we mention that it also had considerable influence on mathematics as a whole, on the concept of a function, on integration theory, on convergence theory for series, and so on (see Ref. [GR7] in App. 1).

Chapter 11 is concerned mainly with Fourier series. However, the underlying ideas can also be extended to *nonperiodic* phenomena. This leads to *Fourier integrals* and *transforms.* A common name for the whole area is **Fourier analysis.**

Chapter 12 deals with the most important **partial differential equations (PDEs)** of physics and engineering. This is the area in which Fourier analysis has its most basic applications, related to boundary and initial value problems of mechanics, heat flow, electrostatics, and other fields.

---

[1]JEAN-BAPTISTE JOSEPH FOURIER (1768–1830), French physicist and mathematician, lived and taught in Paris, accompanied Napoléon in the Egyptian War, and was later made prefect of Grenoble. The beginnings on Fourier series can be found in works by Euler and by Daniel Bernoulli, but it was Fourier who employed them in a systematic and general manner in his main work, *Théorie analytique de la chaleur* (*Analytic Theory of Heat,* Paris, 1822), in which he developed the theory of heat conduction (heat equation; see Sec. 12.5), making these series a most important tool in applied mathematics.



CHAPTER 11

Fourier Series, Integrals, and Transforms

Fourier series (Sec. 11.1) are infinite series designed to represent general periodic functions in terms of simple ones, namely, cosines and sines. They constitute a very important tool, in particular in solving problems that involve ODEs and PDEs.

In this chapter we discuss Fourier series and their engineering use from a practical point of view, in connection with ODEs and with the approximation of periodic functions. Application to PDEs follows in Chap. 12.

The *theory* of Fourier series is complicated, but we shall see that the *application* of these series is rather simple. Fourier series are in a certain sense more universal than the familiar Taylor series in calculus because many *discontinuous* periodic functions of practical interest can be developed in Fourier series but, of course, do not have Taylor series representations.

In the last sections (11.7–11.9) we consider **Fourier integrals** and **Fourier transforms,** which extend the ideas and techniques of Fourier series to nonperiodic functions and have basic applications to PDEs (to be shown in the next chapter).

Prerequisite: Elementary integral calculus (needed for Fourier coefficients)
Sections that may be omitted in a shorter course: 11.4–11.9
References and Answers to Problems: App. 1 Part C, App. 2.

11.1 Fourier Series

Fourier series are the basic tool for representing periodic functions, which play an important role in applications. A function $f(x)$ is called a **periodic function** if $f(x)$ is defined for all real x (perhaps except at some points, such as $x = \pm\pi/2, \pm 3\pi/2, \cdots$ for $\tan x$) and if there is some positive number p, called a **period** of $f(x)$, such that

$$f(x + p) = f(x) \qquad \text{for all } x. \tag{1}$$

The graph of such a function is obtained by periodic repetition of its graph in any interval of length p (Fig. 255).

Familiar periodic functions are the cosine and sine functions. Examples of functions that are not periodic are x, x^2, x^3, e^x, $\cosh x$, and $\ln x$, to mention just a few.

If $f(x)$ has period p, it also has the period $2p$ because (1) implies $f(x + 2p) = f([x + p] + p) = f(x + p) = f(x)$, etc.; thus for any integer $n = 1, 2, 3, \cdots$,

$$f(x + np) = f(x) \qquad \text{for all } x. \tag{2}$$

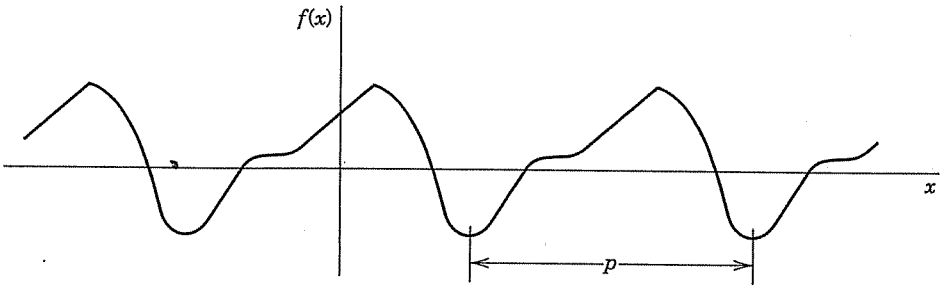


Fig. 255. Periodic function

Furthermore if $f(x)$ and $g(x)$ have period p, then $af(x) + bg(x)$ with any constants a and b also has the period p.

Our problem in the first few sections of this chapter will be the representation of various ***functions f(x) of period 2π*** in terms of the simple functions

$$(3) \quad 1, \quad \cos x, \quad \sin x, \quad \cos 2x, \quad \sin 2x, \cdots, \quad \cos nx, \quad \sin nx, \cdots.$$

All these functions have the period 2π. They form the so-called **trigonometric system.** Figure 256 shows the first few of them (except for the constant 1, which is periodic with any period).

The series to be obtained will be a **trigonometric series,** that is, a series of the form

$$(4) \quad a_0 + a_1 \cos x + b_1 \sin x + a_2 \cos 2x + b_2 \sin 2x + \cdots = a_0 + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \sin nx).$$

$a_0, a_1, b_1, a_2, b_2, \cdots$ are constants, called the **coefficients** of the series. We see that each term has the period 2π. Hence *if the coefficients are such that the series converges, its sum will be a function of period* 2π.

It can be shown that if the series on the left side of (4) converges, then inserting parentheses on the right gives a series that converges and has the same sum as the series on the left. This justifies the equality in (4).

Now suppose that $f(x)$ is a given function of period 2π and is such that it can be **represented** by a series (4), that is, (4) converges and, moreover, has the sum $f(x)$. Then, using the equality sign, we write

$$(5) \quad f(x) = a_0 + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \sin nx)$$

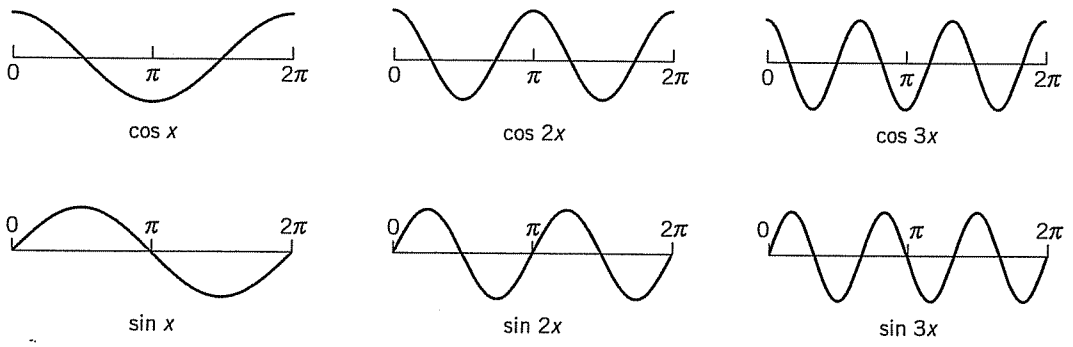


Fig. 256. Cosine and sine functions having the period 2π

and call (5) the **Fourier series** of $f(x)$. We shall prove that in this case the coefficients of (5) are the so-called **Fourier coefficients** of $f(x)$, given by the **Euler formulas**

$$(6)\quad \text{(a)}\quad a_0 = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\,dx$$

$$\text{(b)}\quad a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos nx\,dx \qquad n = 1, 2, \cdots$$

$$\text{(c)}\quad b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin nx\,dx \qquad n = 1, 2, \cdots.$$

The name "Fourier series" is sometimes also used in the exceptional case that (5) with coefficients (6) does not converge or does not have the sum $f(x)$—this may happen but is merely of theoretical interest. (For Euler see footnote 4 in Sec. 2.5.)

A Basic Example

Before we derive the Euler formulas (6), let us become familiar with the application of (5) and (6) in the case of an important example. Since your work for other functions will be quite similar, try to fully understand every detail of the integrations, which because of the n involved differ somewhat from what you have practiced in calculus. Do not just routinely use your software, but make observations: How are continuous functions (cosines and sines) able to represent a given discontinuous function? How does the quality of the approximation increase if you take more and more terms of the series? Why are the approximating functions, called the **partial sums** of the series, always zero at 0 and π? Why is the factor $1/n$ (obtained in the integration) important?

EXAMPLE 1 **Periodic Rectangular Wave (Fig. 257a)**

Find the Fourier coefficients of the periodic function $f(x)$ in Fig. 257a. The formula is

$$(7)\qquad f(x) = \begin{cases} -k & \text{if} \quad -\pi < x < 0 \\ k & \text{if} \quad 0 < x < \pi \end{cases} \qquad \text{and} \qquad f(x + 2\pi) = f(x).$$

Functions of this kind occur as external forces acting on mechanical systems, electromotive forces in electric circuits, etc. (The value of $f(x)$ at a single point does not affect the integral; hence we can leave $f(x)$ undefined at $x = 0$ and $x = \pm\pi$.)

Solution. From (6a) we obtain $a_0 = 0$. This can also be seen without integration, since the area under the curve of $f(x)$ between $-\pi$ and π is zero. From (6b),

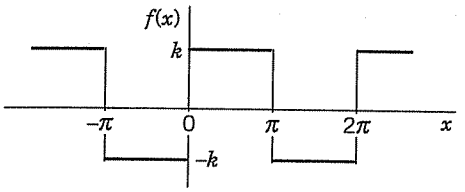
$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos nx\,dx = \frac{1}{\pi}\left[\int_{-\pi}^{0}(-k)\cos nx\,dx + \int_{0}^{\pi} k\cos nx\,dx\right]$$

$$= \frac{1}{\pi}\left[-k\,\frac{\sin nx}{n}\Bigg|_{-\pi}^{0} + k\,\frac{\sin nx}{n}\Bigg|_{0}^{\pi}\right] = 0$$

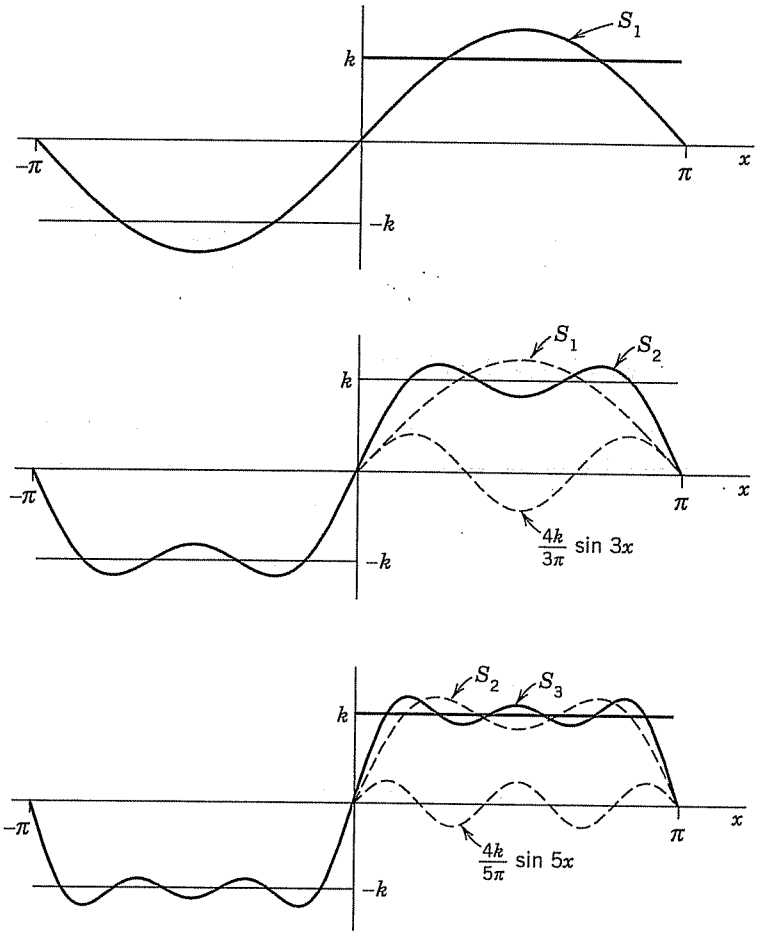
because $\sin nx = 0$ at $-\pi$, 0, and π for all $n = 1, 2, \cdots$. Similarly, from (6c) we obtain

$$b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin nx\,dx = \frac{1}{\pi}\left[\int_{-\pi}^{0}(-k)\sin nx\,dx + \int_{0}^{\pi} k\sin nx\,dx\right]$$

$$= \frac{1}{\pi}\left[k\,\frac{\cos nx}{n}\Bigg|_{-\pi}^{0} - k\,\frac{\cos nx}{n}\Bigg|_{0}^{\pi}\right].$$



(a) The given function $f(x)$ (Periodic rectangular wave)



(b) The first three partial sums of the corresponding Fourier series

Fig. 257. Eample 1

Since $\cos(-\alpha) = \cos\alpha$ and $\cos 0 = 1$, this yields

$$b_n = \frac{k}{n\pi}\left[\cos 0 - \cos(-n\pi) - \cos n\pi + \cos 0\right] = \frac{2k}{n\pi}(1 - \cos n\pi).$$

Now, $\cos\pi = -1$, $\cos 2\pi = 1$, $\cos 3\pi = -1$, etc.; in general,

$$\cos n\pi = \begin{cases} -1 & \text{for odd } n, \\ 1 & \text{for even } n, \end{cases} \qquad \text{and thus} \qquad 1 - \cos n\pi = \begin{cases} 2 & \text{for odd } n, \\ 0 & \text{for even } n. \end{cases}$$

Hence the Fourier coefficients b_n of our function are

$$b_1 = \frac{4k}{\pi}, \qquad b_2 = 0, \qquad b_3 = \frac{4k}{3\pi}, \qquad b_4 = 0, \qquad b_5 = \frac{4k}{5\pi}, \cdots.$$

Since the a_n are zero, the Fourier series of $f(x)$ is

$$\frac{4k}{\pi}\left(\sin x + \frac{1}{3}\sin 3x + \frac{1}{5}\sin 5x + \cdots\right). \tag{8}$$

The partial sums are

$$S_1 = \frac{4k}{\pi}\sin x, \qquad S_2 = \frac{4k}{\pi}\left(\sin x + \frac{1}{3}\sin 3x\right), \qquad \text{etc.,}$$

Their graphs in Fig. 257 seem to indicate that the series is convergent and has the sum $f(x)$, the given function. We notice that at $x = 0$ and $x = \pi$, the points of discontinuity of $f(x)$, all partial sums have the value zero, the arithmetic mean of the limits $-k$ and k of our function, at these points.

Furthermore, assuming that $f(x)$ is the sum of the series and setting $x = \pi/2$, we have

$$f\left(\frac{\pi}{2}\right) = k = \frac{4k}{\pi}\left(1 - \frac{1}{3} + \frac{1}{5} - + \cdots\right).$$

thus

$$1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + - \cdots = \frac{\pi}{4}.$$

This is a famous result obtained by Leibniz in 1673 from geometric considerations. It illustrates that the values of various series with constant terms can be obtained by evaluating Fourier series at specific points. ▩

Derivation of the Euler Formulas (6)

The key to the Euler formulas (6) is the **orthogonality** of (3), a concept of basic importance, as follows.

THEOREM 1

Orthogonality of the Trigonometric System (3)

The trigonometric system (3) *is orthogonal on the interval* $-\pi \leqq x \leqq \pi$ (hence also on $0 \leqq x \leqq 2\pi$ or any other interval of length 2π because of periodicity); *that is, the integral of the product of any two functions in* (3) *over that interval is* 0, *so that for any integers n and m,*

$$\textbf{(a)} \qquad \int_{-\pi}^{\pi} \cos nx \cos mx\, dx = 0 \qquad (n \neq m)$$

$$\textbf{(b)} \qquad \int_{-\pi}^{\pi} \sin nx \sin mx\, dx = 0 \qquad (n \neq m) \tag{9}$$

$$\textbf{(c)} \qquad \int_{-\pi}^{\pi} \sin nx \cos mx\, dx = 0 \qquad (n \neq m \text{ or } n = m).$$

PROOF This follows simply by transforming the integrands trigonometrically from products into sums. In (9a) and (9b), by (11) in App. A3.1,

$$\int_{-\pi}^{\pi} \cos nx \cos mx\, dx = \frac{1}{2}\int_{-\pi}^{\pi} \cos (n+m)x\, dx + \frac{1}{2}\int_{-\pi}^{\pi} \cos (n-m)x\, dx$$

$$\int_{-\pi}^{\pi} \sin nx \sin mx\, dx = \frac{1}{2}\int_{-\pi}^{\pi} \cos (n-m)x\, dx - \frac{1}{2}\int_{-\pi}^{\pi} \cos (n+m)x\, dx.$$

Since $m \neq n$ (integer!), the integrals on the right are all 0. Similarly, in (9c), for all integer m and n (without exception; do you see why?)

$$\int_{-\pi}^{\pi} \sin nx \cos mx \, dx = \frac{1}{2} \int_{-\pi}^{\pi} \sin (n + m)x \, dx + \frac{1}{2} \int_{-\pi}^{\pi} \sin (n - m)x \, dx = 0 + 0. \quad \blacksquare$$

Application of Theorem 1 to the Fourier Series (5)

We prove (6a). Integrating on both sides of (5) from $-\pi$ to π, we get

$$\int_{-\pi}^{\pi} f(x) \, dx = \int_{-\pi}^{\pi} \left[a_0 + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \sin nx) \right] dx.$$

We now assume that termwise integration is allowed. (We shall say in the proof of Theorem 2 when this is true.) Then we obtain

$$\int_{-\pi}^{\pi} f(x) \, dx = a_0 \int_{-\pi}^{\pi} dx + \sum_{n=1}^{\infty} \left(a_n \int_{-\pi}^{\pi} \cos nx \, dx + b_n \int_{-\pi}^{\pi} \sin nx \, dx \right).$$

The first term on the right equals $2\pi a_0$. Integration shows that all the other integrals are 0. Hence division by 2π gives (6a).

We prove (6b). Multiplying (5) on both sides by $\cos mx$ with any ***fixed*** positive integer m and integrating from $-\pi$ to π, we have

$$(10) \quad \int_{-\pi}^{\pi} f(x) \cos mx \, dx = \int_{-\pi}^{\pi} \left[a_0 + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \sin nx) \right] \cos mx \, dx.$$

We now integrate term by term. Then on the right we obtain an integral of $a_0 \cos mx$, which is 0; an integral of $a_n \cos nx \cos mx$, which is $a_m \pi$ for $n = m$ and 0 for $n \neq m$ by (9a); and an integral of $b_n \sin nx \cos mx$, which is 0 for all n and m by (9c). Hence the right side of (10) equals $a_m \pi$. Division by π gives (6b) (with m instead of n).

We finally prove (6c). Multiplying (5) on both sides by $\sin mx$ with any ***fixed*** positive integer m and integrating from $-\pi$ to π, we get

$$(11) \quad \int_{-\pi}^{\pi} f(x) \sin mx \, dx = \int_{-\pi}^{\pi} \left[a_0 + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \sin nx) \right] \sin mx \, dx.$$

Integrating term by term, we obtain on the right an integral of $a_0 \sin mx$, which is 0; an integral of $a_n \cos nx \sin mx$, which is 0 by (9c); and an integral of $b_n \sin nx \sin mx$, which is $b_m \pi$ if $n = m$ and 0 if $n \neq m$, by (9b). This implies (6c) (with n denoted by m). This completes the proof of the Euler formulas (6) for the Fourier coefficients. $\blacksquare$

Convergence and Sum of a Fourier Series

The class of functions that can be represented by Fourier series is surprisingly large and general. Sufficient conditions valid in most applications are as follows.

THEOREM 2

Representation by a Fourier Series

Let $f(x)$ be periodic with period 2π and piecewise continuous (see Sec. 6.1) *in the interval $-\pi \leqq x \leqq \pi$. Furthermore, let $f(x)$ have a left-hand derivative and a right-hand derivative at each point of that interval. Then the Fourier series* (5) *of $f(x)$* [with coefficients (6)] *converges. Its sum is $f(x)$, except at points x_0 where $f(x)$ is discontinuous. There the sum of the series is the average of the left- and right-hand limits*[2] *of $f(x)$ at x_0.*

PROOF We prove convergence in Theorem 2. We prove convergence for a continuous function $f(x)$ having continuous first and second derivatives. Integrating (6b) by parts, we obtain.

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos nx\,dx = \frac{f(x)\sin nx}{n\pi}\bigg|_{-\pi}^{\pi} - \frac{1}{n\pi}\int_{-\pi}^{\pi} f'(x)\sin nx\,dx.$$

The first term on the right is zero. Another integration by parts gives

$$a_n = \frac{f'(x)\cos nx}{n^2\pi}\bigg|_{-\pi}^{\pi} - \frac{1}{n^2\pi}\int_{-\pi}^{\pi} f''(x)\cos nx\,dx.$$

The first term on the right is zero because of the periodicity and continuity of $f'(x)$. Since f'' is continuous in the interval of integration, we have

$$|f''(x)| < M$$

for an appropriate constant M. Furthermore, $|\cos nx| \leqq 1$. It follows that

$$|a_n| = \frac{1}{n^2\pi}\left|\int_{-\pi}^{\pi} f''(x)\cos nx\,dx\right| < \frac{1}{n^2\pi}\int_{-\pi}^{\pi} M\,dx = \frac{2M}{n^2}.$$

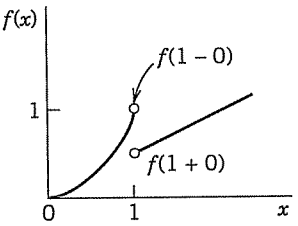


Fig. 258. Left- and right-hand limits

$f(1 - 0) = 1$,

$f(1 + 0) = \frac{1}{2}$

of the function

$$f(x) = \begin{cases} x^2 & \text{if } x < 1 \\ x/2 & \end{cases}$$

[2]The **left-hand limit** of $f(x)$ at x_0 is defined as the limit of $f(x)$ as x approaches x_0 from the left and is commonly denoted by $f(x_0 - 0)$. Thus

$$f(x_0 - 0) = \lim_{h\to 0} f(x_0 - h) \text{ as } h \to 0 \text{ through positive values.}$$

The **right-hand limit** is denoted by $f(x_0 + 0)$ and

$$f(x_0 + 0) = \lim_{h\to 0} f(x_0 + h) \text{ as } h \to 0 \text{ through positive values.}$$

The **left-** and **right-hand derivatives** of $f(x)$ at x_0 are defined as the limits of

$$\frac{f(x_0 - h) - f(x_0 - 0)}{-h} \quad \text{and} \quad \frac{f(x_0 + h) - f(x_0 + 0)}{h},$$

respectively, as $h \to 0$ through positive values. Of course if $f(x)$ is continuous at x_0, the last term in both numerators is simply $f(x_0)$.

Similarly, $|b_n| < 2M/n^2$ for all n. Hence the absolute value of each term of the Fourier series of $f(x)$ is at most equal to the corresponding term of the series

$$|a_0| + 2M\left(1 + 1 + \frac{1}{2^2} + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{3^2} + \cdots\right)$$

which is convergent. Hence that Fourier series converges and the proof is complete. (Readers already familiar with uniform convergence will see that, by the Weierstrass test in Sec. 15.5, under our present assumptions the Fourier series converges uniformly, and our derivation of (6) by integrating term by term is then justified by Theorem 3 of Sec. 15.5.)

The proof of convergence in the case of a piecewise continuous function $f(x)$ and the proof that under the assumptions in the theorem the Fourier series (5) with coefficients (6) represents $f(x)$ are substantially more complicated; see, for instance, Ref. [C12]. ■

EXAMPLE 2 Convergence at a Jump as Indicated in Theorem 2

The rectangular wave in Example 1 has a jump at $x = 0$. Its left-hand limit there is $-k$ and its right-hand limit is k (Fig. 257). Hence the average of these limits is 0. The Fourier series (8) of the wave does indeed converge to this value when $x = 0$ because then all its terms are 0. Similarly for the other jumps. This is in agreement with Theorem 2. ■

Summary. A Fourier series of a given function $f(x)$ of period 2π is a series of the form (5) with coefficients given by the Euler formulas (6). Theorem 2 gives conditions that are sufficient for this series to converge and at each x to have the value $f(x)$, except at discontinuities of $f(x)$, where the series equals the arithmetic mean of the left-hand and right-hand limits of $f(x)$ at that point.

PROBLEM SET 11.1

1. **(Calculus review)** Review integration techniques for integrals as they are likely to arise from the Euler formulas, for instance, definite integrals of $x \cos nx$, $x^2 \sin nx$, $e^{-2x} \cos nx$, etc.

2–3 FUNDAMENTAL PERIOD

The *fundamental period* is the smallest positive period. Find it for

2. $\cos x$, $\sin x$, $\cos 2x$, $\sin 2x$, $\cos \pi x$, $\sin \pi x$, $\cos 2\pi x$, $\sin 2\pi x$

3. $\cos nx$, $\sin nx$, $\cos \dfrac{2\pi x}{k}$, $\sin \dfrac{2\pi x}{k}$, $\cos \dfrac{2\pi nx}{k}$, $\sin \dfrac{2\pi nx}{k}$

4. Show that $f = const$ is periodic with any period but has no fundamental period.

5. If $f(x)$ and $g(x)$ have period p, show that $h(x) = af(x) + bg(x)$ (a, b, constant) has the period p. Thus all functions of period p form a **vector space.**

6. **(Change of scale)** If $f(x)$ has period p, show that $f(ax)$, $a \neq 0$, and $f(x/b)$, $b \neq 0$, are periodic functions of x of periods p/a and bp, respectively. Give examples.

7–12 GRAPHS OF 2π-PERIODIC FUNCTIONS

Sketch or graph $f(x)$, of period 2π, which for $-\pi < x < \pi$ is given as follows.

7. $f(x) = x$

8. $f(x) = e^{-|x|}$

9. $f(x) = \pi - |x|$

10. $f(x) = |\sin 2x|$

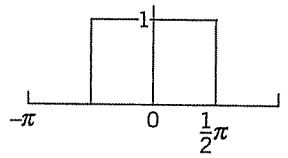
11. $f(x) = \begin{cases} -x^3 & \text{if } -\pi < x < 0 \\ x^3 & \text{if } 0 < x < \pi \end{cases}$

12. $f(x) = \begin{cases} 1 & \text{if } -\pi < x < 0 \\ \cos \frac{1}{2}x & \text{if } 0 < x < \pi \end{cases}$

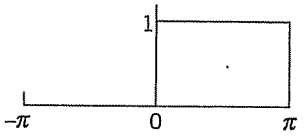
13–24 FOURIER SERIES

Showing the details of your work, find the Fourier series of the given $f(x)$, which is assumed to have the period 2π. Sketch or graph the partial sums up to that including $\cos 5x$ and $\sin 5x$.

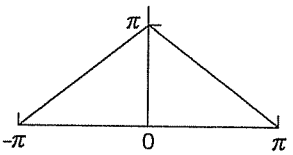
13.



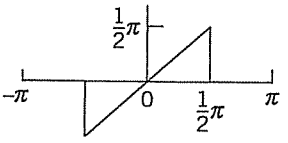
14.



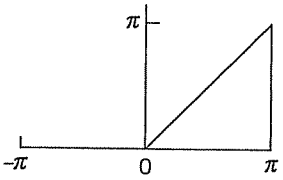
15.



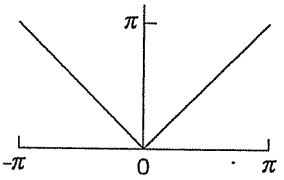
16.



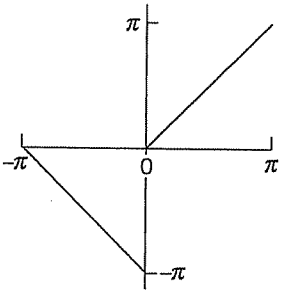
17.



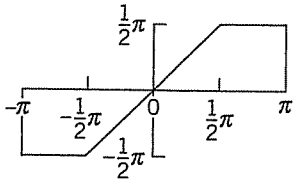
18.



19.



20.



21. $f(x) = x^2 \ (-\pi < x < \pi)$

22. $f(x) = x^2 \ (0 < x < 2\pi)$

23. $f(x) = \begin{cases} x^2 & \text{if } -\frac{1}{2}\pi < x < \frac{1}{2}\pi \\ \frac{1}{4}\pi^2 & \text{if } \frac{1}{2}\pi < x < \frac{3}{2}\pi \end{cases}$

24. $f(x) = \begin{cases} -4x & \text{if } -\pi < x < 0 \\ 4x & \text{if } 0 < x < \pi \end{cases}$

25. (Discontinuities) Verify the last statement in Theorem 2 for the discontinuities of $f(x)$ in Prob. 13.

26. CAS EXPERIMENT. Graphing. Write a program for graphing partial sums of the following series. Guess from the graph what $f(x)$ the series may represent. Confirm or disprove your guess by using the Euler formulas.

(a) $2(\sin x + \frac{1}{3}\sin 3x + \frac{1}{5}\sin 5x + \cdots) - 2(\frac{1}{2}\sin 2x + \frac{1}{4}\sin 4x + \frac{1}{6}\sin 6x \cdots)$

(b) $\frac{1}{2} + \dfrac{4}{\pi^2}(\cos x + \frac{1}{9}\cos 3x + \frac{1}{25}\cos 5x + \cdots)$

(c) $\frac{2}{3}\pi^2 + 4(\cos x - \frac{1}{4}\cos 2x + \frac{1}{9}\cos 3x - \frac{1}{16}\cos 4x + - \cdots)$

27. CAS EXPERIMENT. Order of Fourier Coefficients. The order seems to be $1/n$ if f is discontinous, and $1/n^2$ if f is continuous but $f' = df/dx$ is discontinuous, $1/n^3$ if f and f' are continuous but f'' is discontinuous, etc. Try to verify this for examples. Try to prove it by integrating the Euler formulas by parts. What is the practical significance of this?

28. PROJECT. Euler Formulas in Terms of Jumps Without Integration. Show that for a function whose third derivative is identically zero,

$$a_n = \frac{1}{n\pi}\left[-\sum j_s \sin nx_s - \frac{1}{n}\sum j_s' \cos nx_s + \frac{1}{n^2}\sum j_s'' \sin nx_s\right]$$

$$b_n = \frac{1}{n\pi}\left[\sum j_s \cos nx_s - \frac{1}{n}\sum j_s' \sin nx_s - \frac{1}{n^2}\sum j_s'' \cos nx_s\right]$$

where $n = 1, 2, \cdots$ and we sum over all the jumps j_s, j_s', j_s'' of f, f', f', respectively, located at x_s.

29. Apply the formulas in Project 28 to the function in Prob. 21 and compare the results.

30. CAS EXPERIMENT. Orthogonality. Integrate and graph the integral of the product $\cos mx \cos nx$ (with various integer m and n of your choice) from $-a$ to a as a function of a and conclude orthogonality of $\cos mx$ and $\cos nx$ ($m \neq n$) for $a = \pi$ from the graph. For what m and n will you get orthogonality for $a = \pi/2$, $\pi/3$, $\pi/4$? Other a? Extend the experiment to $\cos mx \sin nx$ and $\sin mx \sin nx$.

11.2 Functions of Any Period $p = 2L$

The functions considered so far had period 2π, for the simplicity of the formulas. Of course, periodic functions in applications will generally have other periods. However, we now show that the transition from period $p = 2\pi$ to a period $2L$ is quite simple. The notation $p = 2L$ is practical because L will be the length of a violin string (Sec. 12.2) or the length of a rod in heat conduction (Sec. 12.5), and so on.

The idea is simply to find and use a ***change of scale*** that gives from a function $g(v)$ of period 2π a function of period $2L$. Now from (5) and (6) in the last section with $g(v)$ instead of $f(x)$ we have the Fourier series

$$g(v) = a_0 + \sum_{n=1}^{\infty} (a_n \cos nv + b_n \sin nv) \tag{1}$$

with coefficients

$$a_0 = \frac{1}{2\pi} \int_{-\pi}^{\pi} g(v)\, dv$$

$$a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} g(v) \cos nv\, dv \tag{2}$$

$$b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} g(v) \sin nv\, dv.$$

We can now write the change of scale as $v = kx$ with k such that the old period $v = 2\pi$ gives for the new variable x the new period $x = 2L$. Thus, $2\pi = k2L$. Hence $k = \pi/L$ and

$$v = kx = \pi x/L. \tag{3}$$

This implies $dv = (\pi/L)\, dx$, which upon substitution into (2) cancels $1/2\pi$ and $1/\pi$ and gives instead the factors $1/2L$ and $1/L$. Writing

$$g(v) = f(x), \tag{4}$$

we thus obtain from (1) the **Fourier series** of the function $f(x)$ of period $2L$

$$f(x) = a_0 + \sum_{n=1}^{\infty} \left(a_n \cos \frac{n\pi}{L} x + b_n \sin \frac{n\pi}{L} x \right) \tag{5}$$

with the **Fourier coefficients** of $f(x)$ given by the **Euler formulas**

$$\text{(a)} \quad a_0 = \frac{1}{2L} \int_{-L}^{L} f(x)\, dx$$

$$\text{(b)} \quad a_n = \frac{1}{L} \int_{-L}^{L} f(x) \cos \frac{n\pi x}{L}\, dx \qquad n = 1, 2, \cdots \tag{6}$$

$$\text{(c)} \quad b_n = \frac{1}{L} \int_{-L}^{L} f(x) \sin \frac{n\pi x}{L}\, dx \qquad n = 1, 2, \cdots.$$

Just as in Sec. 11.1, we continue to call (5) with any coefficients a **trigonometric series.** And we can integrate from 0 to $2L$ or over any other interval of length $p = 2L$.

EXAMPLE 1 **Periodic Rectangular Wave**

Find the Fourier series of the function (Fig. 259)

$$f(x) = \begin{cases} 0 & \text{if } -2 < x < -1 \\ k & \text{if } -1 < x < 1 \\ 0 & \text{if } 1 < x < 2 \end{cases} \qquad p = 2L = 4, \quad L = 2.$$

Solution. From (6a) we obtain $a_0 = k/2$ (verify!). From (6b) we obtain

$$a_n = \frac{1}{2}\int_{-2}^{2} f(x)\cos\frac{n\pi x}{2}\,dx = \frac{1}{2}\int_{-1}^{1} k\cos\frac{n\pi x}{2}\,dx = \frac{2k}{n\pi}\sin\frac{n\pi}{2}.$$

Thus $a_n = 0$ if n is even and

$$a_n = 2k/n\pi \quad \text{if} \quad n = 1, 5, 9, \cdots, \qquad a_n = -2k/n\pi \quad \text{if} \quad n = 3, 7, 11, \cdots.$$

From (6c) we find that $b_n = 0$ for $n = 1, 2, \cdots$. Hence the Fourier series is

$$f(x) = \frac{k}{2} + \frac{2k}{\pi}\left(\cos\frac{\pi}{2}x - \frac{1}{3}\cos\frac{3\pi}{2}x + \frac{1}{5}\cos\frac{5\pi}{2}x - + \cdots\right).$$

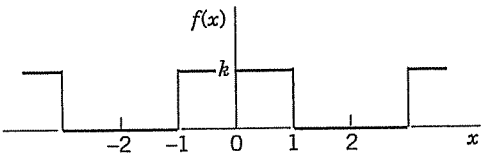


Fig. 259. Example 1

EXAMPLE 2 **Periodic Rectangular Wave**

Find the Fourier series of the function (Fig. 260)

$$f(x) = \begin{cases} -k & \text{if } -2 < x < 0 \\ k & \text{if } 0 < x < 2 \end{cases} \qquad p = 2L = 4, \qquad L = 2.$$

Solution. $a_0 = 0$ from (6a). From (6b), with $1/L = 1/2$,

$$a_n = \frac{1}{2}\left[\int_{-2}^{0}(-k)\cos\frac{n\pi x}{2}\,dx + \int_{0}^{2} k\cos\frac{n\pi x}{2}\,dx\right]$$

$$= \frac{1}{2}\left[-\frac{2k}{n\pi}\sin\frac{n\pi x}{2}\Big|_{-2}^{0} + \frac{2k}{n\pi}\sin\frac{n\pi x}{2}\Big|_{0}^{2}\right] = 0,$$

so that the Fourier series has no cosine terms. From (6c),

$$b_n = \frac{1}{2}\left[\frac{2k}{n\pi}\cos\frac{n\pi x}{2}\Big|_{-2}^{0} - \frac{2k}{n\pi}\cos\frac{n\pi x}{2}\Big|_{0}^{2}\right]$$

$$= \frac{k}{n\pi}(1 - \cos n\pi - \cos n\pi + 1) = \begin{cases} 4k/n\pi & \text{if } n = 1, 3, \cdots \\ 0 & \text{if } n = 2, 4, \cdots. \end{cases}$$

Hence the Fourier series of $f(x)$ is

$$f(x) = \frac{4k}{\pi}\left(\sin\frac{\pi}{2}x + \frac{1}{3}\sin\frac{3\pi}{2}x + \frac{1}{5}\sin\frac{5\pi}{2}x + \cdots\right).$$

It is interesting that we could have derived this from (8) in Sec. 11.1, namely, by the scale change (3). Indeed, writing v instead of x, we have in (8), Sec. 11.1,

$$\frac{4k}{\pi}\left(\sin v + \frac{1}{3}\sin 3v + \frac{1}{5}\sin 5v + \cdots\right).$$

Since the period 2π in v corresponds to $2L = 4$, we have $k = \pi/L = \pi/2$ and $v = kx = \pi x/2$ in (3); hence we obtain the Fourier series of $f(x)$, as before. ■

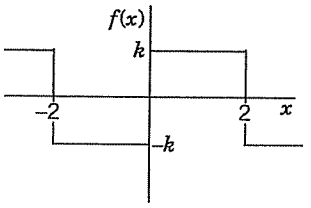


Fig. 260. Example 2

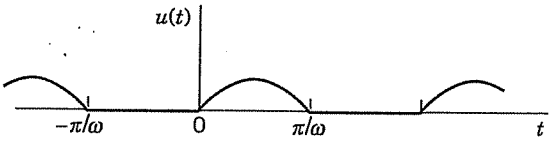


Fig. 261. Half-wave rectifier

EXAMPLE 3 Half-Wave Rectifier

A sinusoidal voltage $E \sin \omega t$, where t is time, is passed through a half-wave rectifier that clips the negative portion of the wave (Fig. 261). Find the Fourier series of the resulting periodic function

$$u(t) = \begin{cases} 0 & \text{if } -L < t < 0, \\ E\sin\omega t & \text{if } 0 < t < L \end{cases} \qquad p = 2L = \frac{2\pi}{\omega}, \qquad L = \frac{\pi}{\omega}.$$

Solution. Since $u = 0$ when $-L < t < 0$, we obtain from (6a), with t instead of x,

$$a_0 = \frac{\omega}{2\pi}\int_0^{\pi/\omega} E\sin\omega t\, dt = \frac{E}{\pi}$$

and from (6b), by using formula (11) in App. A3.1 with $x = \omega t$ and $y = n\omega t$,

$$a_n = \frac{\omega}{\pi}\int_0^{\pi/\omega} E\sin\omega t\cos n\omega t\, dt = \frac{\omega E}{2\pi}\int_0^{\pi/\omega}[\sin(1+n)\omega t + \sin(1-n)\omega t]\, dt.$$

If $n = 1$, the integral on the right is zero, and if $n = 2, 3, \cdots$, we readily obtain

$$a_n = \frac{\omega E}{2\pi}\left[-\frac{\cos(1+n)\omega t}{(1+n)\omega} - \frac{\cos(1-n)\omega t}{(1-n)\omega}\right]_0^{\pi/\omega}$$

$$= \frac{E}{2\pi}\left(\frac{-\cos(1+n)\pi + 1}{1+n} + \frac{-\cos(1-n)\pi + 1}{1-n}\right).$$

If n is odd, this is equal to zero, and for even n we have

$$a_n = \frac{E}{2\pi}\left(\frac{2}{1+n} + \frac{2}{1-n}\right) = -\frac{2E}{(n-1)(n+1)\pi} \qquad (n = 2, 4, \cdots).$$

In a similar fashion we find from (6c) that $b_1 = E/2$ and $b_n = 0$ for $n = 2, 3, \cdots$. Consequently,

$$u(t) = \frac{E}{\pi} + \frac{E}{2}\sin\omega t - \frac{2E}{\pi}\left(\frac{1}{1\cdot 3}\cos 2\omega t + \frac{1}{3\cdot 5}\cos 4\omega t + \cdots\right).$$ ■

PROBLEM SET 11.2

1–11 FOURIER SERIES FOR PERIOD $p = 2L$

Find the Fourier series of the function $f(x)$, of period $p = 2L$, and sketch or graph the first three partial sums. (Show the details of your work.)

1. $f(x) = -1\ (-2 < x < 0),\ f(x) = 1\ (0 < x < 2),\ p = 4$
2. $f(x) = 0\ (-2 < x < 0),\ f(x) = 4\ (0 < x < 2),\ p = 4$
3. $f(x) = x^2 \quad (-1 < x < 1), \quad p = 2$
4. $f(x) = \pi x^3/2 \quad (-1 < x < 1), \quad p = 2$
5. $f(x) = \sin \pi x \quad (0 < x < 1), \quad p = 1$
6. $f(x) = \cos \pi x \quad (-\frac{1}{2} < x < \frac{1}{2}), \quad p = 1$
7. $f(x) = |x| \quad (-1 < x < 1), \quad p = 2$
8. $f(x) = \begin{cases} 1 + x \text{ if } -1 < x < 0 \\ 1 - x \text{ if } \quad 0 < x < 1, \end{cases} \quad p = 2$
9. $f(x) = 1 - x^2 \quad (-1 < x < 1), \quad p = 2$
10. $f(x) = 0\ (-2 < x < 0),\ f(x) = x\ (0 < x < 2),\ p = 4$
11. $f(x) = -x \quad (-1 < x < 0), \quad f(x) = x \quad (0 < x < 1),$ $f(x) = 1 \quad (1 < x < 3), \quad p = 4$

12. **(Rectifier)** Find the Fourier series of the function obtained by passing the voltage $v(t) = V_0 \cos 100\pi t$ through a half-wave rectifier.

13. Show that the familiar identities $\cos^3 x = \frac{3}{4} \cos x + \frac{1}{4} \cos 3x$ and $\sin^3 x = \frac{3}{4} \sin x - \frac{1}{4} \sin 3x$ can be interpreted as Fourier series expansions. Develop $\cos^4 x$.

14. Obtain the series in Prob. 7 from that in Prob. 8.

15. Obtain the series in Prob. 6 from that in Prob. 5.

16. Obtain the series in Prob. 3 from that in Prob. 21 of Problem Set 11.1.

17. Using Prob. 3, show that $1 - \frac{1}{4} + \frac{1}{9} - \frac{1}{16} + - \cdots = \frac{1}{12}\pi^2$.

18. Show that $1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \cdots = \frac{1}{6}\pi^2$.

19. **CAS PROJECT. Fourier Series of $2L$-Periodic Functions. (a)** Write a program for obtaining partial sums of a Fourier series (1).

(b) Apply the program to Probs. 2–5, graphing the first few partial sums of each of the four series on common axes. Choose the first five or more partial sums until they approximate the given function reasonably well. Compare and comment.

20. **CAS EXPERIMENT. Gibbs Phenomenon.** The partial sums $s_n(x)$ of a Fourier series show oscillations near a discontinuity point. These oscillations do not disappear as n increases but instead become sharp "spikes." They were explained mathematically by J. W. Gibbs[3]. Graph $s_n(x)$ in Prob. 10. When $n = 50$, say, you will see those oscillations quite distinctly. Consider other Fourier series of your choice in a similar way. Compare.

11.3 Even and Odd Functions. Half-Range Expansions

The function in Example 1, Sec. 11.2, is *even,* and its Fourier series has only *cosine* terms. The function in Example 2, Sec. 11.2, is *odd*, and its Fourier series has only *sine* terms.

Recall that g is **even** if $g(-x) = g(x)$, so that its graph is symmetric with respect to the vertical axis (Fig. 262). A function h is **odd** if $h(-x) = -h(x)$ (Fig. 263).

Now the cosine terms in the Fourier series (5), Sec. 11.2, are even and the sine terms are odd. So it should not be a surprise that an even function is given by a series of cosine terms and an odd function by a series of sine terms. Indeed, the following holds.

[3]JOSIAH WILLARD GIBBS (1839–1903), American mathematician, professor of mathematical physics at Yale from 1871 on, one of the founders of vector calculus [another being O. Heaviside (see Sec. 6.1)], mathematical thermodynamics, and statistical mechanics. His work was of great importance to the development of mathematical physics.

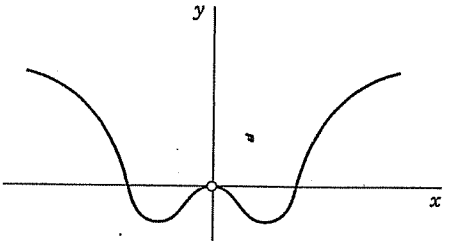


Fig. 262. Even function

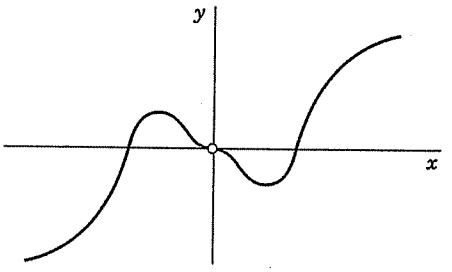


Fig. 263. Odd function

THEOREM 1

Fourier Cosine Series, Fourier Sine Series

The Fourier series of an ***even*** *function of period* $2L$ *is a* **"Fourier cosine series"**

(1) $$f(x) = a_0 + \sum_{n=1}^{\infty} a_n \cos \frac{n\pi}{L} x \qquad (f \text{ even})$$

with coefficients (note: integration from 0 to L only!)

(2) $$a_0 = \frac{1}{L} \int_0^L f(x)\, dx, \qquad a_n = \frac{2}{L} \int_0^L f(x) \cos \frac{n\pi x}{L}\, dx, \qquad n = 1, 2, \cdots.$$

The Fourier series of an ***odd*** *function of period* $2L$ *is a* **"Fourier sine series"**

(3) $$f(x) = \sum_{n=1}^{\infty} b_n \sin \frac{n\pi}{L} x \qquad (f \text{ odd})$$

with coefficients

(4) $$b_n = \frac{2}{L} \int_0^L f(x) \sin \frac{n\pi x}{L}\, dx.$$

PROOF Since the definite integral of a function gives the area under the curve of the function between the limits of integration, we have

$$\int_{-L}^{L} g(x)\, dx = 2 \int_0^L g(x)\, dx \qquad \text{for even } g$$

$$\int_{-L}^{L} h(x)\, dx = 0 \qquad \text{for odd } h$$

as is obvious from the graphs of g and h. (Give a formal proof.) Now let f be even. Then (6a), Sec. 11.2, gives a_0 in (2). Also, the integrand in (6b), Sec. 11.2, is even (a product of even functions is even), so that (6b) gives a_n in (2). Furthermore, the integrand in (6c), Sec. 11.2, is the even f times the odd sine, so that the integrand (the product) is odd, the integral is zero, and there are no sine terms in (1).

Similarly, if f is odd, the integrals for a_0 and a_n in (6a) and (6b), Sec. 11.2, are zero, f times the sine in (6c) is even, (6c) implies (4), and there are no cosine terms in (3). ■

The Case of Period 2π. If $L = \pi$, then $f(x) = a_0 + \sum_{n=1}^{\infty} a_n \cos nx$ (f even) with coefficients

$$(2^*) \qquad a_0 = \frac{1}{\pi}\int_0^{\pi} f(x)\,dx, \qquad a_n = \frac{2}{\pi}\int_0^{\pi} f(x)\cos nx\,dx, \qquad n = 1, 2, \cdots$$

and $f(x) = \sum_{n=1}^{\infty} b_n \sin nx$ (f odd) with coefficients

$$(4^*) \qquad b_n = \frac{2}{\pi}\int_0^{\pi} f(x)\sin nx\,dx, \qquad n = 1, 2, \cdots.$$

For instance, $f(x)$ in Example 1, Sec. 11.1, is odd and is represented by a Fourier sine series.

Further simplifications result from the following property, whose very simple proof is left to the student.

THEOREM 2

Sum and Scalar Multiple

The Fourier coefficients of a sum $f_1 + f_2$ are the sums of the corresponding Fourier coefficients of f_1 and f_2.

The Fourier coefficients of cf are c times the corresponding Fourier coefficients of f.

EXAMPLE 1 **Rectangular Pulse**

The function $f^*(x)$ in Fig. 264 is the sum of the function $f(x)$ in Example 1 of Sec 11.1 and the constant k. Hence, from that example and Theorem 2 we conclude that

$$f^*(x) = k + \frac{4k}{\pi}\left(\sin x + \frac{1}{3}\sin 3x + \frac{1}{5}\sin 5x + \cdots\right). \qquad ■$$

EXAMPLE 2 **Half-Wave Rectifier**

The function $u(t)$ in Example 3 of Sec. 11.2 has a Fourier cosine series plus a single term $v(t) = (E/2)\sin \omega t$. We conclude from this and Theorem 2 that $u(t) - v(t)$ must be an even function. Verify this graphically. (See Fig. 265.) ■

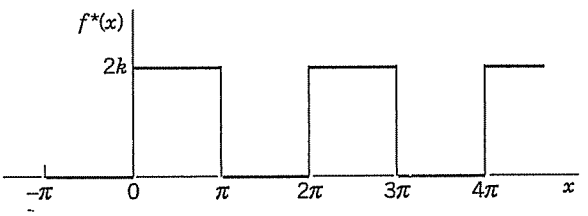


Fig. 264. Example 1

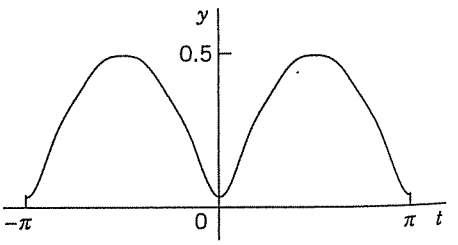
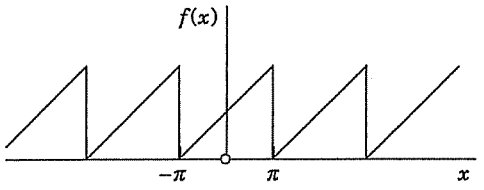


Fig. 265. $u(t) - v(t)$ with $E = 1$, $\omega = 1$

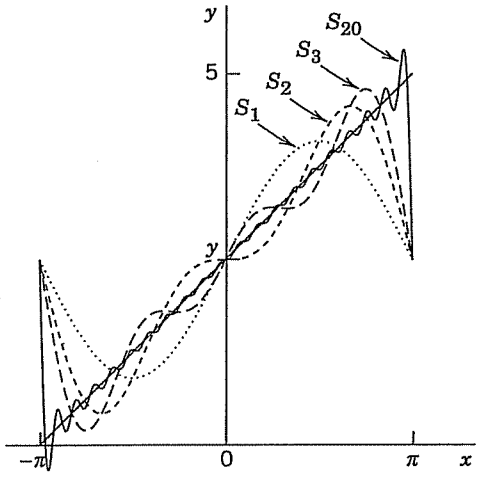
EXAMPLE 3 Sawtooth Wave

Find the Fourier series of the function (Fig. 266)

$$f(x) = x + \pi \quad \text{if} \quad -\pi < x < \pi \qquad \text{and} \qquad f(x + 2\pi) = f(x).$$



(a) The function $f(x)$



(b) Partial sums S_1, S_2, S_3, S_{20}

Fig. 266. Example 3

Solution. We have $f = f_1 + f_2$, where $f_1 = x$ and $f_2 = \pi$. The Fourier coefficients of f_2 are zero, except for the first one (the constant term), which is π. Hence, by Theorem 2, the Fourier coefficients a_n, b_n are those of f_1, except for a_0, which is π. Since f_1 is odd, $a_n = 0$ for $n = 1, 2, \cdots$, and

$$b_n = \frac{2}{\pi}\int_0^{\pi} f_1(x) \sin nx\, dx = \frac{2}{\pi}\int_0^{\pi} x \sin nx\, dx.$$

Integrating by parts, we obtain

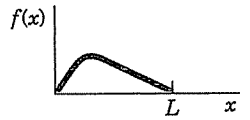
$$b_n = \frac{2}{\pi}\left[\frac{-x\cos nx}{n}\bigg|_0^{\pi} + \frac{1}{n}\int_0^{\pi}\cos nx\, dx\right] = -\frac{2}{n}\cos n\pi.$$

Hence $b_1 = 2$, $b_2 = -2/2$, $b_3 = 2/3$, $b_4 = -2/4$, $\cdots$, and the Fourier series of $f(x)$ is

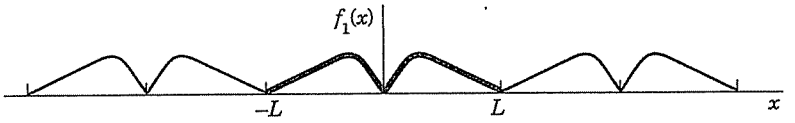
$$f(x) = \pi + 2\left(\sin x - \frac{1}{2}\sin 2x + \frac{1}{3}\sin 3x - + \cdots\right). \qquad \blacksquare$$

Half-Range Expansions

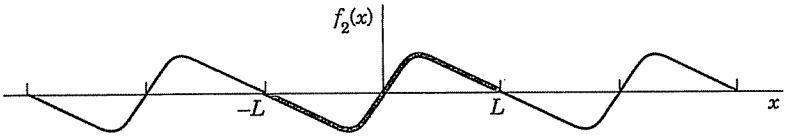
Half-range expansions are Fourier series. The idea is simple and useful. Figure 267 explains it. We want to represent $f(x)$ in Fig. 267a by a Fourier series, where $f(x)$ may be the shape of a distorted violin string or the temperature in a metal bar of length L, for example. (Corresponding problems will be discussed in Chap. 12.) Now comes the idea.



(a) The given function $f(x)$



(b) $f(x)$ extended as an ***even*** periodic function of period $2L$



(c) $f(x)$ extended as an ***odd*** periodic function of period $2L$

Fig. 267. (a) Function $f(x)$ given on an interval $0 \leqq x \leqq L$

(b) **Even extension** to the full "range" (interval) $-L \leqq x \leqq L$ (heavy curve) and the periodic extension of period $2L$ to the x-axis

(c) **Odd extension** to $-L \leqq x \leqq L$ (heavy curve) and the periodic extension of period $2L$ to the x-axis

We could extend $f(x)$ as a function of period L and develop the extended function into a Fourier series. But this series would in general contain *both* cosine *and* sine terms. We can do better and get simpler series. Indeed, for our given f we can calculate Fourier coefficients from (2) or from (4) in Theorem 1. And we have a choice and can take what seems more practical. If we use (2), we get (1). This is the **even periodic extension** f_1 of f in Fig. 267b. If we choose (4) instead, we get (3), the **odd periodic extension** f_2 of f in Fig. 267c.

Both extensions have period $2L$. This motivates the name **half-range expansions:** f is given (and of physical interest) only on half the range, half the interval of periodicity of length $2L$.

Let us illustrate these ideas with an example that we shall also need in Chap. 12.

EXAMPLE 4 **"Triangle" and Its Half-Range Expansions**

Find the two half-range expansions of the function (Fig. 268)

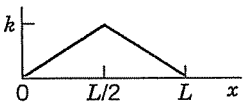


Fig. 268. The given function in Example 4

$$f(x) = \begin{cases} \dfrac{2k}{L}x & \text{if } 0 < x < \dfrac{L}{2} \\ \dfrac{2k}{L}(L - x) & \text{if } \dfrac{L}{2} < x < L. \end{cases}$$

Solution. *(a) Even periodic extension.* From (2) we obtain

$$a_0 = \frac{1}{L}\left[\frac{2k}{L}\int_0^{L/2} x\,dx + \frac{2k}{L}\int_{L/2}^{L}(L - x)\,dx\right] = \frac{k}{2},$$

$$a_n = \frac{2}{L}\left[\frac{2k}{L}\int_0^{L/2} x\cos\frac{n\pi}{L}x\,dx + \frac{2k}{L}\int_{L/2}^{L}(L - x)\cos\frac{n\pi}{L}x\,dx\right].$$

We consider a_n. For the first integral we obtain by integration by parts

$$\int_0^{L/2} x \cos \frac{n\pi}{L} x \, dx = \frac{Lx}{n\pi} \sin \frac{n\pi}{L} x \Big|_0^{L/2} - \frac{L}{n\pi} \int_0^{L/2} \sin \frac{n\pi}{L} x \, dx$$

$$= \frac{L^2}{2n\pi} \sin \frac{n\pi}{2} + \frac{L^2}{n^2\pi^2} \left(\cos \frac{n\pi}{2} - 1 \right).$$

Similarly, for the second integral we obtain

$$\int_{L/2}^{L} (L - x) \cos \frac{n\pi}{L} x \, dx = \frac{L}{n\pi} (L - x) \sin \frac{n\pi}{L} x \Big|_{L/2}^{L} + \frac{L}{n\pi} \int_{L/2}^{L} \sin \frac{n\pi}{L} x \, dx$$

$$= \left(0 - \frac{L}{n\pi} \left(L - \frac{L}{2} \right) \sin \frac{n\pi}{2} \right) - \frac{L^2}{n^2\pi^2} \left(\cos n\pi - \cos \frac{n\pi}{2} \right).$$

We insert these two results into the formula for a_n. The sine terms cancel and so does a factor L^2. This gives

$$a_n = \frac{4k}{n^2\pi^2} \left(2 \cos \frac{n\pi}{2} - \cos n\pi - 1 \right).$$

Thus,

$$a_2 = -16k/(2^2\pi^2), \qquad a_6 = -16k/(6^2\pi^2), \qquad a_{10} = -16k/(10^2\pi^2), \cdots$$

and $a_n = 0$ if $n \neq 2, 6, 10, 14, \cdots$. Hence the first half-range expansion of $f(x)$ is (Fig. 269a)

$$f(x) = \frac{k}{2} - \frac{16k}{\pi^2} \left(\frac{1}{2^2} \cos \frac{2\pi}{L} x + \frac{1}{6^2} \cos \frac{6\pi}{L} x + \cdots \right).$$

This Fourier cosine series represents the even periodic extension of the given function $f(x)$, of period $2L$.

(b) ***Odd periodic extension.*** Similarly, from (4) we obtain

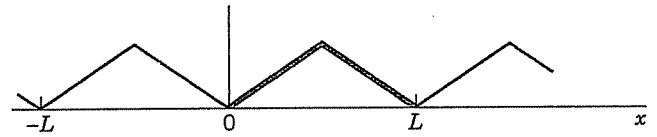
$$b_n = \frac{8k}{n^2\pi^2} \sin \frac{n\pi}{2}. \tag{5}$$

Hence the other half-range expansion of $f(x)$ is (Fig. 269b)

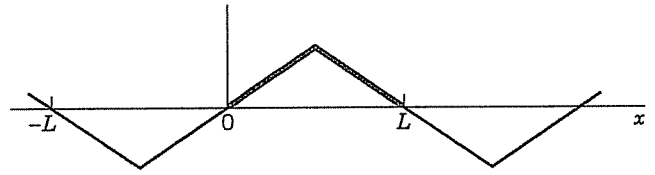
$$f(x) = \frac{8k}{\pi^2} \left(\frac{1}{1^2} \sin \frac{\pi}{L} x - \frac{1}{3^2} \sin \frac{3\pi}{L} x + \frac{1}{5^2} \sin \frac{5\pi}{L} x - + \cdots \right).$$

This series represents the odd periodic extension of $f(x)$, of period $2L$.

Basic applications of these results will be shown in Secs. 12.3 and 12.5.



(a) Even extension



(b) Odd extension

Fig. 269. Periodic extensions of $f(x)$ in Example 4

PROBLEM SET 11.3

1–9 EVEN AND ODD FUNCTIONS

e the following functions even, odd, or neither even nor odd?

. $|x|$, $x^2 \sin nx$, $x + x^2$, $e^{-|x|}$, $\ln x$, $x \cosh x$

. $\sin(x^2)$, $\sin^2 x$, $x \sinh x$, $|x^3|$, $e^{\pi x}$, xe^x, $\tan 2x$, $x/(1 + x^2)$

Are the following functions, which are assumed to be riodic of period 2π, even, odd, or neither even nor odd?

. $f(x) = x^3 \quad (-\pi < x < \pi)$

4. $f(x) = x^2 \quad (-\pi/2 < x < 3\pi/2)$

. $f(x) = e^{-4x} \quad (-\pi < x < \pi)$

. $f(x) = x^3 \sin x \quad (-\pi < x < \pi)$

7. $f(x) = x|x| - x^3 \quad (-\pi < x < \pi)$

. $f(x) = 1 - x + x^3 - x^5 \quad (-\pi < x < \pi)$

. $f(x) = 1/(1 + x^2)$ if $-\pi < x < 0$, $f(x) = -1/(1 + x^2)$ if $0 < x < \pi$

. **PROJECT. Even and Odd Functions.** **(a)** Are the following expressions even or odd? Sums and products of even functions and of odd functions. Products of even times odd functions. Absolute values of odd functions. $f(x) + f(-x)$ and $f(x) - f(-x)$ for arbitrary $f(x)$.

(b) Write e^{kx}, $1/(1 - x)$, $\sin(x + k)$, $\cosh(x + k)$ as sums of an even and an odd function.

(c) Find all functions that are both even and odd.

(d) Is $\cos^3 x$ even or odd? $\sin^3 x$? Find the Fourier series of these functions. Do you recognize familiar identities?

11–16 FOURIER SERIES OF EVEN AND ODD FUNCTIONS

l the given function even or odd? Find its Fourier series. Sketch or graph the function and some partial sums. (Show the details of your work.)

. $f(x) = \pi - |x| \quad (-\pi < x < \pi)$

12. $f(x) = 2x|x| \quad (-1 < x < 1)$

13. $f(x) = \begin{cases} x & \text{if } -\pi/2 < x < \pi/2 \\ \pi - x & \text{if } \pi/2 < x < 3\pi/2 \end{cases}$

14. $f(x) = \begin{cases} \pi e^{-x} & \text{if } -\pi < x < 0 \\ \pi e^{x} & \text{if } 0 < x < \pi \end{cases}$

15. $f(x) = \begin{cases} 2 & \text{if } -2 < x < 0 \\ 0 & \text{if } 0 < x < 2 \end{cases}$

16. $f(x) = \begin{cases} 1 - \frac{1}{2}|x| & \text{if } -2 < x < 2 \\ 0 & \text{if } 2 < x < 6 \end{cases} \quad (p = 8)$

17–25 HALF-RANGE EXPANSIONS

Find (a) the Fourier cosine series, (b) the Fourier sine series. Sketch $f(x)$ and its two periodic extensions. (Show the details of your work.)

17. $f(x) = 1 \quad (0 < x < 2)$

18. $f(x) = x \quad (0 < x < \frac{1}{2})$

19. $f(x) = 2 - x \quad (0 < x < 2)$

20. $f(x) = \begin{cases} 0 & (0 < x < 2) \\ 1 & (2 < x < 4) \end{cases}$

21. $f(x) = \begin{cases} 1 & (0 < x < 1) \\ 2 & (1 < x < 2) \end{cases}$

22. $f(x) = \begin{cases} x & (0 < x < \pi/2) \\ \pi/2 & (\pi/2 < x < \pi) \end{cases}$

23. $f(x) = x \quad (0 < x < L)$

24. $f(x) = x^2 \quad (0 < x < L)$

25. $f(x) = \pi - x \quad (0 < x < \pi)$

26. Illustrate the formulas in the proof of Theorem 1 with examples. Prove the formulas.

11.4 Complex Fourier Series. *Optional*

In this optional section we show that the Fourier series

$$(1) \qquad f(x) = a_0 + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \sin nx)$$

can be written in complex form, which sometimes simplifies calculations (see Example 1, on page 498). This complex form can be obtained because in complex, the exponential function e^{it} and $\cos t$ and $\sin t$ are related by the basic **Euler formula** (see (11) in Sec. 2.2)

$$(2) \qquad e^{it} = \cos t + i \sin t. \qquad \text{Thus} \qquad e^{-it} = \cos t - i \sin t.$$

Conversely, by adding and subtracting these two formulas, we obtain

$$(3) \qquad \text{(a)} \quad \cos t = \frac{1}{2}(e^{it} + e^{-it}), \qquad \text{(b)} \quad \sin t = \frac{1}{2i}(e^{it} - e^{-it}).$$

From (3), using $1/i = -i$ in $\sin t$ and setting $t = nx$ in both formulas, we get

$$a_n \cos nx + b_n \sin nx = \frac{1}{2}\, a_n(e^{inx} + e^{-inx}) + \frac{1}{2i}\, b_n(e^{inx} - e^{-inx})$$

$$= \frac{1}{2}\,(a_n - ib_n)e^{inx} + \frac{1}{2}\,(a_n + ib_n)e^{-inx}.$$

We insert this into (1). Writing $a_0 = c_0$, $\frac{1}{2}(a_n - ib_n) = c_n$, and $\frac{1}{2}(a_n + ib_n) = k_n$, we get from (1)

$$(4) \qquad f(x) = c_0 + \sum_{n=1}^{\infty} (c_n e^{inx} + k_n e^{-inx}).$$

The coefficients $c_1, c_2, \cdots$, and $k_1, k_2, \cdots$ are obtained from (6b), (6c) in Sec. 11.1 and then (2) above with $t = nx$,

$$c_n = \frac{1}{2}\,(a_n - ib_n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)(\cos nx - i \sin nx)\, dx = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-inx}\, dx$$

$$(5)$$

$$k_n = \frac{1}{2}\,(a_n + ib_n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)(\cos nx + i \sin nx)\, dx = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{inx}\, dx.$$

Finally, we can combine (5) into a single formula by the trick of writing $k_n = c_{-n}$. Then (4), (5), and $c_0 = a_0$ in (6a) of Sec. 11.1 give (summation from $-\infty$!)

$$(6) \qquad f(x) = \sum_{n=-\infty}^{\infty} c_n e^{inx},$$

$$c_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-inx}\, dx, \qquad n = 0, \pm 1, \pm 2, \cdots.$$

This is the so-called *complex form of the Fourier series* or, more briefly, the **complex Fourier series,** of $f(x)$. The c_n are called the **complex Fourier coefficients** of $f(x)$.

For a function of period $2L$ our reasoning gives the **complex Fourier series**

$$(7) \qquad f(x) = \sum_{n=-\infty}^{\infty} c_n e^{in\pi x/L},$$

$$c_n = \frac{1}{2L}\int_{-L}^{L} f(x)e^{-in\pi x/L}\, dx, \qquad n = 0, \pm 1, \pm 2, \cdots.$$

EXAMPLE 1 Complex Fourier Series

Find the complex Fourier series of $f(x) = e^x$ if $-\pi < x < \pi$ and $f(x + 2\pi) = f(x)$ and obtain from it the usual Fourier series.

Solution. Since $\sin n\pi = 0$ for integer n, we have

$$e^{\pm in\pi} = \cos n\pi \pm i \sin n\pi = \cos n\pi = (-1)^n.$$

With this we obtain from (6) by integration

$$c_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^x e^{-inx}\,dx = \frac{1}{2\pi}\,\frac{1}{1-in}\,e^{x-inx}\Big|_{x=-\pi}^{\pi} = \frac{1}{2\pi}\,\frac{1}{1-in}\,(e^{\pi} - e^{-\pi})(-1)^n.$$

On the right,

$$\frac{1}{1-in} = \frac{1+in}{(1-in)(1+in)} = \frac{1+in}{1+n^2} \qquad \text{and} \qquad e^{\pi} - e^{-\pi} = 2\sinh \pi.$$

Hence the complex Fourier series is

$$e^x = \frac{\sinh \pi}{\pi}\sum_{n=-\infty}^{\infty}(-1)^n\,\frac{1+in}{1+n^2}\,e^{inx} \qquad (-\pi < x < \pi). \tag{8}$$

From this let us derive the real Fourier series. Using (2) with $t = nx$ and $i^2 = -1$, we have in (8)

$$(1 + in)e^{inx} = (1 + in)(\cos nx + i \sin nx) = (\cos nx - n \sin nx) + i(n \cos nx + \sin nx).$$

Now (8) also has a corresponding term with $-n$ instead of n. Since $\cos(-nx) = \cos nx$ and $\sin(-nx) = -\sin nx$, we obtain in this term

$$(1 - in)e^{-inx} = (1 - in)(\cos nx - i \sin nx) = (\cos nx - n \sin nx) - i(n \cos nx + \sin nx).$$

If we add these two expressions, the imaginary parts cancel. Hence their sum is

$$2(\cos nx - n \sin nx), \qquad n = 1, 2, \cdots.$$

For $n = 0$ we get 1 (not 2) because there is only one term. Hence the real Fourier series is

$$e^x = \frac{2\sinh \pi}{\pi}\left[\frac{1}{2} - \frac{1}{1+1^2}(\cos x - \sin x) + \frac{1}{1+2^2}(\cos 2x - 2\sin 2x) - + \cdots\right]. \tag{9}$$

In Fig. 270 the poor approximation near the jumps at $\pm\pi$ is a case of the Gibbs phenomenon (see CAS Experiment 20 in Problem Set 11.2).

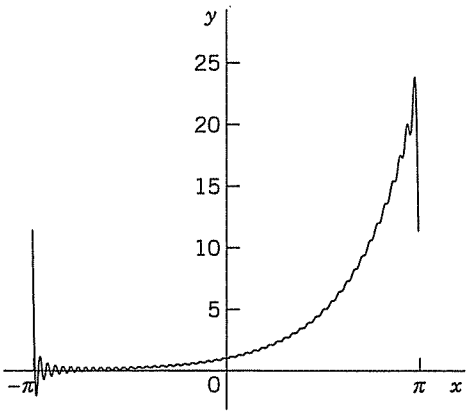


Fig. 270. Partial sum of (9), terms from $n = 0$ to 50

PROBLEM SET 11.4

1. **(Calculus review)** Review complex numbers.
2. **(Even and odd functions)** Show that the complex Fourier coefficients of an even function are real and those of an odd function are pure imaginary.
3. **(Fourier coefficients)** Show that $a_0 = c_0$, $a_n = c_n + c_{-n}$, $b_n = i(c_n - c_{-n})$.
4. Verify the calculations in Example 1.
5. Find further terms in (9) and graph partial sums with your CAS.
6. Obtain the real series in Example 1 directly from the Euler formulas in Sec. 11.

7–13 COMPLEX FOURIER SERIES

Find the complex Fourier series of the following functions. (Show the details of your work.)

7. $f(x) = -1$ if $-\pi < x < 0$, $f(x) = 1$ if $0 < x < \pi$
8. Convert the series in Prob. 7 to real form.
9. $f(x) = x \quad (-\pi < x < \pi)$
10. Convert the series in Prob. 9 to real form.
11. $f(x) = x^2 \quad (-\pi < x < \pi)$
12. Convert the series in Prob. 11 to real form.
13. $f(x) = x \quad (0 < x < 2\pi)$
14. **PROJECT. Complex Fourier Coefficients.** It is very interesting that the c_n in (6) can be derived directly by a method similar to that for a_n and b_n in Sec. 11.1. For this, multiply the series in (6) by e^{-imx} with fixed integer m, and integrate termwise from $-\pi$ to π on both sides (allowed, for instance, in the case of uniform convergence) to get

$$\int_{-\pi}^{\pi} f(x)e^{-imx}\,dx = \sum_{n=-\infty}^{\infty} c_n \int_{-\pi}^{\pi} e^{i(n-m)x}\,dx.$$

Show that the integral on the right equals 2π when $n = m$ and 0 when $n \neq m$ [use (3b)], so that you get the coefficient formula in (6).

11.5 Forced Oscillations

Fourier series have important applications in connection with ODEs and PDEs. We show this for a basic problem modeled by an ODE. Various applications to PDEs will follow in Chap. 12. This will show the enormous usefulness of Euler's and Fourier's ingenious idea of splitting up periodic functions into the simplest ones possible.

From Sec. 2.8 we know that forced oscillations of a body of mass m on a spring of modulus k are governed by the ODE

$$(1) \qquad my'' + cy' + ky = r(t)$$

where $y = y(t)$ is the displacement from rest, c the damping constant, k the spring constant (spring modulus), and $r(t)$ the external force depending on time t. Figure 271 shows the model and Fig. 272 its electrical analog, an *RLC*-circuit governed by

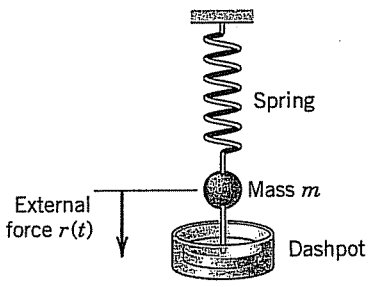


Fig. 271. Vibrating system under consideration

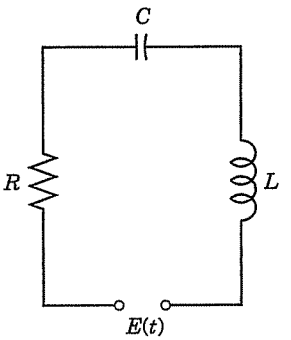


Fig. 272. Electrical analog of the system in Fig. 271 (*RLC*-circuit)

(1*) $$LI'' + RI' + \frac{1}{C} I = E'(t)$$ (Sec. 2.9).

We consider (1). If $r(t)$ is a sine or cosine function and if there is damping ($c > 0$), then the steady-state solution is a harmonic oscillation with frequency equal to that of $r(t)$. However, if $r(t)$ is not a pure sine or cosine function but is any other periodic function, then the steady-state solution will be a superposition of harmonic oscillations with frequencies equal to that of $r(t)$ and integer multiples of the latter. And if one of these frequencies is close to the (practical) resonant frequency of the vibrating system (see Sec. 2.8), then the corresponding oscillation may be the dominant part of the response of the system to the external force. This is what the use of Fourier series will show us. Of course, this is quite surprising to an observer unfamiliar with Fourier series, which are highly important in the study of vibrating systems and resonance. Let us discuss the entire situation in terms of a typical example.

EXAMPLE 1 Forced Oscillations under a Nonsinusoidal Periodic Driving Force

In (1), let $m = 1$ (gm), $c = 0.05$ (gm/sec), and $k = 25$ (gm/sec^2), so that (1) becomes

(2) $$y'' + 0.05y' + 25y = r(t)$$

where $r(t)$ is measured in gm · cm/sec^2. Let (Fig. 273)

$$r(t) = \begin{cases} t + \frac{\pi}{2} & \text{if } -\pi < t < 0, \\ -t + \frac{\pi}{2} & \text{if } 0 < t < \pi, \end{cases} \qquad r(t + 2\pi) = r(t).$$

Find the steady-state solution $y(t)$.

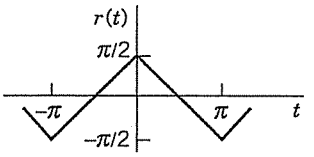


Fig. 273. Force in Example 1

Solution. We represent $r(t)$ by a Fourier series, finding

(3) $$r(t) = \frac{4}{\pi}\left(\cos t + \frac{1}{3^2}\cos 3t + \frac{1}{5^2}\cos 5t + \cdots\right)$$

(take the answer to Prob. 11 in Problem Set 11.3 minus $\frac{1}{2}\pi$ and write t for x). Then we consider the ODE

(4) $$y'' + 0.05y' + 25y = \frac{4}{n^2\pi}\cos nt \qquad (n = 1, 3, \cdots)$$

whose right side is a single term of the series (3). From Sec. 2.8 we know that the steady-state solution $y_n(t)$ of (4) is of the form

(5) $$y_n = A_n \cos nt + B_n \sin nt.$$

By substituting this into (4) we find that

$$(6) \qquad A_n = \frac{4(25 - n^2)}{n^2\pi D_n}, \qquad B_n = \frac{0.2}{n\pi D_n}, \qquad \text{where} \qquad D_n = (25 - n^2)^2 + (0.05n)^2.$$

Since the ODE (2) is linear, we may expect the steady-state solution to be

$$(7) \qquad y = y_1 + y_3 + y_5 + \cdots$$

where y_n is given by (5) and (6). In fact, this follows readily by substituting (7) into (2) and using the Fourier series of $r(t)$, provided that termwise differentiation of (7) is permissible. (Readers already familiar with the notion of uniform convergence [Sec. 15.5] may prove that (7) may be differentiated term by term.)

From (6) we find that the amplitude of (5) is (a factor $\sqrt{D_n}$ cancels out)

$$C_n = \sqrt{A_n^2 + B_n^2} = \frac{4}{n^2\pi\sqrt{D_n}}.$$

Numeric values are

$$C_1 = 0.0531$$

$$C_3 = 0.0088$$

$$C_5 = 0.2037$$

$$C_7 = 0.0011$$

$$C_9 = 0.0003.$$

Figure 274 shows the input (multiplied by 0.1) and the output. For $n = 5$ the quantity D_n is very small, the denominator of C_5 is small, and C_5 is so large that y_5 is the dominating term in (7). Hence the output is almost a harmonic oscillation of five times the frequency of the driving force, a little distorted due to the term y_1, whose amplitude is about 25% of that of y_5. You could make the situation still more extreme by decreasing the damping constant c. Try it. ▩

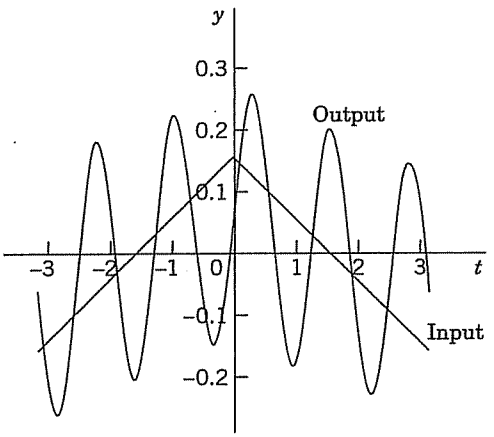


Fig. 274. Input and steady-state output in Example 1

PROBLEM SET 11.5

1. **(Coefficients)** Derive the formula for C_n from A_n and B_n.
2. **(Spring constant)** What would happen to the amplitudes C_n in Example 1 (and thus to the form of the vibration) if we changed the spring constant to the value 9? If we took a stiffer spring with $k = 81$? First guess.
3. **(Damping)** In Example 1 change c to 0.02 and discuss how this changes the output.
4. **(Input)** What would happen in Example 1 if we replaced $r(t)$ with its derivative (the rectangular wave)? What is the ratio of the new C_n to the old ones?

5–11 GENERAL SOLUTION

Find a general solution of the ODE $y'' + \omega^2 y = r(t)$ with $r(t)$ as given. (Show the details of your work.)

5. $r(t) = \cos \omega t$, $\omega = 0.5, 0.8, 1.1, 1.5, 5.0, 10.0$

6. $r(t) = \cos \omega_1 t + \cos \omega_2 t \quad (\omega^2 \neq \omega_1^2, \omega_2^2)$

7. $r(t) = \sum_{n=1}^{N} a_n \cos nt, \quad |\omega| \neq 1, 2, \cdots, N$

8. $r(t) = \sin t + \frac{1}{3} \sin 3t + \frac{1}{5} \sin 5t + \frac{1}{7} \sin 7t$

9. $r(t) = \begin{cases} t + \pi & \text{if } -\pi < t < 0 \\ -t + \pi & \text{if } 0 < t < \pi \end{cases}$

and $r(t + 2\pi) = r(t)$, $|\omega| \neq 0, 1, 3, \cdots$

10. $r(t) = \begin{cases} t & \text{if } -\pi/2 < t < \pi/2 \\ \pi - t & \text{if } \pi/2 < t < 3\pi/2 \end{cases}$

and $r(t + 2\pi) = r(t)$, $|\omega| \neq 1, 3, 5, \cdots$

11. $r(t) = \frac{\pi}{4} |\sin t|$ if $-\pi < t < \pi$ and

$r(t + 2\pi) = r(t)$, $|\omega| \neq 0, 2, 4, \cdots$

12. (CAS Program) Write a program for solving the ODE just considered and for jointly graphing input and output of an initial value problem involving that ODE. Apply the program to Probs. 5 and 9 with initial values of your choice.

13. (Sign of coefficients) Some A_n in Example 1 are positive and some negative. Is this physically understandable?

14–17 STEADY-STATE DAMPED OSCILLATIONS

Find the steady-state oscillation of $y'' + cy' + y = r(t)$ with $c > 0$ and $r(t)$ as given. (Show the details of your work.)

14. $r(t) = a_n \cos nt$

15. $r(t) = \sin 3t$

16. $r(t) = \begin{cases} \pi t & \text{if } -\pi/2 < t < \pi/2 \\ \pi(\pi - t) & \text{if } \pi/2 < t < 3\pi/2 \end{cases}$

and $r(t + 2\pi) = r(t)$

17. $r(t) = \sum_{n=1}^{N} b_n \sin nt$

18. CAS EXPERIMENT. Maximum Output Term. Graph and discuss outputs of $y'' + cy' + ky = r(t)$ with $r(t)$ as in Example 1 for various c and k with emphasis on the maximum C_n and its ratio to the second largest $|C_n|$.

19–20 *RLC*-CIRCUIT

Find the steady-state current $I(t)$ in the *RLC*-circuit in Fig. 272, where $R = 100\ \Omega$, $L = 10$ H, $C = 10^{-2}$ F and $E(t)$ V as follows and periodic with period 2π. Sketch or graph the first four partial sums. Note that the coefficients of the solution decrease rapidly.

19. $E(t) = 200t(\pi^2 - t^2) \quad (-\pi < t < \pi)$

20. $E(t) = \begin{cases} 100(\pi t + t^2) & \text{if } -\pi < t < 0 \\ 100(\pi t - t^2) & \text{if } 0 < t < \pi \end{cases}$

11.6 Approximation by Trigonometric Polynomials

Fourier series play a prominent role in differential equations. Another field in which they have major applications is **approximation theory**, which concerns the approximation of functions by other (usually simpler) functions. In connection with Fourier series the idea is as follows.

Let $f(x)$ be a function on the interval $-\pi \leqq x \leqq \pi$ that can be represented on this interval by a Fourier series. Then the ***N*th partial sum** of the series

$$f(x) \approx a_0 + \sum_{n=1}^{N} (a_n \cos nx + b_n \sin nx) \tag{1}$$

is an approximation of the given $f(x)$. It is natural to ask whether (1) is the "best" approximation of f by a **trigonometric polynomial of degree *N***, that is, by a function of the form

$$F(x) = A_0 + \sum_{n=1}^{N} (A_n \cos nx + B_n \sin nx) \qquad (N \text{ fixed}) \tag{2}$$

where "best" means that the "error" of the approximation is as small as possible.

Of course, we must first define what we mean by the **error** E of such an approximation. We could choose the maximum of $|f - F|$. But in connection with Fourier series it is better to choose a definition that measures the goodness of agreement between f and F *on the whole interval* $-\pi \leqq x \leqq \pi$. This seems preferable, in particular if f has jumps: F in Fig. 275 is a good overall approximation of f, but the maximum of $|f - F|$ (more precisely, the *supremum*) is large (it equals at least half the jump of f at x_0). We choose

$$E = \int_{-\pi}^{\pi} (f - F)^2 \, dx. \tag{3}$$

This is called the **square error** of F relative to the function f on the interval $-\pi \leqq x \leqq \pi$. Clearly, $E \geqq 0$.

N being fixed, we want to determine the coefficients in (2) such that E is minimum. Since $(f - F)^2 = f^2 - 2fF + F^2$, we have

$$E = \int_{-\pi}^{\pi} f^2 \, dx - 2 \int_{-\pi}^{\pi} fF \, dx + \int_{-\pi}^{\pi} F^2 \, dx. \tag{4}$$

We square (2), insert it into the last integral in (4), and evaluate the occurring integrals. This gives integrals of $\cos^2 nx$ and $\sin^2 nx$ ($n \geqq 1$), which equal π, and integrals of $\cos nx$, $\sin nx$, and $(\cos nx)(\sin mx)$, which are zero (just as in Sec. 11.1). Thus

$$\int_{-\pi}^{\pi} F^2 \, dx = \int_{-\pi}^{\pi} \left[A_0 + \sum_{n=1}^{N} (A_n \cos nx + B_n \sin nx) \right]^2 dx$$

$$= \pi(2A_0^2 + A_1^2 + \cdots + A_N^2 + B_1^2 + \cdots + B_N^2).$$

We now insert (2) into the integral of fF in (4). This gives integrals of $f \cos nx$ as well as $f \sin nx$, just as in Euler's formulas, Sec. 11.1, for a_n and b_n (each multiplied by A_n or B_n). Hence

$$\int_{-\pi}^{\pi} fF \, dx = \pi(2A_0 a_0 + A_1 a_1 + \cdots + A_N a_N + B_1 b_1 + \cdots + B_N b_N).$$

With these expressions, (4) becomes

$$E = \int_{-\pi}^{\pi} f^2 \, dx - 2\pi \left[2A_0 a_0 + \sum_{n=1}^{N} (A_n a_n + B_n b_n) \right] + \pi \left[2A_0^2 + \sum_{n=1}^{N} (A_n^2 + B_n^2) \right]. \tag{5}$$

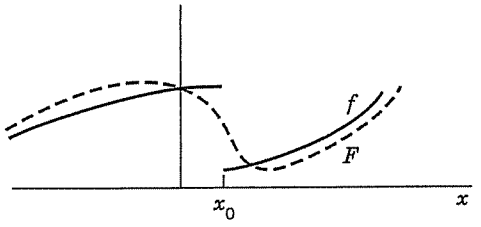


Fig. 275. Error of approximation

We now take $A_n = a_n$ and $B_n = b_n$ in (2). Then in (5) the second line cancels half of the integral-free expression in the first line. Hence for this choice of the coefficients of F the square error, call it E^*, is

$$E^* = \int_{-\pi}^{\pi} f^2\,dx - \pi\left[2a_0^2 + \sum_{n=1}^{N}(a_n^2 + b_n^2)\right]. \tag{6}$$

We finally subtract (6) from (5). Then the integrals drop out and we get terms $A_n^2 - 2A_n a_n + a_n^2 = (A_n - a_n)^2$ and similar terms $(B_n - b_n)^2$:

$$E - E^* = \pi\left\{2(A_0 - a_0)^2 + \sum_{n=1}^{N}\left[(A_n - a_n)^2 + (B_n - b_n)^2\right]\right\}.$$

Since the sum of squares of real numbers on the right cannot be negative,

$$E - E^* \geqq 0, \qquad \text{thus} \qquad E \geqq E^*,$$

and $E = E^*$ if and only if $A_0 = a_0, \cdots, B_N = b_N$. This proves the following fundamental minimum property of the partial sums of Fourier series.

THEOREM 1

Minimum Square Error

The square error of F in (2) (with fixed N) relative to f on the interval $-\pi \leqq x \leqq \pi$ is minimum if and only if the coefficients of F in (2) are the Fourier coefficients of f. This minimum value E^* is given by (6).

From (6) we see that E^* cannot increase as N increases, but may decrease. Hence *with increasing N the partial sums of the Fourier series of f yield better and better approximations to f*, considered from the viewpoint of the square error.

Since $E^* \geqq 0$ and (6) holds for every N, we obtain from (6) the important **Bessel's inequality**

$$2a_0^2 + \sum_{n=1}^{\infty}(a_n^2 + b_n^2) \leqq \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)^2\,dx \tag{7}$$

for the Fourier coefficients of any function f for which integral on the right exists. (For F. W. Bessel see Sec. 5.5.)

It can be shown (see [C12] in App. 1) that for such a function f, **Parseval's theorem** holds; that is, formula (7) holds with the equality sign, so that it becomes **Parseval's identity**[4]

$$2a_0^2 + \sum_{n=1}^{\infty}(a_n^2 + b_n^2) = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)^2\,dx. \tag{8}$$

[4]MARC ANTOINE PARSEVAL (1755–1836), French mathematician. A physical interpretation of the identity follows in the next section.

EXAMPLE 1 Minimum Square Error for the Sawtooth Wave

Compute the minimum square error E^* of $F(x)$ with $N = 1, 2, \cdots, 10, 20, \cdots, 100$ and 1000 relative to

$$f(x) = x + \pi \qquad (-\pi < x < \pi)$$

on the interval $-\pi \leqq x \leqq \pi$.

Solution. $F(x) = \pi + 2(\sin x - \frac{1}{2}\sin 2x + \frac{1}{3}\sin 3x - + \cdots + \frac{(-1)^{N+1}}{N}\sin Nx)$ by Example 3 in Sec. 11.3. From this and (6),

$$E^* = \int_{-\pi}^{\pi} (x + \pi)^2\, dx - \pi\left(2\pi^2 + 4\sum_{n=1}^{N}\frac{1}{n^2}\right).$$

Numeric values are:

N	E^*	N	E^*	N	E^*	N	E^*
1	8.1045	6	1.9295	20	0.6129	70	0.1782
2	4.9629	7	1.6730	30	0.4120	80	0.1561
3	3.5666	8	1.4767	40	0.3103	90	0.1389
4	2.7812	9	1.3216	50	0.2488	100	0.1250
5	2.2786	10	1.1959	60	0.2077	1000	0.0126

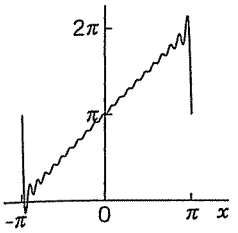


Fig. 276. F with $N = 20$ in Example 1

$F = S_1, S_2, S_3$ are shown in Fig. 266 in Sec. 11.3, and $F = S_{20}$ is shown in Fig. 276. Although $|f(x) - F(x)|$ is large at $\pm\pi$ (how large?), where f is discontinuous, F approximates f quite well on the whole interval, except near $\pm\pi$, where "waves" remain owing to the Gibbs phenomenon (see CAS Experiment 20 in Problem Set 11.2).

Can you think of functions f for which E^* decreases more quickly with increasing N? ■

This is the end of our discussion of Fourier series, which has emphasized the practical aspects of these series, as needed in applications. In the last three sections of this chapter we show how ideas and techniques in Fourier series can be extended to ***nonperiodic*** functions.

PROBLEM SET 11.6

1–9 MINIMUM SQUARE ERROR

Find the trigonometric polynomial $F(x)$ of the form (2) for which the square error with respect to the given $f(x)$ on the interval $-\pi \leqq x \leqq \pi$ is minimum, and compute the minimum value for $N = 1, 2, \cdots, 5$ (or also for larger values if you have a CAS).

1. $f(x) = x \quad (-\pi < x < \pi)$
2. $f(x) = x^2 \quad (-\pi < x < \pi)$
3. $f(x) = |x| \quad (-\pi < x < \pi)$
4. $f(x) = x^3 \quad (-\pi < x < \pi)$
5. $f(x) = |\sin x| \quad (-\pi < x < \pi)$
6. $f(x) = e^{-|x|} \quad (-\pi < x < \pi)$
7. $f(x) = \begin{cases} -1 & \text{if } -\pi < x < 0 \\ 1 & \text{if } 0 < x < \pi \end{cases}$
8. $f(x) = \begin{cases} x & \text{if } -\frac{1}{2}\pi < x < \frac{1}{2}\pi \\ 0 & \text{if } \frac{1}{2}\pi < x < \frac{3}{2}\pi \end{cases}$
9. $f(x) = x(x + \pi)$ if $-\pi < x < 0$, $f(x) = x(-x + \pi)$ if $0 < x < \pi$
10. **CAS EXPERIMENT. Size and Decrease of E^*.** Compare the size of the minimum square error E^* for functions of your choice. Find experimentally the factors on which the decrease of E^* with N depends. For each function considered find the smallest N such that $E^* < 0.1$.
11. **(Monotonicity)** Show that the minimum square error (6) is a monotone decreasing function of N. How can you use this in practice?

12–16 PARSEVAL'S IDENTITY

Using Parseval's identity, prove that the series have the indicated sums. Compute the first few partial sums to see that the convergence is rapid.

12. $1 + \frac{1}{3^4} + \frac{1}{5^4} + \frac{1}{7^4} + \cdots = \frac{\pi^4}{96} = 1.01467\ 8032$

(Use Prob. 15 in Sec. 11.1.)

13. $1 + \frac{1}{3^2} + \frac{1}{5^2} + \cdots = \frac{\pi^2}{8} = 1.23370\ 0550$

(Use Prob. 13 in Sec. 11.1.)

14. $\frac{1}{1^2 \cdot 3^2} + \frac{1}{3^2 \cdot 5^2} + \frac{1}{5^2 \cdot 7^2} + \cdots$

$= \frac{\pi^2}{16} - \frac{1}{2} = 0.11685\ 0275$

(Use Prob. 5, this set.)

15. $1 + \frac{1}{2^4} + \frac{1}{3^4} + \cdots = \frac{\pi^4}{90} = 1.08232\ 3234$

(Use Prob. 21 in Sec. 11.1.)

16. $1 + \frac{1}{3^6} + \frac{1}{5^6} + \frac{1}{7^6} + \cdots = \frac{\pi^6}{960} = 1.00144\ 7078$

(Use Prob. 9, this set.)

11.7 Fourier Integral

Fourier series are powerful tools for problems involving functions that are periodic or are of interest on a finite interval only. Sections 11.3 and 11.5 first illustrated this, and various further applications follow in Chap. 12. Since, of course, many problems involve functions that are **nonperiodic** *and are of interest on the whole x-axis,* we ask what can be done to extend the method of Fourier series to such functions. This idea will lead to "Fourier integrals."

In Example 1 we start from a special function f_L of period $2L$ and see what happens to its Fourier series if we let $L \to \infty$. Then we do the same for an *arbitrary* function f_L of period $2L$. This will motivate and suggest the main result of this section, which is an integral representation given in Theorem 1 (below).

EXAMPLE 1 Rectangular Wave

Consider the periodic rectangular wave $f_L(x)$ of period $2L > 2$ given by

$$f_L(x) = \begin{cases} 0 & \text{if} \quad -L < x < -1 \\ 1 & \text{if} \quad -1 < x < 1 \\ 0 & \text{if} \quad 1 < x < L. \end{cases}$$

The left part of Fig. 277 shows this function for $2L = 4, 8, 16$ as well as the nonperiodic function $f(x)$, which we obtain from f_L if we let $L \to \infty$,

$$f(x) = \lim_{L \to \infty} f_L(x) = \begin{cases} 1 & \text{if } -1 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

We now explore what happens to the Fourier coefficients of f_L as L increases. Since f_L is even, $b_n = 0$ for all n. For a_n the Euler formulas (6), Sec. 11.2, give

$$a_0 = \frac{1}{2L}\int_{-1}^{1} dx = \frac{1}{L}, \qquad a_n = \frac{1}{L}\int_{-1}^{1} \cos\frac{n\pi x}{L}\,dx = \frac{2}{L}\int_{0}^{1} \cos\frac{n\pi x}{L}\,dx = \frac{2}{L}\,\frac{\sin(n\pi/L)}{n\pi/L}.$$

This sequence of Fourier coefficients is called the **amplitude spectrum** of f_L because $|a_n|$ is the maximum amplitude of the wave $a_n \cos(n\pi x/L)$. Figure 277 shows this spectrum for the periods $2L = 4, 8, 16$. We see that for increasing L these amplitudes become more and more dense on the positive w_n-axis, where $w_n = n\pi/L$. Indeed, for $2L = 4, 8, 16$ we have 1, 3, 7 amplitudes per "half-wave" of the function $(2 \sin w_n)/(Lw_n)$ (dashed in the figure). Hence for $2L = 2^k$ we have $2^{k-1} - 1$ amplitudes per half-wave, so that these amplitudes will eventually be everywhere dense on the positive w_n-axis (and will decrease to zero).

The outcome of this example gives an intuitive impression of what about to expect if we turn from our special function to an arbitrary one, as we shall do next. ▩

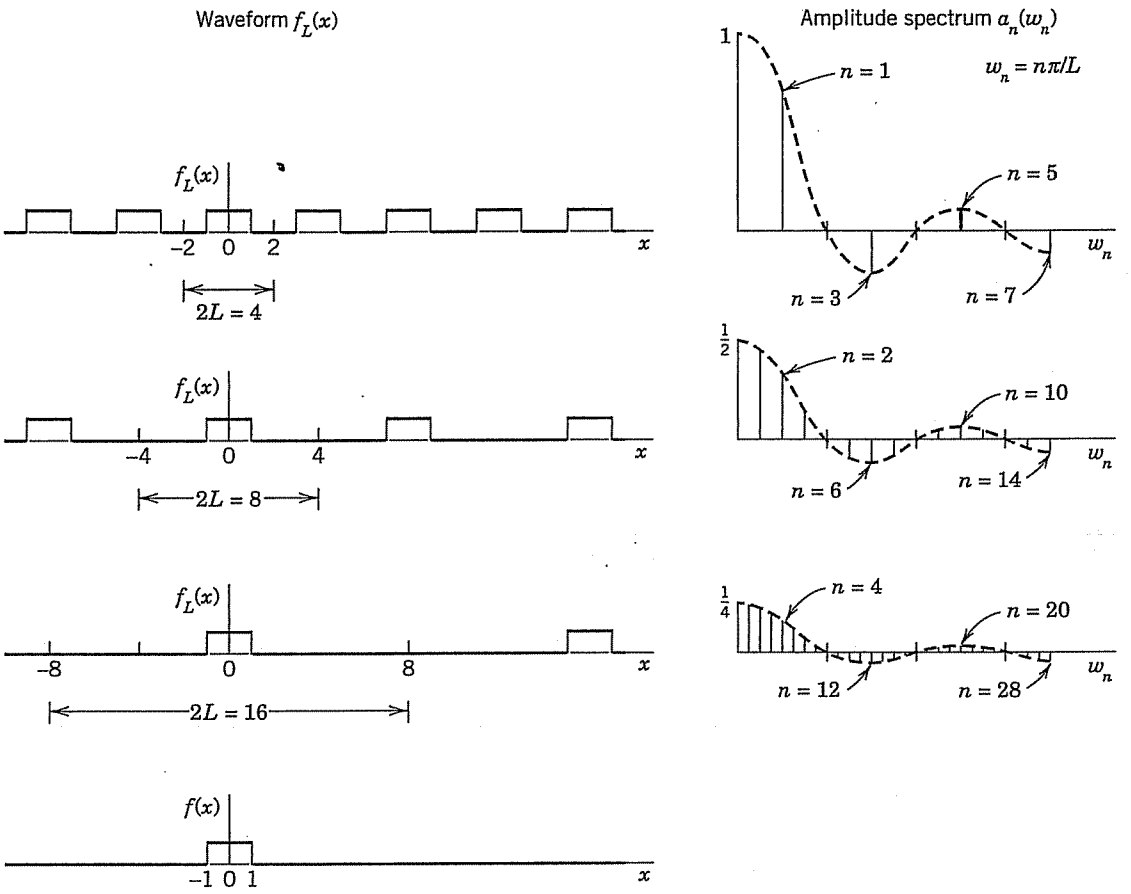


Fig. 277. Waveforms and amplitude spectra in Example 1

From Fourier Series to Fourier Integral

We now consider any periodic function $f_L(x)$ of period $2L$ that can be represented by a Fourier series

$$f_L(x) = a_0 + \sum_{n=1}^{\infty} (a_n \cos w_n x + b_n \sin w_n x), \qquad w_n = \frac{n\pi}{L}$$

and find out what happens if we let $L \to \infty$. Together with Example 1 the present calculation will suggest that we should expect an integral (instead of a series) involving $\cos wx$ and $\sin wx$ with w no longer restricted to integer multiples $w = w_n = n\pi/L$ of π/L but taking *all* values. We shall also see what form such an integral might have.

If we insert a_n and b_n from the Euler formulas (6), Sec. 11.2, and denote the variable of integration by v, the Fourier series of $f_L(x)$ becomes

$$f_L(x) = \frac{1}{2L} \int_{-L}^{L} f_L(v)\, dv + \frac{1}{L} \sum_{n=1}^{\infty} \left[\cos w_n x \int_{-L}^{L} f_L(v) \cos w_n v\, dv + \sin w_n x \int_{-L}^{L} f_L(v) \sin w_n v\, dv \right].$$

We now set

$$\Delta w = w_{n+1} - w_n = \frac{(n+1)\pi}{L} - \frac{n\pi}{L} = \frac{\pi}{L}.$$

Then $1/L = \Delta w/\pi$, and we may write the Fourier series in the form

$$(1) \quad f_L(x) = \frac{1}{2L}\int_{-L}^{L} f_L(v)\,dv + \frac{1}{\pi}\sum_{n=1}^{\infty}\left[(\cos w_n x)\,\Delta w \int_{-L}^{L} f_L(v)\cos w_n v\,dv + (\sin w_n x)\,\Delta w \int_{-L}^{L} f_L(v)\sin w_n v\,dv\right].$$

This representation is valid for any fixed L, arbitrarily large, but finite.

We now let $L \to \infty$ and assume that the resulting nonperiodic function

$$f(x) = \lim_{L\to\infty} f_L(x)$$

is **absolutely integrable** on the x-axis; that is, the following (finite!) limits exist:

$$(2) \quad \lim_{a\to-\infty}\int_{a}^{0}|f(x)|\,dx + \lim_{b\to\infty}\int_{0}^{b}|f(x)|\,dx \quad \left(\text{written } \int_{-\infty}^{\infty}|f(x)|\,dx\right).$$

Then $1/L \to 0$, and the value of the first term on the right side of (1) approaches zero. Also $\Delta w = \pi/L \to 0$ and it seems ***plausible*** that the infinite series in (1) becomes an integral from 0 to ∞, which represents $f(x)$, namely,

$$(3) \quad f(x) = \frac{1}{\pi}\int_{0}^{\infty}\left[\cos wx\int_{-\infty}^{\infty} f(v)\cos wv\,dv + \sin wx\int_{-\infty}^{\infty} f(v)\sin wv\,dv\right]dw.$$

If we introduce the notations

$$(4) \quad A(w) = \frac{1}{\pi}\int_{-\infty}^{\infty} f(v)\cos wv\,dv, \qquad B(w) = \frac{1}{\pi}\int_{-\infty}^{\infty} f(v)\sin wv\,dv$$

we can write this in the form

$$(5) \quad f(x) = \int_{0}^{\infty}[A(w)\cos wx + B(w)\sin wx]\,dw.$$

This is called a representation of $f(x)$ by a **Fourier integral.**

It is clear that our naive approach merely ***suggests*** the representation (5), but by no means establishes it; in fact, the limit of the series in (1) as Δw approaches zero is not the definition of the integral (3). Sufficient conditions for the validity of (5) are as follows.

THEOREM 1

Fourier Integral

If $f(x)$ is piecewise continuous (see Sec. 6.1) *in every finite interval and has a right-hand derivative and a left-hand derivative at every point* (see Sec 11.1) *and if the integral* (2) *exists, then $f(x)$ can be represented by a Fourier integral* (5) *with A and B given by* (4). *At a point where $f(x)$ is discontinuous the value of the Fourier integral equals the average of the left- and right-hand limits of $f(x)$ at that point* (see Sec. 11.1). (Proof in Ref. [C12]; see App. 1.)

Applications of Fourier Integrals

The main application of Fourier integrals is in solving ODEs and PDEs, as we shall see for PDEs in Sec. 12.6. However, we can also use Fourier integrals in integration and in discussing functions defined by integrals, as the next examples (2 and 3) illustrate.

EXAMPLE 2 Single Pulse, Sine Integral

Find the Fourier integral representation of the function

$$f(x) = \begin{cases} 1 & \text{if} \quad |x| < 1 \\ 0 & \text{if} \quad |x| > 1 \end{cases} \qquad \text{(Fig. 278).}$$

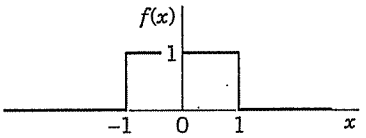


Fig. 278. Example 2

Solution. From (4) we obtain

$$A(w) = \frac{1}{\pi}\int_{-\infty}^{\infty} f(v)\cos wv\,dv = \frac{1}{\pi}\int_{-1}^{1}\cos wv\,dv = \frac{\sin wv}{\pi w}\bigg|_{-1}^{1} = \frac{2\sin w}{\pi w}$$

$$B(w) = \frac{1}{\pi}\int_{-1}^{1}\sin wv\,dv = 0$$

and (5) gives the answer

$$f(x) = \frac{2}{\pi}\int_0^{\infty}\frac{\cos wx\sin w}{w}\,dw. \tag{6}$$

The average of the left- and right-hand limits of $f(x)$ at $x = 1$ is equal to $(1 + 0)/2$, that is, 1/2.

Furthermore, from (6) and Theorem 1 we obtain (multiply by $\pi/2$)

$$\int_0^{\infty}\frac{\cos wx\sin w}{w}\,dw = \begin{cases} \pi/2 & \text{if} \quad 0 \leqq x < 1, \\ \pi/4 & \text{if} \quad x = 1, \\ 0 & \text{if} \quad x > 1. \end{cases} \tag{7}$$

We mention that this integral is called **Dirichlet's discontinous factor.** (For P. L. Dirichlet see Sec. 10.8.)

The case $x = 0$ is of particular interest. If $x = 0$, then (7) gives

$$\int_0^{\infty}\frac{\sin w}{w}\,dw = \frac{\pi}{2} \tag{8*}$$

We see that this integral is the limit of the so-called **sine integral**

$$\mathrm{Si}(u) = \int_0^{u}\frac{\sin w}{w}\,dw \tag{8}$$

as $u \to \infty$. The graphs of $\mathrm{Si}(u)$ and of the integrand are shown in Fig. 279.

In the case of a Fourier series the graphs of the partial sums are approximation curves of the curve of the periodic function represented by the series. Similarly, in the case of the Fourier integral (5), approximations are obtained by replacing ∞ by numbers a. Hence the integral

$$\frac{2}{\pi}\int_0^{a}\frac{\cos wx\sin w}{w}\,dw \tag{9}$$

approximates the right side in (6) and therefore $f(x)$.

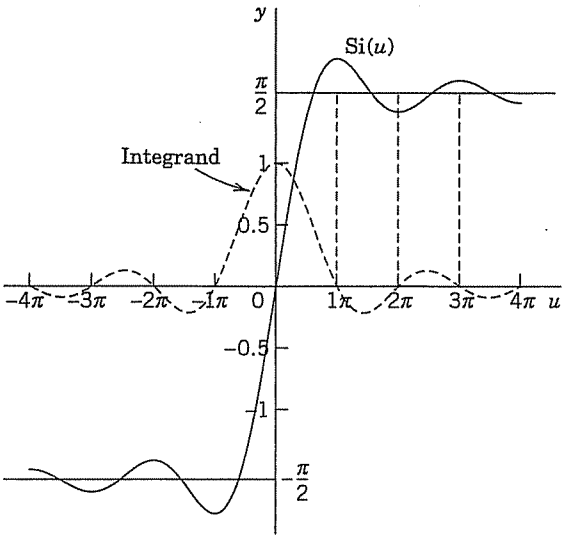


Fig. 279. Sine integral Si(u) and integrand

Figure 280 shows oscillations near the points of discontinuity of $f(x)$. We might expect that these oscillations disappear as a approaches infinity. But this is not true; with increasing a, they are shifted closer to the points $x = \pm 1$. This unexpected behavior, which also occurs in connection with Fourier series, is known as the **Gibbs phenomenon.** (See also Problem Set 11.2.) We can explain it by representing (9) in terms of sine integrals as follows. Using (11) in App. A3.1, we have

$$\frac{2}{\pi}\int_0^a \frac{\cos wx \sin w}{w}\,dw = \frac{1}{\pi}\int_0^a \frac{\sin (w + wx)}{w}\,dw + \frac{1}{\pi}\int_0^a \frac{\sin (w - wx)}{w}\,dw.$$

In the first integral on the right we set $w + wx = t$. Then $dw/w = dt/t$, and $0 \leqq w \leqq a$ corresponds to $0 \leqq t \leqq (x + 1)a$. In the last integral we set $w - wx = -t$. Then $dw/w = dt/t$, and $0 \leqq w \leqq a$ corresponds to $0 \leqq t \leqq (x - 1)a$. Since $\sin(-t) = -\sin t$, we thus obtain

$$\frac{2}{\pi}\int_0^a \frac{\cos wx \sin w}{w}\,dw = \frac{1}{\pi}\int_0^{(x+1)a} \frac{\sin t}{t}\,dt - \frac{1}{\pi}\int_0^{(x-1)a} \frac{\sin t}{t}\,dt.$$

From this and (8) we see that our integral (9) equals

$$\frac{1}{\pi}\,\mathrm{Si}(a[x + 1]) - \frac{1}{\pi}\,\mathrm{Si}(a[x - 1])$$

and the oscillations in Fig. 280 result from those in Fig. 279. The increase of a amounts to a transformation of the scale on the axis and causes the shift of the oscillations (the waves) toward the points of discontinuity -1 and 1.

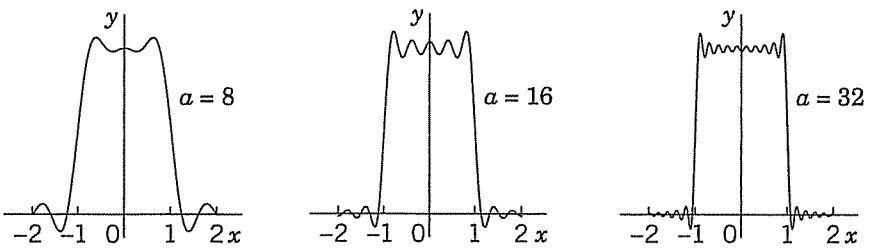


Fig. 280. The integral (9) for $a = 8$, 16, and 32

Fourier Cosine Integral and Fourier Sine Integral

For an even or odd function the Fourier integral becomes simpler. Just as in the case of Fourier series (Sec. 11.3), this is of practical interest in saving work and avoiding errors. The simplifications follow immediately from the formulas just obtained.

Indeed, if $f(x)$ is an ***even*** function, then $B(w) = 0$ in (4) and

$$A(w) = \frac{2}{\pi}\int_0^\infty f(v)\cos wv\,dv. \tag{10}$$

The Fourier integral (5) then reduces to the **Fourier cosine integral**

$$f(x) = \int_0^\infty A(w)\cos wx\,dw \qquad (f \text{ even}). \tag{11}$$

Similarly, if $f(x)$ is ***odd,*** then in (4) we have $A(w) = 0$ and

$$B(w) = \frac{2}{\pi}\int_0^\infty f(v)\sin wv\,dv. \tag{12}$$

The Fourier integral (5) then reduces to the **Fourier sine integral**

$$f(x) = \int_0^\infty B(w)\sin wx\,dw \qquad (f \text{ odd}). \tag{13}$$

Evaluation of Integrals

Earlier in this section we pointed out that the main application of the Fourier integral is in differential equations but that Fourier integral representations also help in evaluating certain integrals. To see this, we show the method for an important case, the Laplace integrals.

EXAMPLE 3 **Laplace Integrals**

We shall derive the Fourier cosine and Fourier sine integrals of $f(x) = e^{-kx}$, where $x > 0$ and $k > 0$ (Fig. 281). The result will be used to evaluate the so-called Laplace integrals.

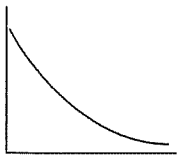


Fig. 281. $f(x)$ in Example 3

Solution. (a) From (10) we have $A(w) = \frac{2}{\pi}\int_0^\infty e^{-kv}\cos wv\,dv$. Now, by integration by parts,

$$\int e^{-kv}\cos wv\,dv = -\frac{k}{k^2+w^2}\,e^{-kv}\left(-\frac{w}{k}\sin wv + \cos wv\right).$$

If $v = 0$, the expression on the right equals $-k/(k^2 + w^2)$. If v approaches infinity, that expression approaches zero because of the exponential factor. Thus

$$A(w) = \frac{2k/\pi}{k^2+w^2}. \tag{14}$$

By substituting this into (11) we thus obtain the Fourier cosine integral representation

$$f(x) = e^{-kx} = \frac{2k}{\pi}\int_0^\infty \frac{\cos wx}{k^2+w^2}\,dw. \qquad (x > 0,\quad k > 0).$$

From this representation we see that

(15) $$\int_0^\infty \frac{\cos wx}{k^2 + w^2}\, dw = \frac{\pi}{2k}\, e^{-kx} \qquad (x > 0, \quad k > 0).$$

(b) Similarly, from (12) we have $B(w) = \frac{2}{\pi}\int_0^\infty e^{-kv} \sin wv\, dv$. By integration by parts,

$$\int e^{-kv} \sin wv\, dv = -\frac{w}{k^2 + w^2}\, e^{-kv}\left(\frac{k}{w}\sin wv + \cos wv\right).$$

This equals $-w/(k^2 + w^2)$ if $v = 0$, and approaches 0 as $v \to \infty$. Thus

(16) $$B(w) = \frac{2w/\pi}{k^2 + w^2}.$$

From (13) we thus obtain the Fourier sine integral representation

$$f(x) = e^{-kx} = \frac{2}{\pi}\int_0^\infty \frac{w \sin wx}{k^2 + w^2}\, dw.$$

From this we see that

(17) $$\int_0^\infty \frac{w \sin wx}{k^2 + w^2}\, dw = \frac{\pi}{2}\, e^{-kx} \qquad (x > 0, \quad k > 0).$$

The integrals (15) and (17) are called the **Laplace integrals.**

PROBLEM SET 11.7

1–6 EVALUATION OF INTEGRALS

Show that the given integral represents the indicated function. *Hint.* Use (5), (11), or (13); the integral tells you which one, and its value tells you what function to consider. (Show the details of your work.)

1. $$\int_0^\infty \frac{\cos xw + w \sin xw}{1 + w^2}\, dw = \begin{cases} 0 & \text{if } x < 0 \\ \pi/2 & \text{if } x = 0 \\ \pi e^{-x} & \text{if } x > 0 \end{cases}$$

2. $$\int_0^\infty \frac{\sin w - w \cos w}{w^2} \sin xw\, dw$$

$$= \begin{cases} \pi x/2 & \text{if } 0 < x < 1 \\ \pi/4 & \text{if } x = 1 \\ 0 & \text{if } x > 1 \end{cases}$$

3. $$\int_0^\infty \frac{\cos xw}{1 + w^2}\, dw = \frac{\pi}{2} e^{-x} \text{ if } x > 0$$

4. $$\int_0^\infty \frac{\sin w}{w} \cos xw\, dw = \begin{cases} \pi/2 & \text{if } 0 \leqq x < 1 \\ \pi/4 & \text{if } x = 1 \\ 0 & \text{if } x > 1 \end{cases}$$

5. $$\int_0^\infty \frac{\cos(\pi w/2)}{1 - w^2} \cos xw\, dw$$

$$= \begin{cases} \frac{\pi}{2}\cos x & \text{if } 0 < |x| < \pi/2 \\ 0 & \text{if } |x| \geqq \pi/2 \end{cases}$$

6. $$\int_0^\infty \frac{\sin \pi w \sin xw}{1 - w^2}\, dw = \begin{cases} \frac{\pi}{2}\sin x & \text{if } 0 \leqq x \leqq \pi \\ 0 & \text{if } x > \pi \end{cases}$$

7–12 FOURIER COSINE INTEGRAL REPRESENTATIONS

Represent $f(x)$ as an integral (11).

7. $$f(x) = \begin{cases} 1 & \text{if } 0 < x < a \\ 0 & \text{if } x > a \end{cases}$$

8. $f(x) = \begin{cases} x^2 & \text{if } 0 < x < a \\ 0 & \text{if } x > a \end{cases}$

9. $f(x) = \begin{cases} x & \text{if } 0 < x < 1 \\ 0 & \text{if } x > 1 \end{cases}$

10. $f(x) = \begin{cases} x/2 & \text{if } 0 < x < 1 \\ 1 - x/2 & \text{if } 1 < x < 2 \\ 0 & \text{if } x > 2 \end{cases}$

11. $f(x) = \begin{cases} \sin x & \text{if } 0 < x < \pi \\ 0 & \text{if } x > \pi \end{cases}$

12. $f(x) = \begin{cases} e^{-x} & \text{if } 0 < x < a \\ 0 & \text{if } x > a \end{cases}$

13. CAS EXPERIMENT. Approximate Fourier Cosine Integrals. Graph the integrals in Prob. 7, 9, and 11 as functions of x. Graph approximations obtained by replacing ∞ with finite upper limits of your choice. Compare the quality of the approximations. Write a short report on your empirical results and observations.

14–19 FOURIER SINE INTEGRAL REPRESENTATIONS

Represent $f(x)$ as an integral (13).

14. $f(x) = \begin{cases} 1 & \text{if } 0 < x < a \\ 0 & \text{if } x > a \end{cases}$

15. $f(x) = \begin{cases} \sin x & \text{if } 0 < x < \pi \\ 0 & \text{if } x > \pi \end{cases}$

16. $f(x) = \begin{cases} 1 - x^2 & \text{if } 0 < x < 1 \\ 0 & \text{if } x > 1 \end{cases}$

17. $f(x) = \begin{cases} \pi - x & \text{if } 0 < x < \pi \\ 0 & \text{if } x > \pi \end{cases}$

18. $f(x) = \begin{cases} \cos x & \text{if } 0 < x < \pi \\ 0 & \text{if } x > \pi \end{cases}$

19. $f(x) = \begin{cases} a - x & \text{if } 0 < x < a \\ 0 & \text{if } x > a \end{cases}$

20. PROJECT. Properties of Fourier Integrals

(a) Fourier cosine integral. Show that (11) implies

(a1) $$f(ax) = \frac{1}{a}\int_0^\infty A\left(\frac{w}{a}\right)\cos xw\, dw$$

$(a > 0)$ *(Scale change)*

(a2) $$xf(x) = \int_0^\infty B^*(w)\sin xw\, dw,$$

$$B^* = -\frac{dA}{dw}, \qquad A \text{ as in (10)}$$

(a3) $$x^2 f(x) = \int_0^\infty A^*(w)\cos xw\, dw,$$

$$A^* = -\frac{d^2A}{dw^2}.$$

(b) Solve Prob. 8 by applying (a3) to the result of Prob. 7.

(c) Verify (a2) for $f(x) = 1$ if $0 < x < a$ and $f(x) = 0$ if $x > a$.

(d) Fourier sine integral. Find formulas for the Fourier sine integral similar to those in (a).

11.8 Fourier Cosine and Sine Transforms

An **integral transform** is a transformation in the form of an integral that produces from given functions new functions depending on a different variable. These transformations are of interest mainly as tools for solving ODEs, PDEs, and integral equations, and they often also help in handling and applying special functions. The **Laplace transform** (Chap. 6) is of this kind and is by far the most important integral transform in engineering.

The next in order of importance are Fourier transforms. We shall see that these transforms can be obtained from the Fourier integral in Sec. 11.7 in a rather simple fashion. In this section we consider two of them, which are real, and in the next section a third one that is complex.

Fourier Cosine Transform

For an ***even*** function $f(x)$, the Fourier integral is the Fourier cosine integral

$$(1)\quad (a)\quad f(x) = \int_0^\infty A(w)\cos wx\,dw, \qquad \text{where} \qquad (b)\quad A(w) = \frac{2}{\pi}\int_0^\infty f(v)\cos wv\,dv$$

[see (10), (11), Sec. 11.7]. We now set $A(w) = \sqrt{2/\pi}\,\hat{f}_c(w)$, where c suggests "cosine." Then from (1b), writing $v = x$, we have

$$(2)\qquad \hat{f}_c(w) = \sqrt{\frac{2}{\pi}}\int_0^\infty f(x)\cos wx\,dx$$

and from (1a),

$$(3)\qquad f(x) = \sqrt{\frac{2}{\pi}}\int_0^\infty \hat{f}_c(w)\cos wx\,dw.$$

ATTENTION! In (2) we integrate with respect to x and in (3) with respect to w. Formula (2) gives from $f(x)$ a new function $\hat{f}_c(w)$, called the **Fourier cosine transform** of $f(x)$. Formula (3) gives us back $f(x)$ from $\hat{f}_c(w)$, and we therefore call $f(x)$ the **inverse Fourier cosine transform** of $\hat{f}_c(w)$.

The process of obtaining the transform $\hat{f}_c$ from a given f is also called the **Fourier cosine transform** or the *Fourier cosine transform method.*

Fourier Sine Transform

Similarly, for an ***odd*** function $f(x)$, the Fourier integral is the Fourier sine integral [see (12), (13), Sec. 11.7]

$$(4)\quad (a)\quad f(x) = \int_0^\infty B(w)\sin wx\,dw, \qquad \text{where} \qquad (b)\quad B(w) = \frac{2}{\pi}\int_0^\infty f(v)\sin wv\,dv.$$

We now set $B(w) = \sqrt{2/\pi}\,\hat{f}_s(w)$, where s suggests "sine." Then from (4b), writing $v = x$, we have

$$(5)\qquad \hat{f}_s(w) = \sqrt{\frac{2}{\pi}}\int_0^\infty f(x)\sin wx\,dx.$$

This is called the **Fourier sine transform** of $f(x)$. Similarly, from (4a) we have

$$(6)\qquad f(x) = \sqrt{\frac{2}{\pi}}\int_0^\infty \hat{f}_s(w)\sin wx\,dw.$$

This is called the **inverse Fourier sine transform** of $\hat{f}_s(w)$. The process of obtaining $\hat{f}_s(w)$ from $f(x)$ is also called the **Fourier sine transform** or the *Fourier sine transform method.*

Other notations are

$$\mathcal{F}_c(f) = \hat{f}_c, \qquad \mathcal{F}_s(f) = \hat{f}_s$$

and $\mathcal{F}_c^{-1}$ and $\mathcal{F}_s^{-1}$ for the inverses of $\mathcal{F}_c$ and $\mathcal{F}_s$, respectively.

EXAMPLE 1 **Fourier Cosine and Fourier Sine Transforms**

Find the Fourier cosine and Fourier sine transforms of the function

$$f(x) = \begin{cases} k & \text{if } 0 < x < a \\ 0 & \text{if } x > a \end{cases} \quad \text{(Fig. 282).}$$

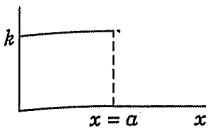


Fig. 282. $f(x)$ in Example 1

Solution. From the definitions (2) and (5) we obtain by integration

$$\hat{f}_c(w) = \sqrt{\frac{2}{\pi}}\, k \int_0^a \cos wx\, dx = \sqrt{\frac{2}{\pi}}\, k \left(\frac{\sin aw}{w}\right)$$

$$\hat{f}_s(w) = \sqrt{\frac{2}{\pi}}\, k \int_0^a \sin wx\, dx = \sqrt{\frac{2}{\pi}}\, k \left(\frac{1 - \cos aw}{w}\right).$$

This agrees with formulas 1 in the first two tables in Sec. 11.10 (where $k = 1$).

Note that for $f(x) = k = const$ $(0 < x < \infty)$, these transforms do not exist. (Why?) ▩

EXAMPLE 2 **Fourier Cosine Transform of the Exponential Function**

Find $\mathcal{F}_c(e^{-x})$.

Solution. By integration by parts and recursion,

$$\mathcal{F}_c(e^{-x}) = \sqrt{\frac{2}{\pi}} \int_0^\infty e^{-x} \cos wx\, dx = \sqrt{\frac{2}{\pi}}\, \frac{e^{-x}}{1 + w^2} (-\cos wx + w \sin wx) \Big|_0^\infty = \frac{\sqrt{2/\pi}}{1 + w^2}.$$

This agrees with formula 3 in Table I, Sec. 11.10, with $a = 1$. See also the next example. ▩

What did we do to introduce the two integral transforms under consideration? Actually not much: We changed the notations A and B to get a "symmetric" distribution of the constant $2/\pi$ in the original formulas (10)–(13), Sec. 11.7. This redistribution is a standard convenience, but it is not essential. One could do without it.

What have we gained? We show next that these transforms have operational properties that permit them to convert differentiations into algebraic operations (just as the Laplace transform does). This is the key to their application in solving differential equations.

Linearity, Transforms of Derivatives

If $f(x)$ is absolutely integrable (see Sec. 11.7) on the positive x-axis and piecewise continuous (see Sec. 6.1) on every finite interval, then the Fourier cosine and sine transforms of f exist.

Furthermore, if f and g have Fourier cosine and sine transforms, so does $af + bg$ for any constants a and b, and by (2),

$$\mathcal{F}_c(af + bg) = \sqrt{\frac{2}{\pi}} \int_0^\infty [af(x) + bg(x)] \cos wx\, dx$$

$$= a\sqrt{\frac{2}{\pi}} \int_0^\infty f(x) \cos wx\, dx + b\sqrt{\frac{2}{\pi}} \int_0^\infty g(x) \cos wx\, dx.$$

The right side is $a\mathcal{F}_c(f) + b\mathcal{F}_c(g)$. Similarly for $\mathcal{F}_s$, by (5). This shows that the Fourier cosine and sine transforms are **linear operations,**

(7)
$$\text{(a)} \quad \mathcal{F}_c(af + bg) = a\mathcal{F}_c(f) + b\mathcal{F}_c(g),$$
$$\text{(b)} \quad \mathcal{F}_s(af + bg) = a\mathcal{F}_s(f) + b\mathcal{F}_s(g).$$

THEOREM 1

Cosine and Sine Transforms of Derivatives

Let $f(x)$ be continuous and absolutely integrable on the x-axis, let $f'(x)$ be piecewise continuous on every finite interval, and let let $f(x) \to 0$ *as* $x \to \infty$. *Then*

(8)

$$\textbf{(a)} \qquad \mathcal{F}_c\{f'(x)\} = w\mathcal{F}_s\{f(x)\} - \sqrt{\frac{2}{\pi}}\, f(0),$$

$$\textbf{(b)} \qquad \mathcal{F}_s\{f'(x)\} = -w\mathcal{F}_c\{f(x)\}.$$

PROOF This follows from the definitions by integration by parts, namely,

$$\begin{aligned}\mathcal{F}_c\{f'(x)\} &= \sqrt{\frac{2}{\pi}} \int_0^\infty f'(x) \cos wx\, dx \\ &= \sqrt{\frac{2}{\pi}} \left[f(x) \cos wx \,\Big|_0^\infty + w \int_0^\infty f(x) \sin wx\, dx \right] \\ &= -\sqrt{\frac{2}{\pi}}\, f(0) + w\mathcal{F}_s\{f(x)\};\end{aligned}$$

and similarly,

$$\begin{aligned}\mathcal{F}_s\{f'(x)\} &= \sqrt{\frac{2}{\pi}} \int_0^\infty f'(x) \sin wx\, dx \\ &= \sqrt{\frac{2}{\pi}} \left[f(x) \sin wx \,\Big|_0^\infty - w \int_0^\infty f(x) \cos wx\, dx \right] \\ &= 0 - w\mathcal{F}_c\{f(x)\}.\end{aligned}$$

∎

Formula (8a) with f' instead of f gives (when f', f'' satisfy the respective assumptions for f, f' in Theorem 1)

$$\mathcal{F}_c\{f''(x)\} = w\mathcal{F}_s\{f'(x)\} - \sqrt{\frac{2}{\pi}}\, f'(0);$$

hence by (8b)

$$\textbf{(9a)} \qquad \mathcal{F}_c\{f''(x)\} = -w^2\mathcal{F}_c\{f(x)\} - \sqrt{\frac{2}{\pi}}\, f'(0).$$

Similarly,

$$\textbf{(9b)} \qquad \mathcal{F}_s\{f''(x)\} = -w^2\mathcal{F}_s\{f(x)\} + \sqrt{\frac{2}{\pi}}\, wf(0).$$

A basic application of (9) to PDEs will be given in Sec. 12.6. For the time being we show how (9) can be used for deriving transforms.

EXAMPLE 3 An Application of the Operational Formula (9)

Find the Fourier cosine transform $\mathcal{F}_c(e^{-ax})$ of $f(x) = e^{-ax}$, where $a > 0$.

Solution. By differentiation, $(e^{-ax})'' = a^2 e^{-ax}$; thus

$$a^2 f(x) = f''(x).$$

From this, (9a), and the linearity (7a),

$$\begin{aligned} a^2 \mathcal{F}_c(f) &= \mathcal{F}_c(f'') \\ &= -w^2 \mathcal{F}_c(f) - \sqrt{\frac{2}{\pi}}\, f'(0) \\ &= -w^2 \mathcal{F}_c(f) + a\sqrt{\frac{2}{\pi}}. \end{aligned}$$

Hence

$$(a^2 + w^2)\mathcal{F}_c(f) = a\sqrt{2/\pi}.$$

The *answer* is (see Table I, Sec. 11.10)

$$\mathcal{F}_c(e^{-ax}) = \sqrt{\frac{2}{\pi}}\left(\frac{a}{a^2 + w^2}\right) \qquad (a > 0).$$

Tables of Fourier cosine and sine transforms are included in Sec. 11.10.

PROBLEM SET 11.8

1–10 FOURIER COSINE TRANSFORM

1. Let $f(x) = -1$ if $0 < x < 1$, $f(x) = 1$ if $1 < x < 2$, $f(x) = 0$ if $x > 2$. Find $\hat{f}_c(w)$.
2. Let $f(x) = x$ if $0 < x < k$, $f(x) = 0$ if $x > k$. Find $\hat{f}_c(w)$.
3. Derive formula 3 in Table 1 of Sec. 11.10 by integration.
4. Find the inverse Fourier cosine transform $f(x)$ from the answer to Prob. 1. *Hint.* Use Prob. 4 in Sec. 11.7.
5. Obtain $\mathcal{F}_c^{-1}(1/(1 + w^2))$ from Prob. 3 in Sec. 11.7.
6. Obtain $\mathcal{F}_c^{-1}(e^{-w})$ by integration.
7. Find $\mathcal{F}_c((1 - x^2)^{-1} \cos(\pi x/2))$. *Hint.* Use Prob. 5 in Sec. 11.7.
8. Let $f(x) = x^2$ if $0 < x < 1$ and 0 if $x > 1$. Find $\mathcal{F}_c(f)$.
9. Does the Fourier cosine transform of $x^{-1} \sin x$ exist? Of $x^{-1} \cos x$? Give reasons.
10. $f(x) = 1$ $(0 < x < \infty)$ has no Fourier cosine or sine transform. Give reasons.

11–20 FOURIER SINE TRANSFORM

11. Find $\mathcal{F}_s(e^{-\pi x})$ by integration.
12. Find the answer to Prob. 11 from (9b).
13. Obtain formula 8 in Table II of Sec. 11.11 from (8b) and a suitable formula in Table I.
14. Let $f(x) = \sin x$ if $0 < x < \pi$ and 0 if $x > \pi$. Find $\mathcal{F}_s(f)$. Compare with Prob. 6 in Sec. 11.7. Comment.
15. In Table II of Sec. 11.10 obtain formula 2 from formula 4, using $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ [(30) in App. 3.1].
16. Show that $\mathcal{F}_s(x^{-1/2}) = w^{-1/2}$ by setting $wx = t^2$ and using $S(\infty) = \sqrt{\pi/8}$ in (38) of App. 3.1.
17. Obtain $\mathcal{F}_s(e^{-ax})$ from (8a) and formula 3 in Table I of Sec. 11.10.
18. Show that $\mathcal{F}_s(x^{-3/2}) = 2w^{1/2}$. *Hint.* Set $wx = t^2$, integrate by parts, and use $C(\infty) = \sqrt{\pi/8}$ in (38) of App. 3.1.
19. **(Scale change)** Using the notation of (5), show that $f(ax)$ has the Fourier sine transform $(1/a)\hat{f}_s(w/a)$.
20. **WRITING PROJECT. Obtaining Fourier Cosine and Sine Transforms.** Write a short report on ways of obtaining these transforms, giving illustrations with examples of your own.

11.9 Fourier Transform. Discrete and Fast Fourier Transforms

The two transforms in the last section are real. We now consider a third one, called the **Fourier transform,** which is complex. We shall obtain this transform from the complex Fourier integral, which we explain first.

Complex Form of the Fourier Integral

The (real) Fourier integral is [see (4), (5), Sec. 11.7]

$$f(x) = \int_0^\infty [A(w) \cos wx + B(w) \sin wx]\, dw$$

where

$$A(w) = \frac{1}{\pi} \int_{-\infty}^{\infty} f(v) \cos wv\, dv, \qquad B(w) = \frac{1}{\pi} \int_{-\infty}^{\infty} f(v) \sin wv\, dv.$$

Substituting A and B into the integral for f, we have

$$f(x) = \frac{1}{\pi} \int_0^\infty \int_{-\infty}^{\infty} f(v) \, [\cos wv \cos wx + \sin wv \sin wx]\, dv\, dw.$$

By the addition formula for the cosine [(6) in App. A3.1] the expression in the brackets $[\cdots]$ equals $\cos (wv - wx)$ or, since the cosine is even, $\cos (wx - wv)$. We thus obtain

$$(1^*) \qquad f(x) = \frac{1}{\pi} \int_0^\infty \left[\int_{-\infty}^{\infty} f(v) \cos (wx - wv)\, dv \right] dw.$$

The integral in brackets is an *even* function of w, call it $F(w)$, because $\cos (wx - wv)$ is an even function of w, the function f does not depend on w, and we integrate with respect to v (not w). Hence the integral of $F(w)$ from $w = 0$ to ∞ is 1/2 times the integral of $F(w)$ from $-\infty$ to ∞. Thus (note the change of the integration limit!)

$$(1) \qquad f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} f(v) \cos (wx - wv)\, dv \right] dw.$$

We claim that the integral of the form (1) with sin instead of cos is zero:

$$(2) \qquad \frac{1}{2\pi} \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} f(v) \sin (wx - wv)\, dv \right] dw = 0.$$

This is true since $\sin (wx - wv)$ is an odd function of w, which makes the integral in brackets an odd function of w, call it $G(w)$. Hence the integral of $G(w)$ from $-\infty$ to ∞ is zero, as claimed.

We now take the integrand of (1) plus i $(= \sqrt{-1})$ times the integrand of (2) and use the **Euler formula** [(11) in Sec. 2.2]

$$(3) \qquad e^{ix} = \cos x + i \sin x.$$

Taking $wx - wv$ instead of x in (3) and multiplying by $f(v)$ gives

$$f(v)\cos(wx - wv) + if(v)\sin(wx - wv) = f(v)e^{i(wx-wv)}.$$

Hence the result of adding (1) plus i times (2), called the **complex Fourier integral,** is

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(v)e^{iw(x-v)}\,dv\,dw \qquad (i = \sqrt{-1}). \tag{4}$$

It is now only a very short step to our present goal, the Fourier transform.

Fourier Transform and Its Inverse

Writing the exponential function in (4) as a product of exponential functions, we have

$$f(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\left[\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(v)e^{-iwv}\,dv\right]e^{iwx}\,dw. \tag{5}$$

The expression in brackets is a function of w, is denoted by $\hat{f}(w)$, and is called the **Fourier transform** of f; writing $v = x$, we have

$$\hat{f}(w) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(x)e^{-iwx}\,dx. \tag{6}$$

With this, (5) becomes

$$f(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \hat{f}(w)e^{iwx}\,dw \tag{7}$$

and is called the **inverse Fourier transform** of $\hat{f}(w)$.

Another notation for the Fourier transform is

$$\hat{f} = \mathcal{F}(f),$$

so that

$$f = \mathcal{F}^{-1}(\hat{f}).$$

The process of obtaining the Fourier transform $\mathcal{F}(f) = \hat{f}$ from a given f is also called the **Fourier transform** or the *Fourier transform method.*

Conditions sufficient for the existence of the Fourier transform (involving concepts defined in Secs. 6.1 and 11.7) are as follows, as we state without proof.

THEOREM 1

Existence of the Fourier Transform

If $f(x)$ is absolutely integrable on the x-axis and piecewise continuous on every finite interval, then the Fourier transform $\hat{f}(w)$ of $f(x)$ given by (6) exists.

EXAMPLE 1 Fourier Transform

Find the Fourier transform of $f(x) = 1$ if $|x| < 1$ and $f(x) = 0$ otherwise.

Solution. Using (6) and integrating, we obtain

$$\hat{f}(w) = \frac{1}{\sqrt{2\pi}} \int_{-1}^{1} e^{-iwx}\, dx = \frac{1}{\sqrt{2\pi}} \cdot \frac{e^{-iwx}}{-iw} \Big|_{-1}^{1} = \frac{1}{-iw\sqrt{2\pi}} (e^{-iw} - e^{iw}).$$

As in (3) we have $e^{iw} = \cos w + i \sin w$, $e^{-iw} = \cos w - i \sin w$, and by subtraction

$$e^{iw} - e^{-iw} = 2i \sin w.$$

Substituting this in the previous formula on the right, we see that i drops out and we obtain the answer

$$\hat{f}(w) = \sqrt{\frac{\pi}{2}}\, \frac{\sin w}{w}. \quad \blacksquare$$

EXAMPLE 2 Fourier Transform

Find the Fourier transform $\mathcal{F}(e^{-ax})$ of $f(x) = e^{-ax}$ if $x > 0$ and $f(x) = 0$ if $x < 0$; here $a > 0$.

Solution. From the definition (6) we obtain by integration

$$\mathcal{F}(e^{-ax}) = \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} e^{-ax} e^{-iwx}\, dx$$

$$= \frac{1}{\sqrt{2\pi}}\, \frac{e^{-(a+iw)x}}{-(a + iw)} \Big|_{x=0}^{\infty} = \frac{1}{\sqrt{2\pi}\,(a + iw)}.$$

This proves formula 5 of Table III in Sec. 11.10. ∎

Physical Interpretation: Spectrum

The nature of the representation (7) of $f(x)$ becomes clear if we think of it as a superposition of sinusoidal oscillations of all possible frequencies, called a **spectral representation**. This name is suggested by optics, where light is such a superposition of colo (frequencies). In (7), the **"spectral density"** $\hat{f}(w)$ measures the intensity of $f(x)$ in the frequency interval between w and $w + \Delta w$ (Δw small, fixed). We claim that in connection with vibrations, the integral

$$\int_{-\infty}^{\infty} |\hat{f}(w)|^2\, dw$$

can be interpreted as the **total energy** of the physical system. Hence an integral of $|\hat{f}(w)|^2$ from a to b gives the contribution of the frequencies w between a and b to the total energy.

To make this plausible, we begin with a mechanical system giving a single frequenc namely, the harmonic oscillator (mass on a spring, Sec. 2.4)

$$my'' + ky = 0.$$

Here we denote time t by x. Multiplication by y' gives $my'y'' + ky'y = 0$. By integration

$$\tfrac{1}{2}mv^2 + \tfrac{1}{2}ky^2 = E_0 = const$$

where $v = y'$ is the velocity. The first term is the kinetic energy, the second the potenti energy, and E_0 the total energy of the system. Now a general solution is (use (3) Sec. 11.4 with $t = x$)

$$y = a_1 \cos w_0 x + b_1 \sin w_0 x = c_1 e^{iw_0 x} + c_{-1} e^{-iw_0 x}, \qquad w_0{}^2 = k/m$$

where $c_1 = (a_1 - ib_1)/2$, $c_{-1} = \bar{c}_1 = (a_1 + ib_1)/2$. We write simply $A = c_1 e^{iw_0 x}$, $B = c_{-1} e^{-iw_0 x}$. Then $y = A + B$. By differentiation, $v = y' = A' + B' = iw_0(A - B)$. Substitution of v and y on the left side of the equation for E_0 gives

$$E_0 = \tfrac{1}{2}mv^2 + \tfrac{1}{2}ky^2 = \tfrac{1}{2}m(iw_0)^2(A - B)^2 + \tfrac{1}{2}k(A + B)^2.$$

Here $w_0{}^2 = k/m$, as just stated; hence $mw_0{}^2 = k$. Also $i^2 = -1$, so that

$$E_0 = \tfrac{1}{2}k[-(A - B)^2 + (A + B)^2] = 2kAB = 2kc_1 e^{iw_0 x} c_{-1} e^{-iw_0 x} = 2kc_1 c_{-1} = 2k|c_1|^2.$$

Hence *the energy is proportional to the square of the amplitude* $|c_1|$.

As the next step, if a more complicated system leads to a periodic solution $y = f(x)$ that can be represented by a Fourier series, then instead of the single energy term $|c_1|^2$ we get a series of squares $|c_n|^2$ of Fourier coefficients c_n given by (6), Sec. 11.4. In this case we have a **"discrete spectrum"** (or **"point spectrum"**) consisting of countably many isolated frequencies (infinitely many, in general), the corresponding $|c_n|^2$ being the contributions to the total energy.

Finally, a system whose solution can be represented by an integral (7) leads to the above integral for the energy, as is plausible from the cases just discussed.

Linearity. Fourier Transform of Derivatives

New transforms can be obtained from given ones by

THEOREM 2

Linearity of the Fourier Transform

The Fourier transform is a ***linear operation;*** *that is, for any functions $f(x)$ and $g(x)$ whose Fourier transforms exist and any constants a and b, the Fourier transform of $af + bg$ exists, and*

(8) $$\mathcal{F}(af + bg) = a\mathcal{F}(f) + b\mathcal{F}(g).$$

PROOF This is true because integration is a linear operation, so that (6) gives

$$\begin{aligned}\mathcal{F}\{af(x) + bg(x)\} &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} [af(x) + bg(x)] e^{-iwx}\, dx \\ &= a \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{-iwx}\, dx + b \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g(x) e^{-iwx}\, dx \\ &= a\mathcal{F}\{f(x)\} + b\mathcal{F}\{g(x)\}.\end{aligned}$$

∎

In applying the Fourier transform to differential equations, the key property is that differentiation of functions corresponds to multiplication of transforms by iw:

THEOREM 3

Fourier Transform of the Derivative of $f(x)$

Let $f(x)$ be continuous on the x-axis and $f(x) \to 0$ as $|x| \to \infty$. Furthermore, let $f'(x)$ be absolutely integrable on the x-axis. Then

$$\mathcal{F}\{f'(x)\} = iw\mathcal{F}\{f(x)\}. \tag{9}$$

PROOF From the definition of the Fourier transform we have

$$\mathcal{F}\{f'(x)\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f'(x)e^{-iwx}\,dx.$$

Integrating by parts, we obtain

$$\mathcal{F}\{f'(x)\} = \frac{1}{\sqrt{2\pi}} \left[f(x)e^{-iwx} \Big|_{-\infty}^{\infty} - (-iw) \int_{-\infty}^{\infty} f(x)e^{-iwx}\,dx \right].$$

Since $f(x) \to 0$ as $|x| \to \infty$, the desired result follows, namely,

$$\mathcal{F}\{f'(x)\} = 0 + iw\mathcal{F}\{f(x)\}.$$

Two successive applications of (9) give

$$\mathcal{F}(f'') = iw\mathcal{F}(f') = (iw)^2\mathcal{F}(f).$$

Since $(iw)^2 = -w^2$, we have for the transform of the second derivative of f

$$\mathcal{F}\{f''(x)\} = -w^2\mathcal{F}\{f(x)\}. \tag{10}$$

Similarly for higher derivatives.

An application of (10) to differential equations will be given in Sec. 12.6. For the time being we show how (9) can be used to derive transforms.

EXAMPLE 3 **Application of the Operational Formula (9)**

Find the Fourier transform of xe^{-x^2} from Table III, Sec 11.10.

Solution. We use (9). By formula 9 in Table III.

$$\begin{aligned}
\mathcal{F}(xe^{-x^2}) &= \mathcal{F}\left\{-\frac{1}{2}\left(e^{-x^2}\right)'\right\} \\
&= -\frac{1}{2}\,\mathcal{F}\left\{\left(e^{-x^2}\right)'\right\} \\
&= -\frac{1}{2}\,iw\mathcal{F}(e^{-x^2}) \\
&= -\frac{1}{2}\,iw\,\frac{1}{\sqrt{2}}\,e^{-w^2/4} \\
&= -\frac{iw}{2\sqrt{2}}\,e^{-w^2/4}.
\end{aligned}$$

Convolution

The **convolution** $f * g$ of functions f and g is defined by

$$h(x) = (f * g)(x) = \int_{-\infty}^{\infty} f(p)g(x-p)\,dp = \int_{-\infty}^{\infty} f(x-p)g(p)\,dp. \tag{11}$$

The purpose is the same as in the case of Laplace transforms (Sec. 6.5): taking the convolution of two functions and then taking the transform of the convolution is the same as multiplying the transforms of these functions (and multiplying them by $\sqrt{2\pi}$):

THEOREM 4

Convolution Theorem

Suppose that $f(x)$ and $g(x)$ are piecewise continuous, bounded, and absolutely integrable on the x-axis. Then

$$\mathcal{F}(f * g) = \sqrt{2\pi}\,\mathcal{F}(f)\mathcal{F}(g). \tag{12}$$

PROOF By the definition,

$$\mathcal{F}(f * g) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(p)g(x-p)\,dp\,e^{-iwx}\,dx.$$

An interchange of the order of integration gives

$$\mathcal{F}(f * g) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(p)g(x-p)e^{-iwx}\,dx\,dp.$$

Instead of x we now take $x - p = q$ as a new variable of integration. Then $x = p + q$ and

$$\mathcal{F}(f * g) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(p)g(q)e^{-iw(p+q)}\,dq\,dp.$$

This double integral can be written as a product of two integrals and gives the desired result

$$\mathcal{F}(f * g) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(p)e^{-iwp}\,dp \int_{-\infty}^{\infty} g(q)e^{-iwq}\,dq$$

$$= \frac{1}{\sqrt{2\pi}} [\sqrt{2\pi}\,\mathcal{F}(f)][\sqrt{2\pi}\,\mathcal{F}(g)] = \sqrt{2\pi}\,\mathcal{F}(f)\mathcal{F}(g). \qquad \blacksquare$$

By taking the inverse Fourier transform on both sides of (12), writing $\hat{f} = \mathcal{F}(f)$ and $\hat{g} = \mathcal{F}(g)$ as before, and noting that $\sqrt{2\pi}$ and $1/\sqrt{2\pi}$ in (12) and (7) cancel each other, we obtain

$$(f * g)(x) = \int_{-\infty}^{\infty} \hat{f}(w)\hat{g}(w)e^{iwx}\,dw, \tag{13}$$

a formula that will help us in solving partial differential equations (Sec. 12.6).

Discrete Fourier Transform (DFT), Fast Fourier Transform (FFT)

In using Fourier series, Fourier transforms, and trigonometric approximations (Sec. 11.6) we have to assume that a function $f(x)$, to be developed or transformed, is given on some interval, over which we integrate in the Euler formulas, etc. Now very often a function $f(x)$ is given only in terms of values at finitely many points, and one is interested in extending Fourier analysis to this case. The main application of such a "discrete Fourier analysis" concerns large amounts of equally spaced data, as they occur in telecommunication, time series analysis, and various simulation problems. In these situations, dealing with sampled values rather than with functions, we can replace the Fourier transform by the so-called **discrete Fourier transform (DFT)** as follows.

Let $f(x)$ be periodic, for simplicity of period 2π. We assume that N measurements of $f(x)$ are taken over the interval $0 \leqq x \leqq 2\pi$ at regularly spaced points

$$x_k = \frac{2\pi k}{N}, \qquad k = 0, 1, \cdots, N-1. \tag{14}$$

We also say that $f(x)$ is being **sampled** at these points. We now want to determine a **complex trigonometric polynomial**

$$q(x) = \sum_{n=0}^{N-1} c_n e^{inx_k} \tag{15}$$

that **interpolates** $f(x)$ at the nodes (14), that is, $q(x_k) = f(x_k)$, written out, with f_k denoting $f(x_k)$,

$$f_k = f(x_k) = q(x_k) = \sum_{n=0}^{N-1} c_n e^{inx_k}, \qquad k = 0, 1, \cdots, N-1. \tag{16}$$

Hence we must determine the coefficients $c_0, \cdots, c_{N-1}$ such that (16) holds. We do this by an idea similar to that in Sec. 11.1 for deriving the Fourier coefficients by using the orthogonality of the trigonometric system. Instead of integrals we now take sums. Namely, we multiply (16) by e^{-imx_k} (note the minus!) and sum over k from 0 to $N-1$. Then we interchange the order of the two summations and insert x_k from (14). This gives

$$\sum_{k=0}^{N-1} f_k e^{-imx_k} = \sum_{k=0}^{N-1} \sum_{n=0}^{N-1} c_n e^{i(n-m)x_k} = \sum_{n=0}^{N-1} c_n \sum_{k=0}^{N-1} e^{i(n-m)2\pi k/N}. \tag{17}$$

Now

$$e^{i(n-m)2\pi k/N} = \left[e^{i(n-m)2\pi/N}\right]^k.$$

We denote $[\cdots]$ by r. For $n = m$ we have $r = e^0 = 1$. The sum of *these* terms over k equals N, the number of these terms. For $n \neq m$ we have $r \neq 1$ and by the formula for a geometric sum [(6) in Sec. 15.1 with $q = r$ and $n = N-1$]

$$\sum_{k=0}^{N-1} r^k = \frac{1 - r^N}{1 - r} = 0$$

because $r^N = 1$; indeed, since k, m, and n are integers,

$$r^N = e^{i(n-m)2\pi k} = \cos 2\pi k(n-m) + i \sin 2\pi k(n-m) = 1 + 0 = 1.$$

This shows that the right side of (17) equals $c_m N$. Writing n for m and dividing by N, we thus obtain the desired coefficient formula

(18*) $$c_n = \frac{1}{N} \sum_{k=0}^{N-1} f_k e^{-inx_k} \qquad f_k = f(x_k), \quad n = 0, 1, \cdots, N-1.$$

Since computation of the c_n (by the fast Fourier transform, below) involves successive halfing of the problem size N, it is practical to drop the factor $1/N$ from c_n and define the **discrete Fourier transform** of the given signal $\mathbf{f} = [f_0 \quad \cdots \quad f_{N-1}]^T$ to be the vector $\hat{\mathbf{f}} = [\hat{f}_0 \quad \cdots \quad \hat{f}_{N-1}]$ with components

(18) $$\hat{f}_n = Nc_n = \sum_{k=0}^{N-1} f_k e^{-inx_k}, \qquad f_k = f(x_k), \quad n = 0, \cdots, N-1.$$

This is the frequency spectrum of the signal.

In vector notation, $\hat{\mathbf{f}} = \mathbf{F}_N \mathbf{f}$, where the $N \times N$ **Fourier matrix** $\mathbf{F}_N = [e_{nk}]$ has the entries [given in (18)]

(19) $$e_{nk} = e^{-inx_k} = e^{-2\pi ink/N} = w^{nk}, \qquad w = w_N = e^{-2\pi i/N},$$

where $n, k = 0, \cdots, N-1$.

EXAMPLE 4 Discrete Fourier Transform (DFT). Sample of $N = 4$ Values

Let $N = 4$ measurements (sample values) be given. Then $w = e^{-2\pi i/N} = e^{-\pi i/2} = -i$ and thus $w^{nk} = (-i)^{nk}$. Let the sample values be, say $\mathbf{f} = [0 \quad 1 \quad 4 \quad 9]^T$. Then by (18) and (19),

(20) $$\hat{\mathbf{f}} = \mathbf{F}_4 \mathbf{f} = \begin{bmatrix} w^0 & w^0 & w^0 & w^0 \\ w^0 & w^1 & w^2 & w^3 \\ w^0 & w^2 & w^4 & w^6 \\ w^0 & w^3 & w^6 & w^9 \end{bmatrix} \mathbf{f} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -i & -1 & i \\ 1 & -1 & 1 & -1 \\ 1 & i & -1 & -i \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 4 \\ 9 \end{bmatrix} = \begin{bmatrix} 14 \\ -4 + 8i \\ -6 \\ -4 - 8i \end{bmatrix}.$$

From the first matrix in (20) it is easy to infer what $\mathbf{F}_N$ looks like for arbitrary N, which in practice may be 1000 or more, for reasons given below. ■

From the DFT (the frequency spectrum) $\hat{\mathbf{f}} = \mathbf{F}_N \mathbf{f}$ we can recreate the given signal $\mathbf{f} = \mathbf{F}_N^{-1} \mathbf{f}$, as we shall now prove. Here $\mathbf{F}_N$ and its complex conjugate $\overline{\mathbf{F}}_N = \frac{1}{N} [\overline{w}^{nk}]$ satisfy

(21a) $$\overline{\mathbf{F}}_N \mathbf{F}_N = \mathbf{F}_N \overline{\mathbf{F}}_N = N\mathbf{I}$$

where $\mathbf{I}$ is the $N \times N$ unit matrix; hence $\mathbf{F}_N$ has the inverse

(21b) $$\mathbf{F}_N^{-1} = \frac{1}{N} \overline{\mathbf{F}}_N.$$

PROOF We prove (21). By the multiplication rule (row times column) the product matrix $\mathbf{G}_N = \overline{\mathbf{F}}_N\mathbf{F}_N = [g_{jk}]$ in (21a) has the entries $g_{jk} = \textit{Row } j \textit{ of } \overline{\mathbf{F}}_N \textit{ times Column } k \textit{ of } \mathbf{F}_N$. That is, writing $W = \overline{w}^j w^k$, we prove that

$$g_{jk} = (\overline{w}^j w^k)^0 + (\overline{w}^j w^k)^1 + \cdots + (\overline{w}^j w^k)^{N-1}$$

$$= W^0 + W^1 + \cdots + W^{N-1} = \begin{cases} 0 & \text{if} \quad j \neq k \\ N & \text{if} \quad j = k. \end{cases}$$

Indeed, when $j = k$, then $\overline{w}^k w^k = (\overline{w}w)^k = (e^{2\pi i/N}e^{-2\pi i/N})^k = 1^k = 1$, so that the sum of *these* N terms equals N; these are the diagonal entries of $\mathbf{G}_N$. Also, when $j \neq k$, then $W \neq 1$ and we have a geometric sum (whose value is given by (6) in Sec. 15.1 with $q = W$ and $n = N - 1$)

$$W^0 + W^1 + \cdots + W^{N-1} = \frac{1 - W^N}{1 - W} = 0$$

because $W^N = (\overline{w}^j w^k)^N = (e^{2\pi i})^j(e^{-2\pi i})^k = 1^j \cdot 1^k = 1$. ∎

We have seen that $\hat{\mathbf{f}}$ is the frequency spectrum of the signal $f(x)$. Thus the components $\hat{f}_n$ of $\hat{\mathbf{f}}$ give a resolution of the 2π-periodic function $f(x)$ into simple (complex) harmonics. Here one should use only n's that are much smaller than $N/2$, to avoid **aliasing.** By this we mean the effect caused by sampling at too few (equally spaced) points, so that, for instance, in a motion picture, rotating wheels appear as rotating too slowly or even in the wrong sense. Hence in applications, N is usually large. But this poses a problem. Eq. (18) requires $O(N)$ operations for any particular n, hence $O(N^2)$ operations for, say, all $n < N/2$. Thus, already for 1000 sample points the straightforward calculation would involve millions of operations. However, this difficulty can be overcome by the so called **fast Fourier transform (FFT),** for which codes are readily available (e.g. in Maple). The FFT is a computational method for the DFT that needs only $O(N)\log_2 N$ operations instead of $O(N^2)$. It makes the DFT a practical tool for large N. Here one chooses $N = 2^p$ (p integer) and uses the special form of the Fourier matrix to break down the given problem into smaller problems. For instance, when $N = 1000$, those operations are reduced by a factor $1000/\log_2 1000 \approx 100$.

The breakdown produces two problems of size $M = N/2$. This breakdown is possible because for $N = 2M$ we have in (19)

$$w_N^2 = w_{2M}^2 = (e^{-2\pi i/N})^2 = e^{-4\pi i/(2M)} = e^{-2\pi i/M} = w_M.$$

The given vector $\mathbf{f} = [f_0 \quad \cdots \quad f_{N-1}]^\mathsf{T}$ is split into two vectors with M components each, namely, $\mathbf{f}_{\text{ev}} = [f_0 \quad f_2 \quad \cdots \quad f_{N-2}]^\mathsf{T}$ containing the even components of $\mathbf{f}$, and $\mathbf{f}_{\text{od}} = [f_1 \quad f_3 \quad \cdots \quad f_{N-1}]^\mathsf{T}$ containing the odd components of $\mathbf{f}$. For $\mathbf{f}_{\text{ev}}$ and $\mathbf{f}_{\text{od}}$ we determine the DFTs

$$\hat{\mathbf{f}}_{\text{ev}} = \left[\hat{f}_{\text{ev},0} \quad \hat{f}_{\text{ev},2} \quad \cdots \quad \hat{f}_{\text{ev},N-2}\right]^\mathsf{T} = \mathbf{F}_M\mathbf{f}_{\text{ev}}$$

and

$$\hat{\mathbf{f}}_{\text{od}} = \left[\hat{f}_{\text{od},1} \quad \hat{f}_{\text{od},3} \quad \cdots \quad \hat{f}_{\text{od},N-1}\right]^\mathsf{T} = \mathbf{F}_M\mathbf{f}_{\text{od}}$$

involving the same $M \times M$ matrix $\mathbf{F}_M$. From these vectors we obtain the components of the DFT of the given vector $\mathbf{f}$ by the formulas

(22)

$$\text{(a)} \quad \hat{f}_n = \hat{f}_{\text{ev},n} + w_N^n\hat{f}_{\text{od},n} \qquad n = 0, \cdots, M-1$$

$$\text{(b)} \quad \hat{f}_{n+M} = \hat{f}_{\text{ev},n} - w_N^n\hat{f}_{\text{od},n} \qquad n = 0, \cdots, M-1.$$

For $N = 2^p$ this breakdown can be repeated $p - 1$ times in order to finally arrive at $N/2$ problems of size 2 each, so that the number of multiplications is reduced as indicated above.

We show the reduction from $N = 4$ to $M = N/2 = 2$ and then prove (22).

EXAMPLE 5 Fast Fourier Transform (FFT). Sample of $N = 4$ Values

When $N = 4$, then $w = w_N = -i$ as in Example 4 and $M = N/2 = 2$, hence $w = w_M = e^{-2\pi i/2} = e^{-\pi i} = -1$. Consequently,

$$\hat{\mathbf{f}}_{ev} = \begin{bmatrix} \hat{f}_0 \\ \hat{f}_2 \end{bmatrix} = \mathbf{F}_2\mathbf{f}_{ev} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} f_0 \\ f_2 \end{bmatrix} = \begin{bmatrix} f_0 + f_2 \\ f_0 - f_2 \end{bmatrix}$$

$$\hat{\mathbf{f}}_{od} = \begin{bmatrix} \hat{f}_1 \\ \hat{f}_3 \end{bmatrix} = \mathbf{F}_2\mathbf{f}_{od} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} f_1 \\ f_3 \end{bmatrix} = \begin{bmatrix} f_1 + f_3 \\ f_1 - f_3 \end{bmatrix}.$$

From this and (22a) we obtain

$$\hat{f}_0 = \hat{f}_{ev,0} + w_N{}^0\hat{f}_{od,0} = (f_0 + f_2) + (f_1 + f_3) = f_0 + f_1 + f_2 + f_3$$
$$\hat{f}_1 = \hat{f}_{ev,1} + w_N{}^1\hat{f}_{od,1} = (f_0 - f_2) - i(f_1 + f_3) = f_0 - if_1 - f_2 + if_3.$$

Similarly, by (22b),

$$\hat{f}_2 = \hat{f}_{ev,0} - w_N{}^0\hat{f}_{od,0} = (f_0 + f_2) - (f_1 + f_3) \qquad = f_0 - f_1 + f_2 - f_3$$
$$\hat{f}_3 = \hat{f}_{ev,1} - w_N{}^1\hat{f}_{od,1} = (f_0 - f_2) - (-i)(f_1 - f_3) = f_0 + if_1 - f_2 - if_3.$$

This agrees with Example 4, as can be seen by replacing 0, 1, 4, 9 with f_0, f_1, f_2, f_3. ■

We prove (22). From (18) and (19) we have for the components of the DFT

$$\hat{f}_n = \sum_{k=0}^{N-1} w_N^{kn} f_k.$$

Splitting into two sums of $M = N/2$ terms each gives

$$\hat{f}_n = \sum_{k=0}^{M-1} w_N^{2kn} f_{2k} + \sum_{k=0}^{M-1} w_N^{(2k+1)n} f_{2k+1}.$$

We now use $w_N{}^2 = w_M$ and pull out $w_N{}^n$ from under the second sum, obtaining

$$\hat{f}_n = \sum_{k=0}^{M-1} w_M^{kn} f_{ev,k} + w_N{}^n \sum_{k=0}^{M-1} w_M^{kn} f_{od,k}. \tag{23}$$

The two sums are $f_{ev,n}$ and $f_{od,n}$, the components of the "half-size" transforms $\mathbf{F}\,\mathbf{f}_{ev}$ and $\mathbf{F}\,\mathbf{f}_{od}$.

Formula (22a) is the same as (23). In (22b) we have $n + M$ instead of n. This causes a sign change in (23), namely $-w_N{}^n$ before the second sum because

$$w_N{}^M = e^{-2\pi iM/N} = e^{-2\pi i/2} = e^{-\pi i} = -1.$$

This gives the minus in (22b) and completes the proof. ■

PROBLEM SET 11.9

1. (Review) Show that $1/i = -i$, $e^{ix} + e^{-ix} = 2\cos x$, $e^{ix} - e^{-ix} = 2i \sin x$.

2–9 FOURIER TRANSFORMS BY INTEGRATION

Find the Fourier transform of $f(x)$ (without using Table III in Sec. 11.10). Show the details.

2. $f(x) = \begin{cases} e^{kx} & \text{if } x < 0 \quad (k > 0) \\ 0 & \text{if } x > 0 \end{cases}$

3. $f(x) = \begin{cases} k & \text{if } 0 < x < b \\ 0 & \text{otherwise} \end{cases}$

4. $f(x) = \begin{cases} e^{2ix} & \text{if } -1 < x < 1 \\ 0 & \text{otherwise} \end{cases}$

5. $f(x) = \begin{cases} k & \text{if } -1 < x < 1 \\ 0 & \text{otherwise} \end{cases}$

6. $f(x) = \begin{cases} x & \text{if } -1 < x < 1 \\ 0 & \text{otherwise} \end{cases}$

7. $f(x) = \begin{cases} x & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$

8. $f(x) = \begin{cases} xe^{-x} & \text{if } -1 < x < 0 \\ 0 & \text{otherwise} \end{cases}$

9. $f(x) = \begin{cases} -1 & \text{if } -1 < x < 0 \\ 1 & \text{if } \quad 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$

OTHER METHODS

10. Find the Fourier transform of $f(x) = xe^{-x}$ if $x > 0$ and 0 if $x < 0$ from formula 5 in Table III and (9) in the text. *Hint:* Consider xe^{-x} and e^{-x}.

11. Obtain $\mathcal{F}(e^{-x^2/2})$ from formula 9 in Table III.

12. Obtain formula 7 in Table III from formula 8.

13. Obtain formula 1 in Table III from formula 2.

14. TEAM PROJECT. Shifting. (a) Show that if $f(x)$ has a Fourier transform, so does $f(x - a)$, and $\mathcal{F}\{f(x - a)\} = e^{-iwa}\mathcal{F}\{f(x)\}$.

(b) Using (a), obtain formula 1 in Table III, Sec. 11.10, from formula 2.

(c) Shifting on the w-Axis. Show that if $\hat{f}(w)$ is the Fourier transform of $f(x)$, then $\hat{f}(w - a)$ is the Fourier transform of $e^{iax}f(x)$.

(d) Using (c), obtain formula 7 in Table III from 1 and formula 8 from 2.

11.10 Tables of Transforms

Table I. Fourier Cosine Transforms

See (2) in Sec. 11.8.

	$f(x)$	$\hat{f}_c(w) = \mathcal{F}_c(f)$	
1	$\begin{cases} 1 & \text{if } 0 < x < a \\ 0 & \text{otherwise} \end{cases}$	$\sqrt{\frac{2}{\pi}}\,\frac{\sin aw}{w}$	
2	x^{a-1} $(0 < a < 1)$	$\sqrt{\frac{2}{\pi}}\,\frac{\Gamma(a)}{w^a}\cos\frac{a\pi}{2}$	($\Gamma(a)$ see App. A3.1.)
3	e^{-ax} $(a > 0)$	$\sqrt{\frac{2}{\pi}}\left(\frac{a}{a^2 + w^2}\right)$	
4	$e^{-x^2/2}$	$e^{-w^2/2}$	
5	e^{-ax^2} $(a > 0)$	$\frac{1}{\sqrt{2a}}\,e^{-w^2/(4a)}$	
6	$x^n e^{-ax}$ $(a > 0)$	$\sqrt{\frac{2}{\pi}}\,\frac{n!}{(a^2 + w^2)^{n+1}}\,\mathrm{Re}\,(a + iw)^{n+1}$	Re = Real part
7	$\begin{cases} \cos x & \text{if } 0 < x < a \\ 0 & \text{otherwise} \end{cases}$	$\frac{1}{\sqrt{2\pi}}\left[\frac{\sin a(1 - w)}{1 - w} + \frac{\sin a(1 + w)}{1 + w}\right]$	
8	$\cos(ax^2)$ $(a > 0)$	$\frac{1}{\sqrt{2a}}\cos\left(\frac{w^2}{4a} - \frac{\pi}{4}\right)$	
9	$\sin(ax^2)$ $(a > 0)$	$\frac{1}{\sqrt{2a}}\cos\left(\frac{w^2}{4a} + \frac{\pi}{4}\right)$	
10	$\frac{\sin ax}{x}$ $(a > 0)$	$\sqrt{\frac{\pi}{2}}\,(1 - u(w - a))$	(See Sec. 6.3.)
11	$\frac{e^{-x}\sin x}{x}$	$\frac{1}{\sqrt{2\pi}}\arctan\frac{2}{w^2}$	
12	$J_0(ax)$ $(a > 0)$	$\sqrt{\frac{2}{\pi}}\,\frac{1}{\sqrt{a^2 - w^2}}\,(1 - u(w - a))$	(See Secs. 5.5, 6.3.)

Table II. Fourier Sine Transforms

See (5) in Sec. 11.8.

	$f(x)$	$\hat{f}_s(w) = \mathcal{F}_s(f)$
1	$\begin{cases} 1 & \text{if } 0 < x < a \\ 0 & \text{otherwise} \end{cases}$	$\sqrt{\frac{2}{\pi}} \left[\frac{1 - \cos aw}{w} \right]$
2	$1/\sqrt{x}$	$1/\sqrt{w}$
3	$1/x^{3/2}$	$2\sqrt{w}$
4	x^{a-1} $(0 < a < 1)$	$\sqrt{\frac{2}{\pi}} \frac{\Gamma(a)}{w^a} \sin \frac{a\pi}{2}$ ($\Gamma(a)$ see App. A3.1.)
5	e^{-ax} $(a > 0)$	$\sqrt{\frac{2}{\pi}} \left(\frac{w}{a^2 + w^2} \right)$
6	$\frac{e^{-ax}}{x}$ $(a > 0)$	$\sqrt{\frac{2}{\pi}} \arctan \frac{w}{a}$
7	$x^n e^{-ax}$ $(a > 0)$	$\sqrt{\frac{2}{\pi}} \frac{n!}{(a^2 + w^2)^{n+1}} \operatorname{Im} (a + iw)^{n+1}$ Im = Imaginary part
8	$xe^{-x^2/2}$	$we^{-w^2/2}$
9	xe^{-ax^2} $(a > 0)$	$\frac{w}{(2a)^{3/2}} e^{-w^2/4a}$
10	$\begin{cases} \sin x & \text{if } 0 < x < a \\ 0 & \text{otherwise} \end{cases}$	$\frac{1}{\sqrt{2\pi}} \left[\frac{\sin a(1 - w)}{1 - w} - \frac{\sin a(1 + w)}{1 + w} \right]$
11	$\frac{\cos ax}{x}$ $(a > 0)$	$\sqrt{\frac{\pi}{2}}\, u(w - a)$ (See Sec. 6.3.)
12	$\arctan \frac{2a}{x}$ $(a > 0)$	$\sqrt{2\pi} \frac{\sinh aw}{w} e^{-aw}$

Table III. Fourier Transforms

See (6) in Sec. 11.9.

	$f(x)$	$\hat{f}(w) = \mathcal{F}(f)$
1	$\begin{cases} 1 & \text{if } -b < x < b \\ 0 & \text{otherwise} \end{cases}$	$\sqrt{\frac{2}{\pi}}\,\frac{\sin bw}{w}$
2	$\begin{cases} 1 & \text{if } b < x < c \\ 0 & \text{otherwise} \end{cases}$	$\frac{e^{-ibw} - e^{-icw}}{iw\sqrt{2\pi}}$
3	$\frac{1}{x^2 + a^2} \quad (a > 0)$	$\sqrt{\frac{\pi}{2}}\,\frac{e^{-a\lvert w\rvert}}{a}$
4	$\begin{cases} x & \text{if } 0 < x < b \\ 2x - b & \text{if } b < x < 2b \\ 0 & \text{otherwise} \end{cases}$	$\frac{-1 + 2e^{ibw} - e^{-2ibw}}{\sqrt{2\pi}\,w^2}$
5	$\begin{cases} e^{-ax} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases} \quad (a > 0)$	$\frac{1}{\sqrt{2\pi}(a + iw)}$
6	$\begin{cases} e^{ax} & \text{if } b < x < c \\ 0 & \text{otherwise} \end{cases}$	$\frac{e^{(a-iw)c} - e^{(a-iw)b}}{\sqrt{2\pi}(a - iw)}$
7	$\begin{cases} e^{iax} & \text{if } -b < x < b \\ 0 & \text{otherwise} \end{cases}$	$\sqrt{\frac{2}{\pi}}\,\frac{\sin b(w - a)}{w - a}$
8	$\begin{cases} e^{iax} & \text{if } b < x < c \\ 0 & \text{otherwise} \end{cases}$	$\frac{i}{\sqrt{2\pi}}\,\frac{e^{ib(a-w)} - e^{ic(a-w)}}{a - w}$
9	$e^{-ax^2} \quad (a > 0)$	$\frac{1}{\sqrt{2a}}\,e^{-w^2/4a}$
10	$\frac{\sin ax}{x} \quad (a > 0)$	$\sqrt{\frac{\pi}{2}}$ if $\lvert w\rvert < a$; 0 if $\lvert w\rvert > a$

CHAPTER 11 REVIEW QUESTIONS AND PROBLEMS

1. What is a Fourier series? A Fourier sine series? A half-range expansion?
2. Can a discontinuous function have a Fourier series? A Taylor series? Explain.
3. Why did we start with period 2π? How did we proceed to functions of any period p?
4. What is the trigonometric system? Its main property by which we obtained the Euler formulas?
5. What do you know about the convergence of a Fourier series?
6. What is the Gibbs phenomenon?
7. What is approximation by trigonometric polynomials? The minimum square error?
8. What is remarkable about the response of a vibrating system to an *arbitrary* periodic force?
9. What do you know about the Fourier integral? Its applications?
10. What is the Fourier sine transform? Give examples.

11–20 FOURIER SERIES

Find the Fourier series of $f(x)$ as given over one period. Sketch $f(x)$. (Show the details of your work.)

11. $f(x) = \begin{cases} -k & \text{if } -1 < x < 0 \\ k & \text{if } 0 < x < 1 \end{cases}$

12. $f(x) = \begin{cases} 0 & \text{if } -\pi/2 < x < \pi/2 \\ 1 & \text{if } \pi/2 < x < 3\pi/2 \end{cases}$

13. $f(x) = x \quad (-2\pi < x < 2\pi)$

14. $f(x) = |x| \quad (-2 < x < 2)$

15. $f(x) = \begin{cases} x & \text{if } -1 < x < 1 \\ 2 - x & \text{if } 1 < x < 3 \end{cases}$

16. $f(x) = \begin{cases} -1 - x & \text{if } -1 < x < 0 \\ 1 - x & \text{if } 0 < x < 1 \end{cases}$

17. $f(x) = |\sin 8\pi x| \quad (-1/8 < x < 1/8)$

18. $f(x) = e^x \quad (-\pi < x < \pi)$

19. $f(x) = x^2 \quad (-\pi/2 < x < \pi/2)$

20. $f(x) = x \quad (0 < x < 2\pi)$

21–23

Using the answers to suitable odd-numbered problems, find the sum of

21. $1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + - \cdots$

22. $\dfrac{1}{1 \cdot 3} + \dfrac{1}{3 \cdot 5} + \dfrac{1}{5 \cdot 7} + \cdots$

23. $1 + \frac{1}{9} + \frac{1}{25} + \cdots$

24. **(Parseval's identity)** Obtain the result of Prob. 23 by applying Parseval's identity to Prob. 12.

25. What are the sum of the cosine terms and the sum of the sine terms in a Fourier series whose sum is $f(x)$? Give two examples.

26. **(Half-range expansion)** Find the half-range sine series of $f(x) = 0$ if $0 < x < \pi/2$, $f(x) = 1$ if $\pi/2 < x < \pi$. Compare with Prob. 12.

27. **(Half-range cosine series)** Find the half-range cosine series of $f(x) = x$ $(0 < x < 2\pi)$. Compare with Prob. 20.

28–29 MINIMUM SQUARE ERROR

Compute the minimum square errors for the trigonometric polynomials of degree $N = 1, \cdots, 8$:

28. For $f(x)$ in Prob. 12.

29. For $f(x) = x$ $(-\pi < x < \pi)$.

30–31 GENERAL SOLUTION

Solve $y'' + \omega^2 y = r(t)$, where $|\omega| \neq 0, 1, 2, \cdots$, $r(t)$ is 2π-periodic and:

30. $r(t) = t(\pi^2 - t^2) \quad (-\pi < t < \pi)$

31. $r(t) = t^2 \quad (-\pi < t < \pi)$

32–37 FOURIER INTEGRALS AND TRANSFORMS

Sketch the given function and represent it as indicated. If you have a CAS, graph approximate curves obtained by replacing ∞ with finite limits; also look for Gibbs phenomena.

32. $f(x) = 1$ if $1 < x < 2$ and 0 otherwise, by a Fourier integral

33. $f(x) = x$ if $0 < x < 1$ and 0 otherwise, by a Fourier integral

34. $f(x) = 1 + x/2$ if $-2 < x < 0$, $f(x) = 1 - x/2$ if $0 < x < 2$, $f(x) = 0$ otherwise, by a Fourier cosine integral

35. $f(x) = -1 - x/2$ if $-2 < x < 0$, $f(x) = 1 - x/2$ if $0 < x < 2$, $f(x) = 0$ otherwise, by a Fourier sine integral

36. $f(x) = -4 + x^2$ if $-2 < x < 0$, $f(x) = 4 - x^2$ if $0 < x < 2$, $f(x) = 0$ otherwise, by a Fourier sine integral

37. $f(x) = 4 - x^2$ if $-2 < x < 2$, $f(x) = 0$ otherwise, by a Fourier cosine integral

38. Find the Fourier transform of $f(x) = k$ if $a < x < b$, $f(x) = 0$ otherwise.

39. Find the Fourier cosine transform of $f(x) = e^{-2x}$ if $x > 0$, $f(x) = 0$ if $x < 0$.

40. Find $\mathcal{F}_c(e^{-2x})$ and $\mathcal{F}_s(e^{-2x})$ by formulas involving second derivatives.

SUMMARY OF CHAPTER 11

Fourier Series, Integrals, Transforms

Fourier series concern **periodic functions** $f(x)$ of period $p = 2L$, that is, by definition $f(x + p) = f(x)$ for all x and some fixed $p > 0$; thus, $f(x + np) = f(x)$ for any integer n. These series are of the form

$$(1) \qquad f(x) = a_0 + \sum_{n=1}^{\infty} \left(a_n \cos \frac{n\pi}{L} x + b_n \sin \frac{n\pi}{L} x \right) \qquad \text{(Sec. 11.2)}$$

with coefficients, called the **Fourier coefficients** of $f(x)$, given by the Euler formulas (Sec. 11.2)

$$(2) \qquad a_0 = \frac{1}{2L} \int_{-L}^{L} f(x)\, dx, \qquad a_n = \frac{1}{L} \int_{-L}^{L} f(x) \cos \frac{n\pi x}{L}\, dx$$

$$b_n = \frac{1}{L} \int_{-L}^{L} f(x) \sin \frac{n\pi x}{L}\, dx$$

where $n = 1, 2, \cdots$. For period 2π we simply have (Sec. 11.1)

$$(1^*) \qquad f(x) = a_0 + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \sin nx)$$

with the *Fourier coefficients* of $f(x)$ (Sec. 11.1)

$$a_0 = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x)\, dx, \quad a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos nx\, dx, \quad b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin nx\, dx.$$

Fourier series are fundamental in connection with periodic phenomena, particularly in models involving differential equations (Sec. 11.5, Chap. 12). If $f(x)$ is even $[f(-x) = f(x)]$ or odd $[f(-x) = -f(x)]$, they reduce to **Fourier cosine** or **Fourier sine series,** respectively (Sec. 11.3). If $f(x)$ is given for $0 \leqq x \leqq L$ only, it has two **half-range expansions** of period $2L$, namely, a cosine and a sine series (Sec. 11.3).

The set of cosine and sine functions in (1) is called the **trigonometric system.** Its most basic property is its **orthogonality** on an interval of length $2L$; that is, for all integers m and $n \neq m$ we have

$$\int_{-L}^{L} \cos \frac{m\pi x}{L} \cos \frac{n\pi x}{L}\, dx = 0, \qquad \int_{-L}^{L} \sin \frac{m\pi x}{L} \sin \frac{n\pi x}{L}\, dx = 0$$

and for all integers m and n,

$$\int_{-L}^{L} \cos \frac{m\pi x}{L} \sin \frac{n\pi x}{L}\, dx = 0.$$

This orthogonality was crucial in deriving the Euler formulas (2).

Partial sums of Fourier series minimize the **square error** (Sec. 11.6).

Ideas and techniques of Fourier series extend to nonperiodic functions $f(x)$ defined on the entire real line; this leads to the **Fourier integral**

$$(3) \qquad f(x) = \int_0^{\infty} [A(w) \cos wx + B(w) \sin wx]\, dw \qquad \text{(Sec. 11.7)}$$

where

$$(4) \qquad A(w) = \frac{1}{\pi} \int_{-\infty}^{\infty} f(v) \cos wv\, dv, \qquad B(w) = \frac{1}{\pi} \int_{-\infty}^{\infty} f(v) \sin wv\, dv$$

or, in complex form (Sec. 11.9),

$$(5) \qquad f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{f}(w) e^{iwx}\, dw \qquad (i = \sqrt{-1})$$

where

$$(6) \qquad \hat{f}(w) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{-iwx}\, dx.$$

Formula (6) transforms $f(x)$ into its **Fourier transform** $\hat{f}(w)$, and (5) is the inverse transform.

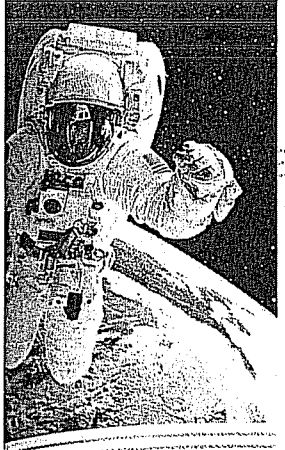
Related to this are the **Fourier cosine transform** (Sec. 11.8)

$$(7) \qquad \hat{f}_c(w) = \sqrt{\frac{2}{\pi}} \int_0^{\infty} f(x) \cos wx\, dx$$

and the **Fourier sine transform** (Sec. 11.8)

$$(8) \qquad \hat{f}_s(w) = \sqrt{\frac{2}{\pi}} \int_0^{\infty} f(x) \sin wx\, dx.$$

The **discrete Fourier transform (DFT)** and a practical method of computing it, called the **fast Fourier transform (FFT),** are discussed in Sec. 11.9.



CHAPTER 12

Partial Differential Equations (PDEs)

PDEs are models of various physical and geometrical problems, arising when the unknown functions (the solutions) depend on two or more variables, usually on time t and one or several space variables. It is fair to say that only the simplest physical systems can be modeled by ODEs, whereas most problems in dynamics, elasticity, heat transfer, electromagnetic theory, and quantum mechanics require PDEs. Indeed, the range of applications of PDEs is enormous, compared to that of ODEs.

In this chapter we concentrate on the most important PDEs of applied mathematics, the wave equations governing the vibrating string (Sec. 12.2) and the vibrating membrane (Sec. 12.7), the heat equation (Sec. 12.5), and the Laplace equation (Secs. 12.5, 12.10). We derive these PDEs from physics and consider methods for solving **initial and boundary value problems,** that is, methods of obtaining solutions satisfying conditions that are given by the physical situation.

In Secs. 12.6 and 12.11 we show that PDEs can also be solved by Fourier and Laplace transform methods.

COMMENT. ***Numerics for PDEs*** is explained in Secs. 21.4–21.7.

Prerequisites: Linear ODEs (Chap. 2), Fourier series (Chap. 11)
Sections that may be omitted in a shorter course: 12.6, 12.9–12.11
References and Answers to Problems: App. 1 Part C, App. 2

12.1 Basic Concepts

A **partial differential equation (PDE)** is an equation involving one or more partial derivatives of an (unknown) function, call it u, that depends on two or more variables, often time t and one or several variables in space. The order of the highest derivative is called the **order** of the PDE. As for ODEs, second-order PDEs will be the most important ones in applications.

Just as for ordinary differential equations (ODEs) we say that a PDE is **linear** if it is of the first degree in the unknown function u and its partial derivatives. Otherwise we call it **nonlinear.** Thus, all the equations in Example 1 on p. 536 are linear. We call a *linear* PDE **homogeneous** if each of its terms contains either u or one of its partial derivatives. Otherwise we call the equation **nonhomogeneous.** Thus, (4) in Example 1 (with f not identically zero) is nonhomogeneous, whereas the other equations are homogeneous.

EXAMPLE 1 **Important Second-Order PDEs**

(1) $$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}$$ *One-dimensional wave equation*

(2) $$\frac{\partial u}{\partial t} = c^2 \frac{\partial^2 u}{\partial x^2}$$ *One-dimensional heat equation*

(3) $$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$$ *Two-dimensional Laplace equation*

(4) $$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x, y)$$ *Two-dimensional Poisson equation*

(5) $$\frac{\partial^2 u}{\partial t^2} = c^2 \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right)$$ *Two-dimensional wave equation*

(6) $$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = 0$$ *Three-dimensional Laplace equation*

Here c is a positive constant, t is time, x, y, z are Cartesian coordinates, and *dimension* is the number of these coordinates in the equation. ▦

A **solution** of a PDE in some region R of the space of the independent variables is a function that has all the partial derivatives appearing in the PDE in some domain D (definition in Sec. 9.6) containing R, and satisfies the PDE everywhere in R.

Often one merely requires that the function is continuous on the boundary of R, has those derivatives in the interior of R, and satisfies the PDE in the interior of R. Letting R lie in D simplifies the situation regarding derivatives on the boundary of R, which is then the same on the boundary as it is in the interior of R.

In general, the totality of solutions of a PDE is very large. For example, the functions

(7) $$u = x^2 - y^2, \qquad u = e^x \cos y, \qquad u = \sin x \cosh y, \qquad u = \ln (x^2 + y^2)$$

which are entirely different from each other, are solutions of (3), as you may verify. We shall see later that the unique solution of a PDE corresponding to a given physical problem will be obtained by the use of **additional conditions** arising from the problem. For instance, this may be the condition that the solution u assume given values on the boundary of the region R (**"boundary conditions"**). Or, when time t is one of the variables, u (or $u_t = \partial u/\partial t$ or both) may be prescribed at $t = 0$ (**"initial conditions"**).

We know that if an ODE is linear and homogeneous, then from known solutions we can obtain further solutions by superposition. For PDEs the situation is quite similar:

THEOREM 1

Fundamental Theorem on Superposition

If u_1 and u_2 are solutions of a ***homogeneous linear*** *PDE in some region R, then*

$$u = c_1 u_1 + c_2 u_2$$

with any constants c_1 and c_2 is also a solution of that PDE in the region R.

The simple proof of this important theorem is quite similar to that of Theorem 1 in Sec. 2.1 and is left to the student.

Verification of solutions in Probs. 14–25 proceeds as for ODEs. Problems 1–12 concern PDEs solvable like ODEs. To help the student with them, we consider two typical examples.

EXAMPLE 2 **Solving $u_{xx} - u = 0$ Like an ODE**

Find solutions u of the PDE $u_{xx} - u = 0$ depending on x and y.

Solution. Since no y-derivatives occur, we can solve this PDE like $u'' - u = 0$. In Sec. 2.2 we would have obtained $u = Ae^{x} + Be^{-x}$ with constant A and B. Here A and B may be functions of y, so that the answer is

$$u(x, y) = A(y)e^{x} + B(y)e^{-x}$$

with arbitrary functions A and B. We thus have a great variety of solutions. Check the result by differentiation. ■

EXAMPLE 3 **Solving $u_{xy} = -u_x$ Like an ODE**

Find solutions $u = u(x, y)$ of this PDE.

Solution. Setting $u_x = p$, we have $p_y = -p$, $p_y/p = -1$, $\ln p = -y + \tilde{c}(x)$, $p = c(x)e^{-y}$ and by integration with respect to x,

$$u(x, y) = f(x)e^{-y} + g(y) \qquad \text{where} \qquad f(x) = \int c(x)\,dx;$$

here, $f(x)$ and $g(y)$ are arbitrary. ■

PROBLEM SET 12.1

1–12 PDEs SOLVABLE AS ODEs

This happens if a PDE involves derivatives with respect to one variable only (or can be transformed to such a form), so that the other variable(s) can be treated as parameter(s). Solve for $u = u(x, y)$:

1. $u_{yy} + 16u = 0$
2. $u_{xx} = u$
3. $u_{yy} = 0$
4. $u_y + 2yu = 0$
5. $u_y + u = e^{xy}$
6. $u_{xx} = 4y^2u$
7. $u_y = (\cosh x)yu$
8. $u_y = 2xyu$
9. $y^2u_{yy} + 2yu_y - 2u = 0$
10. $u_{yy} = 4xu_y$
11. $u_{xy} = u_x$
12. $u_{yy} + 10u_y + 25u = e^{-5y}$

13. **(Fundamental Theorem)** Prove Fundamental Theorem 1 for second-order PDEs in two and three independent variables.

14–25 VERIFICATION OF SOLUTIONS

Verify (by substitution) that the given function is a solution of the indicated PDE. Sketch or graph the solution as a surface in space.

14–17 **Wave Equation (1)** with suitable c

14. $u = 4x^2 + t^2$
15. $u = \sin 8x \cos 2t$
16. $u = \sin 3x \sin 18t$
17. $u = \sin kx \cos kct$

18–21 **Heat Equation (2)** with suitable c

18. $u = e^{-2kt} \cos 8x$
19. $u = e^{-\pi^2 t} \sin 4x$
20. $u = e^{-4\omega^2 t} \sin \omega x$
21. $u = e^{-\omega^2 c^2 t} \cos \omega x$

22–25 **Laplace Equation (3)**

22. u in (7) in the text
23. $u = \cos 2y \sinh 2x$
24. $u = \arctan(y/x)$
25. $u = e^{x^2 - y^2} \sin 2xy$

26. **TEAM PROJECT. Verification of Solutions**
(a) Wave equation. Verify that $u(x, t) = v(x + ct) + w(x - ct)$ with any twice differentiable functions v and w satisfies (1).
(b) Poisson equation. Verify that each u satisfies (4) with $f(x, y)$ as indicated.

$u = x^4 + y^4$	$f = 12(x^2 + y^2)$
$u = \cos x \sin y$	$f = -2 \cos x \sin y$
$u = y/x$	$f = 2y/x^3$

(c) Laplace equation. Verify that $u = 1/\sqrt{x^2 + y^2 + z^2}$ satisfies (6) and $u = \ln(x^2 + y^2)$ satisfies (3). Is $u = 1/\sqrt{x^2 + y^2}$ a solution of (3)? Of what Poisson equation?

(d) Verify that u with any (sufficiently often differentiable) v and w satisfies the given PDE.

$u = v(x) + w(y)$ $\quad u_{xy} = 0$

$u = v(x)w(y)$ $\quad uu_{xy} = u_x u_y$

$u = v(x + 3t) + w(x - 3t)$ $\quad u_{tt} = 9u_{xx}$

27. (Boundary value problem) Verify that the function $u(x, y) = a \ln(x^2 + y^2) + b$ satisfies Laplace's equation (3) and determine a and b so that u satisfies the boundary conditions $u = 110$ on the circle $x^2 + y^2 = 1$ and $u = 0$ on the circle $x^2 + y^2 = 100$.

28–30 SYSTEMS OF PDEs

Solve

28. $u_x = 0,\ u_y = 0$

29. $u_{xx} = 0,\ u_{xy} = 0$

30. $u_{xx} = 0,\ u_{yy} = 0$

12.2 Modeling: Vibrating String, Wave Equation

As a first important PDE let us derive the equation modeling small transverse vibrations of an elastic string, such as a violin string. We place the string along the x-axis, stretch it to length L, and fasten it at the ends $x = 0$ and $x = L$. We then distort the string, and at some instant, call it $t = 0$, we release it and allow it to vibrate. The problem is to determine the vibrations of the string, that is, to find its deflection $u(x, t)$ at any point x and at any time $t > 0$; see Fig. 283.

$u(x, t)$ will be the solution of a PDE that is the model of our physical system to be derived. This PDE should not be too complicated, so that we can solve it. Reasonable simplifying assumptions (just as for ODEs modeling vibrations in Chap. 2) are as follows.

Physical Assumptions

1. The mass of the string per unit length is constant ("homogeneous string"). The string is perfectly elastic and does not offer any resistance to bending.
2. The tension caused by stretching the string before fastening it at the ends is so large that the action of the gravitational force on the string (trying to pull the string down a little) can be neglected.
3. The string performs small transverse motions in a vertical plane; that is, every particle of the string moves strictly vertically and so that the deflection and the slope at every point of the string always remain small in absolute value.

Under these assumptions we may expect solutions $u(x, t)$ that describe the physical reality sufficiently well.

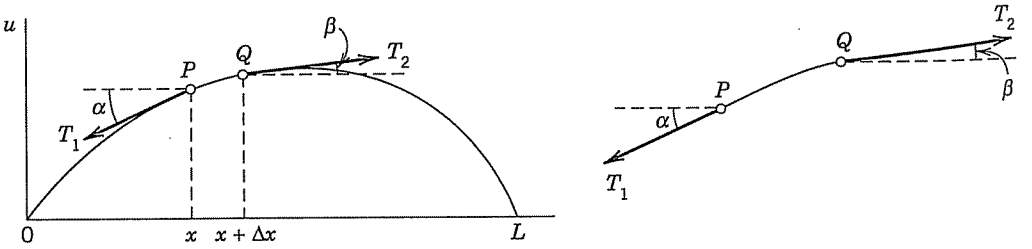


Fig. 283. Deflected string at fixed time t. Explanation on p. 539

Derivation of the PDE of the Model ("Wave Equation") from Forces

The model of the vibrating string will consist of a PDE ("wave equation") and additional conditions. To obtain the PDE, we consider the ***forces acting on a small portion of the string*** (Fig. 283). This method is typical of modeling in mechanics and elsewhere.

Since the string offers no resistance to bending, the tension is tangential to the curve of the string at each point. Let T_1 and T_2 be the tension at the endpoints P and Q of that portion. Since the points of the string move vertically, there is no motion in the horizontal direction. Hence the horizontal components of the tension must be constant. Using the notation shown in Fig. 283, we thus obtain

$$T_1 \cos \alpha = T_2 \cos \beta = T = const. \tag{1}$$

In the vertical direction we have two forces, namely, the vertical components $-T_1 \sin \alpha$ and $T_2 \sin \beta$ of T_1 and T_2; here the minus sign appears because the component at P is directed downward. By **Newton's second law** the resultant of these two forces is equal to the mass $\rho \Delta x$ of the portion times the acceleration $\partial^2 u/\partial t^2$, evaluated at some point between x and $x + \Delta x$; here ρ is the mass of the undeflected string per unit length, and Δx is the length of the portion of the undeflected string. (Δ is generally used to denote small quantities; this has nothing to do with the Laplacian ∇^2, which is sometimes also denoted by Δ.) Hence

$$T_2 \sin \beta - T_1 \sin \alpha = \rho \Delta x \frac{\partial^2 u}{\partial t^2}.$$

Using (1), we can divide this by $T_2 \cos \beta = T_1 \cos \alpha = T$, obtaining

$$\frac{T_2 \sin \beta}{T_2 \cos \beta} - \frac{T_1 \sin \alpha}{T_1 \cos \alpha} = \tan \beta - \tan \alpha = \frac{\rho \Delta x}{T} \frac{\partial^2 u}{\partial t^2}. \tag{2}$$

Now $\tan \alpha$ and $\tan \beta$ are the slopes of the string at x and $x + \Delta x$:

$$\tan \alpha = \left(\frac{\partial u}{\partial x}\right)\bigg|_x \qquad \text{and} \qquad \tan \beta = \left(\frac{\partial u}{\partial x}\right)\bigg|_{x+\Delta x}.$$

Here we have to write *partial* derivatives because u depends also on time t. Dividing (2) by Δx, we thus have

$$\frac{1}{\Delta x}\left[\left(\frac{\partial u}{\partial x}\right)\bigg|_{x+\Delta x} - \left(\frac{\partial u}{\partial x}\right)\bigg|_x\right] = \frac{\rho}{T} \frac{\partial^2 u}{\partial t^2}.$$

If we let Δx approach zero, we obtain the linear PDE

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}, \qquad c^2 = \frac{T}{\rho}. \tag{3}$$

This is called the **one-dimensional wave equation.** We see that it is homogeneous and of the second order. The physical constant T/ρ is denoted by c^2 (instead of c) to indicate

that this constant is *positive,* a fact that will be essential to the form of the solutions. "One-dimensional" means that the equation involves only one space variable, x. In the next section we shall complete setting up the model and then show how to solve it by a general method that is probably the most important one for PDEs in engineering mathematics.

12.3 Solution by Separating Variables. Use of Fourier Series

The model of a vibrating elastic string (a violin string, for instance) consists of the **one-dimensional wave equation**

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2} \qquad c^2 = \frac{T}{\rho} \tag{1}$$

for the unknown deflection $u(x, t)$ of the string, a PDE that we have just obtained, and some ***additional conditions,*** which we shall now derive.

Since the string is fastened at the ends $x = 0$ and $x = L$ (see Sec. 12.2), we have the two **boundary conditions**

$$\text{(a)}\quad u(0, t) = 0, \qquad \text{(b)}\quad u(L, t) = 0 \qquad \text{for all } t. \tag{2}$$

Furthermore, the form of the motion of the string will depend on its *initial deflection* (deflection at time $t = 0$), call it $f(x)$, and on its *initial velocity* (velocity at $t = 0$), call it $g(x)$. We thus have the two **initial conditions**

$$\text{(a)}\quad u(x, 0) = f(x), \qquad \text{(b)}\quad u_t(x, 0) = g(x) \qquad (0 \leqq x \leqq L) \tag{3}$$

where $u_t = \partial u/\partial t$. We now have to find a solution of the PDE (1) satisfying the conditions (2) and (3). This will be the solution of our problem. We shall do this in three steps, as follows.

Step 1. By the **"method of separating variables"** or *product method,* setting $u(x, t) = F(x)G(t)$, we obtain from (1) two ODEs, one for $F(x)$ and the other one for $G(t)$.

Step 2. We determine solutions of these ODEs that satisfy the boundary conditions (2).

Step 3. Finally, using **Fourier series,** we compose the solutions gained in Step 2 to obtain a solution of (1) satisfying both (2) and (3), that is, the solution of our model of the vibrating string.

Step 1. Two ODEs from the Wave Equation (1)

In the **method of separating variables,** or *product method,* we determine solutions of the wave equation (1) of the form

$$u(x, t) = F(x)G(t) \tag{4}$$

which are a product of two functions, each depending only on one of the variables x and t. This is a powerful general method that has various applications in engineering mathematics, as we shall see in this chapter. Differentiating (4), we obtain

$$\frac{\partial^2 u}{\partial t^2} = F\ddot{G} \qquad \text{and} \qquad \frac{\partial^2 u}{\partial x^2} = F''G$$

where dots denote derivatives with respect to t and primes derivatives with respect to x. By inserting this into the wave equation (1) we have

$$F\ddot{G} = c^2 F''G.$$

Dividing by c^2FG and simplifying gives

$$\frac{\ddot{G}}{c^2 G} = \frac{F''}{F}.$$

The variables are now separated, the left side depending only on t and the right side only on x. Hence both sides must be constant because if they were variable, then changing t or x would affect only one side, leaving the other unaltered. Thus, say,

$$\frac{\ddot{G}}{c^2 G} = \frac{F''}{F} = k.$$

Multiplying by the denominators gives immediately two ***ordinary*** DEs

(5) $$F'' - kF = 0$$

and

(6) $$\ddot{G} - c^2 kG = 0.$$

Here, the **separation constant** k is still arbitrary.

Step 2. Satisfying the Boundary Conditions (2)

We now determine solutions F and G of (5) and (6) so that $u = FG$ satisfies the boundary conditions (2), that is,

(7) $$u(0, t) = F(0)G(t) = 0, \qquad u(L, t) = F(L)G(t) = 0 \qquad \text{for all } t.$$

We first solve (5). If $G \equiv 0$, then $u = FG \equiv 0$, which is of no interest. Hence $G \not\equiv 0$ and then by (7),

(8) $$\text{(a)} \quad F(0) = 0, \qquad \text{(b)} \quad F(L) = 0.$$

We show that k must be negative. For $k = 0$ the general solution of (5) is $F = ax + b$, and from (8) we obtain $a = b = 0$, so that $F \equiv 0$ and $u = FG \equiv 0$, which is of no interest. For positive $k = \mu^2$ a general solution of (5) is

$$F = Ae^{\mu x} + Be^{-\mu x}$$

and from (8) we obtain $F \equiv 0$ as before (verify!). Hence we are left with the possibility of choosing k negative, say, $k = -p^2$. Then (5) becomes $F'' + p^2F = 0$ and has as a general solution

$$F(x) = A \cos px + B \sin px.$$

From this and (8) we have

$$F(0) = A = 0 \qquad \text{and then} \qquad F(L) = B \sin pL = 0.$$

We must take $B \neq 0$ since otherwise $F \equiv 0$. Hence $\sin pL = 0$. Thus

$$pL = n\pi, \qquad \text{so that} \qquad p = \frac{n\pi}{L} \qquad (n \text{ integer}). \tag{9}$$

Setting $B = 1$, we thus obtain infinitely many solutions $F(x) = F_n(x)$, where

$$F_n(x) = \sin \frac{n\pi}{L} x \qquad (n = 1, 2, \cdots). \tag{10}$$

These solutions satisfy (8). [For negative integer n we obtain essentially the same solutions, except for a minus sign, because $\sin(-\alpha) = -\sin \alpha$.]

We now solve (6) with $k = -p^2 = -(n\pi/L)^2$ resulting from (9), that is,

$$\ddot{G} + \lambda_n{}^2 G = 0 \qquad \text{where} \qquad \lambda_n = cp = \frac{cn\pi}{L}. \tag{11*}$$

A general solution is

$$G_n(t) = B_n \cos \lambda_n t + B_n{}^* \sin \lambda_n t.$$

Hence solutions of (1) satisfying (2) are $u_n(x, t) = F_n(x)G_n(t) = G_n(t)F_n(x)$, written out

$$u_n(x, t) = (B_n \cos \lambda_n t + B_n{}^* \sin \lambda_n t) \sin \frac{n\pi}{L} x \qquad (n = 1, 2, \cdots). \tag{11}$$

These functions are called the **eigenfunctions,** or *characteristic functions,* and the values $\lambda_n = cn\pi/L$ are called the **eigenvalues,** or *characteristic values,* of the vibrating string. The set $\{\lambda_1, \lambda_2, \cdots\}$ is called the **spectrum.**

Discussion of Eigenfunctions. We see that each u_n represents a harmonic motion having the **frequency** $\lambda_n/2\pi = cn/2L$ cycles per unit time. This motion is called the nth **normal mode** of the string. The first normal mode is known as the *fundamental mode* ($n = 1$), and the others are known as *overtones;* musically they give the octave, octave plus fifth, etc. Since in (11)

$$\sin \frac{n\pi x}{L} = 0 \qquad \text{at} \qquad x = \frac{L}{n}, \frac{2L}{n}, \cdots, \frac{n-1}{n} L,$$

the nth normal mode has $n - 1$ **nodes,** that is, points of the string that do not move (in addition to the fixed endpoints); see Fig. 284.

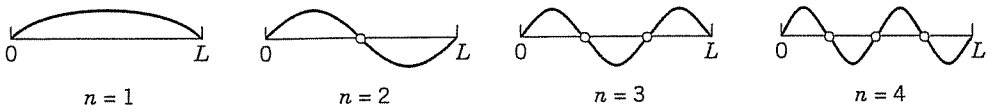


Fig. 284. Normal modes of the vibrating string

Figure 285 shows the second normal mode for various values of t. At any instant the string has the form of a sine wave. When the left part of the string is moving down, the other half is moving up, and conversely. For the other modes the situation is similar.

Tuning is done by changing the tension T. Our formula for the frequency $\lambda_n/2\pi = cn/2L$ of u_n with $c = \sqrt{T/\rho}$ [see (3), Sec. 12.2] confirms that effect because it shows that the frequency is proportional to the tension. T cannot be increased indefinitely, but can you see what to do to get a string with a high fundamental mode? (Think of both L and ρ.) Why is a violin smaller than a double-bass?

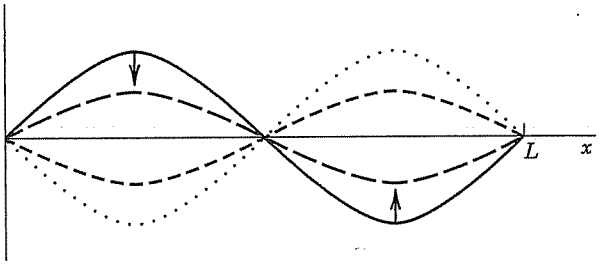


Fig. 285. Second normal mode for various values of t

Step 3. Solution of the Entire Problem. Fourier Series

The eigenfunctions (11) satisfy the wave equation (1) and the boundary conditions (2) (string fixed at the ends). A single u_n will generally not satisfy the initial conditions (3). But since the wave equation (1) is linear and homogeneous, it follows from Fundamental Theorem 1 in Sec. 12.1 that the sum of finitely many solutions u_n is a solution of (1). To obtain a solution that also satisfies the initial conditions (3), we consider the infinite series (with $\lambda_n = cn\pi/L$ as before)

$$\textbf{(12)} \qquad u(x, t) = \sum_{n=1}^{\infty} u_n(x, t) = \sum_{n=1}^{\infty} (B_n \cos \lambda_n t + B_n^* \sin \lambda_n t) \sin \frac{n\pi}{L} x.$$

Satisfying Initial Condition (3a) (Given Initial Displacement). From (12) and (3a) we obtain

$$(13) \qquad u(x, 0) = \sum_{n=1}^{\infty} B_n \sin \frac{n\pi}{L} x = f(x).$$

Hence we must choose the B_n's so that $u(x, 0)$ becomes the **Fourier sine series** of $f(x)$. Thus, by (4) in Sec. 11.3,

$$\textbf{(14)} \qquad B_n = \frac{2}{L} \int_0^L f(x) \sin \frac{n\pi x}{L} dx, \qquad n = 1, 2, \cdots.$$

Satisfying Initial Condition (3b) (Given Initial Velocity). Similarly, by differentiating (12) with respect to t and using (3b), we obtain

$$\left.\frac{\partial u}{\partial t}\right|_{t=0} = \left[\sum_{n=1}^{\infty} (-B_n\lambda_n \sin \lambda_n t + B_n{}^*\lambda_n \cos \lambda_n t) \sin \frac{n\pi x}{L}\right]_{t=0}$$

$$= \sum_{n=1}^{\infty} B_n{}^*\lambda_n \sin \frac{n\pi x}{L} = g(x).$$

Hence we must choose the $B_n{}^*$'s so that for $t = 0$ the derivative $\partial u/\partial t$ becomes the Fourier sine series of $g(x)$. Thus, again by (4) in Sec. 11.3,

$$B_n{}^*\lambda_n = \frac{2}{L}\int_0^L g(x) \sin \frac{n\pi x}{L}\, dx.$$

Since $\lambda_n = cn\pi/L$, we obtain by division

(15)
$$B_n{}^* = \frac{2}{cn\pi}\int_0^L g(x) \sin \frac{n\pi x}{L}\, dx, \qquad n = 1, 2, \cdots.$$

Result. Our discussion shows that $u(x, t)$ given by (12) with coefficients (14) and (15) is a solution of (1) that satisfies all the conditions in (2) and (3), provided the series (12) converges and so do the series obtained by differentiating (12) twice termwise with respect to x and t and have the sums $\partial^2 u/\partial x^2$ and $\partial^2 u/\partial t^2$, respectively, which are continuous.

Solution (12) Established. According to our derivation the solution (12) is at first a purely formal expression, but we shall now establish it. For the sake of simplicity we consider only the case when the initial velocity $g(x)$ is identically zero. Then the $B_n{}^*$ are zero, and (12) reduces to

(16)
$$u(x, t) = \sum_{n=1}^{\infty} B_n \cos \lambda_n t \sin \frac{n\pi x}{L}, \qquad \lambda_n = \frac{cn\pi}{L}.$$

It is possible to ***sum this series,*** that is, to write the result in a closed or finite form. For this purpose we use the formula [see (11), App. A3.1]

$$\cos \frac{cn\pi}{L} t \sin \frac{n\pi}{L} x = \frac{1}{2}\left[\sin\left\{\frac{n\pi}{L}(x - ct)\right\} + \sin\left\{\frac{n\pi}{L}(x + ct)\right\}\right].$$

Consequently, we may write (16) in the form

$$u(x, t) = \frac{1}{2}\sum_{n=1}^{\infty} B_n \sin\left\{\frac{n\pi}{L}(x - ct)\right\} + \frac{1}{2}\sum_{n=1}^{\infty} B_n \sin\left\{\frac{n\pi}{L}(x + ct)\right\}.$$

These two series are those obtained by substituting $x - ct$ and $x + ct$, respectively, for the variable x in the Fourier sine series (13) for $f(x)$. Thus

(17)
$$u(x, t) = \tfrac{1}{2}[f^*(x - ct) + f^*(x + ct)]$$

where f^* is the odd periodic extension of f with the period $2L$ (Fig. 286). Since the initial deflection $f(x)$ is continuous on the interval $0 \leqq x \leqq L$ and zero at the endpoints, it follows from (17) that $u(x, t)$ is a continuous function of both variables x and t for all values of the variables. By differentiating (17) we see that $u(x, t)$ is a solution of (1), provided $f(x)$ is twice differentiable on the interval $0 < x < L$, and has one-sided second derivatives at $x = 0$ and $x = L$, which are zero. Under these conditions $u(x, t)$ is established as a solution of (1), satisfying (2) and (3) with $g(x) \equiv 0$. ■

Fig. 286. Odd periodic extension of $f(x)$

Generalized Solution. If $f'(x)$ and $f''(x)$ are merely piecewise continuous (see Sec. 6.1), or if those one-sided derivatives are not zero, then for each t there will be finitely many values of x at which the second derivatives of u appearing in (1) do not exist. Except at these points the wave equation will still be satisfied. We may then regard $u(x, t)$ as a **"generalized solution,"** as it is called, that is, as a solution in a broader sense. For instance, a triangular initial deflection as in Example 1 (below) leads to a generalized solution.

Physical Interpretation of the Solution (17). The graph of $f^*(x - ct)$ is obtained from the graph of $f^*(x)$ by shifting the latter ct units to the right (Fig. 287). This means that $f^*(x - ct)$ $(c > 0)$ represents a wave that is traveling to the right as t increases. Similarly, $f^*(x + ct)$ represents a wave that is traveling to the left, and $u(x, t)$ is the superposition of these two waves.

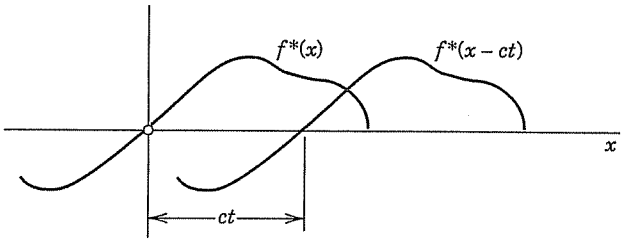


Fig. 287. Interpretation of (17)

EXAMPLE 1 **Vibrating String if the Initial Deflection Is Triangular**

Find the solution of the wave equation (1) corresponding to the triangular initial deflection

$$f(x) = \begin{cases} \dfrac{2k}{L}x & \text{if} \quad 0 < x < \dfrac{L}{2} \\ \dfrac{2k}{L}(L - x) & \text{if} \quad \dfrac{L}{2} < x < L \end{cases}$$

and initial velocity zero. (Figure 288 shows $f(x) = u(x, 0)$ at the top.)

Solution. Since $g(x) \equiv 0$, we have $B_n^* = 0$ in (12), and from Example 4 in Sec. 11.3 we see that the B_n are given by (5), Sec. 11.3. Thus (12) takes the form

$$u(x, t) = \frac{8k}{\pi^2}\left[\frac{1}{1^2}\sin\frac{\pi}{L}x\cos\frac{\pi c}{L}t - \frac{1}{3^2}\sin\frac{3\pi}{L}x\cos\frac{3\pi c}{L}t + - \cdots\right].$$

For graphing the solution we may use $u(x, 0) = f(x)$ and the above interpretation of the two functions in the representation (17). This leads to the graph shown in Fig. 288. ■

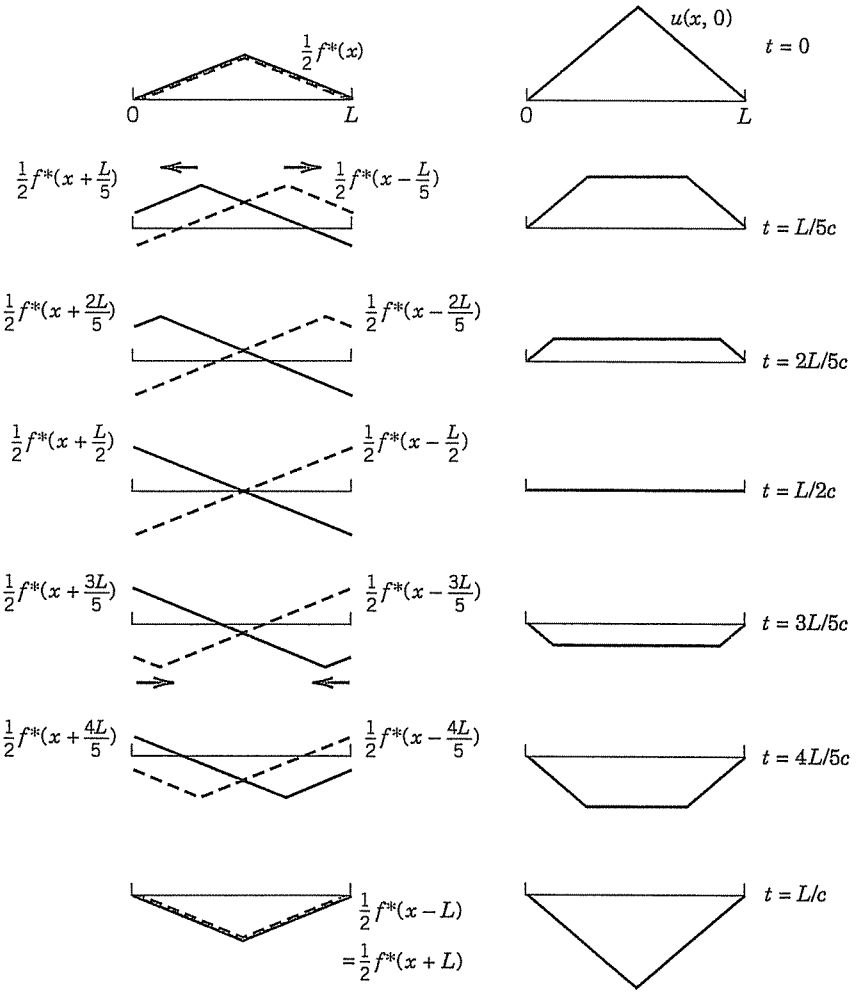


Fig. 288. Solution $u(x, t)$ in Example 1 for various values of t (right part of the figure) obtained as the superposition of a wave traveling to the right (dashed) and a wave traveling to the left (left part of the figure)

PROBLEM SET 12.3

1–10 DEFLECTION OF THE STRING

Find $u(x, t)$ for the string of length $L = 1$ and $c^2 = 1$ when the initial velocity is zero and the initial deflection with small k (say, 0.01) is as follows. Sketch or graph $u(x, t)$ as in Fig. 288.

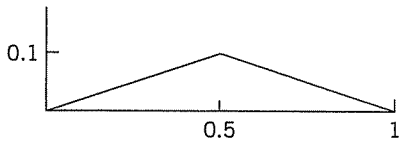
1. $k \sin 2\pi x$

2. $k(\sin \pi x - \frac{1}{3} \sin 3\pi x)$

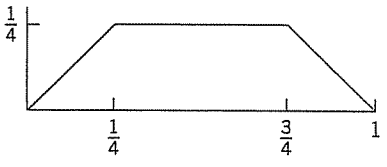
3. $kx(1 - x)$

4. $kx(1 - x^2)$

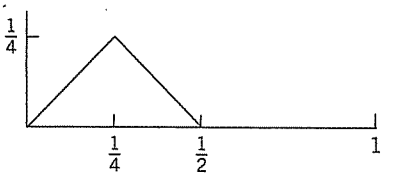
5.



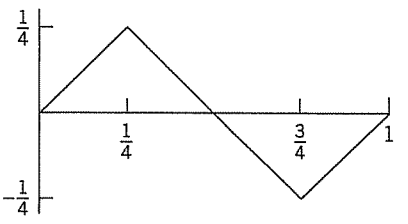
6.



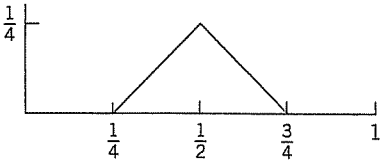
7.



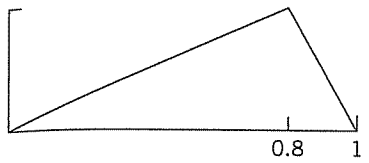
8.



9.



10.



11. (Frequency) How does the frequency of the fundamental mode of the vibrating string depend on the length of the string? On the mass per unit length? What happens to the string if we double the tension? Why is a contrabass larger than a violin?

12. (Nonzero initial velocity) Find the deflection $u(x, t)$ of the string of length $L = \pi$ and $c^2 = 1$ for zero initial displacement and "triangular" initial velocity $u_t(x, 0) = 0.01x$ if $0 \leqq x \leqq \frac{1}{2}\pi$, $u_t(x, 0) = 0.01(\pi - x)$ if $\frac{1}{2}\pi \leqq x \leqq \pi$. (Initial conditions with $u_t(x, 0) \neq 0$ are hard to realize experimentally.)

13. CAS PROJECT. Graphing Normal Modes. Write a program for graphing u_n with $L = \pi$ and c^2 of your choice similarly as in Fig. 284. Apply the program to u_2, u_3, u_4. Also graph these solutions as surfaces over the xt-plane. Explain the connection between these two kinds of graphs.

14. TEAM PROJECT. Forced Vibrations of an Elastic String. Show the following.

(a) Substitution of

$$(17) \qquad u(x, t) = \sum_{n=1}^{\infty} G_n(t) \sin \frac{n\pi x}{L}$$

(L = length of the string) into the wave equation (1) governing free vibrations leads to [see (10*)]

$$(18) \qquad \ddot{G}_n + {\lambda_n}^2 G = 0, \qquad \lambda_n = \frac{cn\pi}{L}.$$

(b) Forced vibrations of the string under an external force $P(x, t)$ per unit length acting normal to the string are governed by the PDE

$$(19) \qquad u_{tt} = c^2 u_{xx} + \frac{P}{\rho}.$$

(c) For a sinusoidal force $P = A\rho \sin \omega t$ we obtain

$$(20) \qquad \frac{P}{\rho} = A \sin \omega t = \sum_{n=1}^{\infty} k_n(t) \sin \frac{n\pi x}{L},$$

$$k_n(t) = \begin{cases} (4A/n\pi) \sin \omega t & (n \text{ odd}) \\ 0 & (n \text{ even}). \end{cases}$$

Substituting (17) and (20) into (19) gives

$$\ddot{G}_n + {\lambda_n}^2 G_n = \frac{2A}{n\pi} (1 - \cos n\pi) \sin \omega t.$$

If ${\lambda_n}^2 \neq \omega^2$, the solution is

$$G_n(t) = B_n \cos \lambda_n t + B_n^* \sin \lambda_n t + \frac{2A(1 - \cos n\pi)}{n\pi({\lambda_n}^2 - \omega^2)} \sin \omega t.$$

Determine B_n and B_n^* so that u satisfies the initial conditions $u(x, 0) = f(x)$, $u_t(x, 0) = 0$.

(d) (Resonance) Show that if $\lambda_n = \omega$, then

$$G_n(t) = B_n \cos \omega t + B_n^* \sin \omega t - \frac{A}{n\pi\omega} (1 - \cos n\pi) t \cos \omega t.$$

(e) (Reduction of boundary conditions) Show that a problem (1)–(3) with more complicated boundary conditions $u(0, t) = 0$, $u(L, t) = h(t)$, can be reduced to a problem for a new function v satisfying conditions $v(0, t) = v(L, t) = 0$, $v(x, 0) = f_1(x)$, $v_t(x, 0) = g_1(x)$ but a nonhomogeneous wave equation. *Hint:* Set $u = v + w$ and determine w suitably.

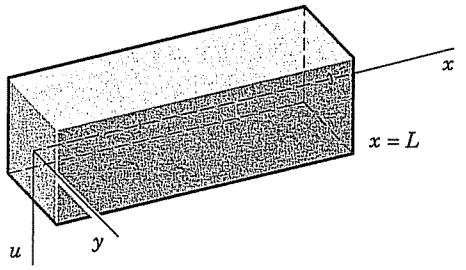


Fig. 289. Elastic beam

15–20 SEPARATION OF A FOURTH-ORDER PDE. VIBRATING BEAM

By the principles used in modeling the string it can be shown that small free vertical vibrations of a uniform elastic beam (Fig. 289) are modeled by the fourth-order PDE

$$(21) \qquad \frac{\partial^2 u}{\partial t^2} = -c^2 \frac{\partial^4 u}{\partial x^4} \qquad \text{(Ref. [C11])}$$

where $c^2 = EI/\rho A$ (E = Young's modulus of elasticity, I = moment of intertia of the cross section with respect to the y-axis in the figure, ρ = density, A = cross-sectional area). (*Bending* of a beam under a load is discussed in Sec. 3.3.)

15. Substituting $u = F(x)G(t)$ into (21), show that

$$F^{(4)}/F = -\ddot{G}/c^2 G = \beta^4 = const,$$

$$F(x) = A \cos \beta x + B \sin \beta x + C \cosh \beta x + D \sinh \beta x,$$

$$G(t) = a \cos c\beta^2 t + b \sin c\beta^2 t.$$

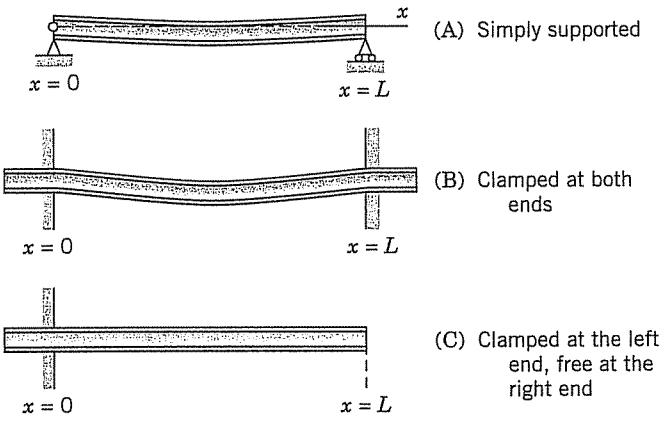


Fig. 290. Supports of a beam

16. **(Simply supported beam in Fig. 290A)** Find solutions $u_n = F_n(x)G_n(t)$ of (21) corresponding to zero initial velocity and satisfying the boundary conditions (see Fig. 290A)

$$u(0, t) = 0, \quad u(L, t) = 0$$

(ends simply supported for all times t),

$$u_{xx}(0, t) = 0, \quad u_{xx}(L, t) = 0$$

(zero moments, hence zero curvature, at the ends).

17. Find the solution of (21) that satisfies the conditions in Prob. 16 as well as the initial condition

$$u(x, 0) = f(x) = x(L - x).$$

18. Compare the results of Probs. 17 and 3. What is the basic difference between the frequencies of the normal modes of the vibrating string and the vibrating beam?

19. **(Clamped beam in Fig. 290B)** What are the boundary conditions for the clamped beam in Fig. 290B? Show that F in Prob. 15 satisfies these conditions if βL is a solution of the equation

$$\cosh \beta L \cos \beta L = 1. \tag{22}$$

Determine approximate solutions of (22), for instance, graphically from the intersections of the curves of $\cos \beta L$ and $1/\cosh \beta L$.

20. **(Clamped-free beam in Fig. 290C)** If the beam is clamped at the left and free at the right (Fig. 290C), the boundary conditions are

$$u(0, t) = 0, \qquad u_x(0, t) = 0,$$
$$u_{xx}(L, t) = 0, \qquad u_{xxx}(L, t) = 0.$$

Show that F in Prob. 15 satisfies these conditions if βL is a solution of the equation

$$\cosh \beta L \cos \beta L = -1. \tag{23}$$

Find approximate solutions of (18).

12.4 D'Alembert's Solution of the Wave Equation. Characteristics

It is interesting that the solution (17), Sec. 12.3, of the wave equation

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}, \qquad c^2 = \frac{T}{\rho}, \tag{1}$$

can be immediately obtained by transforming (1) in a suitable way, namely, by introducing the new independent variables

$$v = x + ct, \qquad w = x - ct. \tag{2}$$

Then u becomes a function of v and w. The derivatives in (1) can now be expressed in terms of derivatives with respect to v and w by the use of the chain rule in Sec. 9.6. Denoting partial derivatives by subscripts, we see from (2) that $v_x = 1$ and $w_x = 1$. For simplicity let us denote $u(x, t)$, as a function of v and w, by the same letter u. Then

$$u_x = u_v v_x + u_w w_x = u_v + u_w.$$

We now apply the chain rule to the right side of this equation. We assume that all the partial derivatives involved are continuous, so that $u_{wv} = u_{vw}$. Since $v_x = 1$ and $w_x = 1$, we obtain

$$u_{xx} = (u_v + u_w)_x = (u_v + u_w)_v v_x + (u_v + u_w)_w w_x = u_{vv} + 2u_{vw} + u_{ww}.$$

Transforming the other derivative in (1) by the same procedure, we find

$$u_{tt} = c^2(u_{vv} - 2u_{vw} + u_{ww}).$$

By inserting these two results in (1) we get (see footnote 2 in App. A3.2)

$$u_{vw} \equiv \frac{\partial^2 u}{\partial w\, \partial v} = 0. \tag{3}$$

The point of the present method is that (3) can be readily solved by two successive integrations, first with respect to w and then with respect to v. This gives

$$\frac{\partial u}{\partial v} = h(v) \qquad \text{and} \qquad u = \int h(v)\, dv + \psi(w).$$

Here $h(v)$ and $\psi(w)$ are arbitrary functions of v and w, respectively. Since the integral is a function of v, say, $\phi(v)$, the solution is of the form $u = \phi(v) + \psi(w)$. In terms of x and t, by (2), we thus have

$$u(x, t) = \phi(x + ct) + \psi(x - ct). \tag{4}$$

This is known as **d'Alembert's solution**[1] of the wave equation (1).

Its derivation was much more elegant than the method in Sec. 12.3, but d'Alembert's method is special, whereas the use of Fourier series applies to various equations, as we shall see.

D'Alembert's Solution Satisfying the Initial Conditions

(5) (a) $u(x, 0) = f(x)$, (b) $u_t(x, 0) = g(x)$.

These are the same as (3) in Sec. 12.3. By differentiating (4) we have

$$u_t(x, t) = c\phi'(x + ct) - c\psi'(x - ct) \tag{6}$$

---

[1]JEAN LE ROND D'ALEMBERT (1717–1783), French mathematician, also known for his important work in mechanics.

We mention that the general theory of PDEs provides a systematic way for finding the transformation (2) that simplifies (1). See Ref. [C8] in App. 1.

where primes denote derivatives with respect to the *entire* arguments $x + ct$ and $x - ct$, respectively, and the minus sign comes from the chain rule. From (4)–(6) we have

$$u(x, 0) = \phi(x) + \psi(x) = f(x), \tag{7}$$

$$u_t(x, 0) = c\phi'(x) - c\psi'(x) = g(x). \tag{8}$$

Dividing (8) by c and integrating with respect to x, we obtain

$$\phi(x) - \psi(x) = k(x_0) + \frac{1}{c}\int_{x_0}^{x} g(s)\,ds, \qquad k(x_0) = \phi(x_0) - \psi(x_0). \tag{9}$$

If we add this to (7), then ψ drops out and division by 2 gives

$$\phi(x) = \frac{1}{2}\,f(x) + \frac{1}{2c}\int_{x_0}^{x} g(s)\,ds + \frac{1}{2}\,k(x_0). \tag{10}$$

Similarly, subtraction of (9) from (7) and division by 2 gives

$$\psi(x) = \frac{1}{2}\,f(x) - \frac{1}{2c}\int_{x_0}^{x} g(s)\,ds - \frac{1}{2}\,k(x_0). \tag{11}$$

In (10) we replace x by $x + ct$; we then get an integral from x_0 to $x + ct$. In (11) we replace x by $x - ct$ and get minus an integral from x_0 to $x - ct$ or plus an integral from $x - ct$ to x_0. Hence addition of $\phi(x + ct)$ and $\psi(x - ct)$ gives $u(x, t)$ [see (4)] in the form

$$\mathbf{u(x, t) = \frac{1}{2}\,[f(x + ct) + f(x - ct)] + \frac{1}{2c}\int_{x-ct}^{x+ct} g(s)\,ds.} \tag{12}$$

If the initial velocity is zero, we see that this reduces to

$$u(x, t) = \tfrac{1}{2}[f(x + ct) + f(x - ct)], \tag{13}$$

in agreement with (17) in Sec. 12.3. You may show that because of the boundary conditions (2) in that section the function f must be odd and must have the period $2L$.

Our result shows that the two initial conditions [the functions $f(x)$ and $g(x)$ in (5)] determine the solution uniquely.

The solution of the wave equation by the Laplace transform method will be shown in Sec. 12.11.

Characteristics. Types and Normal Forms of PDEs

The idea of d'Alembert's solution is just a special instance of the **method of characteristics.** This concerns PDEs of the form

$$\mathbf{Au_{xx} + 2Bu_{xy} + Cu_{yy} = F(x, y, u, u_x, u_y)} \tag{14}$$

(as well as PDEs in more than two variables). Equation (14) is called **quasilinear** because it is linear in the highest derivatives (but may be arbitrary otherwise). There are three types of PDEs (14), depending on the discriminant $AC - B^2$, as follows.

Type	Defining Condition	Example in Sec. 12.1
Hyperbolic	$AC - B^2 < 0$	Wave equation (1)
Parabolic	$AC - B^2 = 0$	Heat equation (2)
Elliptic	$AC - B^2 > 0$	Laplace equation (3)

Note that (1) and (2) in Sec. 12.1 involve t, but to have y as in (14), we set $y = ct$ in (1), obtaining $u_{tt} - c^2u_{xx} = c^2(u_{yy} - u_{xx}) = 0$. And in (2) we set $y = c^2t$, so that $u_t - c^2u_{xx} = c^2(u_y - u_{xx})$.

A, B, C may be functions of x, y, so that a PDE may be **of mixed type,** that is, of different type in different regions of the xy-plane. An important mixed-type PDE is the **Tricomi equation** (see Prob. 10).

Transformation of (14) to Normal Form. The normal forms of (14) and the corresponding transformations depend on the type of the PDE. They are obtained by solving the **characteristic equation** of (14), which is the ODE

$$\textbf{(15)} \qquad Ay'^2 - 2By' + C = 0$$

where $y' = dy/dx$ (note $-2B$, not $+2B$). The solutions of (15) are called the **characteristics** of (14), and we write them in the form $\Phi(x, y) = const$ and $\Psi(x, y) = const$. Then the transformations giving new variables v, w instead of x, y and the normal forms of (14) are as follows.

Type	New Variables		Normal Form
Hyperbolic	$v = \Phi$	$w = \Psi$	$u_{vw} = F_1$
Parabolic	$v = x$	$w = \Phi = \Psi$	$u_{ww} = F_2$
Elliptic	$v = \frac{1}{2}(\Phi + \Psi)$	$w = \dfrac{1}{2i}(\Phi - \Psi)$	$u_{vv} + u_{ww} = F_3$

Here, $\Phi = \Phi(x, y)$, $\Psi = \Psi(x, y)$, $F_1 = F_1(v, w, u, u_v, u_w)$, etc., and we denote u as function of v, w again by u, for simplicity. We see that the normal form of a hyperbolic PDE is as in d'Alembert's solution. In the parabolic case we get just one family of solutions $\Phi = \Psi$. In the elliptic case, $i = \sqrt{-1}$, and the characteristics are complex and are of minor interest. For derivation, see Ref. [GR3] in App. 1.

EXAMPLE 1 D'Alembert's Solution Obtained Systematically

The theory of characteristics gives d'Alembert's solution in a systematic fashion. To see this, we write the wave equation $u_{tt} - c^2u_{xx} = 0$ in the form (14) by setting $y = ct$. By the chain rule, $u_t = u_y y_t = cu_y$ and $u_{tt} = c^2u_{yy}$. Division by c^2 gives $u_{xx} - u_{yy} = 0$, as stated before. Hence the characteristic equation is $y'^2 - 1 = (y' + 1)(y' - 1) = 0$. The two families of solutions (characteristics) are $\Phi(x, y) = y + x = const$ and $\Psi(x, y) = y - x = const$. This gives the new variables $v = \Phi = y + x = ct + x$ and $w = \Psi = y - x = ct - x$ and d'Alembert's solution $u = f_1(x + ct) + f_2(x - ct)$. ■

PROBLEM SET 12.4

1. Show that c is the speed of each of the two waves given by (4).
2. Show that because of the boundary conditions (2), Sec. 12.3, the function f in (13) of this section must be odd and of period $2L$.
3. If a steel wire 2 m in length weighs 0.9 nt (about 0.20 lb) and is stretched by a tensile force of 300 nt (about 67.4 lb), what is the corresponding speed of transverse waves?
4. What are the frequencies of the eigenfunctions in Prob. 3?
5. **Longitudinal Vibrations of an Elastic Bar or Rod.** These vibrations in the direction of the x-axis are modeled by the wave equation $u_{tt} = c^2 u_{xx}$, $c^2 = E/\rho$ (see Tolstov [C9], p. 275). If the rod is fastened at one end, $x = 0$, and free at the other, $x = L$, we have $u(0, t) = 0$ and $u_x(L, t) = 0$. Show that the motion corresponding to initial displacement $u(x, 0) = f(x)$ and initial velocity zero is

$$u = \sum_{n=0}^{\infty} A_n \sin p_n x \cos p_n c t,$$

$$A_n = \frac{2}{L}\int_0^L f(x) \sin p_n x \, dx, \qquad p_n = \frac{(2n+1)\pi}{2L}.$$

6–9 GRAPHING SOLUTIONS

Using (13), sketch or graph a figure (similar to Fig. 288 in Sec. 12.3) of the deflection $u(x, t)$ of a vibrating string (length $L = 1$, ends fixed, $c = 1$) starting with initial velocity 0 and initial deflection (k small, say, $k = 0.01$).

6. $f(x) = k \sin \pi x$
7. $f(x) = k(1 - \cos 2\pi x)$
8. $f(x) = kx(1 - x)$
9. $f(x) = k(x - x^3)$
10. **(Tricomi and Airy equations[2])** Show that the *Tricomi equation* $yu_{xx} + u_{yy} = 0$ is of mixed type. Obtain the **Airy equation** $G'' - yG = 0$ from the Tricomi equation by separation. (For solutions, see p. 446 of Ref. [GR1] listed in App. 1.)

11–20 NORMAL FORMS

Find the type, transform to normal form, and solve. (Show the details of your work.)

11. $u_{xy} - u_{yy} = 0$
12. $u_{xx} - 2u_{xy} + u_{yy} = 0$
13. $u_{xx} + 9u_{yy} = 0$
14. $u_{xx} + u_{xy} - 2u_{yy} = 0$
15. $u_{xx} + 2u_{xy} + u_{yy} = 0$
16. $xu_{xy} - yu_{yy} = 0$
17. $u_{xx} - 4u_{xy} + 4u_{yy} = 0$
18. $u_{xx} + 2u_{xy} + 5u_{yy} = 0$
19. $xu_{xx} - yu_{xy} = 0$
20. $u_{xx} - 4u_{xy} + 3u_{yy} = 0$

12.5 Heat Equation: Solution by Fourier Series

From the wave equation we now turn to the next "big" PDE, the **heat equation**

$$\frac{\partial u}{\partial t} = c^2 \nabla^2 u, \qquad c^2 = \frac{K}{\sigma\rho},$$

which gives the temperature $u(x, y, z, t)$ in a body of homogeneous material. Here c^2 is the thermal diffusivity, K the thermal conductivity, σ the specific heat, and ρ the density of the material of the body. $\nabla^2 u$ is the Laplacian of u, and with respect to Cartesian coordinates x, y, z,

$$\nabla^2 u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}.$$

The heat equation was derived in Sec. 10.8. It is also called the **diffusion equation.**

As an important application, let us first consider the temperature in a long thin metal bar or wire of constant cross section and homogeneous material, which is oriented along the x-axis (Fig. 291) and is perfectly insulated laterally, so that heat flows in the x-direction

[2]SIR GEORGE BIDELL AIRY (1801–1892), English mathematician, known for his work in elasticity. FRANCESCO TRICOMI (1897–1978), Italian mathematician, who worked in integral equations.

Fig. 291. Bar under consideration

only. Then u depends only on x and time t, and the heat equation becomes the **one-dimensional heat equation**

$$\frac{\partial u}{\partial t} = c^2 \frac{\partial^2 u}{\partial x^2}. \tag{1}$$

This seems to differ only very little from the wave equation, which has a term u_{tt} instead of u_t, but we shall see that this will make the solutions of (1) behave quite differently from those of the wave equation.

We shall solve (1) for some important types of boundary and initial conditions. We begin with the case in which the ends $x = 0$ and $x = L$ of the bar are kept at temperature zero, so that we have the **boundary conditions**

$$u(0, t) = 0, \qquad u(L, t) = 0 \qquad \text{for all } t. \tag{2}$$

Furthermore, the initial temperature in the bar at time $t = 0$ is given, say, $f(x)$, so that we have the **initial condition**

$$u(x, 0) = f(x) \qquad [f(x) \text{ given}]. \tag{3}$$

Here we must have $f(0) = 0$ and $f(L) = 0$ because of (2).

We shall determine a solution $u(x, t)$ of (1) satisfying (2) and (3)—one initial condition will be enough, as opposed to two initial conditions for the wave equation. Technically, our method will parallel that for the wave equation in Sec. 12.3: a separation of variables, followed by the use of Fourier series. You may find a step-by-step comparison worthwhile.

Step 1. **Two ODEs from the heat equation (1).** Substitution of a product $u(x, t) = F(x)G(t)$ into (1) gives $F\dot{G} = c^2F''G$ with $\dot{G} = dG/dt$ and $F'' = d^2F/dx^2$. To separate the variables, we divide by c^2FG, obtaining

$$\frac{\dot{G}}{c^2G} = \frac{F''}{F}. \tag{4}$$

The left side depends only on t and the right side only on x, so that both sides must equal a constant k (as in Sec. 12.3). You may show that for $k = 0$ or $k > 0$ the only solution $u = FG$ satisfying (2) is $u \equiv 0$. For negative $k = -p^2$ we have from (4)

$$\frac{\dot{G}}{c^2G} = \frac{F''}{F} = -p^2.$$

Multiplication by the denominators gives immediately the two ODEs

$$F'' + p^2F = 0 \tag{5}$$

and

$$\dot{G} + c^2p^2G = 0. \tag{6}$$

***Step 2.* Satisfying the boundary conditions (2).** We first solve (5). A general solution is

$$F(x) = A \cos px + B \sin px. \tag{7}$$

From the boundary conditions (2) it follows that

$$u(0, t) = F(0)G(t) = 0 \qquad \text{and} \qquad u(L, t) = F(L)G(t) = 0.$$

Since $G \equiv 0$ would give $u \equiv 0$, we require $F(0) = 0$, $F(L) = 0$ and get $F(0) = A = 0$ by (7) and then $F(L) = B \sin pL = 0$, with $B \neq 0$ (to avoid $F \equiv 0$); thus,

$$\sin pL = 0, \qquad \text{hence} \qquad p = \frac{n\pi}{L}, \qquad n = 1, 2, \cdots.$$

Setting $B = 1$, we thus obtain the following solutions of (5) satisfying (2):

$$F_n(x) = \sin \frac{n\pi x}{L}, \qquad n = 1, 2, \cdots.$$

(As in Sec. 12.3, we need not consider *negative* integral values of n.)

All this was literally the same as in Sec. 12.3. From now on it differs since (6) differs from (6) in Sec. 12.3. We now solve (6). For $p = n\pi/L$, as just obtained, (6) becomes

$$\dot{G} + \lambda_n{}^2G = 0 \qquad \text{where} \qquad \lambda_n = \frac{cn\pi}{L}.$$

It has the general solution

$$G_n(t) = B_n e^{-\lambda_n{}^2 t}, \qquad n = 1, 2, \cdots$$

where B_n is a constant. Hence the functions

$$u_n(x, t) = F_n(x)G_n(t) = B_n \sin \frac{n\pi x}{L} e^{-\lambda_n{}^2 t} \qquad (n = 1, 2, \cdots) \tag{8}$$

are solutions of the heat equation (1), satisfying (2). These are the **eigenfunctions** of the problem, corresponding to the **eigenvalues** $\lambda_n = cn\pi/L$.

***Step 3.* Solution of the entire problem. Fourier series.** So far we have solutions (8) satisfying the boundary conditions (2). To obtain a solution that also satisfies the initial condition (3), we consider a series of these eigenfunctions,

$$u(x, t) = \sum_{n=1}^{\infty} u_n(x, t) = \sum_{n=1}^{\infty} B_n \sin \frac{n\pi x}{L} e^{-\lambda_n{}^2 t} \qquad \left(\lambda_n = \frac{cn\pi}{L}\right). \tag{9}$$

From this and (3) we have

$$u(x, 0) = \sum_{n=1}^{\infty} B_n \sin \frac{n\pi x}{L} = f(x).$$

Hence for (9) to satisfy (3), the B_n's must be the coefficients of the **Fourier sine series,** as given by (4) in Sec. 11.3; thus

(10)
$$B_n = \frac{2}{L} \int_0^L f(x) \sin \frac{n\pi x}{L} \, dx \qquad (n = 1, 2, \cdots).$$

The solution of our problem can be established, assuming that $f(x)$ is piecewise continuous (see Sec. 6.1) on the interval $0 \leqq x \leqq L$ and has one-sided derivatives (see Sec. 11.1) at all interior points of that interval; that is, under these assumptions the series (9) with coefficients (10) is the solution of our physical problem. A proof requires knowledge of uniform convergence and will be given at a later occasion (Probs. 19, 20 in Problem Set 15.5).

Because of the exponential factor, all the terms in (9) approach zero as t approaches infinity. The rate of decay increases with n.

EXAMPLE 1 Sinusoidal Initial Temperature

Find the temperature $u(x, t)$ in a laterally insulated copper bar 80 cm long if the initial temperature is $100 \sin (\pi x/80)$ °C and the ends are kept at 0°C. How long will it take for the maximum temperature in the bar to drop to 50°C? First guess, then calculate. *Physical data for copper:* density 8.92 gm/cm^3, specific heat 0.092 cal/(gm °C), thermal conductivity 0.95 cal/(cm sec °C).

Solution. The initial condition gives

$$u(x, 0) = \sum_{n=1}^{\infty} B_n \sin \frac{n\pi x}{80} = f(x) = 100 \sin \frac{\pi x}{80}.$$

Hence, by inspection or from (9) we get $B_1 = 100$, $B_2 = B_3 = \cdots = 0$. In (9) we need $\lambda_1^2 = c^2\pi^2/L^2$, where $c^2 = K/(\sigma\rho) = 0.95/(0.092 \cdot 8.92) = 1.158$ [cm^2/sec]. Hence we obtain

$$\lambda_1^2 = 1.158 \cdot 9.870/80^2 = 0.001785 \ [\text{sec}^{-1}].$$

The solution (9) is

$$u(x, t) = 100 \sin \frac{\pi x}{80} \, e^{-0.001785t}.$$

Also, $100e^{-0.001785t} = 50$ when $t = (\ln 0.5)/(-0.001785) = 388$ [sec] ≈ 6.5 [min]. Does your guess, or at least its order of magnitude, agree with this result? ■

EXAMPLE 2 Speed of Decay

Solve the problem in Example 1 when the initial temperature is $100 \sin (3\pi x/80)$ °C and the other data are as before.

Solution. In (9), instead of $n = 1$ we now have $n = 3$, and $\lambda_3^2 = 3^2\lambda_1^2 = 9 \cdot 0.001785 = 0.01607$, so that the solution now is

$$u(x, t) = 100 \sin \frac{3\pi x}{80} \, e^{-0.01607t}.$$

Hence the maximum temperature drops to 50°C in $t = (\ln 0.5)/(-0.01607) \approx 43$ [seconds], which is much faster (9 times as fast as in Example 1; why?).

Had we chosen a bigger n, the decay would have been still faster, and in a sum or series of such terms, each term has its own rate of decay, and terms with large n are practically 0 after a very short time. Our next example is of this type, and the curve in Fig. 292 corresponding to $t = 0.5$ looks almost like a sine curve; that is, it is practically the graph of the first term of the solution.

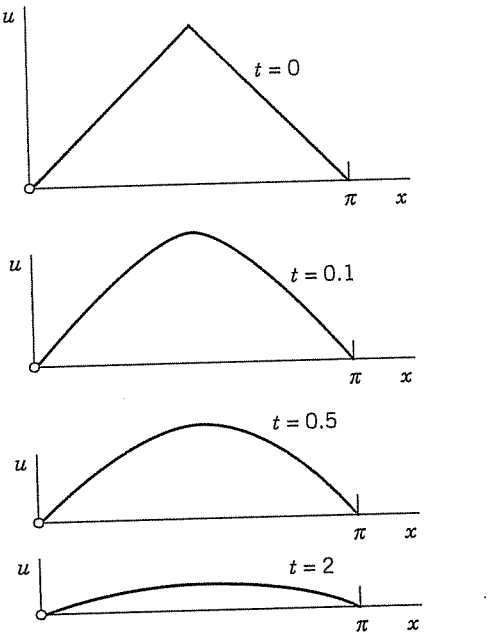


Fig. 292. Example 3. Decrease of temperature with time t for $L = \pi$ and $c = 1$

EXAMPLE 3 **"Triangular" Initial Temperature in a Bar**

Find the temperature in a laterally insulated bar of length L whose ends are kept at temperature 0, assuming that the initial temperature is

$$f(x) = \begin{cases} x & \text{if} \quad 0 < x < L/2, \\ L - x & \text{if} \quad L/2 < x < L. \end{cases}$$

(The uppermost part of Fig. 292 shows this function for the special $L = \pi$.)

Solution. From (10) we get

$$(10^*) \qquad B_n = \frac{2}{L}\left(\int_0^{L/2} x \sin\frac{n\pi x}{L}\,dx + \int_{L/2}^{L} (L - x)\sin\frac{n\pi x}{L}\,dx\right).$$

Integration gives $B_n = 0$ if n is even,

$$B_n = \frac{4L}{n^2\pi^2} \quad (n = 1, 5, 9, \cdots) \qquad \text{and} \qquad B_n = -\frac{4L}{n^2\pi^2} \quad (n = 3, 7, 11, \cdots).$$

(see also Example 4 in Sec. 11.3 with $k = L/2$). Hence the solution is

$$u(x, t) = \frac{4L}{\pi^2}\left[\sin\frac{\pi x}{L}\exp\left[-\left(\frac{c\pi}{L}\right)^2 t\right] - \frac{1}{9}\sin\frac{3\pi x}{L}\exp\left[-\left(\frac{3c\pi}{L}\right)^2 t\right] + - \cdots\right].$$

Figure 292 shows that the temperature decreases with increasing t, because of the heat loss due to the cooling of the ends.

Compare Fig. 292 and Fig. 288 in Sec. 12.3 and comment.

EXAMPLE 4 **Bar with Insulated Ends. Eigenvalue 0**

Find a solution formula of (1), (3) with (2) replaced by the condition that both ends of the bar are insulated.

Solution. Physical experiments show that the rate of heat flow is proportional to the gradient of the temperature. Hence if the ends $x = 0$ and $x = L$ of the bar are insulated, so that no heat can flow through the ends, we have grad $u = u_x = \partial u/\partial x$ and the boundary conditions

$$(2^*) \qquad u_x(0, t) = 0, \qquad u_x(L, t) = 0 \qquad \text{for all } t.$$

Since $u(x, t) = F(x)G(t)$, this gives $u_x(0, t) = F'(0)G(t) = 0$ and $u_x(L, t) = F'(L)G(t) = 0$. Differentiating (7), we have $F'(x) = -Ap \sin px + Bp \cos px$, so that

$$F'(0) = Bp = 0 \qquad \text{and then} \qquad F'(L) = -Ap \sin pL = 0.$$

The second of these conditions gives $p = p_n = n\pi/L$, $(n = 0, 1, 2, \cdots)$. From this and (7) with $A = 1$ and $B = 0$ we get $F_n(x) = \cos(n\pi x/L)$, $(n = 0, 1, 2, \cdots)$. With G_n as before, this yields the eigenfunctions

$$(11) \qquad u_n(x, t) = F_n(x)G_n(t) = A_n \cos\frac{n\pi x}{L}\, e^{-\lambda_n^2 t} \qquad (n = 0, 1, \cdots)$$

corresponding to the eigenvalues $\lambda_n = cn\pi/L$. The latter are as before, but we now have the additional eigenvalue $\lambda_0 = 0$ and eigenfunction $u_0 = const$, which is the solution of the problem if the initial temperature $f(x)$ is constant. This shows the remarkable fact that ***a separation constant can very well be zero, and zero can be an eigenvalue.***

Furthermore, whereas (8) gave a Fourier sine series, we now get from (11) a Fourier cosine series

$$(12) \qquad u(x, t) = \sum_{n=0}^{\infty} u_n(x, t) = \sum_{n=0}^{\infty} A_n \cos\frac{n\pi x}{L}\, e^{-\lambda_n^2 t} \qquad \left(\lambda_n = \frac{cn\pi}{L}\right).$$

Its coefficients result from the initial condition (3),

$$u(x, 0) = \sum_{n=0}^{\infty} A_n \cos\frac{n\pi x}{L} = f(x),$$

in the form (2), Sec. 11.3, that is,

$$(13) \qquad A_0 = \frac{1}{L}\int_0^L f(x)\, dx, \qquad A_n = \frac{2}{L}\int_0^L f(x) \cos\frac{n\pi x}{L}\, dx, \qquad n = 1, 2, \cdots. \qquad \blacksquare$$

EXAMPLE 5 **"Triangular" Initial Temperature in a Bar with Insulated Ends**

Find the temperature in the bar in Example 3, assuming that the ends are insulated (instead of being kept at temperature 0).

Solution. For the triangular initial temperature, (13) gives $A_0 = L/4$ and (see also Example 4 in Sec. 11.3 with $k = L/2$)

$$A_n = \frac{2}{L}\left[\int_0^{L/2} x \cos\frac{n\pi x}{L}\, dx + \int_{L/2}^{L} (L - x) \cos\frac{n\pi x}{L}\, dx\right] = \frac{2L}{n^2\pi^2}\left(2 \cos\frac{n\pi}{2} - \cos n\pi - 1\right).$$

Hence the solution (12) is

$$u(x, t) = \frac{L}{4} - \frac{8L}{\pi^2}\left\{\frac{1}{2^2} \cos\frac{2\pi x}{L} \exp\left[-\left(\frac{2c\pi}{L}\right)^2 t\right] + \frac{1}{6^2} \cos\frac{6\pi x}{L} \exp\left[-\left(\frac{6c\pi}{L}\right)^2 t\right] + \cdots\right\}.$$

We see that the terms decrease with increasing t, and $u \to L/4$ as $t \to \infty$; this is the mean value of the initial temperature. This is plausible because no heat can escape from this totally insulated bar. In contrast, the cooling of the ends in Example 3 led to heat loss and $u \to 0$, the temperature at which the ends were kept. $\blacksquare$

Steady Two-Dimensional Heat Problems. Laplace's Equation

We shall now extend our discussion from one to two space dimensions and consider the two-dimensional heat equation

$$\frac{\partial u}{\partial t} = c^2\nabla^2 u = c^2\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right)$$

for **steady** (that is, *time-independent*) problems. Then $\partial u/\partial t = 0$ and the heat equation reduces to **Laplace's equation**

$$\nabla^2 u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \tag{14}$$

(which has already occurred in Sec. 10.8 and will be considered further in Secs. 12.7–12.10). A heat problem then consists of this PDE to be considered in some region R of the xy-plane and a given boundary condition on the boundary curve C of R. This is a **boundary value problem (BVP).** One calls it:

First BVP or **Dirichlet Problem** if u is prescribed on C (**"Dirichlet boundary condition"**)

Second BVP or **Neumann Problem** if the normal derivative $u_n = \partial u/\partial n$ is prescribed on C (**"Neumann boundary condition"**)

Third BVP, Mixed BVP, or **Robin Problem** if u is prescribed on a portion of C and u_n on the rest of C (**"Mixed boundary condition"**).

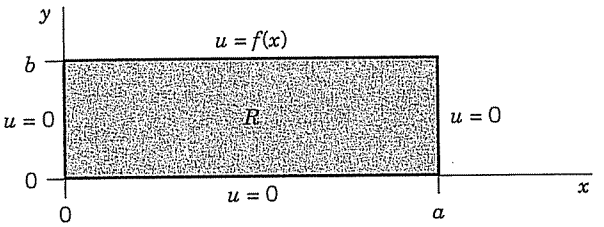


Fig. 293. Rectangle R and given boundary values

Dirichlet Problem in a Rectangle R (Fig. 293). We consider a Dirichlet problem for Laplace's equation (14) in a rectangle R, assuming that the temperature $u(x, y)$ equals a given function $f(x)$ on the upper side and 0 on the other three sides of the rectangle.

We solve this problem by separating variables. Substituting $u(x, y) = F(x)G(y)$ into (14) written as $u_{xx} = -u_{yy}$, dividing by FG, and equating both sides to a negative constant, we obtain

$$\frac{1}{F} \cdot \frac{d^2F}{dx^2} = -\frac{1}{G} \cdot \frac{d^2G}{dy^2} = -k.$$

From this we get

$$\frac{d^2F}{dx^2} + kF = 0,$$

and the left and right boundary conditions imply

$$F(0) = 0, \qquad \text{and} \qquad F(a) = 0.$$

This gives $k = (n\pi/a)^2$ and corresponding nonzero solutions

$$F(x) = F_n(x) = \sin \frac{n\pi}{a} x, \qquad n = 1, 2, \cdots. \tag{15}$$

The ODE for G with $k = (n\pi/a)^2$ then becomes

$$\frac{d^2G}{dy^2} - \left(\frac{n\pi}{a}\right)^2 G = 0.$$

Solutions are

$$G(y) = G_n(y) = A_n e^{n\pi y/a} + B_n e^{-n\pi y/a}.$$

Now the boundary condition $u = 0$ on the lower side of R implies that $G_n(0) = 0$; that is, $G_n(0) = A_n + B_n = 0$ or $B_n = -A_n$. This gives

$$G_n(y) = A_n(e^{n\pi y/a} - e^{-n\pi y/a}) = 2A_n \sinh \frac{n\pi y}{a}.$$

From this and (15), writing $2A_n = A_n^*$, we obtain as the **eigenfunctions** of our problem

$$u_n(x, y) = F_n(x)G_n(y) = A_n^* \sin \frac{n\pi x}{a} \sinh \frac{n\pi y}{a}. \tag{16}$$

These solutions satisfy the boundary condition $u = 0$ on the left, right, and lower sides.

To get a solution also satisfying the boundary condition $u(x, b) = f(x)$ on the upper side, we consider the infinite series

$$u(x, y) = \sum_{n=1}^{\infty} u_n(x, y).$$

From this and (16) with $y = b$ we obtain

$$u(x, b) = f(x) = \sum_{n=1}^{\infty} A_n^* \sin \frac{n\pi x}{a} \sinh \frac{n\pi b}{a}.$$

We can write this in the form

$$u(x, b) = \sum_{n=1}^{\infty} \left(A_n^* \sinh \frac{n\pi b}{a}\right) \sin \frac{n\pi x}{a}.$$

This shows that the expressions in the parentheses must be the Fourier coefficients b_n of $f(x)$; that is, by (4) in Sec. 11.3,

$$b_n = A_n^* \sinh \frac{n\pi b}{a} = \frac{2}{a} \int_0^a f(x) \sin \frac{n\pi x}{a}\, dx.$$

From this and (16) we see that the solution of our problem is

$$\textbf{(17)} \qquad u(x, y) = \sum_{n=1}^{\infty} u_n(x, y) = \sum_{n=1}^{\infty} A_n^* \sin \frac{n\pi x}{a} \sinh \frac{n\pi y}{a}$$

where

$$\textbf{(18)} \qquad A_n^* = \frac{2}{a \sinh (n\pi b/a)} \int_0^a f(x) \sin \frac{n\pi x}{a}\, dx.$$

We have obtained this solution formally, neither considering convergence nor showing that the series for u, u_{xx}, and u_{yy} have the right sums. This can be proved if one assumes that f and f' are continuous and f'' is piecewise continuous on the interval $0 \leqq x \leqq a$. The proof is somewhat involved and relies on uniform convergence. It can be found in [C4] listed in App. 1.

Unifying Power of Methods. Electrostatics, Elasticity

The Laplace equation (14) also governs the electrostatic potential of electrical charges in any region that is free of these charges. Thus our steady-state heat problem can also be interpreted as an electrostatic potential problem. Then (17), (18) is the potential in the rectangle R when the upper side of R is at potential $f(x)$ and the other three sides are grounded.

Actually, in the steady-state case, the two-dimensional wave equation (to be considered in Secs. 12.7, 12.8) also reduces to (14). Then (17), (18) is the displacement of a rectangular elastic membrane (rubber sheet, drumhead) that is fixed along its boundary, with three sides lying in the xy-plane and the fourth side given the displacement $f(x)$.

This is another impressive demonstration of the ***unifying power*** of mathematics. It illustrates that ***entirely different physical systems may have the same mathematical model*** and can thus be treated by the same mathematical methods.

PROBLEM SET 12.5

1. **WRITING PROJECT. Wave and Heat Equations.** Compare the two PDEs with respect to type, general behavior of eigenfunctions, and kind of boundary and initial conditions and resulting practical problems. Also discuss the difference between Figs. 288 in Sec. 12.3 and 292.

2. **(Eigenfunctions)** Sketch (or graph) and compare the first three eigenfunctions (8) with $B_n = 1$, $c = 1$, $L = \pi$ for $t = 0, 0.2, 0.4, 0.6, 0.8, 1.0$.

3. **(Decay)** How does the rate of decay of (8) with fixed n depend on the specific heat, the density, and the thermal conductivity of the material?

4. If the first eigenfunction (8) of the bar decreases to half its value within 10 sec, what is the value of the diffusivity?

5–9 LATERALLY INSULATED BAR

A laterally insulated bar of length 10 cm and constant cross-sectional area 1 cm^2, of density 10.6 gm/cm^3, thermal conductivity 1.04 cal/(cm sec °C), and specific heat 0.056 cal/(gm °C) (this corresponds to silver, a good heat conductor) has initial temperature $f(x)$ and is kept at 0°C at the ends $x = 0$ and $x = 10$. Find the temperature $u(x, t)$ at later times. Here, $f(x)$ equals:

5. $f(x) = \sin 0.4\pi x$
6. $f(x) = \sin 0.1\pi x + \frac{1}{2} \sin 0.2\pi x$
7. $f(x) = 0.2x$ if $0 < x < 5$ and 0 otherwise
8. $f(x) = 1 - 0.2|x - 5|$
9. $f(x) = x$ if $0 < x < 2.5$, $f(x) = 2.5$ if $2.5 < x < 7.5$, $f(x) = 10 - x$ if $7.5 < x < 10$
10. **(Arbitrary temperatures at ends)** If the ends $x = 0$ and $x = L$ of the bar in the text are kept at constant temperatures U_1 and U_2, respectively, what is the temperature $u_I(x)$ in the bar after a long time (theoretically, as $t \to \infty$)? First guess, then calculate.
11. In Prob. 10 find the temperature at any time.
12. **(Changing end temperatures)** Assume that the ends of the bar in Probs. 5–9 have been kept at 100°C for a long time. Then at some instant, call it $t = 0$, the temperature at $x = L$ is suddenly changed to 0°C and kept at 0°C, whereas the temperature at $x = 0$ is kept at 100°C. Find the temperature in the middle of the bar at $t = 1, 2, 3, 10, 50$ sec. First guess, then calculate.

BAR UNDER ADIABATIC CONDITIONS

"Adiabatic" means no heat exchange with the neighborhood, because the bar is completely insulated, also at the ends. *Physical Information:* The heat flux at the ends is proportional to the value of $\partial u/\partial x$ there.

13. Show that for the completely insulated bar, $u_x(0, t) = 0$, $u_x(L, t) = 0$, $u(x, t) = f(x)$ and separation of variables gives the following solution, with A_n given by (2) in Sec. 11.3.

$$u(x, t) = A_0 + \sum_{n=1}^{\infty} A_n \cos \frac{n\pi x}{L} e^{-(cn\pi/L)^2 t}$$

14–19 Find the temperature in Prob. 13 with $L = \pi$, $c = 1$, and

14. $f(x) = x$
15. $f(x) = 1$
16. $f(x) = 0.5 \cos 4x$
17. $f(x) = \pi^2 - x^2$
18. $f(x) = \frac{1}{2}\pi - |x - \frac{1}{2}\pi|$
19. $f(x) = (x - \frac{1}{2}\pi)^2$
20. Find the temperature of the bar in Prob. 13 if the left end is kept at 0°C, the right end is insulated, and the initial temperature is $U_0 = const$.
21. The **boundary condition of heat transfer**

$$(19) \qquad -u_x(\pi, t) = k[u(\pi, t) - u_0]$$

applies when a bar of length π with $c = 1$ is laterally insulated, the left end $x = 0$ is kept at 0°C, and at the right end heat is flowing into air of constant temperature u_0. Let $k = 1$ for simplicity, and $u_0 = 0$. Show that a solution is $u(x, t) = \sin px\, e^{-p^2 t}$, where p is a solution of $\tan p\pi = -p$. Show graphically that this equation has infinitely many positive solutions $p_1, p_2, p_3, \cdots$, where $p_n > n - \frac{1}{2}$ and $\lim_{n\to\infty} (p_n - n + \frac{1}{2}) = 0$. (Formula (19) is also known as **radiation boundary condition,** but this is misleading; see Ref. [C3], p. 19.)
22. **(Discontinuous f)** Solve (1), (2), (3) with $L = \pi$ and $f(x) = U_0 = const$ ($\neq 0$) if $0 < x < \pi/2$, $f(x) = 0$ if $\pi/2 < x < \pi$.
23. **(Heat flux)** The *heat flux* of a solution $u(x, t)$ across $x = 0$ is defined by $\phi(t) = -Ku_x(0, t)$. Find $\phi(t)$ for the solution (9). Explain the name. Is it physically understandable that ϕ goes to 0 as $t \to \infty$?

OTHER HEAT EQUATIONS

24. **(Bar with heat generation)** If heat is generated at a constant rate throughout a bar of length $L = \pi$ with initial temperature $f(x)$ and the ends at $x = 0$ and π are kept at temperature 0, the heat equation is $u_t = c^2 u_{xx} + H$ with constant $H > 0$. Solve this problem. *Hint.* Set $u = v - Hx(x - \pi)/(2c^2)$.
25. **(Convection)** If heat in the bar in the text is free to flow through an end into the surrounding medium kept at 0°C, the PDE becomes $v_t = c^2 v_{xx} - \beta v$. Show that it can be reduced to the form (1) by setting $v(x, t) = u(x, t)w(t)$.
26. Consider $v_t = c^2 v_{xx} - v$ $(0 < x < L, t > 0)$, $v(0, t) = 0$, $v(L, t) = 0$, $v(x, 0) = f(x)$, where the term $-v$ models heat transfer to the surrounding medium kept at temperature 0. Reduce this PDE by setting $v(x, t) = u(x, t)w(t)$ with w such that u is given by (9), (10).
27. **(Nonhomogeneous heat equation)** Show that the problem modeled by

$$u_t - c^2 u_{xx} = Ne^{-\alpha x}$$

and (2), (3) can be reduced to a problem for the homogeneous heat equation by setting

$$u(x, t) = v(x, t) + w(x)$$

and determining w so that v satisfies the homogeneous PDE and the conditions $v(0, t) = v(L, t) = 0$, $v(x, 0) = f(x) - w(x)$. (The term $Ne^{-\alpha x}$ may represent heat loss due to radioactive decay in the bar.)

28–35 TWO-DIMENSIONAL PROBLEMS

28. (Laplace equation) Find the potential in the rectangle $0 \leqq x \leqq 20$, $0 \leqq y \leqq 40$ whose upper side is kept at potential 220 V and whose other sides are grounded.

29. Find the potential in the square $0 \leqq x \leqq 2$, $0 \leqq y \leqq 2$ if the upper side is kept at the potential $\sin \frac{1}{2}\pi x$ and the other sides are grounded.

30. CAS PROJECT. Isotherms. Find the steady-state solutions (temperatures) in the square plate in Fig. 294 with $a = 2$ satisfying the following boundary conditions. Graph isotherms.

(a) $u = \sin \pi x$ on the upper side, 0 on the others.

(b) $u = 0$ on the vertical sides, assuming that the other sides are perfectly insulated.

(c) Boundary conditions of your choice (such that the solution is not identically zero).

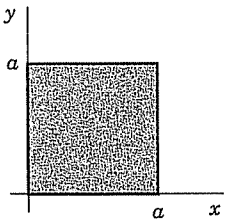


Fig. 294. Square plate

31. (Heat flow in a plate) The faces of the thin square plate in Fig. 294 with side $a = 24$ are perfectly insulated. The upper side is kept at 20°C and the other sides are kept at 0°C. Find the steady-state temperature $u(x, y)$ in the plate.

32. Find the steady-state temperature in the plate in Prob. 31 if the lower side is kept at U_0°C, the upper side at U_1°C, and the other sides are kept at 0°C. *Hint:* Split into two problems in which the boundary temperature is 0 on three sides for each problem.

33. (Mixed boundary value problem) Find the steady-state temperature in the plate in Prob. 31 with the upper and lower sides perfectly insulated, the left side kept at 0°C, and the right side kept at $f(y)$°C.

34. (Radiation) Find steady-state temperatures in the rectangle in Fig. 293 with the upper and left sides perfectly insulated and the right side radiating into a medium at 0°C according to $u_x(a, y) + hu(a, y) = 0$, $h > 0$ constant. (You will get many solutions since no condition on the lower side is given.)

35. Find formulas similar to (17), (18) for the temperature in the rectangle R of the text when the lower side of R is kept at temperature $f(x)$ and the other sides are kept at 0°C.

12.6 Heat Equation: Solution by Fourier Integrals and Transforms

Our discussion of the heat equation

$$\frac{\partial u}{\partial t} = c^2 \frac{\partial^2 u}{\partial x^2} \tag{1}$$

in the last section extends to bars of infinite length, which are good models of very long bars or wires (such as a wire of length, say, 300 ft). Then the role of Fourier series in the solution process will be taken by **Fourier integrals** (Sec. 11.7).

Let us illustrate the method by solving (1) for a bar that extends to infinity on both sides (and is laterally insulated as before). Then we do not have boundary conditions, but only the **initial condition**

$$u(x, 0) = f(x) \qquad (-\infty < x < \infty) \tag{2}$$

where $f(x)$ is the given initial temperature of the bar.

To solve this problem, we start as in the last section, substituting $u(x, t) = F(x)G(t)$ into (1). This gives the two ODEs

$$F'' + p^2 F = 0 \qquad \text{[see (5), Sec. 12.5]} \tag{3}$$

and

$$\dot{G} + c^2p^2G = 0 \qquad \text{[see (6), Sec. 12.5].} \tag{4}$$

Solutions are

$$F(x) = A\cos px + B\sin px \qquad \text{and} \qquad G(t) = e^{-c^2p^2t},$$

respectively, where A and B are any constants. Hence a solution of (1) is

$$u(x, t; p) = FG = (A\cos px + B\sin px)e^{-c^2p^2t}. \tag{5}$$

Here we had to choose the separation constant k negative, $k = -p^2$, because positive values of k would lead to an increasing exponential function in (5), which has no physical meaning.

Use of Fourier Integrals

Any series of functions (5), found in the usual manner by taking p as multiples of a fixed number, would lead to a function that is periodic in x when $t = 0$. However, since $f(x)$ in (2) is not assumed to be periodic, it is natural to use **Fourier integrals** instead of Fourier series. Also, A and B in (5) are arbitrary and we may regard them as functions of p, writing $A = A(p)$ and $B = B(p)$. Now, since the heat equation (1) is linear and homogeneous, the function

$$\textbf{(6)} \qquad u(x, t) = \int_0^\infty u(x, t; p)\,dp = \int_0^\infty [A(p)\cos px + B(p)\sin px]\,e^{-c^2p^2t}\,dp$$

is then a solution of (1), provided this integral exists and can be differentiated twice with respect to x and once with respect to t.

Determination of $A(p)$ and $B(p)$ from the Initial Condition. From (6) and (2) we get

$$u(x, 0) = \int_0^\infty [A(p)\cos px + B(p)\sin px]\,dp = f(x). \tag{7}$$

This gives $A(p)$ and $B(p)$ in terms of $f(x)$; indeed, from (4) in Sec. 11.7 we have

$$\textbf{(8)} \qquad A(p) = \frac{1}{\pi}\int_{-\infty}^{\infty} f(v)\cos pv\,dv, \qquad B(p) = \frac{1}{\pi}\int_{-\infty}^{\infty} f(v)\sin pv\,dv.$$

According to (1*), Sec. 11.9, our Fourier integral (7) with these $A(p)$ and $B(p)$ can be written

$$u(x, 0) = \frac{1}{\pi}\int_0^\infty \left[\int_{-\infty}^{\infty} f(v)\cos(px - pv)\,dv\right] dp.$$

Similarly, (6) in this section becomes

$$u(x, t) = \frac{1}{\pi}\int_0^\infty \left[\int_{-\infty}^{\infty} f(v)\cos(px - pv)\,e^{-c^2p^2t}\,dv\right] dp.$$

Assuming that we may reverse the order of integration, we obtain

$$u(x, t) = \frac{1}{\pi} \int_{-\infty}^{\infty} f(v) \left[\int_0^{\infty} e^{-c^2 p^2 t} \cos(px - pv)\, dp \right] dv. \tag{9}$$

Then we can evaluate the inner integral by using the formula

$$\int_0^{\infty} e^{-s^2} \cos 2bs\, ds = \frac{\sqrt{\pi}}{2} e^{-b^2}. \tag{10}$$

[A derivation of (10) is given in Problem Set 16.4 (Team Project 28).] This takes the form of our inner integral if we choose $p = s/(c\sqrt{t})$ as a new variable of integration and set

$$b = \frac{x - v}{2c\sqrt{t}}.$$

Then $2bs = (x - v)p$ and $ds = c\sqrt{t}\, dp$, so that (10) becomes

$$\int_0^{\infty} e^{-c^2 p^2 t} \cos(px - pv)\, dp = \frac{\sqrt{\pi}}{2c\sqrt{t}} \exp\left\{ -\frac{(x - v)^2}{4c^2 t} \right\}.$$

By inserting this result into (9) we obtain the representation

$$u(x, t) = \frac{1}{2c\sqrt{\pi t}} \int_{-\infty}^{\infty} f(v) \exp\left\{ -\frac{(x - v)^2}{4c^2 t} \right\} dv. \tag{11}$$

Taking $z = (v - x)/(2c\sqrt{t})$ as a variable of integration, we get the alternative form

$$u(x, t) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} f(x + 2cz\sqrt{t})\, e^{-z^2} dz. \tag{12}$$

If $f(x)$ is bounded for all values of x and integrable in every finite interval, it can be shown (see Ref. [C10]) that the function (11) or (12) satisfies (1) and (2). Hence this function is the required solution in the present case.

EXAMPLE 1 Temperature in an Infinite Bar

Find the temperature in the infinite bar if the initial temperature is (Fig. 295)

$$f(x) = \begin{cases} U_0 = const & \text{if } |x| < 1, \\ 0 & \text{if } |x| > 1. \end{cases}$$

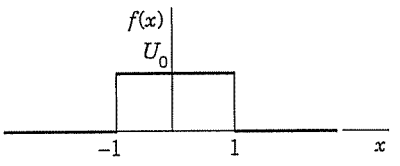


Fig. 295. Initial temperature in Example 1

Solution. From (11) we have

$$u(x, t) = \frac{U_0}{2c\sqrt{\pi t}} \int_{-1}^{1} \exp\left\{-\frac{(x - v)^2}{4c^2 t}\right\} dv.$$

If we introduce the above variable of integration z, then the integration over v from -1 to 1 corresponds to the integration over z from $(-1 - x)/(2c\sqrt{t})$ to $(1 - x)/(2c\sqrt{t})$, and

$$u(x, t) = \frac{U_0}{\sqrt{\pi}} \int_{-(1+x)/(2c\sqrt{t})}^{(1-x)/(2c\sqrt{t})} e^{-z^2}\, dz \qquad (t > 0). \tag{13}$$

We mention that this integral is not an elementary function, but can be expressed in terms of the error function, whose values have been tabulated. (Table A4 in App. 5 contains a few values; larger tables are listed in Ref. [GR1] in App. 1. See also CAS Project 10, p. 568.) Figure 296 shows $u(x, t)$ for $U_0 = 100°C$, $c^2 = 1\ \text{cm}^2/\text{sec}$, and several values of t. ■

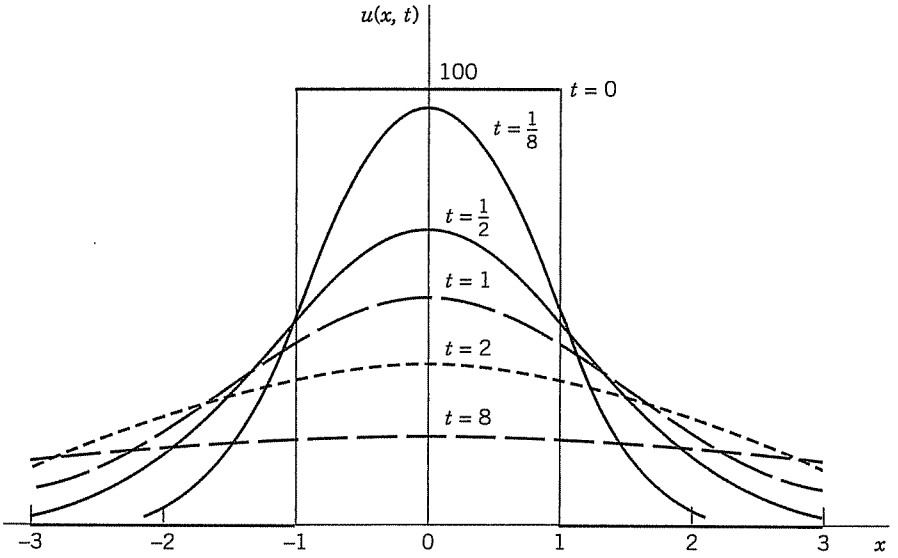


Fig. 296. Solution $u(x, t)$ in Example 1 for $U_0 = 100°C$, $c^2 = 1\ \text{cm}^2/\text{sec}$, and several values of t

Use of Fourier Transforms

The Fourier transform is closely related to the Fourier integral, from which we obtained the transform in Sec. 11.9. And the transition to the Fourier cosine and sine transform in Sec. 11.8 was even simpler. (You may perhaps wish to review this before going on.) Hence it should not surprise you that we can use these transforms for solving our present or similar problems. The Fourier transform applies to problems concerning the entire axis, and the Fourier cosine and sine transforms to problems involving the positive half-axis. Let us explain these transform methods by typical applications that fit our present discussion.

EXAMPLE 2 **Temperature in the Infinite Bar in Example 1**

Solve Example 1 using the Fourier transform.

Solution. The problem consists of the heat equation (1) and the initial condition (2), which in this example is

$$f(x) = U_0 = const \quad \text{if } |x| < 1 \qquad \text{and 0 otherwise.}$$

Our strategy is to take the Fourier transform with respect to x and then to solve the resulting ***ordinary*** DE in t. The details are as follows.

Let $\hat{u} = \mathcal{F}(u)$ denote the Fourier transform of u, ***regarded as a function of x.*** From (10) in Sec. 11.9 we see that the heat equation (1) gives

$$\mathcal{F}(u_t) = c^2\mathcal{F}(u_{xx}) = c^2(-w^2)\mathcal{F}(u) = -c^2w^2\hat{u}.$$

On the left, assuming that we may interchange the order of differentiation and integration, we have

$$\mathcal{F}(u_t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} u_t e^{-iwx}\,dx = \frac{1}{\sqrt{2\pi}}\frac{\partial}{\partial t}\int_{-\infty}^{\infty} u e^{-iwx}\,dx = \frac{\partial \hat{u}}{\partial t}.$$

Thus

$$\frac{\partial \hat{u}}{\partial t} = -c^2w^2\hat{u}.$$

Since this equation involves only a derivative with respect to t but none with respect to w, this is a first-order ***ordinary DE,*** with t as the independent variable and w as a parameter. By separating variables (Sec. 1.3) we get the general solution

$$\hat{u}(w, t) = C(w)e^{-c^2w^2t}$$

with the arbitrary "constant" $C(w)$ depending on the parameter w. The initial condition (2) yields the relationship $\hat{u}(w, 0) = C(w) = \hat{f}(w) = \mathcal{F}(f)$. Our intermediate result is

$$\hat{u}(w, t) = \hat{f}(w)e^{-c^2w^2t}.$$

The inversion formula (7), Sec. 11.9, now gives the solution

$$u(x, t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \hat{f}(w)e^{-c^2w^2t}e^{iwx}\,dw. \tag{14}$$

In this solution we may insert the Fourier transform

$$\hat{f}(w) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(v)e^{-ivw}\,dv.$$

Assuming that we may invert the order of integration, we then obtain

$$u(x, t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} f(v)\left[\int_{-\infty}^{\infty} e^{-c^2w^2t}e^{i(wx-wv)}\,dw\right]dv.$$

By the Euler formula (3). Sec. 11.9, the integrand of the inner integral equals

$$e^{-c^2w^2t}\cos(wx - wv) + ie^{-c^2w^2t}\sin(wx - wv).$$

We see that its imaginary part is an odd function of w, so that its integral is 0. (More precisely, this is the principal part of the integral; see Sec. 16.4.) The real part is an even function of w, so that its integral from $-\infty$ to ∞ equals twice the integral from 0 to ∞:

$$u(x, t) = \frac{1}{\pi}\int_{-\infty}^{\infty} f(v)\left[\int_{0}^{\infty} e^{-c^2w^2t}\cos(wx - wv)\,dw\right]dv.$$

This agrees with (9) (with $p = w$) and leads to the further formulas (11) and (13). ■

EXAMPLE 3 Solution in Example 1 by the Method of Convolution

Solve the heat problem in Example 1 by the method of convolution.

Solution. The beginning is as in Example 2 and leads to (14), that is,

$$u(x, t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \hat{f}(w)e^{-c^2w^2t}e^{iwx}\,dw. \tag{15}$$

Now comes the crucial idea. We recognize that this is of the form (13) in Sec. 11.9, that is,

$$u(x, t) = (f * g)(x) = \int_{-\infty}^{\infty} \hat{f}(w)\hat{g}(w)e^{iwx}\, dw \tag{16}$$

where

$$\hat{g}(w) = \frac{1}{\sqrt{2\pi}}\, e^{-c^2w^2t}. \tag{17}$$

Since, by the definition of convolution [(11), Sec. 11.9],

$$(f * g)(x) = \int_{-\infty}^{\infty} f(p)g(x - p)\, dp, \tag{18}$$

as our next and last step we must determine the inverse Fourier transform g of $\hat{g}$. For this we can use formula 9 in Table III of Sec. 11.10,

$$\mathcal{F}(e^{-ax^2}) = \frac{1}{\sqrt{2a}}\, e^{-w^2/(4a)}$$

with a suitable a. With $c^2t = 1/(4a)$ or $a = 1/(4c^2t)$, using (17) we obtain

$$\mathcal{F}(e^{-x^2/(4c^2t)}) = \sqrt{2c^2t}\, e^{-c^2w^2t} = \sqrt{2c^2t}\,\sqrt{2\pi}\,\hat{g}(w).$$

Hence $\hat{g}$ has the inverse

$$\frac{1}{\sqrt{2c^2t}\,\sqrt{2\pi}}\, e^{-x^2/(4c^2t)}.$$

Replacing x with $x - p$ and substituting this into (18) we finally have

$$u(x, t) = (f * g)(x) = \frac{1}{2c\sqrt{\pi t}} \int_{-\infty}^{\infty} f(p) \exp\left\{-\frac{(x - p)^2}{4c^2t}\right\} dp. \tag{19}$$

This solution formula of our problem agrees with (11). We wrote $(f * g)(x)$, without indicating the parameter t with respect to which we did not integrate. ▩

EXAMPLE 4 Fourier Sine Transform Applied to the Heat Equation

If a laterally insulated bar extends from $x = 0$ to infinity, we can use the Fourier sine transform. We let the initial temperature be $u(x, 0) = f(x)$ and impose the boundary condition $u(0, t) = 0$. Then from the heat equation and (9b) in Sec. 11.8, since $f(0) = u(0, 0) = 0$, we obtain

$$\mathcal{F}_s(u_t) = \frac{\partial \hat{u}_s}{\partial t} = c^2\mathcal{F}_s(u_{xx}) = -c^2w^2\mathcal{F}_s(u) = -c^2w^2\hat{u}_s(w, t).$$

This is a first-order ODE $\partial\hat{u}_s/\partial t + c^2w^2\hat{u}_s = 0$. Its solution is

$$\hat{u}_s(w, t) = C(w)e^{-c^2w^2t}.$$

From the initial condition $u(x, 0) = f(x)$ we have $\hat{u}_s(w, 0) = \hat{f}_s(w) = C(w)$. Hence

$$\hat{u}_s(w, t) = \hat{f}_s(w)e^{-c^2w^2t}.$$

Taking the inverse Fourier sine transform and substituting

$$\hat{f}_s(w) = \sqrt{\frac{2}{\pi}} \int_0^{\infty} f(p) \sin wp\, dp$$

on the right, we obtain the solution formula

$$u(x, t) = \frac{2}{\pi} \int_0^\infty \int_0^\infty f(p) \sin wp \, e^{-c^2 w^2 t} \sin wx \, dp \, dw. \tag{20}$$

Figure 297 shows (20) with $c = 1$ for $f(x) = 1$ if $0 \leqq x \leqq 1$ and 0 otherwise, graphed over the xt-plane for $0 \leqq x \leqq 2$, $0.01 \leqq t \leqq 1.5$. Note that the curves of $u(x, t)$ for constant t resemble those in Fig. 296 on p. 565.

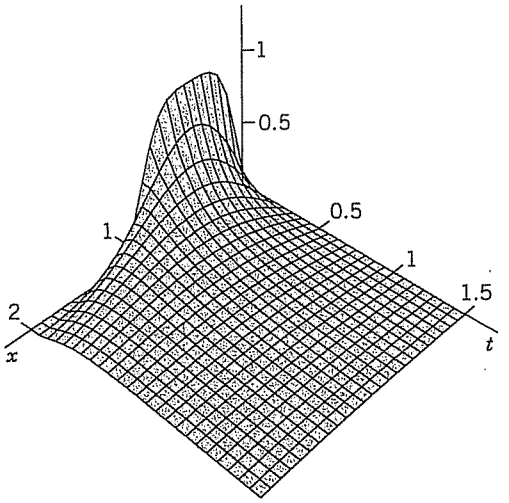


Fig. 297. Solution (20) in Example 4

PROBLEM SET 12.6

1–7 SOLUTION IN INTEGRAL FORM

Using (6), obtain the solution of (1) in integral form satisfying the initial condition $u(x, 0) = f(x)$, where

1. $f(x) = 1$ if $|x| < a$ and 0 otherwise
2. $f(x) = e^{-k|x|}$ $(k > 0)$
3. $f(x) = 1/(1 + x^2)$. [Use (15) in Sec. 11.7.]
4. $f(x) = (\sin x)/x$. [Use Prob. 4 in Sec. 11.7.]
5. $f(x) = (\sin \pi x)/x$. [Use Prob. 4 in Sec. 11.7.]
6. $f(x) = x$ if $|x| < 1$ and 0 otherwise
7. $f(x) = |x|$ if $|x| < 1$ and 0 otherwise.

8. Verify that u in Prob. 5 satisfies the initial condition.
9. **CAS PROJECT. Heat Flow.** (a) Graph the basic Fig. 296.

   (b) In (a) apply animation to "see" the heat flow in terms of the decrease of temperature.

   (c) Graph $u(x, t)$ with $c = 1$ as a surface over the upper xt-half-plane.
10. **CAS PROJECT. Error Function**

$$\text{erf } x = \frac{2}{\sqrt{\pi}} \int_0^x e^{-w^2} \, dw \tag{21}$$

This function is important in applied mathematics and physics (probability theory and statistics, thermodynamics, etc.) and fits our present discussion. Regarding it as a typical case of a special function defined by an integral that cannot be evaluated as in elementary calculus, do the following.

(a) Sketch or graph the **bell-shaped curve** [the curve of the integrand in (21)]. Show that erf x is odd. Show that

$$\int_a^b e^{-w^2} \, dw = \frac{\sqrt{\pi}}{2} (\text{erf } b - \text{erf } a),$$

$$\int_{-b}^b e^{-w^2} \, dw = \sqrt{\pi} \, \text{erf } b.$$

(b) Obtain the Maclaurin series of erf x from that of the integrand. Use that series to compute a table of erf x for $x = 0(0.01)3$ (meaning $x = 0, 0.01, 0.02, \cdots, 3$).

(c) Obtain the values required in (b) by an integration command of your CAS. Compare accuracy.

(d) It can be shown that erf $(\infty) = 1$. Confirm this experimentally by computing erf x for large x.

(e) Let $f(x) = 1$ when $x > 0$ and 0 when $x < 0$. Using $\text{erf}(\infty) = 1$, show that (12) then gives

$$u(x, t) = \frac{1}{\sqrt{\pi}} \int_{-x/(2c\sqrt{t})}^{\infty} e^{-z^2}\, dz = \frac{1}{2} - \frac{1}{2}\,\text{erf}\left(-\frac{x}{2c\sqrt{t}}\right) \qquad (t > 0).$$

(f) Express the temperature (13) in terms of the error function.

(g) Show that $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-s^2/2}\, ds = \frac{1}{2} + \frac{1}{2}\,\text{erf}\left(\frac{x}{\sqrt{2}}\right)$.

Here, the integral is the definition of the "distribution function of the normal distribution" to be discussed in Sec. 24.8.

12.7 Modeling: Membrane, Two-Dimensional Wave Equation

The vibrating string in Sec. 12.2 is a basic one-dimensional vibrational problem. Equally important is its two-dimensional analog, namely, the motion of an elastic membrane, such as a drumhead, that is stretched and then fixed along its edge. Indeed, setting up the model will proceed almost as in Sec. 12.2.

Physical Assumptions

1. The mass of the membrane per unit area is constant ("homogeneous membrane"). The membrane is perfectly flexible and offers no resistance to bending.
2. The membrane is stretched and then fixed along its entire boundary in the xy-plane. The tension per unit length T caused by stretching the membrane is the same at all points and in all directions and does not change during the motion.
3. The deflection $u(x, y, t)$ of the membrane during the motion is small compared to the size of the membrane, and all angles of inclination are small.

Although these assumptions cannot be realized exactly, they hold relatively accurately for small transverse vibrations of a thin elastic membrane, so that we shall obtain a good model, for instance, of a drumhead.

Derivation of the PDE of the Model ("Two-Dimensional Wave Equation") from Forces. As in Sec. 12.2 the model will consist of a PDE and additional conditions. The PDE will be obtained by the same method as in Sec. 12.2, namely, by considering the forces acting on a small portion of the physical system, the membrane in Fig. 298 on the next page, as it is moving up and down.

Since the deflections of the membrane and the angles of inclination are small, the sides of the portion are approximately equal to Δx and Δy. The tension T is the force per unit length. Hence the forces acting on the sides of the portion are approximately $T\,\Delta x$ and $T\,\Delta y$. Since the membrane is perfectly flexible, these forces are tangent to the moving membrane at every instant.

Horizontal Components of the Forces. We first consider the horizontal components of the forces. These components are obtained by multiplying the forces by the cosines of the angles of inclination. Since these angles are small, their cosines are close to 1. Hence

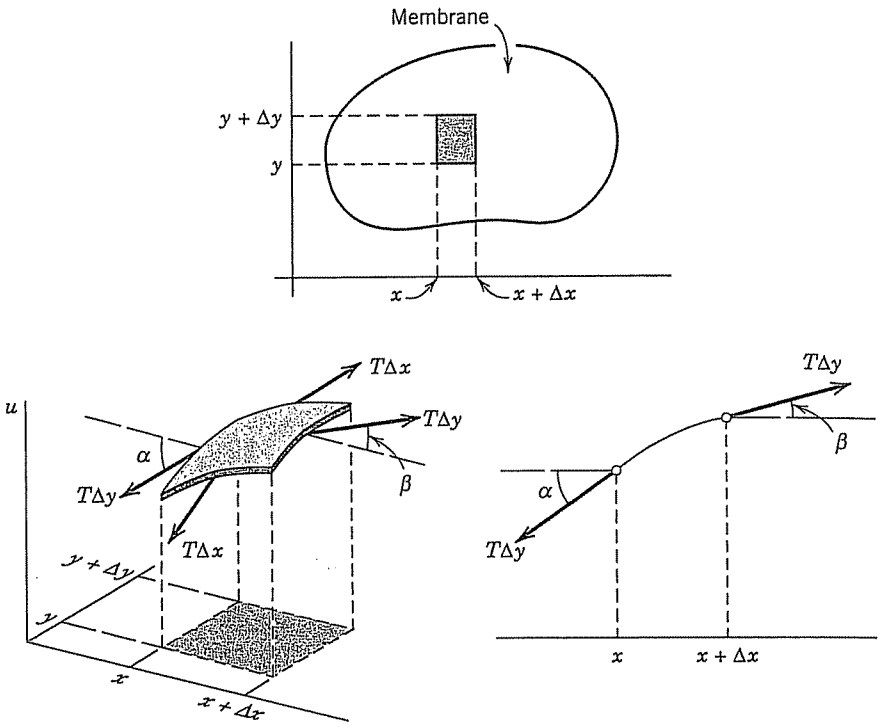


Fig. 298. Vibrating membrane

the horizontal components of the forces at opposite sides are approximately equal. Therefore, the motion of the particles of the membrane in a horizontal direction will be negligibly small. From this we conclude that we may regard the motion of the membrane as transversal; that is, each particle moves vertically.

Vertical Components of the Forces. These components along the right side and the left side are (Fig. 298), respectively,

$$T\,\Delta y \sin \beta \qquad \text{and} \qquad -T\,\Delta y \sin \alpha.$$

Here α and β are the values of the angle of inclination (which varies slightly along the edges) in the middle of the edges, and the minus sign appears because the force on the left side is directed downward. Since the angles are small, we may replace their sines by their tangents. Hence the resultant of those two vertical components is

$$\begin{aligned} T\,\Delta y\,(\sin \beta - \sin \alpha) &\approx T\,\Delta y\,(\tan \beta - \tan \alpha) \\ &= T\,\Delta y\,[u_x(x + \Delta x, y_1) - u_x(x, y_2)] \end{aligned} \tag{1}$$

where subscripts x denote partial derivatives and y_1 and y_2 are values between y and $y + \Delta y$. Similarly, the resultant of the vertical components of the forces acting on the other two sides of the portion is

$$T\,\Delta x\,[u_y(x_1, y + \Delta y) - u_y(x_2, y)] \tag{2}$$

where x_1 and x_2 are values between x and $x + \Delta x$.

Newton's Second Law Gives the PDE of the Model. By Newton's second law (see Sec. 2.4) the sum of the forces given by (1) and (2) is equal to the mass $\rho \Delta A$ of that small

portion times the acceleration $\partial^2 u/\partial t^2$; here ρ is the mass of the undeflected membrane per unit area, and $\Delta A = \Delta x\, \Delta y$ is the area of that portion when it is undeflected. Thus

$$\rho\, \Delta x\, \Delta y\, \frac{\partial^2 u}{\partial t^2} = T\, \Delta y\, [u_x(x + \Delta x, y_1) - u_x(x, y_2)] + T\, \Delta x\, [u_y(x_1, y + \Delta y) - u_y(x_2, y)]$$

where the derivative on the left is evaluated at some suitable point $(\tilde{x}, \tilde{y})$ corresponding to that portion. Division by $\rho\, \Delta x\, \Delta y$ gives

$$\frac{\partial^2 u}{\partial t^2} = \frac{T}{\rho} \left[\frac{u_x(x + \Delta x, y_1) - u_x(x, y_2)}{\Delta x} + \frac{u_y(x_1, y + \Delta y) - u_y(x_2, y)}{\Delta y} \right].$$

If we let Δx and Δy approach zero, we obtain the PDE of the model

(3) $$\frac{\partial^2 u}{\partial t^2} = c^2 \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right) \qquad c^2 = \frac{T}{\rho}.$$

This PDE is called the **two-dimensional wave equation.** The expression in parentheses is the Laplacian $\nabla^2 u$ of u (Sec. 10.8). Hence (3) can be written

(3′) $$\frac{\partial^2 u}{\partial t^2} = c^2 \nabla^2 u.$$

Solutions of the wave equation (3) will be obtained and discussed in the next section.

12.8 Rectangular Membrane. Double Fourier Series

The model of the vibrating membrane for obtaining the displacement $u(x, y, t)$ of a point (x, y) of the membrane from rest ($u = 0$) at time t is

(1) $$\frac{\partial^2 u}{\partial t^2} = c^2 \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right)$$

(2) $$u = 0 \text{ on the boundary}$$

(3a) $$u(x, y, 0) = f(x, y)$$

(3b) $$u_t(x, y, 0) = g(x, y).$$

Here (1) is the **two-dimensional wave equation** with $c^2 = T/\rho$ just derived, (2) is the **boundary condition** (membrane fixed along the boundary in the xy-plane for all times $t \geqq 0$), and (3) are the **initial conditions** at $t = 0$, consisting of the given *initial displacement* (initial shape) $f(x, y)$ and the given *initial velocity* $g(x, y)$, where $u_t = \partial u/\partial t$. We see that these conditions are quite similar to those for the string in Sec. 12.2.

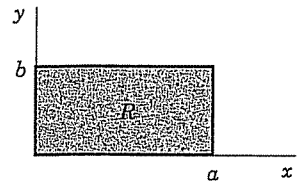


Fig. 299. Rectangular membrane

As a first important model, let us consider the **rectangular membrane** R in Fig. 299, which is simpler than the circular drumhead to follow. Then the boundary in (2) is the rectangle in Fig. 299. We shall solve this problem in three steps:

Step 1. By separating variables, setting $u(x, y, t) = F(x, y)G(t)$ and later $F(x, y) = H(x)Q(y)$ we obtain from (1) an ODE (4) for G and later from a PDE (5) for F two ODEs (6) and (7) for H and Q.

Step 2. From the solutions of those ODEs we determine solutions (13) of (1) ("**eigenfunctions**" u_{mn}) that satisfy the boundary condition (2).

Step 3. We compose the u_{mn} into a double series (14) solving the whole model (1), (2), (3).

Step 1. Three ODEs From the Wave Equation (1)

To obtain ODEs from (1), we apply two successive separations of variables. In the first separation we set $u(x, y, t) = F(x, y)G(t)$. Substitution into (1) gives

$$F\ddot{G} = c^2(F_{xx}G + F_{yy}G)$$

where subscripts denote partial derivatives and dots denote derivatives with respect to t. To separate the variables, we divide both sides by c^2FG:

$$\frac{\ddot{G}}{c^2G} = \frac{1}{F}(F_{xx} + F_{yy}).$$

Since the left side depends only on t, whereas the right side is independent of t, both sides must equal a constant. By a simple investigation we see that only negative values of that constant will lead to solutions that satisfy (2) without being identically zero; this is similar to Sec. 12.3. Denoting that negative constant by $-\nu^2$, we have

$$\frac{\ddot{G}}{c^2G} = \frac{1}{F}(F_{xx} + F_{yy}) = -\nu^2.$$

This gives two equations: for the **"time function"** $G(t)$ we have the ODE

$$\ddot{G} + \lambda^2 G = 0 \qquad \text{where } \lambda = c\nu, \tag{4}$$

and for the **"amplitude function"** $F(x, y)$ a PDE, called the *two-dimensional* **Helmholtz[3] equation**

$$F_{xx} + F_{yy} + \nu^2 F = 0. \tag{5}$$

[3]HERMANN VON HELMHOLTZ (1821–1894), German physicist, known for his basic work in thermodynamics, fluid flow, and acoustics.

Separation of the Helmholtz equation is achieved if we set $F(x, y) = H(x)Q(y)$. By substitution of this into (5) we obtain

$$\frac{d^2H}{dx^2}\,Q = -\left(H\,\frac{d^2Q}{dy^2} + \nu^2 HQ\right).$$

To separate the variables, we divide both sides by HQ, finding

$$\frac{1}{H}\,\frac{d^2H}{dx^2} = -\frac{1}{Q}\left(\frac{d^2Q}{dy^2} + \nu^2 Q\right).$$

Both sides must equal a constant, by the usual argument. This constant must be negative, say, $-k^2$, because only negative values will lead to solutions that satisfy (2) without being identically zero. Thus

$$\frac{1}{H}\,\frac{d^2H}{dx^2} = -\frac{1}{Q}\left(\frac{d^2Q}{dy^2} + \nu^2 Q\right) = -k^2.$$

This yields two ODEs for H and Q, namely,

(6) $$\frac{d^2H}{dx^2} + k^2 H = 0$$

and

(7) $$\frac{d^2Q}{dy^2} + p^2 Q = 0 \qquad \text{where } p^2 = \nu^2 - k^2.$$

Step 2. Satisfying the Boundary Condition

General solutions of (6) and (7) are

$$H(x) = A\cos kx + B\sin kx \qquad \text{and} \qquad Q(y) = C\cos py + D\sin py$$

with constant A, B, C, D. From $u = FG$ and (2) it follows that $F = HQ$ must be zero on the boundary, that is, on the edges $x = 0$, $x = a$, $y = 0$, $y = b$; see Fig. 299. This gives the conditions

$$H(0) = 0, \qquad H(a) = 0, \qquad Q(0) = 0, \qquad Q(b) = 0.$$

Hence $H(0) = A = 0$ and then $H(a) = B\sin ka = 0$. Here we must take $B \neq 0$ since otherwise $H(x) \equiv 0$ and $F(x, y) \equiv 0$. Hence $\sin ka = 0$ or $ka = m\pi$, that is,

$$k = \frac{m\pi}{a} \quad (m \text{ integer}).$$

In precisely the same fashion we conclude that $C = 0$ and p must be restricted to the values $p = n\pi/b$ where n is an integer. We thus obtain the solutions $H = H_m$, $Q = Q_n$, where

$$H_m(x) = \sin \frac{m\pi x}{a} \qquad \text{and} \qquad Q_n(y) = \sin \frac{n\pi y}{b}, \qquad \begin{matrix} m = 1, 2, \cdots, \\ n = 1, 2, \cdots. \end{matrix}$$

As in the case of the vibrating string, it is not necessary to consider $m, n = -1, -2, \cdots$ since the corresponding solutions are essentially the same as for positive m and n, except for a factor -1. Hence the functions

$$\textbf{(8)} \qquad F_{mn}(x, y) = H_m(x)Q_n(y) = \sin \frac{m\pi x}{a} \sin \frac{n\pi y}{b}, \qquad \begin{matrix} m = 1, 2, \cdots, \\ n = 1, 2, \cdots, \end{matrix}$$

are solutions of the Helmholtz equation (5) that are zero on the boundary of our membrane.

Eigenfunctions and Eigenvalues. Having taken care of (5), we turn to (4). Since $p^2 = \nu^2 - k^2$ in (7) and $\lambda = c\nu$ in (4), we have

$$\lambda = c\sqrt{k^2 + p^2}.$$

Hence to $k = m\pi/a$ and $p = n\pi/b$ there corresponds the value

$$\textbf{(9)} \qquad \lambda = \lambda_{mn} = c\pi \sqrt{\frac{m^2}{a^2} + \frac{n^2}{b^2}}, \qquad \begin{matrix} m = 1, 2, \cdots, \\ n = 1, 2, \cdots, \end{matrix}$$

in the ODE (4). A corresponding general solution of (4) is

$$G_{mn}(t) = B_{mn} \cos \lambda_{mn}t + B^*_{mn} \sin \lambda_{mn}t.$$

It follows that the functions $u_{mn}(x, y, t) = F_{mn}(x, y)G_{mn}(t)$, written out

$$\textbf{(10)} \qquad u_{mn}(x, y, t) = (B_{mn} \cos \lambda_{mn}t + B^*_{mn} \sin \lambda_{mn}t) \sin \frac{m\pi x}{a} \sin \frac{n\pi y}{b}$$

with λ_{mn} according to (9), are solutions of the wave equation (1) that are zero on the boundary of the rectangular membrane in Fig. 299. These functions are called the **eigenfunctions** or *characteristic functions,* and the numbers λ_{mn} are called the **eigenvalues** or *characteristic values* of the vibrating membrane. The frequency of u_{mn} is $\lambda_{mn}/2\pi$.

Discussion of Eigenfunctions. It is very interesting that, depending on a and b, several functions F_{mn} may correspond to the same eigenvalue. Physically this means that there may exist vibrations having the same frequency but entirely different **nodal lines** (curves of points on the membrane that do not move). Let us illustrate this with the following example.

EXAMPLE 1 **Eigenvalues and Eigenfunctions of the Square Membrane**

Consider the square membrane with $a = b = 1$. From (9) we obtain its eigenvalues

$$\lambda_{mn} = c\pi\sqrt{m^2 + n^2}. \tag{11}$$

Hence $\lambda_{mn} = \lambda_{nm}$, but for $m \neq n$ the corresponding functions

$$F_{mn} = \sin m\pi x \sin n\pi y \qquad \text{and} \qquad F_{nm} = \sin n\pi x \sin m\pi y$$

are certainly different. For example, to $\lambda_{12} = \lambda_{21} = c\pi\sqrt{5}$ there correspond the two functions

$$F_{12} = \sin \pi x \sin 2\pi y \qquad \text{and} \qquad F_{21} = \sin 2\pi x \sin \pi y.$$

Hence the corresponding solutions

$$u_{12} = (B_{12} \cos c\pi\sqrt{5}t + B_{12}^* \sin c\pi\sqrt{5}t)F_{12} \qquad \text{and} \qquad u_{21} = (B_{21} \cos c\pi\sqrt{5}t + B_{21}^* \sin c\pi\sqrt{5}t)F_{21}$$

have the nodal lines $y = \frac{1}{2}$ and $x = \frac{1}{2}$, respectively (see Fig. 300). Taking $B_{12} = 1$ and $B_{12}^* = B_{21}^* = 0$, we obtain

$$u_{12} + u_{21} = \cos c\pi\sqrt{5}t\,(F_{12} + B_{21}F_{21}) \tag{12}$$

which represents another vibration corresponding to the eigenvalue $c\pi\sqrt{5}$. The nodal line of this function is the solution of the equation

$$F_{12} + B_{21}F_{21} = \sin \pi x \sin 2\pi y + B_{21} \sin 2\pi x \sin \pi y = 0$$

or, since $\sin 2\alpha = 2 \sin \alpha \cos \alpha$,

$$\sin \pi x \sin \pi y\,(\cos \pi y + B_{21} \cos \pi x) = 0. \tag{13}$$

This solution depends on the value of B_{21} (see Fig. 301).

From (11) we see that even more than two functions may correspond to the same numerical value of λ_{mn}. For example, the four functions F_{18}, F_{81}, F_{47}, and F_{74} correspond to the value

$$\lambda_{18} = \lambda_{81} = \lambda_{47} = \lambda_{74} = c\pi\sqrt{65}, \qquad \text{because} \qquad 1^1 + 8^2 = 4^2 + 7^2 = 65.$$

This happens because 65 can be expressed as the sum of two squares of positive integers in several ways. According to a theorem by Gauss, this is the case for every sum of two squares among whose prime factors there are at least two different ones of the form $4n + 1$ where n is a positive integer. In our case we have $65 = 5 \cdot 13 = (4 + 1)(12 + 1)$. ■

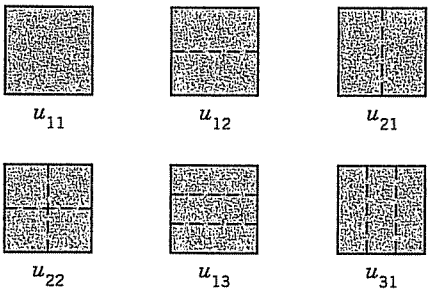


Fig. 300. Nodal lines of the solutions u_{11}, u_{12}, u_{21}, u_{22}, u_{13}, u_{31} in the case of the square membrane

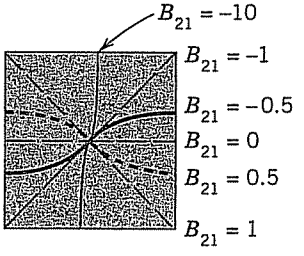


Fig. 301. Nodal lines of the solution (12) for some values of B_{21}

Step 3. Solution of the Model (1), (2), (3). Double Fourier Series

So far we have solutions (10) satisfying (1) and (2) only. To obtain the solution that also satisfies (3), we proceed as in Sec. 12.3. We consider the double series

$$\begin{aligned} u(x, y, t) &= \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} u_{mn}(x, y, t) \\ &= \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} (B_{mn} \cos \lambda_{mn} t + B_{mn}^{*} \sin \lambda_{mn} t) \sin \frac{m\pi x}{a} \sin \frac{n\pi y}{b} \end{aligned} \tag{14}$$

(without discussing convergence and uniqueness). From (14) and (3a), setting $t = 0$, we have

$$u(x, y, 0) = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} B_{mn} \sin \frac{m\pi x}{a} \sin \frac{n\pi y}{b} = f(x, y). \tag{15}$$

Suppose that $f(x, y)$ can be represented by (15). (Sufficient for this is the continuity of f, $\partial f/\partial x$, $\partial f/\partial y$, $\partial^2 f/\partial x \partial y$ in R.) Then (15) is called the **double Fourier series** of $f(x, y)$. Its coefficients can be determined as follows. Setting

$$K_m(y) = \sum_{n=1}^{\infty} B_{mn} \sin \frac{n\pi y}{b} \tag{16}$$

we can write (15) in the form

$$f(x, y) = \sum_{m=1}^{\infty} K_m(y) \sin \frac{m\pi x}{a}.$$

For fixed y this is the Fourier sine series of $f(x, y)$, considered as a function of x. From (4) in Sec. 11.3 we see that the coefficients of this expansion are

$$K_m(y) = \frac{2}{a} \int_0^a f(x, y) \sin \frac{m\pi x}{a} \, dx. \tag{17}$$

Furthermore, (16) is the Fourier sine series of $K_m(y)$, and from (4) in Sec. 11.3 it follows that the coefficients are

$$B_{mn} = \frac{2}{b} \int_0^b K_m(y) \sin \frac{n\pi y}{b} \, dy.$$

From this and (17) we obtain the **generalized Euler formula**

$$B_{mn} = \frac{4}{ab} \int_0^b \int_0^a f(x, y) \sin \frac{m\pi x}{a} \sin \frac{n\pi y}{b} \, dx \, dy \qquad \begin{matrix} m = 1, 2, \cdots \\ n = 1, 2, \cdots \end{matrix} \tag{18}$$

for the **Fourier coefficients** of $f(x, y)$ in the double Fourier series (15).

The B_{mn} in (14) are now determined in terms of $f(x, y)$. To determine the B^*_{mn}, we differentiate (14) termwise with respect to t; using (3b), we obtain

$$\left.\frac{\partial u}{\partial t}\right|_{t=0} = \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} B^*_{mn}\lambda_{mn} \sin\frac{m\pi x}{a} \sin\frac{n\pi y}{b} = g(x, y).$$

Suppose that $g(x, y)$ can be developed in this double Fourier series. Then, proceeding as before, we find that the coefficients are

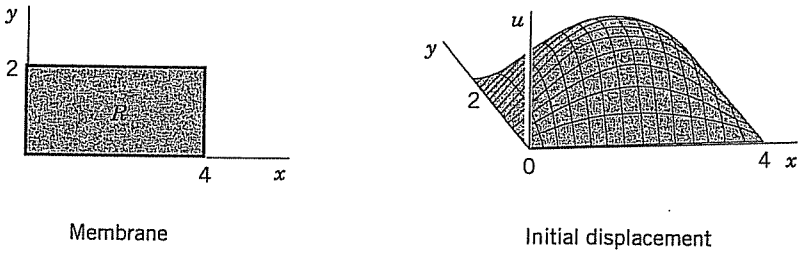
$$\textbf{(19)} \qquad B^*_{mn} = \frac{4}{ab\lambda_{mn}} \int_0^b\int_0^a g(x, y) \sin\frac{m\pi x}{a} \sin\frac{n\pi y}{b}\, dx\, dy \qquad \begin{matrix} m = 1, 2, \cdots \\ n = 1, 2, \cdots. \end{matrix}$$

Result. *If f and g in* (3) *are such that u can be represented by* (14), *then* (14) *with coefficients* (18) *and* (19) *is the solution of the model* (1), (2), (3).

EXAMPLE 2 **Vibration of a Rectangular Membrane**

Find the vibrations of a rectangular membrane of sides $a = 4$ ft and $b = 2$ ft (Fig. 302) if the tension is 12.5 lb/ft, the density is 2.5 slugs/ft^2 (as for light rubber), the initial velocity is 0, and the initial displacement is

$$(20) \qquad f(x, y) = 0.1(4x - x^2)(2y - y^2) \text{ ft}.$$



Membrane Initial displacement

Fig. 302. Example 2

Solution. $c^2 = T/\rho = 12.5/2.5 = 5$ [ft^2/sec^2]. Also, $B^*_{mn} = 0$ from (19). From (18) and (20),

$$B_{mn} = \frac{4}{4 \cdot 2}\int_0^2\int_0^4 0.1(4x - x^2)(2y - y^2) \sin\frac{m\pi x}{4} \sin\frac{n\pi y}{2}\, dx\, dy$$

$$= \frac{1}{20}\int_0^4 (4x - x^2) \sin\frac{m\pi x}{4}\, dx \int_0^2 (2y - y^2) \sin\frac{n\pi y}{2}\, dy.$$

Two integrations by parts give for the first integral on the right

$$\frac{128}{m^3\pi^3}\left[1 - (-1)^m\right] = \frac{256}{m^3\pi^3} \qquad (m \text{ odd})$$

and for the second integral

$$\frac{16}{n^3\pi^3}\left[1 - (-1)^n\right] = \frac{32}{n^3\pi^3} \qquad (n \text{ odd}).$$

For even m or n we get 0. Together with the factor 1/20 we thus have $B_{mn} = 0$ if m or n is even and

$$B_{mn} = \frac{256 \cdot 32}{20m^3n^3\pi^6} \approx \frac{0.426\,050}{m^3n^3} \qquad (m \text{ and } n \text{ both odd}).$$

From this, (9), and (14) we obtain the answer

$$u(x, y, t) = 0.426\,050 \sum_{m,n \text{ odd}} \sum \frac{1}{m^3 n^3} \cos\left(\frac{\sqrt{5}\pi}{4}\sqrt{m^2 + 4n^2}\right) t \sin\frac{m\pi x}{4} \sin\frac{n\pi y}{2}$$

$$(21) \quad = 0.426\,050 \left(\cos\frac{\sqrt{5}\pi\sqrt{5}}{4} t \sin\frac{\pi x}{4} \sin\frac{\pi y}{2} + \frac{1}{27}\cos\frac{\sqrt{5}\pi\sqrt{37}}{4} t \sin\frac{\pi x}{4}\sin\frac{3\pi y}{2}\right.$$

$$\left. + \frac{1}{27}\cos\frac{\sqrt{5}\pi\sqrt{13}}{4} t \sin\frac{3\pi x}{4}\sin\frac{\pi y}{2} + \frac{1}{729}\cos\frac{\sqrt{5}\pi\sqrt{45}}{4} t \sin\frac{3\pi x}{4}\sin\frac{3\pi y}{2} + \cdots\right).$$

To discuss this solution, we note that the first term is very similar to the initial shape of the membrane, has no nodal lines, and is by far the dominating term because the coefficients of the next terms are much smaller. The second term has two horizontal nodal lines ($y = 2/3, 4/3$), the third term two vertical ones ($x = 4/3, 8/3$), the fourth term two horizontal and two vertical ones, and so on. ■

PROBLEM SET 12.8

1. **(Frequency)** How does the frequency of the eigenfunctions of the rectangular membrane change if (a) we double the tension, (b) we take a membrane of half the mass of the original one, (c) we double the sides of the membrane? (Give reason.)

SQUARE MEMBRANE

2. Determine and sketch the nodal lines of the eigenfunctions of the square membrane for $m = 1, 2, 3, 4$ and $n = 1, 2, 3, 4$.

3–8 **Double Fourier Series.** Represent $f(x, y)$ by a series (15), where $0 < x < 1$, $0 < y < 1$.

3. $f(x, y) = 1$
4. $f(x, y) = x$
5. $f(x, y) = y$
6. $f(x, y) = x + y$
7. $f(x, y) = xy$
8. $f(x, y) = xy(1 - x)(1 - y)$

9. **CAS PROJECT. Double Fourier Series. (a)** Write a program that gives and graphs partial sums of (15). Apply it to Probs. 4 and 5. Do the graphs show that those partial sums satisfy the boundary condition (3a)? Explain why. Why is the convergence rapid?

   (b) Do the tasks in (a) for Prob. 3. Graph a portion, say, $0 < x < \frac{1}{2}$, $0 < y < \frac{1}{2}$, of several partial sums on common axes, so that you can see how they differ. (See Fig. 303.)

   (c) Do the tasks in (b) for functions of your choice.

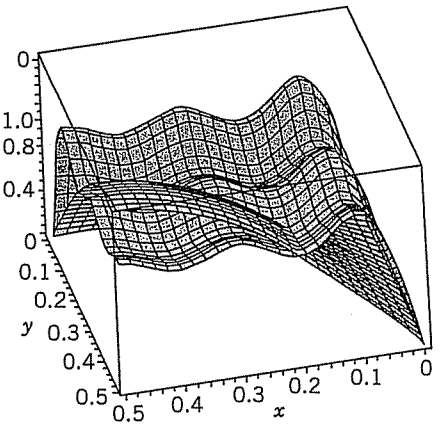


Fig. 303. Partial sums $S_{2,2}$ and $S_{10,10}$ in CAS Project 9b

10. **CAS EXPERIMENT. Quadruples of F_{mn}.** Write a program that gives you four numerically equal λ_{mn} in Example 1, so that four different F_{mn} correspond to it. Sketch the nodal lines of F_{18}, F_{81}, F_{47}, F_{74} in Example 1 and similarly for further F_{mn} that you will find.

11–13 **Deflection.** Find the deflection $u(x, y, t)$ of the square membrane of side π and $c^2 = 1$ if the initial velocity is 0 and the initial deflection is

11. $k \sin 2x \sin 5y$
12. $0.1 \sin x \sin y$
13. $0.1xy(\pi - x)(\pi - y)$

RECTANGULAR MEMBRANE

14. Verify the discussion of the terms of (21) in Example 2.
15. Repeat the task of Prob. 2 when $a = 4$ and $b = 1$.

16. Verify the calculation of B_{mn} in Example 2 by integration by parts.

17. Find eigenvalues of the rectangular membrane of sides $a = 2$ and $b = 1$ to which there correspond two or more different (independent) eigenfunctions.

18. **(Minimum property)** Show that among all rectangular membranes of the same area $A = ab$ and the same c the square membrane is that for which u_{11} [see (10)] has the lowest frequency.

19–22 **Double Fourier Series.** Represent $f(x, y)$ $(0 < x < a, 0 < y < b)$ by a double Fourier series (15).

19. $f(x, y) = k$

20. $f(x, y) = 0.25xy$

21. $f(x, y) = xy(a^2 - x^2)(b^2 - y^2)$

22. $f(x, y) = xy(a - x)(b - y)$

23. **(Deflection)** Find the deflection of the membrane of sides a and b with $c^2 = 1$ for the initial deflection $f(x, y) = \sin \frac{3\pi x}{a} \sin \frac{4\pi y}{b}$ and initial velocity 0.

24. Repeat the task in Prob. 23 with $c^2 = 1$, for $f(x, y)$ as in Prob. 22 and initial velocity 0.

25. **(Forced vibrations)** Show that forced vibrations of a membrane are modeled by the PDE $u_{tt} = c^2\nabla^2 u + P/\rho$, where $P(x, y, t)$ is the external force per unit area acting perpendicular to the xy-plane.

12.9 Laplacian in Polar Coordinates. Circular Membrane. Fourier–Bessel Series

In boundary value problems for PDEs it is a ***general principle*** to use coordinates in which the formula for the boundary is as simple as possible. Since we want to discuss circular membranes (drumheads), we first transform the Laplacian in the wave equation (1), Sec. 12.8,

$$u_{tt} = c^2\nabla^2 u = c^2(u_{xx} + u_{yy}) \tag{1}$$

(subscripts denoting partial derivatives) into **polar coordinates**

$$r = \sqrt{x^2 + y^2}, \qquad \theta = \arctan \frac{y}{x}.$$

Hence $x = r\cos\theta$, $y = r\sin\theta$. By the chain rule (Sec. 9.6) we obtain

$$u_x = u_r r_x + u_\theta \theta_x.$$

Differentiating once more with respect to x and using the product rule and then again the chain rule gives

$$\begin{aligned} u_{xx} &= (u_r r_x)_x + (u_\theta \theta_x)_x \\ &= (u_r)_x r_x + u_r r_{xx} + (u_\theta)_x \theta_x + u_\theta \theta_{xx} \\ &= (u_{rr} r_x + u_{r\theta}\theta_x) r_x + u_r r_{xx} + (u_{\theta r} r_x + u_{\theta\theta}\theta_x)\theta_x + u_\theta \theta_{xx}. \end{aligned} \tag{2}$$

Also, by differentiation of r and θ we find

$$r_x = \frac{x}{\sqrt{x^2 + y^2}} = \frac{x}{r}, \qquad \theta_x = \frac{1}{1 + (y/x)^2}\left(-\frac{y}{x^2}\right) = -\frac{y}{r^2}.$$

Differentiating these two formulas again, we obtain

$$r_{xx} = \frac{r - xr_x}{r^2} = \frac{1}{r} - \frac{x^2}{r^3} = \frac{y^2}{r^3}, \qquad \theta_{xx} = -y\left(-\frac{2}{r^3}\right) r_x = \frac{2xy}{r^4}.$$

We substitute all these expressions into (2). Assuming continuity of the first and second partial derivatives, we have $u_{r\theta} = u_{\theta r}$, and by simplifying,

$$u_{xx} = \frac{x^2}{r^2}\, u_{rr} - 2\,\frac{xy}{r^3}\, u_{r\theta} + \frac{y^2}{r^4}\, u_{\theta\theta} + \frac{y^2}{r^3}\, u_r + 2\,\frac{xy}{r^4}\, u_\theta. \tag{3}$$

In a similar fashion it follows that

$$u_{yy} = \frac{y^2}{r^2}\, u_{rr} + 2\,\frac{xy}{r^3}\, u_{r\theta} + \frac{x^2}{r^4}\, u_{\theta\theta} + \frac{x^2}{r^3}\, u_r - 2\,\frac{xy}{r^4}\, u_\theta. \tag{4}$$

By adding (3) and (4) we see that the **Laplacian of u in polar coordinates** is

$$\nabla^2 u = \frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\,\frac{\partial u}{\partial r} + \frac{1}{r^2}\,\frac{\partial^2 u}{\partial \theta^2}. \tag{5}$$

Circular Membrane

Circular membranes occur in drums, pumps, microphones, telephones, and so on. This accounts for their great importance in engineering. Whenever a circular membrane is plane and its material is elastic, but offers no resistance to bending (this excludes thin metallic membranes!), its vibrations are modeled by the **two-dimensional wave equation in polar coordinates** obtained from (1) with $\nabla^2 u$ given by (5), that is,

$$\frac{\partial^2 u}{\partial t^2} = c^2\left(\frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\,\frac{\partial u}{\partial r} + \frac{1}{r^2}\,\frac{\partial^2 u}{\partial \theta^2}\right) \qquad c^2 = \frac{T}{\rho}. \tag{6}$$

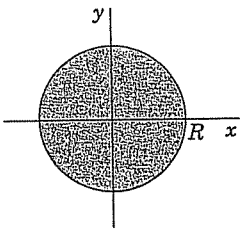


Fig. 304. Circular membrane

We shall consider a membrane of radius R (Fig. 304) and determine solutions $u(r, t)$ that are radially symmetric. (Solutions also depending on the angle θ will be discussed in the problem set.) Then $u_{\theta\theta} = 0$ in (6) and the model of the problem (the analog of (1), (2), (3) in Sec. 12.8) is

$$\frac{\partial^2 u}{\partial t^2} = c^2\left(\frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\,\frac{\partial u}{\partial r}\right) \tag{7}$$

$$u(R, t) = 0 \text{ for all } t \geqq 0 \tag{8}$$

$$u(r, 0) = f(r) \tag{9a}$$

$$u_t(r, 0) = g(r). \tag{9b}$$

Here (8) means that the membrane is fixed along the boundary circle $r = R$. The initial deflection $f(r)$ and the initial velocity $g(r)$ depend only on r, not on θ, so that we can expect radially symmetric solutions $u(r, t)$.

Step 1. Two ODEs From the Wave Equation (7). Bessel's Equation

Using the **method of separation of variables,** we first determine solutions $u(r, t) = W(r)\,G(t)$. (We write W, not F because W depends on r, whereas F, used before, depended on x.) Substituting $u = WG$ and its derivatives into (7) and dividing the result by c^2WG, we get

$$\frac{\ddot{G}}{c^2G} = \frac{1}{W}\left(W'' + \frac{1}{r}\,W'\right)$$

where dots denote derivatives with respect to t and primes denote derivatives with respect to r. The expressions on both sides must equal a constant. This constant must be negative, say, $-k^2$, in order to obtain solutions that satisfy the boundary condition without being identically zero. Thus,

$$\frac{\ddot{G}}{c^2G} = \frac{1}{W}\left(W'' + \frac{1}{r}\,W'\right) = -k^2.$$

This gives the two linear ODEs

(10) $$\ddot{G} + \lambda^2 G = 0 \qquad \text{where } \lambda = ck$$

and

(11) $$W'' + \frac{1}{r}\,W' + k^2W = 0.$$

We can reduce (11) to Bessel's equation (Sec. 5.5) if we set $s = kr$. Then $1/r = k/s$ and, retaining the notation W for simplicity, we obtain by the chain rule

$$W' = \frac{dW}{dr} = \frac{dW}{ds}\,\frac{ds}{dr} = \frac{dW}{ds}\,k \qquad \text{and} \qquad W'' = \frac{d^2W}{ds^2}\,k^2.$$

By substituting this into (11) and omitting the common factor k^2 we have

(12) $$\frac{d^2W}{ds^2} + \frac{1}{s}\,\frac{dW}{ds} + W = 0.$$

This is **Bessel's equation** (1), Sec. 5.5, with parameter $\nu = 0$.

Step 2. Satisfying the Boundary Condition (8)

Solutions of (12) are the Bessel functions J_0 and Y_0 of the first and second kind (see Secs. 5.5, 5.6). But Y_0 becomes infinite at 0, so that we cannot use it because the deflection of the membrane must always remain finite. This leaves us with

(13) $$W(r) = J_0(s) = J_0(kr) \qquad (s = kr).$$

On the boundary $r = R$ we get $W(R) = J_0(kR) = 0$ from (8) (because $G \equiv 0$ would imply $u \equiv 0$). We can satisfy this condition because J_0 has (infinitely many) positive zeros, $s = \alpha_1, \alpha_2, \cdots$ (see Fig. 305), with numerical values

$$\alpha_1 = 2.4048, \quad \alpha_2 = 5.5201, \quad \alpha_3 = 8.6537, \quad \alpha_4 = 11.7915, \quad \alpha_5 = 14.9309$$

and so on. (For further values, consult your CAS or Ref. [GR1] in App. 1.) These zeros are slightly irregularly spaced, as we see. Equation (13) now implies

$$(14) \qquad kR = \alpha_m \qquad \text{thus} \qquad k = k_m = \frac{\alpha_m}{R}, \qquad m = 1, 2, \cdots.$$

Hence the functions

$$(15) \qquad W_m(r) = J_0(k_m r) = J_0\left(\frac{\alpha_m}{R}\, r\right), \qquad m = 1, 2, \cdots$$

are solutions of (11) that are zero on the boundary circle $r = R$.

Eigenfunctions and Eigenvalues. For W_m in (15), a corresponding general solution of (10) with $\lambda = \lambda_m = ck_m = c\alpha_m/R$ is

$$G_m(t) = A_m \cos \lambda_m t + B_m \sin \lambda_m t.$$

Hence the functions

$$(16) \qquad u_m(r, t) = W_m(r)G_m(t) = (A_m \cos \lambda_m t + B_m \sin \lambda_m t) J_0(k_m r)$$

with $m = 1, 2, \cdots$ are solutions of the wave equation (7) satisfying the boundary condition (8). These are the **eigenfunctions** of our problem. The corresponding **eigenvalues** are λ_m.

The vibration of the membrane corresponding to u_m is called the mth **normal mode;** it has the frequency $\lambda_m/2\pi$ cycles per unit time. Since the zeros of the Bessel function J_0 are not regularly spaced on the axis (in contrast to the zeros of the sine functions appearing in the case of the vibrating string), the sound of a drum is entirely different from that of a violin. The forms of the normal modes can easily be obtained from Fig. 305 and are shown in Fig. 306. For $m = 1$, all the points of the membrane move up (or down) at the same time. For $m = 2$, the situation is as follows. The function $W_2(r) = J_0(\alpha_2 r/R)$ is zero for $\alpha_2 r/R = \alpha_1$, thus $r = \alpha_1 R/\alpha_2$. The circle $r = \alpha_1 R/\alpha_2$ is, therefore, **nodal line,** and when at some instant the central part of the membrane moves up, the outer part ($r > \alpha_1 R/\alpha_2$) moves down, and conversely. The solution $u_m(r, t)$ has $m - 1$ nodal lines, which are circles (Fig. 306).

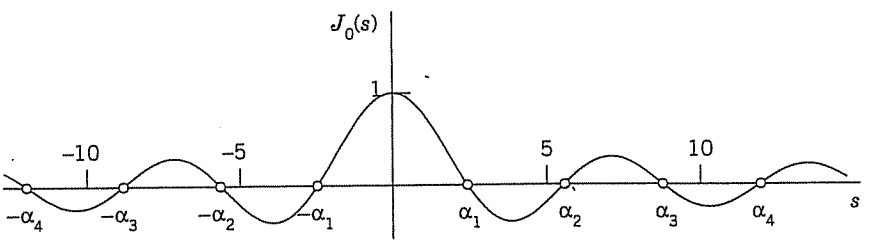


Fig. 305. Bessel function $J_0(s)$

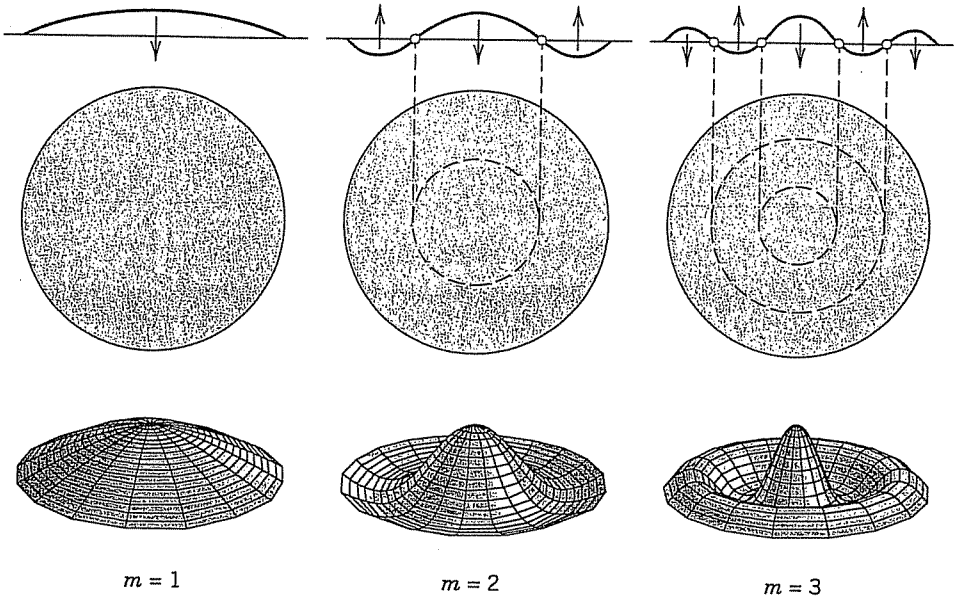


Fig. 306. Normal modes of the circular membrane in the case of vibrations independent of the angle

Step 3. Solution of the Entire Problem

To obtain a solution $u(r, t)$ that also satisfies the initial conditions (9), we may proceed as in the case of the string. That is, we consider the series

$$(17) \qquad u(r, t) = \sum_{m=1}^{\infty} W_m(r)G_m(t) = \sum_{m=1}^{\infty} (A_m \cos \lambda_m t + B_m \sin \lambda_m t)\, J_0\left(\frac{\alpha_m}{R} r\right)$$

(leaving aside the problems of convergence and uniqueness). Setting $t = 0$ and using (9a), we obtain

$$\textbf{(18)} \qquad u(r, 0) = \sum_{m=1}^{\infty} A_m J_0\left(\frac{\alpha_m}{R} r\right) = f(r).$$

Thus for the series (17) to satisfy the condition (9a), the constants A_m must be the coefficients of the **Fourier–Bessel series** (18) that represents $f(r)$ in terms of $J_0(\alpha_m r/R)$; that is [see (10) in Sec. 5.8 with $n = 0$, $\alpha_{0,m} = \alpha_m$, and $x = r$],

$$\textbf{(19)} \qquad A_m = \frac{2}{R^2 J_1^2(\alpha_m)} \int_0^R r f(r) J_0\left(\frac{\alpha_m}{R} r\right) dr \qquad (m = 1, 2, \cdots).$$

Differentiability of $f(r)$ in the interval $0 \leqq r \leqq R$ is sufficient for the existence of the development (18); see Ref. [A13]. The coefficients B_m in (17) can be determined from (9b) in a similar fashion. Numeric values of A_m and B_m may be obtained from a CAS or by a numeric integration method, using tables of J_0 and J_1. However, numeric integration can sometimes be ***avoided,*** as the following example shows.

EXAMPLE 1 Vibrations of a Circular Membrane

Find the vibrations of a circular drumhead of radius 1 ft and density 2 slugs/ft^2 if the tension is 8 lb/ft, the initial velocity is 0, and the initial displacement is

$$f(r) = 1 - r^2 \text{ [ft]}.$$

Solution. $c^2 = T/\rho = 8/2 = 4$ [ft^2/sec^2]. Also $B_m = 0$, since the initial velocity is 0. From (19) and Example 3 in Sec. 5.8, since $R = 1$, we obtain

$$A_m = \frac{2}{J_1^2(\alpha_m)} \int_0^1 r(1 - r^2)\, J_0(\alpha_m r)\, dr$$

$$= \frac{4J_2(\alpha_m)}{\alpha_m^2 J_1^2(\alpha_m)}$$

$$= \frac{8}{\alpha_m^3 J_1(\alpha_m)}$$

where the last equality follows from (24c), Sec. 5.5, with $\nu = 1$, that is,

$$J_2(\alpha_m) = \frac{2}{\alpha_m} J_1(\alpha_m) - J_0(\alpha_m) = \frac{2}{\alpha_m} J_1(\alpha_m).$$

Table 9.5 on p. 409 of [GR1] gives α_m and $J_0'(\alpha_m)$. From this we get $J_1(\alpha_m) = -J_0'(\alpha_m)$ by (24b), Sec. 5.5, with $\nu = 0$, and compute the coefficients A_m:

m	α_m	$J_1(\alpha_m)$	$J_2(\alpha_m)$	A_m
1	2.40483	0.51915	0.43176	1.10801
2	5.52008	−0.34026	−0.12328	−0.13978
3	8.65373	0.27145	0.06274	0.04548
4	11.79153	−0.23246	−0.03943	−0.02099
5	14.93092	0.20655	0.02767	0.01164
6	18.07106	−0.18773	−0.02078	−0.00722
7	21.21164	0.17327	0.01634	0.00484
8	24.35247	−0.16170	−0.01328	−0.00343
9	27.49348	0.15218	0.01107	0.00253
10	30.63461	−0.14417	−0.00941	−0.00193

Thus

$$f(r) = 1.108J_0(2.4048r) - 0.140J_0(5.5201r) + 0.045J_0(8.6537r) - \cdots.$$

We see that the coefficients decrease relatively slowly. The sum of the explicitly given coefficients in the table is 0.99915. The sum of *all* the coefficients should be 1. (Why?) Hence by the Leibniz test in App. A3.3 the partial sum of those terms gives about three correct decimals of the amplitude $f(r)$.

Since

$$\lambda_m = ck_m = c\alpha_m/R = 2\alpha_m,$$

from (17) we thus obtain the solution (with r measured in feet and t in seconds)

$$u(r, t) = 1.108J_0(2.4048r)\cos 4.8097t - 0.140J_0(5.5201r)\cos 11.0402t + 0.045J_0(8.6537r)\cos 17.3075t - \cdots.$$

In Fig. 306, $m = 1$ gives an idea of the motion of the first term of our series, $m = 2$ of the second term, and $m = 3$ of the third term, so that we can "see" our result about as well as for a violin string in Sec. 12.3. ■

PROBLEM SET 12.9

1. Why did we use polar coordinates in this section?

2. Work out the details of the calculation leading to the Laplacian in polar coordinates.

3. If u is independent of θ, then (5) reduces to $\nabla^2 u = u_{rr} + u_r/r$. Derive this directly from the Laplacian in Cartesian coordinates.

4. An **alternative form** of (5) is $\nabla^2 u = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial u}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2}$. Derive this from (5).

5. **(Radial solution)** Show that the only solution of $\nabla^2 u = 0$ depending only on $r = \sqrt{x^2 + y^2}$ is $u = a \ln r + b$ with constant a and b.

6. **TEAM PROJECT. Series for Dirichlet and Neumann Problems**

(a) Show that $u_n = r^n \cos n\theta$, $u_n = r^n \sin n\theta$, $n = 0, 1, \cdots$, are solutions of Laplace's equation $\nabla^2 u = 0$ with $\nabla^2 u$ given by (5). (What would u_n be in Cartesian coordinates? Experiment with small n.)

(b) Dirichlet problem (See Sec. 12.5) Assuming that termwise differentiation is permissible, show that a solution of the Laplace equation in the disk $r < R$ satisfying the boundary condition $u(R, \theta) = f(\theta)$ (f given) is

$$u(r, \theta) = a_0 + \sum_{n=1}^{\infty}\left[a_n\left(\frac{r}{R}\right)^n \cos n\theta + b_n\left(\frac{r}{R}\right)^n \sin n\theta\right] \tag{20}$$

where a_n, b_n are the Fourier coefficients of f (see Sec. 11.1).

(c) Dirichlet problem Solve the Dirichlet problem using (20) if $R = 1$ and the boundary values are $u(\theta) = -100$ volts if $-\pi < \theta < 0$, $u(\theta) = 100$ volts if $0 < \theta < \pi$. (Sketch this disk, indicate the boundary values.)

(d) Neumann problem Show that the solution of the Neumann problem $\nabla^2 u = 0$ if $r < R$, $u_N(R, \theta) = f(\theta)$ (where $u_N = \partial u/\partial N$ is the directional derivative in the direction of the outer normal) is

$$u(r, \theta) = A_0 + \sum_{n=1}^{\infty} r^n (A_n \cos n\theta + B_n \sin n\theta)$$

with arbitrary A_0 and

$$A_n = \frac{1}{\pi n R^{n-1}}\int_{-\pi}^{\pi} f(\theta)\cos n\theta \, d\theta,$$

$$B_n = \frac{1}{\pi n R^{n-1}}\int_{-\pi}^{\pi} f(\theta)\sin n\theta \, d\theta.$$

(e) Compatibility condition Show that (9), Sec. 10.4, imposes on $f(\theta)$ in (d) the *"compatibility condition"*

$$\int_{-\pi}^{\pi} f(\theta)\, d\theta = 0.$$

(f) Neumann problem Solve $\nabla^2 u = 0$ in the annulus $1 < r < 3$ if $u_r(1, \theta) = \sin\theta$, $u_r(3, \theta) = 0$.

7–12 ELECTROSTATIC POTENTIAL. STEADY-STATE HEAT PROBLEMS

The electrostatic potential satisfies Laplace's equation $\nabla^2 u = 0$ in any region free of charges. Also the heat equation $u_t = c^2 \nabla^2 u$ (Sec. 12.5) reduces to Laplace's equation if the temperature u is time-independent **("steady-state case").** Using (20), find the potential (equivalently: the steady-state temperature) in the disk $r < 1$ if the boundary values are (sketch them, to see what is going on).

7. $u(1, \theta) = 40 \cos^3 \theta$
8. $u(1, \theta) = 800 \sin^3 \theta$
9. $u(1, \theta) = 110$ if $-\frac{1}{2}\pi < \theta < \frac{1}{2}\pi$ and 0 otherwise
10. $u(1, \theta) = \theta$ if $-\frac{1}{2}\pi < \theta < \frac{1}{2}\pi$ and 0 otherwise
11. $u(1, \theta) = |\theta|$ if $-\pi < \theta < \pi$
12. $u(1, \theta) = \theta^2$ if $-\pi < \theta < \pi$

13. **CAS EXPERIMENT. Equipotential Lines.** Guess what the equipotential lines $u(r, \theta) = const$ in Probs. 9 and 11 may look like. Then graph some of them, using partial sums of the series.

14. **(Semidisk)** Find the electrostatic potential in the semidisk $r < 1$, $0 < \theta < \pi$ which equals $110\,\theta(\pi - \theta)$ on the semicircle $r = 1$ and 0 on the segment $-1 < x < 1$.

15. **(Semidisk)** Find the steady-state temperature in a semicircular thin plate $r < a$, $0 < \theta < \pi$ with the semicircle $r = a$ kept at constant temperature u_0 and the segment $-a < x < a$ at 0.

16. **(Invariance)** Show that $\nabla^2 u$ is invariant under translations $x^* = x + a$, $y^* = y + b$ and under rotations $x^* = x\cos\alpha - y\sin\alpha$, $y^* = x\sin\alpha + y\cos\alpha$.

CIRCULAR MEMBRANE

17. **(Frequency)** What happens to the frequency of an eigenfunction of a drum if you double the tension?

18. **(Size of a drum)** A small drum should have a higher fundamental frequency than a large one, tension and density being the same. How does this follow from our formulas?

19. **(Tension)** Find a formula for the tension required to produce a desired fundamental frequency f_1 of a drum.

20. **CAS PROJECT. Normal Modes. (a)** Graph the normal modes u_4, u_5, u_6 as in Fig. 306.

(b) Write a program for calculating the A_m's in Example 1 and extend the table to $m = 15$. Verify numerically that $\alpha_m \approx (m - \frac{1}{4})\pi$ and compute the error for $m = 1, \cdots, 10$.

(c) Graph the initial deflection $f(r)$ in Example 1 as well as the first three partial sums of the series. Comment on accuracy.

(d) Compute the radii of the nodal lines of u_2, u_3, u_4 when $R = 1$. How do these values compare to those of the nodes of the vibrating string of length 1? Can you establish any empirical laws by experimentation with further u_m?

21. **(Nodal lines)** Is it possible that for fixed c and R two or more u_m [see (16)] with different nodal lines correspond to the same eigenvalue? (Give a reason.)

22. Why is $A_1 + A_2 + \cdots = 1$ in Example 1? Compute the first few partial sums until you get 3-digit accuracy. What does this problem mean in the field of music?

23. **(Nonzero initial velocity)** Show that for (17) to satisfy (9b) we must have

$$B_m = \frac{2}{c\alpha_m R J_1^2(\alpha_m)} \times \times \int_0^R rg(r) J_0(\alpha_m r/R)\, dr. \tag{21}$$

VIBRATIONS OF A CIRCULAR MEMBRANE DEPENDING ON BOTH r AND θ

24. **(Separations)** Show that substitution of $u = F(r, \theta)G(t)$ into the wave equation (6), that is,

$$u_{tt} = c^2 \left(u_{rr} + \frac{1}{r} u_r + \frac{1}{r^2} u_{\theta\theta} \right) \tag{22}$$

gives an ODE and a PDE

$$\ddot{G} + \lambda^2 G = 0, \qquad \text{where } \lambda = ck, \tag{23}$$

$$F_{rr} + \frac{1}{r} F_r + \frac{1}{r^2} F_{\theta\theta} + k^2 F = 0. \tag{24}$$

Show that the PDE can now be separated by substituting $F = W(r)Q(\theta)$, giving

$$Q'' + n^2 Q = 0, \tag{25}$$

$$r^2 W'' + rW' + (k^2 r^2 - n^2) W = 0. \tag{26}$$

25. **(Periodicity)** Show that $Q(\theta)$ must be periodic with period 2π and, therefore, $n = 0, 1, 2, \cdots$ in (25) and (26). Show that this yields the solutions $Q_n = \cos n\theta$, $Q_n^* = \sin n\theta$, $W_n = J_n(kr)$, $n = 0, 1, \cdots$.

26. **(Boundary condition)** Show that the boundary condition

$$u(R, \theta, t) = 0 \tag{27}$$

leads to $k = k_{mn} = \alpha_{mn}/R$, where $s = \alpha_{mn}$ is the mth positive zero of $J_n(s)$.

27. **(Solutions depending on both r and θ)** Show that solutions of (22) satisfying (27) are (see Fig. 307)

$$u_{mn} = (A_{mn} \cos ck_{mn}t + B_{mn} \sin ck_{mn}t) \times \times J_n(k_{mn}r) \cos n\theta$$

$$u_{mn}^* = (A_{mn}^* \cos ck_{mn}t + B_{mn}^* \sin ck_{mn}t) \times \times J_n(k_{mn}r) \sin n\theta \tag{28}$$

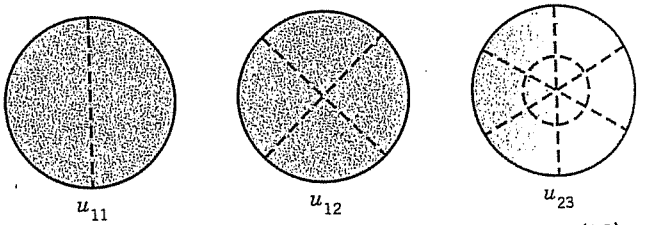


Fig. 307. Nodal lines of some of the solutions (28)

28. **(Initial condition)** Show that $u_t(r, \theta, 0) = 0$ gives $B_{mn} = 0$, $B_{mn}^* = 0$ in (28).

29. Show that $u_{m0}^* = 0$ and u_{m0} is identical with (16) in the current section.

30. **(Semicircular membrane)** Show that u_{11} represents the fundamental mode of a semicircular membrane and find the corresponding frequency when $c^2 = 1$ and $R = 1$.

12.10 Laplace's Equation in Cylindrical and Spherical Coordinates. Potential

Laplace's equation

$$\nabla^2 u = u_{xx} + u_{yy} + u_{zz} = 0 \tag{1}$$

is one of the most important PDEs in physics and its engineering applications. Here, x, y, z are Cartesian coordinates in space (Fig. 165 in Sec. 9.1), $u_{xx} = \partial^2 u/\partial x^2$, etc. The expression $\nabla^2 u$ is called the **Laplacian** of u. The theory of the solutions of (1) is called **potential theory.** Solutions of (1) that have ***continuous*** second partial derivatives are known as **harmonic functions.**

Laplace's equation occurs mainly in **gravitation, electrostatics** (see Theorem 3, Sec. 9.7). steady-state **heat flow** (Sec. 12.5), and **fluid flow** (to be discussed in Chap. 18.4).

Recall from Sec. 9.7 that the gravitational **potential** $u(x, y, z)$ at a point (x, y, z) resulting from a single mass located at a point (X, Y, Z) is

$$u(x, y, z) = \frac{c}{r} = \frac{c}{\sqrt{(x - X)^2 + (y - Y)^2 + (z - Z)^2}} \qquad (r > 0) \tag{2}$$

and u satisfies (1). Similarly, if mass is distributed in a region T in space with density $\rho(X, Y, Z)$, its potential at a point (x, y, z) not occupied by mass is

$$u(x, y, z) = k \iiint_T \frac{\rho(X, Y, Z)}{r}\, dX\, dY\, dZ. \tag{3}$$

It satisfies (1) because $\nabla^2(1/r) = 0$ (Sec. 9.7) and ρ is not a function of x, y, z.

Practical problems involving Laplace's equation are boundary value problems in a region T in space with boundary surface S. Such a problem is called (see also Sec. 12.5 for the two-dimensional case):

(I) First boundary value problem or **Dirichlet problem** if u is prescribed on S.

(II) Second boundary value problem or **Neumann problem** if the normal derivative $u_n = \partial u/\partial n$ is prescribed on S.

(III) Third or **mixed boundary value problem** or **Robin problem** if u is prescribed on a portion of S and u_n on the remaining portion of S.

Laplacian in Cylindrical Coordinates

The first step in solving a boundary value problem is generally the introduction of coordinates in which the boundary surface S has a simple representation. Cylindrical symmetry (a cylinder as a region T) calls for cylindrical coordinates r, θ, z related to x, y, z by

$$x = r\cos\theta, \qquad y = r\sin\theta, \qquad z = z \tag{4}$$

(Fig. 308, p. 588).

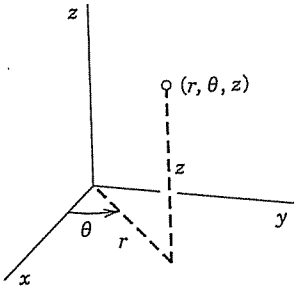


Fig. 308. Cylindrical coordinates

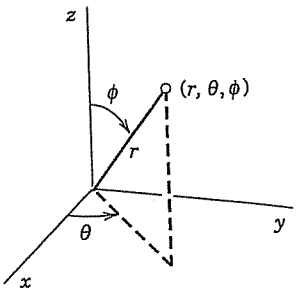


Fig. 309. Spherical coordinates

For these we get $\nabla^2 u$ immediately by adding u_{zz} to (5) in Sec. 12.9; thus,

$$\text{(5)} \qquad \nabla^2 u = \frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2} + \frac{\partial^2 u}{\partial z^2}.$$

Laplacian in Spherical Coordinates

Spherical symmetry (a ball as region T bounded by a sphere S) requires **spherical coordinates** r, θ, ϕ related to x, y, z by

$$\text{(6)} \qquad x = r\cos\theta\sin\phi, \qquad y = r\sin\theta\sin\phi, \qquad z = r\cos\phi \qquad \text{(Fig. 309).}$$

Using the chain rule (as in Sec. 12.9), we obtain $\nabla^2 u$ in spherical coordinates

$$\text{(7)} \qquad \nabla^2 u = \frac{\partial^2 u}{\partial r^2} + \frac{2}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2}\frac{\partial^2 u}{\partial \phi^2} + \frac{\cot\phi}{r^2}\frac{\partial u}{\partial \phi} + \frac{1}{r^2\sin^2\phi}\frac{\partial^2 u}{\partial \theta^2}.$$

We leave the details as an exercise. It is sometimes practical to write (7) in the form

$$\text{(7')} \qquad \nabla^2 u = \frac{1}{r^2}\left[\frac{\partial}{\partial r}\left(r^2\frac{\partial u}{\partial r}\right) + \frac{1}{\sin\phi}\frac{\partial}{\partial \phi}\left(\sin\phi\frac{\partial u}{\partial \phi}\right) + \frac{1}{\sin^2\phi}\frac{\partial^2 u}{\partial \theta^2}\right].$$

Remark on Notation. Equation (6) is used in calculus and extends the familiar notation for polar coordinates. Unfortunately, some books use θ and ϕ interchanged, an extension of the notation $x = r\cos\phi$, $y = r\sin\phi$ for polar coordinates (used in some European countries).

Boundary Value Problem in Spherical Coordinates

We shall solve the following **Dirichlet problem** in spherical coordinates:

$$\text{(8)} \qquad \nabla^2 u = \frac{1}{r^2}\left[\frac{\partial}{\partial r}\left(r^2\frac{\partial u}{\partial r}\right) + \frac{1}{\sin\phi}\frac{\partial}{\partial \phi}\left(\sin\phi\frac{\partial u}{\partial \phi}\right)\right] = 0.$$

$$\text{(9)} \qquad u(R, \phi) = f(\phi)$$

$$\text{(10)} \qquad \lim_{r\to\infty} u(r, \phi) = 0.$$

The PDE (8) follows from (7) by assuming that the solution u will not depend on θ because the Dirichlet condition (9) is independent of θ. This may be an electrostatic potential (or a temperature) $f(\phi)$ at which the sphere S: $r = R$ is kept. Condition (10) means that the potential at infinity will be zero.

Separating Variables by substituting $u(r, \phi) = G(r)H(\phi)$ into (8). Multiplying (8) by r^2, making the substitution and then dividing by GH, we obtain

$$\frac{1}{G}\frac{d}{dr}\left(r^2\frac{dG}{dr}\right) = -\frac{1}{H\sin\phi}\frac{d}{d\phi}\left(\sin\phi\frac{dH}{d\phi}\right).$$

By the usual argument both sides must be equal to a constant k. Thus we get the two ODEs

$$\text{(11)}\qquad \frac{1}{G}\frac{d}{dr}\left(r^2\frac{dG}{dr}\right) = k \qquad \text{or} \qquad r^2\frac{d^2G}{dr^2} + 2r\frac{dG}{dr} = kG$$

and

$$\text{(12)}\qquad \frac{1}{\sin\phi}\frac{d}{d\phi}\left(\sin\phi\frac{dH}{d\phi}\right) + kH = 0.$$

The solutions of (11) will take a simple form if we set $k = n(n + 1)$. Then, writing $G' = dG/dr$, etc., we obtain

$$\text{(13)}\qquad r^2G'' + 2rG' - n(n + 1)G = 0.$$

This is an **Euler–Cauchy equation.** From Sec. 2.5 we know that it has solutions $G = r^a$. Substituting this and dropping the common factor r^a gives

$$a(a - 1) + 2a - n(n + 1) = 0. \qquad \text{The roots are} \qquad a = n \quad \text{and} \quad -n - 1.$$

Hence solutions are

$$\text{(14)}\qquad G_n(r) = r^n \qquad \text{and} \qquad G_n^*(r) = \frac{1}{r^{n+1}}.$$

We now solve (12). Setting $\cos\phi = w$, we have $\sin^2\phi = 1 - w^2$ and

$$\frac{d}{d\phi} = \frac{d}{dw}\frac{dw}{d\phi} = -\sin\phi\frac{d}{dw}.$$

Consequently, (12) with $k = n(n + 1)$ takes the form

$$\text{(15)}\qquad \frac{d}{dw}\left[(1 - w^2)\frac{dH}{dw}\right] + n(n + 1)H = 0.$$

This is **Legendre's equation** (see Sec. 5.3), written out

$$(15') \qquad (1 - w^2)\frac{d^2H}{dw^2} - 2w\frac{dH}{dw} + n(n+1)H = 0.$$

For integer $n = 0, 1, \cdots$ the Legendre polynomials

$$H = P_n(w) = P_n(\cos\phi) \qquad n = 0, 1, \cdots,$$

are solutions of Legendre's equation (15). We thus obtain the following two sequences of solution $u = GH$ of Laplace's equation (8), with constant A_n and B_n, where $n = 0, 1, \cdots$,

$$(16) \qquad \text{(a)} \quad u_n(r, \phi) = A_n r^n P_n(\cos\phi), \qquad \text{(b)} \quad u_n^*(r, \phi) = \frac{B_n}{r^{n+1}} P_n(\cos\phi)$$

Use of Fourier–Legendre Series

Interior Problem: Potential Within the Sphere S. We consider a series of terms from (16a),

$$(17) \qquad u(r, \phi) = \sum_{n=0}^{\infty} A_n r^n P_n(\cos\phi) \qquad (r \leqq R).$$

Since S is given by $r = R$, for (17) to satisfy the Dirichlet condition (9) on the sphere S, we must have

$$(18) \qquad u(R, \phi) = \sum_{n=0}^{\infty} A_n R^n P_n(\cos\phi) = f(\phi);$$

that is, (18) must be the **Fourier–Legendre series** of $f(\phi)$. From (7) in Sec. 5.8 we get the coefficients

$$(19^*) \qquad A_n R^n = \frac{2n+1}{2}\int_{-1}^{1} \tilde{f}(w) P_n(w)\, dw$$

where $\tilde{f}(w)$ denotes $f(\phi)$ as a function of $w = \cos\phi$. Since $dw = -\sin\phi\, d\phi$, and the limits of integration -1 and 1 correspond to $\phi = \pi$ and $\phi = 0$, respectively, we also obtain

$$(19) \qquad A_n = \frac{2n+1}{2R^n}\int_0^{\pi} f(\phi) P_n(\cos\phi)\sin\phi\, d\phi, \qquad n = 0, 1, \cdots.$$

If $f(\phi)$ and $f'(\phi)$ are piecewise continuous on the interval $0 \leqq \phi \leqq \pi$, then the series (17) with coefficients (19) solves our problem for points inside the sphere because it can be shown that under these continuity assumptions the series (17) with coefficients (19) gives the derivatives occurring in (8) by termwise differentiation, thus justifying our derivation.

Exterior Problem: Potential Outside the Sphere S. Outside the sphere we cannot use the functions u_n in (16a) because they do not satisfy (10). But we can use the u_n^* in (16b), which do satisfy (10) (but could not be used inside S; why?). Proceeding as before leads to the solution of the exterior problem

$$u(r, \phi) = \sum_{n=0}^{\infty} \frac{B_n}{r^{n+1}} P_n(\cos \phi) \qquad (r \geqq R) \tag{20}$$

satisfying (8), (9), (10), with coefficients

$$B_n = \frac{2n+1}{2} R^{n+1} \int_0^{\pi} f(\phi) P_n(\cos \phi) \sin \phi \, d\phi. \tag{21}$$

The next example illustrates all this for a sphere of radius 1 consisting of two hemispheres that are separated by a small strip of insulating material along the equator, so that these hemispheres can be kept at different potentials (110 V and 0 V).

EXAMPLE 1 Spherical Capacitor

Find the potential inside and outside a spherical capacitor consisting of two metallic hemispheres of radius 1 ft separated by a small slit for reasons of insulation, if the upper hemisphere is kept at 110 V and the lower is grounded (Fig. 310).

Solution. The given boundary condition is (recall Fig. 309)

$$f(\phi) = \begin{cases} 110 & \text{if} \quad 0 \leqq \phi < \pi/2 \\ 0 & \text{if} \quad \pi/2 < \phi \leqq \pi. \end{cases}$$

Since $R = 1$, we thus obtain from (19)

$$A_n = \frac{2n+1}{2} \cdot 110 \int_0^{\pi/2} P_n(\cos \phi) \sin \phi \, d\phi$$

$$= \frac{2n+1}{2} \cdot 110 \int_0^1 P_n(w) \, dw$$

where $w = \cos \phi$. Hence $P_n(\cos \phi) \sin \phi \, d\phi = -P_n(w) \, dw$, we integrate from 1 to 0, and we finally get rid of the minus by integrating from 0 to 1. You can evaluate this integral by your CAS or continue by using (11) in Sec. 5.3, obtaining

$$A_n = 55(2n+1) \sum_{m=0}^{M} (-1)^m \frac{(2n-2m)!}{2^n m!(n-m)!(n-2m)!} \int_0^1 w^{n-2m} \, dw$$

where $M = n/2$ for even n and $M = (n-1)/2$ for odd n. The integral equals $1/(n-2m+1)$. Thus

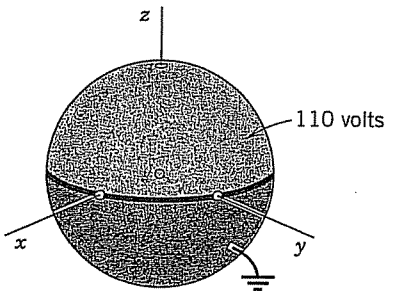


Fig. 310. Spherical capacitor in Example 1

(22) $$A_n = \frac{55(2n+1)}{2^n} \sum_{m=0}^{M} (-1)^m \frac{(2n-2m)!}{m!(n-m)!(n-2m+1)!}.$$

Taking $n = 0$, we get $A_0 = 55$ (since $0! = 1$). For $n = 1, 2, 3, \cdots$ we get

$$A_1 = \frac{165}{2} \cdot \frac{2!}{0!1!2!} = \frac{165}{2},$$

$$A_2 = \frac{275}{4} \left(\frac{4!}{0!2!3!} - \frac{2!}{1!1!1!} \right) = 0,$$

$$A_3 = \frac{385}{8} \left(\frac{6!}{0!3!4!} - \frac{4!}{1!2!2!} \right) = -\frac{385}{8}, \quad \text{etc.}$$

Hence the ***potential* (17) *inside the sphere*** is (since $P_0 = 1$)

(23) $$u(r, \phi) = 55 + \frac{165}{2} r P_1(\cos\phi) - \frac{385}{8} r^3 P_3(\cos\phi) + \cdots$$ (Fig. 311)

with P_1, P_3, $\cdots$ given by (11′), Sec. 5.3. Since $R = 1$, we see from (19) and (21) in this section that $B_n = A_n$, and (20) thus gives the ***potential outside the sphere***

(24) $$u(r, \phi) = \frac{55}{r} + \frac{165}{2r^2} P_1(\cos\phi) - \frac{385}{8r^4} P_3(\cos\phi) + \cdots.$$

Partial sums of these series can now be used for computing approximate values of the inner and outer potential. Also, it is interesting to see that far away from the sphere the potential is approximately that of a point charge, namely, $55/r$. (Compare with Theorem 3 in Sec. 9.7.) ▩

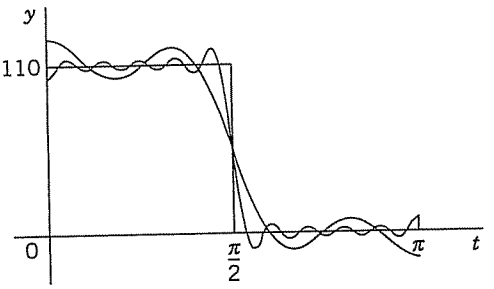


Fig. 311. Partial sums of the first 4, 6, and 11 nonzero terms of (23) for $r = R = 1$

EXAMPLE 2 **Simpler Cases. Help with Problems**

The technicalities occurring in cases like that of Example 1 can often be avoided. For instance, find the potential inside the sphere S: $r = R = 1$ when S is kept at the potential $f(\phi) = \cos 2\phi$. (Can you see the potential on S? What is it at the North Pole? The equator? The South Pole?)

Solution. $w = \cos\phi$, $\cos 2\phi = 2\cos^2\phi - 1 = 2w^2 - 1 = \frac{4}{3}P_2(w) - \frac{1}{3} = \frac{4}{3}(\frac{3}{2}w^2 - \frac{1}{2}) - \frac{1}{3}$. Hence the potential in the interior of the sphere is

$$u = \tfrac{4}{3}r^2 P_2(w) - \tfrac{1}{3} = \tfrac{4}{3}r^2 P_2(\cos\phi) - \tfrac{1}{3} = \tfrac{2}{3}r^2(3\cos^2\phi - 1) - \tfrac{1}{3}.$$

▩

PROBLEM SET 12.10

1. Derive (7) from $\nabla^2 u$ in Cartesian coordinates. (Show the details.)
2. Find the surfaces on which the functions u_1, u_2, u_3 are zero.
3. Sketch the functions $P_n(\cos\phi)$ for $n = 0, 1, 2$ (see (11′) in Sec. 5.3).
4. Sketch the functions $P_3(\cos\phi)$ and $P_4(\cos\phi)$.
5. Verify that u_n and u_n^* in (16) are solutions of (8).

6–11 POTENTIALS DEPENDING ONLY ON r

6. **(Dimension 3)** Show that the only solution of the Laplace equation depending only on $r = \sqrt{x^2 + y^2 + z^2}$ is $u = c/r + k$ with constant c and k.
7. **(Dimension 3)** Verify that $u = c/r$, $r = \sqrt{x^2 + y^2 + z^2}$, satisfies Laplace's equation in spherical coordinates.
8. **(Dirichlet problem).** Find the electrostatic potential between two concentric spheres of radii $r_1 = 10$ cm and $r_2 = 20$ cm kept at potentials $U_1 = 260$ V and $U_2 = 110$ V, respectively.
9. **(Dimension 2, logarithmic potential)** Show that the only solution of the two-dimensional Laplace equation depending only on $r = \sqrt{x^2 + y^2}$ is $u = c \ln r + k$ with constant c and k.
10. **(Logarithmic potential)** Find the electrostatic potential between two coaxial cylinders of radii $r_1 = 10$ cm and $r_2 = 20$ cm kept at potentials $U_1 = 260$ V and $U_2 = 110$ V, respectively. Compare with Prob. 8. Comment.
11. **(Heat problem)** If the surface of the ball $r^2 = x^2 + y^2 + z^2 \leqq R^2$ is kept at temperature zero and the initial temperature in the ball is $f(r)$, show that the temperature $u(r, t)$ in the ball is a solution of $u_t = c^2(u_{rr} + 2u_r/r)$ satisfying the conditions $u(R, t) = 0$, $u(r, 0) = f(r)$. Show that setting $v = ru$ gives $v_t = c^2 v_{rr}$, $v(R, t) = 0$, $v(r, 0) = rf(r)$. Include the condition $v(0, t) = 0$ (which holds because u must be bounded at $r = 0$), and solve the resulting problem by separating variables.
12. **(Two-dimensional potential problems)** Show that the functions $x^2 - y^2$, xy, $x/(x^2 + y^2)$, $e^x \cos y$, $e^x \sin y$, $\cos x \cosh y$, $\ln(x^2 + y^2)$, and $\arctan(y/x)$ satisfy Laplace's equation $u_{xx} + u_{yy} = 0$. (Two-dimensional potential problems are best solved by ***complex analysis,*** as we shall see in Chap. 18.)

13–17 BOUNDARY VALUE PROBLEMS IN SPHERICAL COORDINATES r, θ, ϕ

Find the potential in the interior of the sphere S: $r = R = 1$ if this interior is free of charges and the potential on S is:

13. $f(\phi) = 100$
14. $f(\phi) = \cos\phi$
15. $f(\phi) = \cos 3\phi$
16. $f(\phi) = \sin^2\phi$
17. $f(\phi) = 35\cos 4\phi + 20\cos 2\phi + 9$
18. Show that in Prob. 13 the potential exterior to the sphere is the same as that of a point charge at the origin. Is this physically plausible?
19. Sketch the intersection of the equipotential surfaces in Prob. 14 with the xz-plane.
20. Find the potential exterior to the sphere in Example 2 of the text and in Prob. 15.
21. What is the temperature in a ball of radius 1 and of homogeneous material if its lower boundary hemisphere is kept at 0°C and its upper at 100°C?
22. **(Reflection in a sphere)** Let r, θ, ϕ be spherical coordinates. If $u(r, \theta, \phi)$ satisfies $\nabla^2 u = 0$, show that $v(r, \theta, \phi) = u(1/r, \theta, \phi)/r$ satisfies $\nabla^2 v = 0$. What does this give for (16)?
23. **(Reflection in a circle)** Let r, θ be polar coordinates. If $u(r, \theta)$ satisfies $\nabla^2 u = 0$, show that the function $v(r, \theta) = u(1/r, \theta)$ satisfies $\nabla^2 v = 0$. What are $u = r\cos\theta$ and v in terms of x and y? Answer the same question for $u = r^2 \cos\theta \sin\theta$ and v.
24. **TEAM PROJECT. Transmission Line and Related PDEs.** Consider a long cable or telephone wire (Fig. 312) that is imperfectly insulated, so that leaks occur along the entire length of the cable. The source S of the current $i(x, t)$ in the cable is at $x = 0$, the receiving end T at $x = l$. The current flows from S to T, through the load, and returns to the ground. Let the constants R, L, C, and G denote the resistance, inductance, capacitance to ground, and conductance to ground, respectively, of the cable per unit length.

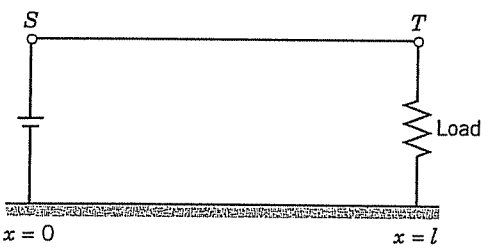


Fig. 312. Transmission line

(a) Show that (**"first transmission line equation"**)

$$-\frac{\partial u}{\partial x} = Ri + L\frac{\partial i}{\partial t}$$

where $u(x, t)$ is the potential in the cable. *Hint:* Apply Kirchhoff's voltage law to a small portion of the cable between x and $x + \Delta x$ (difference of the potentials at x and $x + \Delta x$ = resistive drop + inductive drop).

(b) Show that for the cable in (a) (**"second transmission line equation"**),

$$-\frac{\partial i}{\partial x} = Gu + C\frac{\partial u}{\partial t}$$

Hint: Use Kirchhoff's current law (difference of the currents at x and $x + \Delta x$ = loss due to leakage to ground + capacitive loss).

(c) **Second-order PDEs.** Show that elimination of i or u from the transmission line equations leads to

$$u_{xx} = LCu_{tt} + (RC + GL)u_t + RGu,$$

$$i_{xx} = LCi_{tt} + (RC + GL)i_t + RGi.$$

(d) **Telegraph equations.** For a submarine cable, G is negligible and the frequencies are low. Show that this leads to the so-called *submarine cable equations* or **telegraph equations**

$$u_{xx} = RCu_t, \qquad i_{xx} = RCi_t.$$

Find the potential in a submarine cable with ends ($x = 0, x = l$) grounded and initial voltage distribution $U_0 = const.$

(e) **High-frequency line equations.** Show that in the case of alternating currents of high frequencies the equations in (c) can be approximated by the so-called **high-frequency line equations**

$$u_{xx} = LCu_{tt}, \qquad i_{xx} = LCi_{tt}.$$

Solve the first of them, assuming that the initial potential is

$$U_0 \sin(\pi x/l),$$

and $u_t(x, 0) = 0$ and $u = 0$ at the ends $x = 0$ and $x = l$ for all t.

12.11 Solution of PDEs by Laplace Transforms

Readers familiar with Chap. 6 may wonder whether Laplace transforms can also be used for solving *partial* differential equations. The answer is yes, particularly if one of the independent variables ranges over the positive axis. The steps to obtain a solution are similar to those in Chap. 6. For a PDE in two variables they are as follows.

1. Take the Laplace transform with respect to one of the two variables, usually t. This gives an ***ODE for the transform*** of the unknown function. This is so since the derivatives of this function with respect to the other variable slip into the transformed equation. The latter also incorporates the given boundary and initial conditions.
2. Solving that ODE, obtain the transform of the unknown function.
3. Taking the inverse transform, obtain the solution of the given problem.

If the coefficients of the given equation do not depend on t, the use of Laplace transforms will simplify the problem.

We explain the method in terms of a typical example.

EXAMPLE 1 Semi-Infinite String

Find the displacement $w(x, t)$ of an elastic string subject to the following conditions. (We write w since we need u to denote the unit step function.)

(i) The string is initially at rest on the x-axis from $x = 0$ to ∞ (*"semi-infinite string"*).

(ii) For $t > 0$ the left end of the string ($x = 0$) is moved in a given fashion, namely, according to a single sine wave

$$w(0, t) = f(t) = \begin{cases} \sin t & \text{if } 0 \leqq t \leqq 2\pi \\ 0 & \text{otherwise} \end{cases}$$ (Fig. 313).

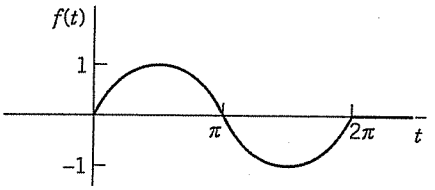


Fig. 313. Motion of the left end of the string in Example 1 as a function of time t

(iii) Furthermore, $\lim_{x\to\infty} w(x, t) = 0$ for $t \geqq 0$.

Of course there is no infinite string, but our model describes a long string or rope (of negligible weight) with its right end fixed far out on the x-axis.

Solution. We have to solve the wave equation (Sec. 12.2)

$$\frac{\partial^2 w}{\partial t^2} = c^2 \frac{\partial^2 w}{\partial x^2}, \qquad c^2 = \frac{T}{\rho} \tag{1}$$

for positive x and t, subject to the "boundary conditions"

$$w(0, t) = f(t), \qquad \lim_{x\to\infty} w(x, t) = 0 \qquad (t \geqq 0) \tag{2}$$

with f as given above, and the initial conditions

$$\text{(a)}\quad w(x, 0) = 0, \qquad \text{(b)}\quad w_t(x, 0) = 0. \tag{3}$$

We take the Laplace transform ***with respect to t.*** By (2) in Sec. 6.2,

$$\mathcal{L}\left\{\frac{\partial^2 w}{\partial t^2}\right\} = s^2\mathcal{L}\{w\} - sw(x, 0) - w_t(x, 0) = c^2\mathcal{L}\left\{\frac{\partial^2 w}{\partial x^2}\right\}.$$

The expression $-sw(x, 0) - w_t(x, 0)$ drops out because of (3). On the right we assume that we may interchange integration and differentiation. Then

$$\mathcal{L}\left\{\frac{\partial^2 w}{\partial x^2}\right\} = \int_0^\infty e^{-st} \frac{\partial^2 w}{\partial x^2}\, dt = \frac{\partial^2}{\partial x^2} \int_0^\infty e^{-st} w(x, t)\, dt = \frac{\partial^2}{\partial x^2}\, \mathcal{L}\{w(x, t)\}.$$

Writing $W(x, s) = \mathcal{L}\{w(x, t)\}$, we thus obtain

$$s^2 W = c^2 \frac{\partial^2 W}{\partial x^2}, \qquad \text{thus} \qquad \frac{\partial^2 W}{\partial x^2} - \frac{s^2}{c^2} W = 0.$$

Since this equation contains only a derivative with respect to x, it may be regarded as an ***ordinary differential equation*** for $W(x, s)$ considered as a function of x. A general solution is

$$W(x, s) = A(s)e^{sx/c} + B(s)e^{-sx/c}. \tag{4}$$

From (2) we obtain, writing $F(s) = \mathcal{L}\{f(t)\}$,

$$W(0, s) = \mathcal{L}\{w(0, t)\} = \mathcal{L}\{f(t)\} = F(s).$$

Assuming that we can interchange integration and taking the limit, we have

$$\lim_{x\to\infty} W(x, s) = \lim_{x\to\infty} \int_0^\infty e^{-st} w(x, t)\, dt = \int_0^\infty e^{-st} \lim_{x\to\infty} w(x, t)\, dt = 0.$$

This implies $A(s) = 0$ in (4) because $c > 0$, so that for every fixed positive s the function $e^{sx/c}$ increases as x increases. Note that we may assume $s > 0$ since a Laplace transform generally exists for *all* s greater than some fixed k (Sec. 6.2). Hence we have

$$W(0, s) = B(s) = F(s),$$

so that (4) becomes

$$W(x, s) = F(s)e^{-sx/c}.$$

From the second shifting theorem (Sec. 6.3) with $a = x/c$ we obtain the inverse transform

(5) $$w(x, t) = f\left(t - \frac{x}{c}\right) u\left(t - \frac{x}{c}\right)$$ (Fig. 314)

that is,

$$w(x, t) = \sin\left(t - \frac{x}{c}\right) \quad \text{if} \quad \frac{x}{c} < t < \frac{x}{c} + 2\pi \quad \text{or} \quad ct > x > (t - 2\pi)c$$

and zero otherwise. This is a single sine wave traveling to the right with speed c. Note that a point x remains at rest until $t = x/c$, the time needed to reach that x if one starts at $t = 0$ (start of the motion of the left end) and travels with speed c. The result agrees with our physical intuition. Since we proceeded formally, we must verify that (5) satisfies the given conditions. We leave this to the student. ■

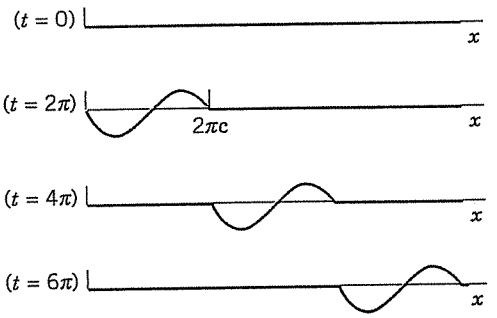


Fig. 314. Traveling wave in Example 1

This is the end of Chap. 12, in which we concentrated on the most important partial differential equations (PDEs) in physics and engineering. This is also the end of Part C on Fourier analysis and PDEs.

We have seen that PDEs have various basic engineering applications, which make them the subject of many ongoing research projects.

Numerics for PDEs follows in Secs. 21.4–21.7, which are independent of the other sections in Part E on numerics.

In the next part, Part D on **complex analysis,** we turn to an area of a different nature that is also highly important to the engineer, as our examples and problems will show. This will include another approach to the (two-dimensional) **Laplace equation** and its applications in Chap. 18.

PROBLEM SET 12.4

1. Sketch a figure similar to Fig. 314 if $c = 1$ and f is "triangular" as in Example 1, Sec. 12.3.
2. How does the speed of the wave in Example 1 depend on the tension and on the mass of the string?
3. Verify the solution in Example 1. What traveling wave do we obtain in Example 1 in the case of a (nonterminating) sinusoidal motion of the left end starting at $t = 0$?

4–6 SOLVE BY LAPLACE TRANSFORMS

4. $\dfrac{\partial w}{\partial x} + x\dfrac{\partial w}{\partial t} = x,\ w(x, 0) = 1,\ w(0, t) = 1$

5. $x\dfrac{\partial w}{\partial x} + \dfrac{\partial w}{\partial t} = xt, \quad w(x, 0) = 0 \text{ if } x \geqq 0,$
$w(0, t) = 0 \text{ if } t \geqq 0$

6. $\dfrac{\partial^2 w}{\partial x^2} = 100\dfrac{\partial^2 w}{\partial t^2} + 100\dfrac{\partial w}{\partial t} + 25w,$
$w(x, 0) = 0 \text{ if } x \geqq 0, \quad w_t(x, 0) = 0 \text{ if } t \geqq 0,$
$w(0, t) = \sin t \text{ if } t \geqq 0$

7. Solve Prob. 5 by another method.

8–10 HEAT PROBLEM

Find the temperature $w(x, t)$ in a semi-infinite laterally insulated bar extending from $x = 0$ along the x-axis to infinity, assuming that the initial temperature is 0, $w(x, t) \to 0$ as $x \to \infty$ for every fixed $t \geqq 0$, and $w(0, t) = f(t)$. Proceed as follows.

8. Set up the model and show that the Laplace transform leads to

$$sW = c^2 \frac{\partial^2 W}{\partial x^2} \quad (W = \mathcal{L}\{w\})$$

and

$$W = F(s)e^{-\sqrt{s}x/c} \quad (F = \mathcal{L}\{f\}).$$

Applying the convolution theorem, show that

$$w(x, t) = \frac{x}{2c\sqrt{\pi}} \int_0^t f(t - \tau)\tau^{-3/2}e^{-x^2/(4c^2\tau)}\, d\tau.$$

9. Let $w(0, t) = f(t) = u(t)$ (Sec. 6.3). Denote the corresponding w, W, and F by w_0, W_0, and F_0. Show that then in Prob. 8,

$$w_0(x, t) = \frac{x}{2c\sqrt{\pi}} \int_0^t \tau^{-3/2}e^{-x^2/(4c^2\tau)}\, d\tau$$
$$= 1 - \operatorname{erf}\left(\frac{x}{2c\sqrt{t}}\right)$$

with the error function erf as defined in Problem Set 12.6.

10. **(Duhamel's formula[4])** Show that in Prob. 9,

$$W_0(x, s) = \frac{1}{s}\, e^{-\sqrt{s}x/c}$$

and the convolution theorem gives *Duhamel's formula*

$$w(x, t) = \int_0^t f(t - \tau)\, \frac{\partial w_0}{\partial \tau}\, d\tau.$$

CHAPTER 12 REVIEW QUESTIONS AND PROBLEMS

1. Write down the three probably most important PDEs from memory and state their main applications.
2. What is the method of separating variables for PDEs? Give an example from memory.
3. What is the superposition principle? Give a typical application.
4. What role did Fourier series play in this chapter? Fourier integrals?
5. What are the eigenfunctions and their frequencies of the vibrating string? Of the heat equation?
6. What additional conditions did we consider for the wave equation? For the heat equation?
7. Name and explain the three kinds of boundary conditions.
8. What do you know about types of PDEs? About transformation to normal forms?
9. What is d'Alembert's method? To what PDE does it apply?
10. When and why did we use polar coordinates? Spherical coordinates?
11. When and why did Legendre's equation occur in this chapter? Bessel's equation?
12. What are the eigenfunctions of the circular membrane? How do their frequencies differ in principle from those of the eigenfunctions of the vibrating string?
13. Explain mathematically (not physically) why we got exponential functions in separating the heat equation, but not for the wave equation.
14. What is the error function? Why did it occur and where?
15. Explain the idea of using Laplace transform methods for PDEs. Give an example from memory.
16. For what k and m are $x^4 + kx^2y^2 + y^4$ and $\sin mx \sinh y$ solutions of Laplace's equation?
17. Verify that $(x^2 - y^2)/(x^2 + y^2)^2$ satisfies Laplace's equation.

18–21 Solve for $u = u(x, y)$:

18. $u_{yy} + 16u = 0$
19. $u_{xx} + u_x - 2u = 0$
20. $u_{xy} + u_y + x + y + 1 = 0$
21. $u_{yy} + u_y = 0, \; u(x, 0) = f(x), \; u_y(x, 0) = g(x)$
22. Find all solution $u(x, y) = F(x)G(y)$ of Laplace's equation in two variables.

[4]JEAN-MARIE CONSTANT DUHAMEL (1797–1872), French mathematician.

23–26 Find and sketch or graph (as in Fig. 285 in Sec. 12.3) the deflection $u(x, t)$ of a vibrating string of length π, extending from $x = 0$ to $x = \pi$, and $c^2 = T/\rho = 1$, starting with velocity 0 and deflection

23. $f(x) = \sin x - \frac{1}{2}\sin 2x$

24. $f(x) = \frac{1}{2}\pi - |x - \frac{1}{2}\pi|$

25. $f(x) = \sin^3 x$

26. $f(x) = x(\pi - x)$

27–30 Find the temperature distribution in a laterally insulated thin copper bar ($c^2 = K/\rho\sigma = 1.158\ \text{cm}^2/\text{sec}$), 50 cm long and of constant cross section with endpoints at $x = 0$ and 50 kept at 0°C and initial temperature

27. $f(x) = \sin(\pi x/50)$

28. $f(x) = x(50 - x)$

29. $f(x) = 25 - |25 - x|$

30. $f(x) = 4\sin^3(\pi x/10)$

31–33 Find the temperature $u(x, t)$ in a laterally insulated bar of length π, extending from $x = 0$ to $x = \pi$, with $c^2 = 1$ for adiabatic boundary condition (see Problem Set 12.5) and initial temperature

31. $100\cos 4x$

32. $3x^2$

33. $\pi - 2|x - \frac{1}{2}\pi|$

34. Using partial sums, graph $u(x, t)$ in Prob. 33 for several constant t on common axes. Do these graphs agree with your physical intuition?

35. Let $f(x, y) = u(x, y, 0)$ be the initial temperature in a thin square plate of side π with edges kept at 0°C and faces perfectly insulated. Separating variables, obtain from $u_t = c^2\nabla^2 u$ the solution

$$u(x, y, t) = \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} B_{mn}\sin mx\sin ny\, e^{-c^2(m^2+n^2)t}$$

where

$$B_{mn} = \frac{4}{\pi^2}\int_0^{\pi}\int_0^{\pi} f(x, y)\sin mx\sin ny\, dx\, dy.$$

36. Find the temperature in Prob. 35 if $f(x, y) = x(\pi - x)y(\pi - y)$.

37–42 Transform to normal form and solve (showing the details!)

37. $u_{xy} = u_{xx}$

38. $u_{xx} + 4u_{xy} + 4u_{yy} = 0$

39. $u_{xx} + 4u_{yy} = 0$

40. $2u_{xx} + 5u_{xy} + 2u_{yy} = 0$

41. $u_{xx} + 2u_{xy} + u_{yy} = 0$

42. $u_{yy} + u_{xy} - 2u_{xx} = 0$

43–47 Show that the following membranes of area 1 with $c^2 = 1$ have the frequencies of the fundamental mode as given (4-decimal values). Compare.

43. Circle: $\alpha_1/(2\sqrt{\pi}) = 0.6784$

44. Square: $1\sqrt{2} = 0.7071$

45. Rectangle (sides 1:2): $\sqrt{5/8} = 0.7906$

46. Semicircle: $3.832/\sqrt{8\pi} = 0.7644$

47. Quadrant of circle: $\alpha_{12}/(4\sqrt{\pi}) = 0.7244$ ($\alpha_{12} = 5.13562$ = first positive zero of J_2)

48–50 Find the electrostatic potential in the following (charge-free) regions:

48. Between two concentric spheres of radii r_0 and r_1 kept at the potentials u_0 and u_1, respectively.

49. Between two coaxial circular cylinders of radii r_0 and r_1 kept at the potential u_0 and u_1, respectively. (Compare with Prob. 48.)

50. In the interior of a sphere of radius 1 kept at the potential $f(\phi) = \cos 3\phi + 3\cos\phi$ (referred to our usual spherical coordinates).

SUMMARY OF CHAPTER 12
Partial Differential Equations (PDEs)

Whereas ODEs (Chaps. 1–6) serve as models of problems involving only *one* independent variable, problems involving *two or more* independent variables (space variables or time t and one or several space variables) lead to PDEs. This accounts for the enormous importance of PDEs to the engineer and physicist. Most important are:

(1) $u_{tt} = c^2 u_{xx}$ One-dimensional wave equation (Secs. 12.2–12.4)

(2) $u_{tt} = c^2(u_{xx} + u_{yy})$ Two-dimensional wave equation (Secs. 12.7–12.9)

(3) $u_t = c^2 u_{xx}$ One-dimensional heat equation (Secs. 12.5, 12.6)

(4) $\nabla^2 u = u_{xx} + u_{yy} = 0$ Two-dimensional Laplace equation (Secs. 12.5, 12.9)

(5) $\nabla^2 u = u_{xx} + u_{yy} + u_{zz} = 0$ Three-dimensional Laplace equation (Sec. 12.10).

Equations (1) and (2) are hyperbolic, (3) is parabolic, (4) and (5) are elliptic.

In practice, one is interested in obtaining the solution of such an equation in a given region satisfying given additional conditions, such as **initial conditions** (conditions at time $t = 0$) or **boundary conditions** (prescribed values of the solution u or some of its derivatives on the boundary surface S, or boundary curve C, of the region) or both. For (1) and (2) one prescribes two initial conditions (initial displacement and initial velocity). For (3) one prescribes the initial temperature distribution. For (4) and (5) one prescribes a boundary condition and calls the resulting problem a (see Sec. 12.5)

Dirichlet problem if u is prescribed on S,
Neumann problem if $u_n = \partial u/\partial n$ is prescribed on S,
Mixed problem if u is prescribed on one part of S and u_n on the other.

A general method for solving such problems is the method of **separating variables** or **product method,** in which one assumes solutions in the form of products of functions each depending on one variable only. Thus equation (1) is solved by setting $u(x, t) = F(x)G(t)$; see Sec. 12.3; similarly for (3) (see Sec. 12.5). Substitution into the given equation yields ***ordinary*** differential equations for F and G, and from these one gets infinitely many solutions $F = F_n$ and $G = G_n$ such that the corresponding functions

$$u_n(x, t) = F_n(x)G_n(t)$$

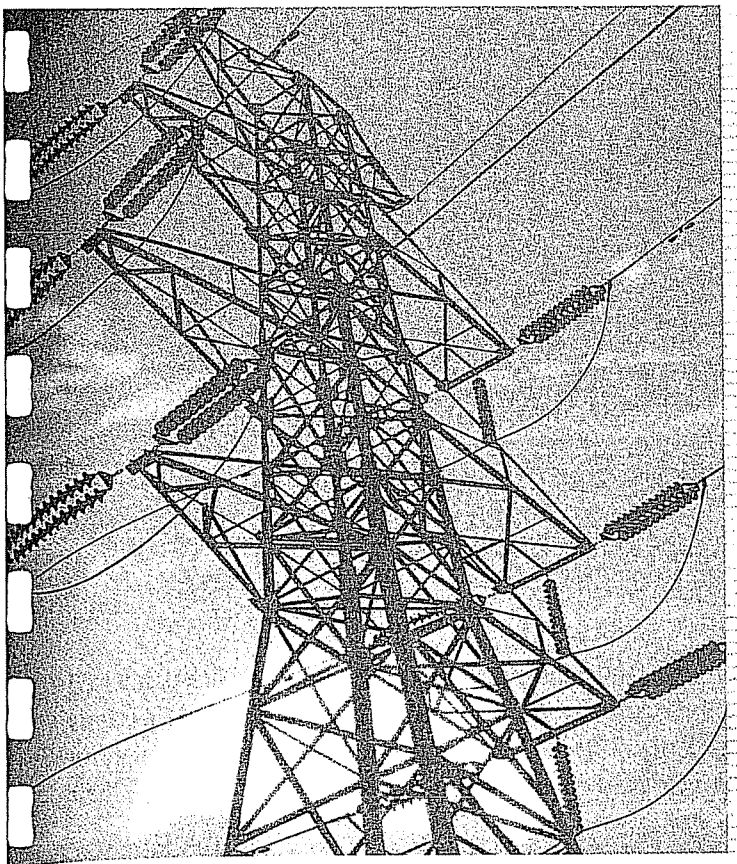
are solutions of the PDE satisfying the given boundary conditions. These are the **eigenfunctions** of the problem, and the corresponding **eigenvalues** determine the frequency of the vibration (or the rapidity of the decrease of temperature in the case of the heat equation, etc.). To satisfy also the initial condition (or conditions), one must consider infinite series of the u_n, whose coefficients turn out to be the Fourier coefficients of the functions f and g representing the given initial conditions (Secs. 12.3, 12.5). Hence **Fourier series** (and *Fourier integrals*) are of basic importance here (Secs. 12.3, 12.5, 12.6, 12.8).

Steady-state problems are problems in which the solution does not depend on time t. For these, the heat equation $u_t = c^2 \nabla^2 u$ becomes the Laplace equation.

Before solving an initial or boundary value problem, one often transforms the PDE into coordinates in which the boundary of the region considered is given by simple formulas. Thus in polar coordinates given by $x = r \cos \theta$, $y = r \sin \theta$, the **Laplacian** becomes (Sec. 12.9)

$$(6) \qquad \nabla^2 u = u_{rr} + \frac{1}{r} u_r + \frac{1}{r^2} u_{\theta\theta};$$

for spherical coordinates see Sec. 12.10. If one now separates the variables, one gets **Bessel's equation** from (2) and (6) (vibrating circular membrane, Sec. 12.9) and **Legendre's equation** from (5) transformed into spherical coordinates (Sec. 12.10).



PART D

Complex Analysis

CHAPTER 13 Complex Numbers and Functions

CHAPTER 14 Complex Integration

CHAPTER 15 Power Series, Taylor Series

CHAPTER 16 Laurent Series. Residue Integration

CHAPTER 17 Conformal Mapping

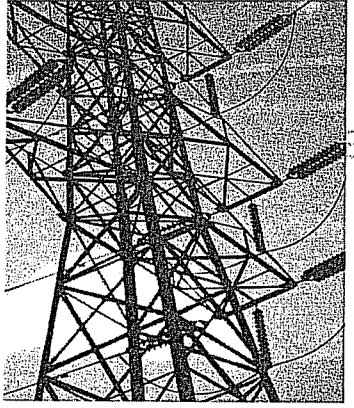
CHAPTER 18 Complex Analysis and Potential Theory

Many engineering problems can be modeled, investigated, and solved by functions of a complex variable. For simpler problems, some acquaintance with complex numbers will suffice. This is true for simpler electric circuits and mechanical vibrating systems. For more complicated problems in heat conduction, fluid flow, electrostatics, etc., one needs *the theory of complex analytic functions,* briefly called **complex analysis.** The importance of the latter in applied mathematics has three main reasons:

1. Most importantly, the real and imaginary parts of an analytic function satisfy Laplace's equation in two real variables. Hence two-dimensional potential problems can be solved by methods for analytic functions, and this is often simpler than working in real.

2. Many complicated real and complex integrals in applications can be evaluated by the elegant methods of complex integration.

3. Most functions in engineering mathematics are analytic functions, and their study as functions of a complex variable leads to a deeper understanding of their properties and to interrelations in complex that have no analog in real calculus.



CHAPTER 13

Complex Numbers and Functions

Complex numbers and their geometric representation in the **complex plane** are discussed in Secs. 13.1 and 13.2. Complex analysis is concerned with complex **analytic functions** as defined in Sec. 13.3. Checking for analyticity is done by the **Cauchy–Riemann equations** (Sec. 13.4). These equations are of basic importance, also because of their relation to **Laplace's equation**.

The remaining sections of the chapter are devoted to elementary complex functions (exponential, trigonometric, hyperbolic, and logarithmic functions). These generalize the familiar real functions of calculus. Their detailed knowledge is an absolute necessity in practical work, just as that of their real counterparts is in calculus.

Prerequisite: Elementary calculus.
References and Answers to Problems: App. 1 Part D, App. 2.

13.1 Complex Numbers. Complex Plane

Equations without *real* solutions, such as $x^2 = -1$ or $x^2 - 10x + 40 = 0$, were observed early in history and led to the introduction of complex numbers.[1] By definition, a **complex number** z is an ordered pair (x, y) of real numbers x and y, written

$$z = (x, y).$$

x is called the **real part** and y the **imaginary part** of z, written

$$x = \operatorname{Re} z, \qquad y = \operatorname{Im} z.$$

By definition, two complex numbers are **equal** if and only if their real parts are equal and their imaginary parts are equal.

(0, 1) is called the **imaginary unit** and is denoted by i,

(1) $$i = (0, 1).$$

[1]First to use complex numbers for this purpose was the Italian mathematician GIROLAMO CARDANO (1501–1576), who found the formula for solving cubic equations. The term "complex number" was introduced by CARL FRIEDRICH GAUSS (see the footnote in Sec. 5.4), who also paved the way for a general use of complex numbers.

Addition, Multiplication. Notation $z = x + iy$

Addition of two complex numbers $z_1 = (x_1, y_1)$ and $z_2 = (x_2, y_2)$ is defined by

$$z_1 + z_2 = (x_1, y_1) + (x_2, y_2) = (x_1 + x_2, \quad y_1 + y_2). \tag{2}$$

Multiplication is defined by

$$z_1 z_2 = (x_1, y_1)(x_2, y_2) = (x_1 x_2 - y_1 y_2, \quad x_1 y_2 + x_2 y_1). \tag{3}$$

In particular, these two definitions imply that

$$(x_1, 0) + (x_2, 0) = (x_1 + x_2, 0)$$

and

$$(x_1, 0)(x_2, 0) = (x_1 x_2, 0)$$

as for real numbers x_1, x_2. Hence the complex numbers ***"extend"*** the real numbers. We can thus write

$$(x, 0) = x. \qquad \text{Similarly,} \qquad (0, y) = iy$$

because by (1) and the definition of multiplication we have

$$iy = (0, 1)y = (0, 1)(y, 0) = (0 \cdot y - 1 \cdot 0, \quad 0 \cdot 0 + 1 \cdot y) = (0, y).$$

Together we have by addition $(x, y) = (x, 0) + (0, y) = x + iy$:

In practice, complex numbers $z = (x, y)$ are written

$$z = x + iy \tag{4}$$

or $z = x + yi$, e.g., $17 + 4i$ (instead of $i4$).

Electrical engineers often write j instead if i because they need i for the current.

If $x = 0$, then $z = iy$ and is called **pure imaginary.** Also, (1) and (3) give

$$i^2 = -1 \tag{5}$$

because by the definition of multiplication, $i^2 = ii = (0, 1)(0, 1) = (-1, 0) = -1$.

For **addition** the standard notation (4) gives [see (2)]

$$(x_1 + iy_1) + (x_2 + iy_2) = (x_1 + x_2) + i(y_1 + y_2).$$

For **multiplication** the standard notation gives the following very simple recipe. Multiply each term by each other term and use $i^2 = -1$ when it occurs [see (3)]:

$$\begin{aligned}(x_1 + iy_1)(x_2 + iy_2) &= x_1 x_2 + i x_1 y_2 + i y_1 x_2 + i^2 y_1 y_2 \\ &= (x_1 x_2 - y_1 y_2) + i(x_1 y_2 + x_2 y_1).\end{aligned}$$

This agrees with (3). And it shows that $x + iy$ is a more practical notation for complex numbers than (x, y).

If you know vectors, you see that (2) is vector addition, whereas the multiplication (3) has no counterpart in the usual vector algebra.

EXAMPLE 1 **Real Part, Imaginary Part, Sum and Product of Complex Numbers**

Let $z_1 = 8 + 3i$ and $z_2 = 9 - 2i$, Then Re $z_1 = 8$, Im $z_1 = 3$, Re $z_2 = 9$, Im $z_2 = -2$ and

$$z_1 + z_2 = (8 + 3i) + (9 - 2i) = 17 + i,$$

$$z_1 z_2 = (8 + 3i)(9 - 2i) = 72 + 6 + i(-16 + 27) = 78 + 11i.$$

Subtraction, Division

Subtraction and **division** are defined as the inverse operations of addition and multiplication, respectively. Thus the **difference** $z = z_1 - z_2$ is the complex number z for which $z_1 = z + z_2$. Hence by (2),

(6) $$z_1 - z_2 = (x_1 - x_2) + i(y_1 - y_2).$$

The **quotient** $z = z_1/z_2\ (z_2 \neq 0)$ is the complex number z for which $z_1 = zz_2$. If we equate the real and the imaginary parts on both sides of this equation, setting $z = x + iy$, we obtain $x_1 = x_2x - y_2y$, $y_1 = y_2x + x_2y$. The solution is

(7*) $$z = \frac{z_1}{z_2} = x + iy, \qquad x = \frac{x_1x_2 + y_1y_2}{x_2^2 + y_2^2}, \qquad y = \frac{x_2y_1 - x_1y_2}{x_2^2 + y_2^2}.$$

The ***practical rule*** used to get this is by multiplying numerator and denominator of z_1/z_2 by $x_2 - iy_2$ and simplifiying:

(7) $$z = \frac{x_1 + iy_1}{x_2 + iy_2} = \frac{(x_1 + iy_1)(x_2 - iy_2)}{(x_2 + iy_2)(x_2 - iy_2)} = \frac{x_1x_2 + y_1y_2}{x_2^2 + y_2^2} + i\,\frac{x_2y_1 - x_1y_2}{x_2^2 + y_2^2}.$$

EXAMPLE 2 **Difference and Quotient of Complex Numbers**

For $z_1 = 8 + 3i$ and $z_2 = 9 - 2i$ we get $z_1 - z_2 = (8 + 3i) - (9 - 2i) = -1 + 5i$ and

$$\frac{z_1}{z_2} = \frac{8 + 3i}{9 - 2i} = \frac{(8 + 3i)(9 + 2i)}{(9 - 2i)(9 + 2i)} = \frac{66 + 43i}{81 + 4} = \frac{66}{85} + \frac{43}{85}\,i.$$

Check the division by multiplication to get $8 + 3i$.

Complex numbers satisfy the same commutative, associative, and distributive laws as real numbers (see the problem set).

Complex Plane

This was algebra. Now comes geometry: the geometrical representation of complex numbers as points in the plane. This is of great practical importance. The idea is quite simple and natural. We choose two perpendicular coordinate axes, the horizontal x-axis, called the **real axis,** and the vertical y-axis, called the **imaginary axis.** On both axes we choose the same unit of length (Fig. 315). This is called a **Cartesian coordinate system.**

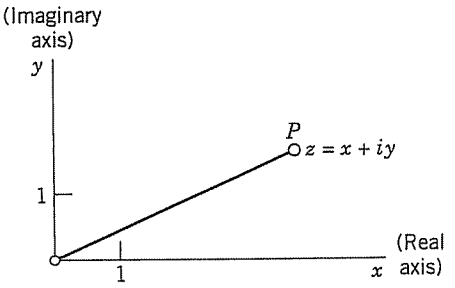


Fig. 315. The complex plane

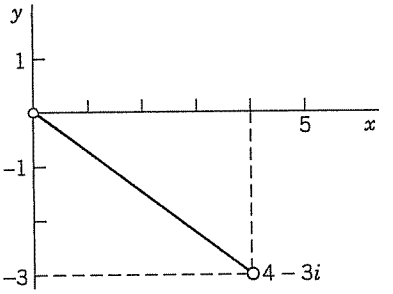


Fig. 316. The number $4 - 3i$ in the complex plane

We now plot a given complex number $z = (x, y) = x + iy$ as the point P with coordinates x, y. The xy-plane in which the complex numbers are represented in this way is called the **complex plane.**[2] Figure 316 shows an example.

Instead of saying "the point represented by z in the complex plane" we say briefly and simply ***"the point z in the complex plane."*** This will cause no misunderstandings.

Addition and subtraction can now be visualized as illustrated in Figs. 317 and 318.

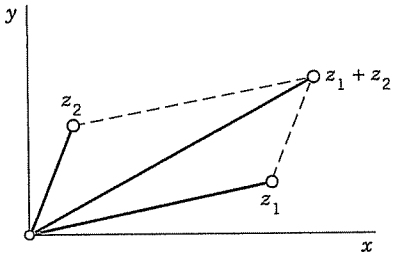


Fig. 317. Addition of complex numbers

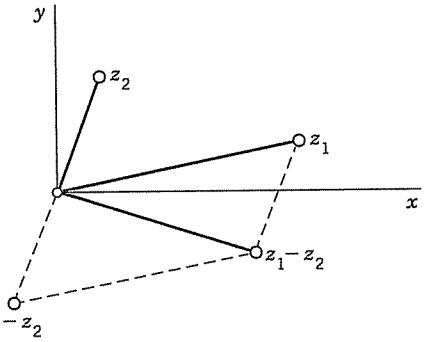


Fig. 318. Subtraction of complex numbers

Complex Conjugate Numbers

The complex conjugate $\bar{z}$ of a complex number $z = x + iy$ is defined by

$$\bar{z} = x - iy.$$

It is obtained geometrically by reflecting the point z in the real axis. Figure 319 shows this for $z = 5 + 2i$ and its conjugate $\bar{z} = 5 - 2i$.

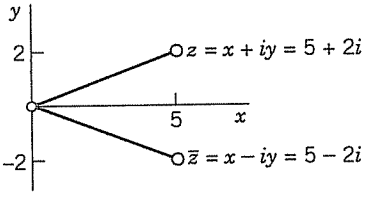


Fig. 319. Complex conjugate numbers

[2]Sometimes called the **Argand diagram,** after the French mathematician JEAN ROBERT ARGAND (1768–1822), born in Geneva and later librarian in Paris. His paper on the complex plane appeared in 1806, nine years after a similar memoir by the Norwegian mathematician CASPAR WESSEL (1745–1818), a surveyor of the Danish Academy of Science.

The complex conjugate is important because it permits us to switch from complex to real. Indeed, by multiplication, $z\bar{z} = x^2 + y^2$ (verify!). By addition and subtraction, $z + \bar{z} = 2x$, $z - \bar{z} = 2iy$. We thus obtain for the real part x and the imaginary part y (not iy!) of $z = x + iy$ the important formulas

$$\text{(8)} \qquad \operatorname{Re} z = x = \frac{1}{2}(z + \bar{z}), \qquad \operatorname{Im} z = y = \frac{1}{2i}(z - \bar{z}).$$

If z is real, $z = x$, then $\bar{z} = z$ by the definition of $\bar{z}$, and conversely. Working with conjugates is easy, since we have

$$\text{(9)} \qquad \overline{(z_1 + z_2)} = \bar{z}_1 + \bar{z}_2, \qquad \overline{(z_1 - z_2)} = \bar{z}_1 - \bar{z}_2,$$
$$\overline{(z_1 z_2)} = \bar{z}_1 \bar{z}_2, \qquad \overline{\left(\frac{z_1}{z_2}\right)} = \frac{\bar{z}_1}{\bar{z}_2}.$$

EXAMPLE 3 Illustration of (8) and (9)

Let $z_1 = 4 + 3i$ and $z_2 = 2 + 5i$. Then by (8),

$$\operatorname{Im} z_1 = \frac{1}{2i}[(4 + 3i) - (4 - 3i)] = \frac{3i + 3i}{2i} = 3.$$

Also, the multiplication formula in (9) is verified by

$$\overline{(z_1 z_2)} = \overline{(4 + 3i)(2 + 5i)} = \overline{(-7 + 26i)} = -7 - 26i,$$
$$\bar{z}_1 \bar{z}_2 = (4 - 3i)(2 - 5i) = -7 - 26i. \quad \blacksquare$$

PROBLEM SET 13.2

1. **(Powers of i)** Show that $i^2 = -1$, $i^3 = -i$, $i^4 = 1$, $i^5 = i, \cdots$ and $1/i = -i$, $1/i^2 = -1$, $1/i^3 = i, \cdots$.
2. **(Rotation)** Multiplication by i is geometrically a counterclockwise rotation through $\pi/2$ (90°). Verify this by graphing z and iz and the angle of rotation for $z = 2 + 2i$, $z = -1 - 5i$, $z = 4 - 3i$.
3. **(Division)** Verify the calculation in (7).
4. **(Multiplication)** If the product of two complex numbers is zero, show that at least one factor must be zero.
5. Show that $z = x + iy$ is pure imaginary if and only if $\bar{z} = -z$.
6. **(Laws for conjugates)** Verify (9) for $z_1 = 24 + 10i$, $z_2 = 4 + 6i$.

7–15 COMPLEX ARITHMETIC

Let $z_1 = 2 + 3i$ and $z_2 = 4 - 5i$. Showing the details of your work, find (in the form $x + iy$):

7. $(5z_1 + 3z_2)^2$
8. $\bar{z}_1 \bar{z}_2$
9. $\operatorname{Re}(1/z_1^2)$
10. $\operatorname{Re}(z_2^2)$, $(\operatorname{Re} z_2)^2$
11. z_2/z_1
12. $\bar{z}_1/\bar{z}_2$, $\overline{(z_1/z_2)}$
13. $(4z_1 - z_2)^2$
14. $\bar{z}_1/z_1$, $z_1/\bar{z}_1$
15. $(z_1 + z_2)/(z_1 - z_2)$

16–19 Let $z = x + iy$. Find:

16. $\operatorname{Im} z^3$, $(\operatorname{Im} z)^3$
17. $\operatorname{Re}(1/\bar{z})$
18. $\operatorname{Im}[(1 + i)^8 z^2]$
19. $\operatorname{Re}(1/\bar{z}^2)$

20. **(Laws of addition and multiplication)** Derive the following laws for complex numbers from the corresponding laws for real numbers.

$$z_1 + z_2 = z_2 + z_1, \quad z_1 z_2 = z_2 z_1 \quad \textit{(Commutative laws)}$$
$$(z_1 + z_2) + z_3 = z_1 + (z_2 + z_3), \quad \textit{(Associative laws)}$$
$$(z_1 z_2) z_3 = z_1 (z_2 z_3)$$
$$z_1(z_2 + z_3) = z_1 z_2 + z_1 z_3 \quad \textit{(Distributive law)}$$
$$0 + z = z + 0 = z,$$
$$z + (-z) = (-z) + z = 0, \qquad z \cdot 1 = z.$$

13.2 Polar Form of Complex Numbers. Powers and Roots

The complex plane becomes even more useful and gives further insight into the arithmetic operations for complex numbers if besides the xy-coordinates we also employ the usual polar coordinates r, θ defined by

$$x = r\cos\theta, \qquad y = r\sin\theta. \tag{1}$$

We see that then $z = x + iy$ takes the so-called **polar form**

$$z = r(\cos\theta + i\sin\theta). \tag{2}$$

r is called the **absolute value** or **modulus** of z and is denoted by $|z|$. Hence

$$|z| = r = \sqrt{x^2 + y^2} = \sqrt{z\bar{z}}\,. \tag{3}$$

Geometrically, $|z|$ is the distance of the point z from the origin (Fig. 320). Similarly, $|z_1 - z_2|$ is the distance between z_1 and z_2 (Fig. 321).

θ is called the **argument** of z and is denoted by $\arg z$. Thus (Fig. 320)

$$\theta = \arg z = \arctan\frac{y}{x} \qquad (z \neq 0). \tag{4}$$

Geometrically, θ is the directed angle from the positive x-axis to OP in Fig. 320. Here, as in calculus, all ***angles are measured in radians and positive in the counterclockwise sense.***

For $z = 0$ this angle θ is undefined. (Why?) For a given $z \neq 0$ it is determined only up to integer multiples of 2π since cosine and sine are periodic with period 2π. But one often wants to specify a unique value of $\arg z$ of a given $z \neq 0$. For this reason one defines the **principal value** $\operatorname{Arg} z$ (with capital A!) of $\arg z$ by the double inequality

$$-\pi < \operatorname{Arg} z \leqq \pi. \tag{5}$$

Then we have $\operatorname{Arg} z = 0$ for positive real $z = x$, which is practical, and $\operatorname{Arg} z = \pi$ (not $-\pi$!) for negative real z, e.g., for $z = -4$. The principal value (5) will be important in connection with roots, the complex logarithm (Sec. 13.7), and certain integrals. Obviously, for a given $z \neq 0$ the other values of $\arg z$ are $\arg z = \operatorname{Arg} z \pm 2n\pi$ $(n = \pm 1, \pm 2, \cdots)$.

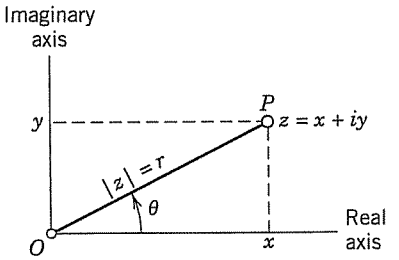


Fig. 320. Complex plane, polar form of a complex number

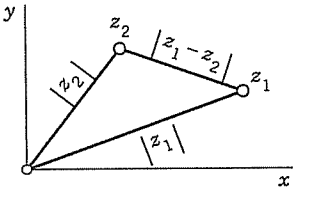


Fig. 321. Distance between two points in the complex plane

EXAMPLE 1 **Polar Form of Complex Numbers. Principal Value Arg z**

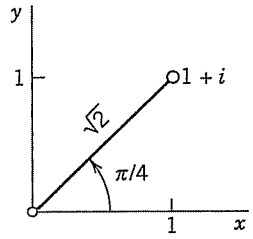


Fig. 322. Example 1

$z = 1 + i$ (Fig. 322) has the polar form $z = \sqrt{2}\,(\cos \frac{1}{4}\pi + i \sin \frac{1}{4}\pi)$. Hence we obtain

$$|z| = \sqrt{2}, \qquad \arg z = \tfrac{1}{4}\pi \pm 2n\pi \ (n = 0, 1, \cdots), \qquad \text{and} \qquad \operatorname{Arg} z = \tfrac{1}{4}\pi \quad \text{(the principal value).}$$

Similarly, $z = 3 + 3\sqrt{3}i = 6\,(\cos \frac{1}{3}\pi + i \sin \frac{1}{3}\pi)$, $|z| = 6$, and $\operatorname{Arg} z = \frac{1}{3}\pi$. ■

CAUTION! In using (4), we must pay attention to the quadrant in which z lies, since $\tan \theta$ has period π, so that the arguments of z and $-z$ have the same tangent. *Example:* for $\theta_1 = \arg (1 + i)$ and $\theta_2 = \arg (-1 - i)$ we have $\tan \theta_1 = \tan \theta_2 = 1$.

Triangle Inequality

Inequalities such as $x_1 < x_2$ make sense for *real* numbers, but not in complex because *there is no natural way of ordering complex numbers*. However, inequalities between absolute values (which are real!), such as $|z_1| < |z_2|$ (meaning that z_1 is closer to the origin than z_2) are of great importance. The daily bread of the complex analyst is the **triangle inequality**

$$|z_1 + z_2| \leqq |z_1| + |z_2| \tag{6}$$

(Fig. 323)

which we shall use quite frequently. This inequality follows by noting that the three points 0, z_1, and $z_1 + z_2$ are the vertices of a triangle (Fig. 323) with sides $|z_1|$, $|z_2|$, and $|z_1 + z_2|$, and one side cannot exceed the sum of the other two sides. A formal proof is left to the reader (Prob. 35). (The triangle degenerates if z_1 and z_2 lie on the same straight line through the origin.)

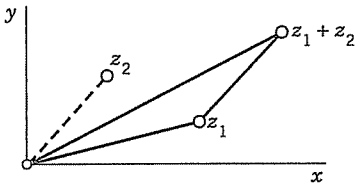


Fig. 323. Triangle inequality

By induction we obtain from (6) the **generalized triangle inequality**

$$|z_1 + z_2 + \cdots + z_n| \leqq |z_1| + |z_2| + \cdots + |z_n|; \tag{6*}$$

that is, *the absolute value of a sum cannot exceed the sum of the absolute values of the terms.*

EXAMPLE 2 **Triangle Inequality**

If $z_1 = 1 + i$ and $z_2 = -2 + 3i$, then (sketch a figure!)

$$|z_1 + z_2| = |-1 + 4i| = \sqrt{17} = 4.123 < \sqrt{2} + \sqrt{13} = 5.020.$$

■

Multiplication and Division in Polar Form

This will give us a "geometrical" understanding of multiplication and division. Let

$$z_1 = r_1(\cos \theta_1 + i \sin \theta_1) \qquad \text{and} \qquad z_2 = r_2(\cos \theta_2 + i \sin \theta_2).$$

Multiplication. By (3) in Sec. 13.1 the product is at first

$$z_1z_2 = r_1r_2[(\cos\theta_1\cos\theta_2 - \sin\theta_1\sin\theta_2) + i(\sin\theta_1\cos\theta_2 + \cos\theta_1\sin\theta_2)].$$

The addition rules for the sine and cosine [(6) in App. A3.1] now yield

$$z_1z_2 = r_1r_2[\cos(\theta_1 + \theta_2) + i\sin(\theta_1 + \theta_2)]. \tag{7}$$

Taking absolute values on both sides of (7), we see that *the absolute value of a product equals the* ***product*** *of the absolute values of the factors,*

$$|z_1z_2| = |z_1||z_2|. \tag{8}$$

Taking arguments in (7) shows that *the argument of a product equals the* ***sum*** *of the arguments of the factors,*

$$\arg(z_1z_2) = \arg z_1 + \arg z_2 \qquad \text{(up to multiples of } 2\pi). \tag{9}$$

Division. We have $z_1 = (z_1/z_2)z_2$. Hence $|z_1| = |(z_1/z_2)z_2| = |z_1/z_2||z_2|$ and by division by $|z_2|$

$$\left|\frac{z_1}{z_2}\right| = \frac{|z_1|}{|z_2|} \qquad (z_2 \neq 0). \tag{10}$$

Similarly, $\arg z_1 = \arg[(z_1/z_2)z_2] = \arg(z_1/z_2) + \arg z_2$ and by subtraction of $\arg z_2$

$$\arg\frac{z_1}{z_2} = \arg z_1 - \arg z_2 \qquad \text{(up to multiples of } 2\pi). \tag{11}$$

Combining (10) and (11) we also have the analog of (7),

$$\frac{z_1}{z_2} = \frac{r_1}{r_2}[\cos(\theta_1 - \theta_2) + i\sin(\theta_1 - \theta_2)]. \tag{12}$$

To comprehend this formula, note that it is the polar form of a complex number of absolute value r_1/r_2 and argument $\theta_1 - \theta_2$. But these are the absolute value and argument of z_1/z_2, as we can see from (10), (11), and the polar forms of z_1 and z_2.

EXAMPLE 3 Illustration of Formulas (8)–(11)

Let $z_1 = -2 + 2i$ and $z_2 = 3i$. Then $z_1z_2 = -6 - 6i$, $z_1/z_2 = 2/3 + (2/3)i$. Hence (make a sketch)

$$|z_1z_2| = 6\sqrt{2} = 3\sqrt{8} = |z_1||z_2|, \qquad |z_1/z_2| = 2\sqrt{2}/3 = |z_1|/|z_2|,$$

and for the arguments we obtain $\operatorname{Arg} z_1 = 3\pi/4$, $\operatorname{Arg} z_2 = \pi/2$,

$$\operatorname{Arg}(z_1z_2) = -\frac{3\pi}{4} = \operatorname{Arg} z_1 + \operatorname{Arg} z_2 - 2\pi, \qquad \operatorname{Arg}(z_1/z_2) = \frac{\pi}{4} = \operatorname{Arg} z_1 - \operatorname{Arg} z_2. \quad \blacksquare$$

EXAMPLE 4 Integer Powers of z. De Moivre's Formula

From (8) and (9) with $z_1 = z_2 = z$ we obtain by induction for $n = 0, 1, 2, \cdots$

$$z^n = r^n(\cos n\theta + i \sin n\theta). \tag{13}$$

Similarly, (12) with $z_1 = 1$ and $z_2 = z^n$ gives (13) for $n = -1, -2, \cdots$. For $|z| = r = 1$, formula (13) becomes **De Moivre's formula**[3]

$$(\cos \theta + i \sin \theta)^n = \cos n\theta + i \sin n\theta. \tag{13*}$$

We can use this to express $\cos n\theta$ and $\sin n\theta$ in terms of powers of $\cos \theta$ and $\sin \theta$. For instance, for $n = 2$ we have on the left $\cos^2 \theta + 2i \cos \theta \sin \theta - \sin^2 \theta$. Taking the real and imaginary parts on both sides of (13*) with $n = 2$ gives the familiar formulas

$$\cos 2\theta = \cos^2 \theta - \sin^2 \theta, \qquad \sin 2\theta = 2 \cos \theta \sin \theta.$$

This shows that *complex* methods often simplify the derivation of *real* formulas. Try $n = 3$. ■

Roots

If $z = w^n$ $(n = 1, 2, \cdots)$, then to each value of w there corresponds *one* value of z. We shall immediately see that, conversely, to a given $z \neq 0$ there correspond precisely n distinct values of w. Each of these values is called an ***n*th root** of z, and we write

$$w = \sqrt[n]{z}. \tag{14}$$

Hence this symbol is ***multivalued,*** namely, *n-valued.* The n values of $\sqrt[n]{z}$ can be obtained as follows. We write z and w in polar form

$$z = r(\cos \theta + i \sin \theta) \qquad \text{and} \qquad w = R(\cos \phi + i \sin \phi).$$

Then the equation $w^n = z$ becomes, by De Moivre's formula (with ϕ instead of θ)

$$w^n = R^n(\cos n\phi + i \sin n\phi) = z = r(\cos \theta + i \sin \theta).$$

The absolute values on both sides must be equal; thus, $R^n = r$, so that $R = \sqrt[n]{r}$, where $\sqrt[n]{r}$ is positive real (an absolute value must be nonnegative!) and thus uniquely determined. Equating the arguments $n\phi$ and θ and recalling that θ is determined only up to integer multiples of 2π, we obtain

$$n\phi = \theta + 2k\pi, \qquad \text{thus} \qquad \phi = \frac{\theta}{n} + \frac{2k\pi}{n}$$

where k is an integer. For $k = 0, 1, \cdots, n - 1$ we get n *distinct* values of w. Further integers of k would give values already obtained. For instance, $k = n$ gives $2k\pi/n = 2\pi$,

[3]ABRAHAM DE MOIVRE (1667–1754), French mathematician, who pioneered the use of complex numbers in trigonometry and also contributed to probability theory (see Sec. 24.8).

hence the w corresponding to $k = 0$, etc. Consequently, $\sqrt[n]{z}$, for $z \neq 0$, has the n distinct values

$$\sqrt[n]{z} = \sqrt[n]{r}\left(\cos\frac{\theta + 2k\pi}{n} + i\sin\frac{\theta + 2k\pi}{n}\right) \tag{15}$$

where $k = 0, 1, \cdots, n-1$. These n values lie on a circle of radius $\sqrt[n]{r}$ with center at the origin and constitute the vertices of a regular polygon of n sides. The value of $\sqrt[n]{z}$ obtained by taking the principal value of $\arg z$ and $k = 0$ in (15) is called the **principal value** of $w = \sqrt[n]{z}$.

Taking $z = 1$ in (15), we have $|z| = r = 1$ and $\operatorname{Arg} z = 0$. Then (15) gives

$$\sqrt[n]{1} = \cos\frac{2k\pi}{n} + i\sin\frac{2k\pi}{n}, \qquad k = 0, 1, \cdots, n-1. \tag{16}$$

These n values are called the ***n*th roots of unity.** They lie on the circle of radius 1 and center 0, briefly called the **unit circle** (and used quite frequently!). Figures 324–326 show $\sqrt[3]{1} = 1, -\frac{1}{2} \pm \frac{1}{2}\sqrt{3}\,i$, $\sqrt[4]{1} = \pm 1, \pm i$, and $\sqrt[5]{1}$.

If ω denotes the value corresponding to $k = 1$ in (16), then the n values of $\sqrt[n]{1}$ can be written as

$$1, \omega, \omega^2, \cdots, \omega^{n-1}.$$

More generally, if w_1 is any nth root of an arbitrary complex number z ($\neq 0$), then the n values of $\sqrt[n]{z}$ in (15) are

$$w_1, \quad w_1\omega, \quad w_1\omega^2, \quad \cdots, \quad w_1\omega^{n-1} \tag{17}$$

because multiplying w_1 by ω^k corresponds to increasing the argument of w_1 by $2k\pi/n$. Formula (17) motivates the introduction of roots of unity and shows their usefulness.

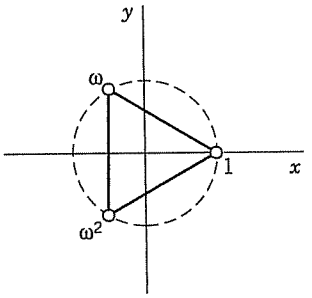


Fig. 324. $\sqrt[3]{1}$

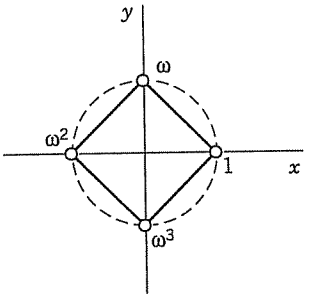


Fig. 325. $\sqrt[4]{1}$

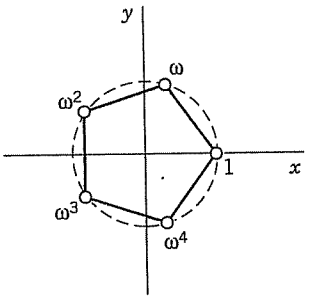


Fig. 326. $\sqrt[5]{1}$

PROBLEM SET 13.2

1–8 POLAR FORM

Do these problems very carefully since polar forms will be needed frequently. Represent in polar form and graph in the complex plane as in Fig. 322 on p. 608. (Show the details of your work.)

1. $3 - 3i$
2. $2i, -2i$
3. -5
4. $\frac{1}{2} + \frac{1}{4}\pi i$
5. $\dfrac{1 + i}{1 - i}$
6. $\dfrac{3\sqrt{2} + 2i}{-\sqrt{2} - (2/3)i}$

7. $\dfrac{-6 + 5i}{3i}$ 8. $\dfrac{2 + 3i}{5 + 4i}$

9–15 PRINCIPAL ARGUMENT

Determine the principal value of the argument.

9. $-1 - i$
10. $-20 + i, -20 - i$
11. $4 \pm 3i$
12. $-\pi^2$
13. $7 \pm 7i$
14. $(1 + i)^{12}$
15. $(9 + 9i)^3$

16–20 CONVERSION TO $x + iy$

Represent in the form $x + iy$ and graph it in the complex plane.

16. $\cos \frac{1}{2}\pi + i \sin (\pm\frac{1}{2}\pi)$
17. $3(\cos 0.2 + i \sin 0.2)$
18. $4(\cos \frac{1}{3}\pi \pm i \sin \frac{1}{3}\pi)$
19. $\cos (-1) + i \sin (-1)$
20. $12(\cos \frac{3}{2}\pi + i \sin \frac{3}{2}\pi)$

21–25 ROOTS

Find and graph all roots in the complex plane.

21. $\sqrt{-i}$
22. $\sqrt[8]{1}$
23. $\sqrt[4]{-1}$
24. $\sqrt[3]{3 + 4i}$
25. $\sqrt[5]{-1}$

26. **TEAM PROJECT. Square Root.** **(a)** Show that $w = \sqrt{z}$ has the values

$$w_1 = \sqrt{r}\left[\cos\frac{\theta}{2} + i \sin\frac{\theta}{2}\right],$$

$$(18)\quad w_2 = \sqrt{r}\left[\cos\left(\frac{\theta}{2} + \pi\right) + i \sin\left(\frac{\theta}{2} + \pi\right)\right] = -w_1.$$

(b) Obtain from (18) the often more practical formula

$$(19)\quad \sqrt{z} = \pm\left[\sqrt{\tfrac{1}{2}(|z| + x)} + (\operatorname{sign} y)\, i\sqrt{\tfrac{1}{2}(|z| + x)}\right]$$

where $\operatorname{sign} y = 1$ if $y \geqq 0$, $\operatorname{sign} y = -1$ if $y < 0$, and all square roots of positive numbers are taken with positive sign. *Hint:* Use (10) in App. A3.1 with $x = \theta/2$.

(c) Find the square roots of $4i$, $16 - 30i$, and $9 + 8\sqrt{7}i$ by both (18) and (19) and comment on the work involved.

(d) Do some further examples of your own and apply a method of checking your results.

27–30 EQUATIONS

Solve and graph all solutions, showing the details:

27. $z^2 - (8 - 5i)z + 40 - 20i = 0$ (Use (19).)
28. $z^4 + (5 - 14i)z^2 - (24 + 10i) = 0$
29. $8z^2 - (36 - 6i)z + 42 - 11i = 0$
30. $z^4 + 16 = 0$. Then use the solutions to factor $z^4 + 16$ into quadratic factors with *real* coefficients.

31. **CAS PROJECT. Roots of Unity and Their Graphs.** Write a program for calculating these roots and for graphing them as points on the unit circle. Apply the program to $z^n = 1$ with $n = 2, 3, \cdots, 10$. Then extend the program to one for arbitrary roots, using an idea near the end of the text, and apply the program to examples of your choice.

32–35 INEQUALITIES AND AN EQUATION

Verify or prove as indicated.

32. **(Re and Im)** Prove $|\operatorname{Re} z| \leqq |z|$, $|\operatorname{Im} z| \leqq |z|$.
33. **(Parallelogram equality)** Prove

$$|z_1 + z_2|^2 + |z_1 - z_2|^2 = 2(|z_1|^2 + |z_2|^2).$$

Explain the name.

34. **(Triangle inequality)** Verify (6) for $z_1 = 4 + 7i$, $z_2 = 5 + 2i$.
35. **(Triangle inequality)** Prove (6).

13.3 Derivative. Analytic Function

Our study of complex functions will involve point sets in the complex plane. Most important will be the following ones.

Circles and Disks. Half-Planes

The **unit circle** $|z| = 1$ (Fig. 327) has already occurred in Sec. 13.2. Figure 328 shows a general circle of radius ρ and center a. Its equation is

$$|z - a| = \rho$$

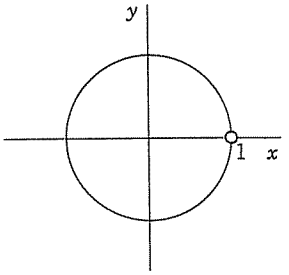


Fig. 327. Unit circle

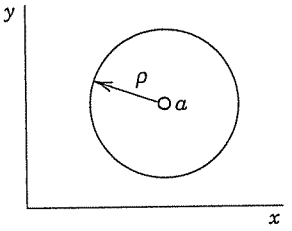


Fig. 328. Circle in the complex plane

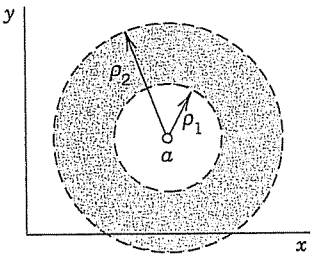


Fig. 329. Annulus in the complex plane

because it is the set of all z whose distance $|z - a|$ from the center a equals ρ. Accordingly, its interior (**"open circular disk"**) is given by $|z - a| < \rho$, its interior plus the circle itself (**"closed circular disk"**) by $|z - a| \leqq \rho$, and its exterior by $|z - a| > \rho$. As an example, sketch this for $a = 1 + i$ and $\rho = 2$, to make sure that you understand these inequalities.

An open circular disk $|z - a| < \rho$ is also called a **neighborhood** of a or, more precisely, a *ρ-neighborhood* of a. And a has infinitely many of them, one for each value of ρ (> 0), and a is a point of each of them, by definition!

In modern literature *any set* containing a ρ-neighborhood of a is also called a *neighborhood* of a.

Figure 329 shows an **open annulus** (circular ring) $\rho_1 < |z - a| < \rho_2$, which we shall need later. This is the set of all z whose distance $|z - a|$ from a is greater than ρ_1 but less than ρ_2. Similarly, the **closed annulus** $\rho_1 \leqq |z - a| \leqq \rho_2$ includes the two circles.

Half-Planes. By the (open) *upper* **half-plane** we mean the set of all points $z = x + iy$ such that $y > 0$. Similarly, the condition $y < 0$ defines the *lower half-plane,* $x > 0$ the *right half-plane,* and $x < 0$ the *left half-plane.*

For Reference: Concepts on Sets in the Complex Plane

To our discussion of special sets let us add some general concepts related to sets that we shall need throughout Chaps. 13–18; keep in mind that you can find them here.

By a **point set** in the complex plane we mean any sort of collection of finitely many or infinitely many points. Examples are the solutions of a quadratic equation, the points of a line, the points in the interior of a circle as well as the sets discussed just before.

A set S is called **open** if every point of S has a neighborhood consisting entirely of points that belong to S. For example, the points in the interior of a circle or a square form an open set, and so do the points of the right half-plane $\operatorname{Re} z = x > 0$.

A set S is called **connected** if any two of its points can be joined by a broken line of finitely many straight-line segments all of whose points belong to S. An open and connected set is called a **domain.** Thus an open disk and an open annulus are domains. An open square with a diagonal removed is not a domain since this set is not connected. (Why?)

The **complement** of a set S in the complex plane is the set of all points of the complex plane that ***do not belong*** to S. A set S is called **closed** if its complement is open. For example, the points on and inside the unit circle form a closed set ("closed unit disk") since its complement $|z| > 1$ is open.

A **boundary point** of a set S is a point every neighborhood of which contains both points that belong to S and points that do not belong to S. For example, the boundary

points of an annulus are the points on the two bounding circles. Clearly, if a set S is open, then no boundary point belongs to S; if S is closed, then every boundary point belongs to S. The set of all boundary points of a set S is called the **boundary** of S.

A **region** is a set consisting of a domain plus, perhaps, some or all of its boundary points. **WARNING!** "Domain" is the *modern* term for an open connected set. Nevertheless, some authors still call a domain a "region" and others make no distinction between the two terms.

Complex Function

Complex analysis is concerned with complex functions that are differentiable in some domain. Hence we should first say what we mean by a complex function and then define the concepts of limit and derivative in complex. This discussion will be similar to that in calculus. Nevertheless it needs great attention because it will show interesting basic differences between real and complex calculus.

Recall from calculus that a *real* function f defined on a set S of real numbers (usually an interval) is a rule that assigns to every x in S a real number $f(x)$, called the *value* of f at x. Now in complex, S is a set of *complex* numbers. And a **function** f defined on S is a rule that assigns to every z in S a complex number w, called the *value* of f at z. We write

$$w = f(z).$$

Here z varies in S and is called a **complex variable.** The set S is called the *domain of definition* of f or, briefly, the **domain** of f. (In most cases S will be open and connected, thus a domain as defined just before.)

Example: $w = f(z) = z^2 + 3z$ is a complex function defined for all z; that is, its domain S is the whole complex plane.

The set of all values of a function f is called the **range** *of* f.

w is complex, and we write $w = u + iv$, where u and v are the real and imaginary parts, respectively. Now w depends on $z = x + iy$. Hence u becomes a real function of x and y, and so does v. We may thus write

$$w = f(z) = u(x, y) + iv(x, y).$$

This shows that a *complex* function $f(z)$ is equivalent to a *pair* of *real* functions $u(x, y)$ and $v(x, y)$, each depending on the two real variables x and y.

EXAMPLE 1 **Function of a Complex Variable**

Let $w = f(z) = z^2 + 3z$. Find u and v and calculate the value of f at $z = 1 + 3i$.

Solution. $u = \text{Re}\, f(z) = x^2 - y^2 + 3x$ and $v = 2xy + 3y$. Also,

$$f(1 + 3i) = (1 + 3i)^2 + 3(1 + 3i) = 1 - 9 + 6i + 3 + 9i = -5 + 15i.$$

This shows that $u(1, 3) = -5$ and $v(1, 3) = 15$. Check this by using the expressions for u and v. ■

EXAMPLE 2 **Function of a Complex Variable**

Let $w = f(z) = 2iz + 6\bar{z}$. Find u and v and the value of f at $z = \frac{1}{2} + 4i$.

Solution. $f(z) = 2i(x + iy) + 6(x - iy)$ gives $u(x, y) = 6x - 2y$ and $v(x, y) = 2x - 6y$. Also,

$$f(\tfrac{1}{2} + 4i) = 2i(\tfrac{1}{2} + 4i) + 6(\tfrac{1}{2} - 4i) = i - 8 + 3 - 24i = -5 - 23i.$$

Check this as in Example 1. ■

Remarks on Notation and Terminology

1. Strictly speaking, $f(z)$ denotes the value of f at z, but it is a convenient abuse of language to talk about *the function* $f(z)$ (instead of *the function* f), thereby exhibiting the notation for the independent variable.

2. We assume all functions to be ***single-valued relations,*** as usual: to each z in S there corresponds but *one* value $w = f(z)$ (but, of course, several z may give the same value $w = f(z)$, just as in calculus). Accordingly, we shall *not use* the term "multivalued function" (used in some books on complex analysis) for a multivalued relation, in which to a z there corresponds more than one w.

Limit, Continuity

A function $f(z)$ is said to have the **limit** l as z approaches a point z_0, written

$$\lim_{z \to z_0} f(z) = l, \tag{1}$$

if f is defined in a neighborhood of z_0 (except perhaps at z_0 itself) and if the values of f are "close" to l for all z "close" to z_0; in precise terms, if for every positive real ϵ we can find a positive real δ such that for all $z \neq z_0$ in the disk $|z - z_0| < \delta$ (Fig. 330) we have

$$|f(z) - l| < \epsilon; \tag{2}$$

geometrically, if for every $z \neq z_0$ in that δ-disk the value of f lies in the disk (2).

Formally, this definition is similar to that in calculus, but there is a big difference. Whereas in the real case, x can approach an x_0 only along the real line, here, by definition, z may approach z_0 ***from any direction*** in the complex plane. This will be quite essential in what follows.

If a limit exists, it is unique. (See Team Project 26.)

A function $f(z)$ is said to be **continuous** at $z = z_0$ if $f(z_0)$ is defined and

$$\lim_{z \to z_0} f(z) = f(z_0). \tag{3}$$

Note that by definition of a limit this implies that $f(z)$ is defined in some neighborhood of z_0.

$f(z)$ is said to be *continuous in a domain* if it is continuous at each point of this domain.

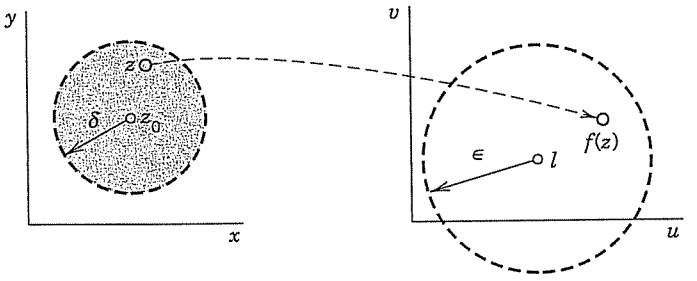


Fig. 330. Limit

Derivative

The **derivative** of a complex function f at a point z_0 is written $f'(z_0)$ and is defined by

$$f'(z_0) = \lim_{\Delta z \to 0} \frac{f(z_0 + \Delta z) - f(z_0)}{\Delta z} \tag{4}$$

provided this limit exists. Then f is said to be **differentiable** at z_0. If we write $\Delta z = z - z_0$, we have $z = z_0 + \Delta z$ and (4) takes the form

$$f'(z_0) = \lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0}. \tag{4'}$$

Now comes an ***important point.*** Remember that, by the definition of limit, $f(z)$ is defined in a neighborhood of z_0 and z in (4′) may approach z_0 from any direction in the complex plane. Hence differentiability at z_0 means that, along whatever path z approaches z_0, the quotient in (4′) always approaches a certain value and all these values are equal. This is important and should be kept in mind.

EXAMPLE 3 **Differentiability. Derivative**

The function $f(z) = z^2$ is differentiable for all z and has the derivative $f'(z) = 2z$ because

$$f'(z) = \lim_{\Delta z \to 0} \frac{(z + \Delta z)^2 - z^2}{\Delta z} = \lim_{\Delta z \to 0} \frac{z^2 + 2z\Delta z + (\Delta z)^2 - z^2}{\Delta z} = \lim_{\Delta z \to 0} (2z + \Delta z) = 2z. \qquad \blacksquare$$

*The **differentiation rules** are the same as in real calculus,* since their proofs are literally the same. Thus for any analytic functions f and g and constants c we have

$$(cf)' = cf', \quad (f + g)' = f' + g', \quad (fg)' = f'g + fg', \quad \left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2}$$

as well as the chain rule and the power rule $(z^n)' = nz^{n-1}$ (n integer).

Also, if $f(z)$ is differentiable at z_0, it is continuous at z_0. (See Team Project 26.)

EXAMPLE 4 **$\bar{z}$ not Differentiable**

It may come as a surprise that there are many complex functions that do not have a derivative at any point. For instance, $f(z) = \bar{z} = x - iy$ is such a function. To see this, we write $\Delta z = \Delta x + i\Delta y$ and obtain

$$\frac{f(z + \Delta z) - f(z)}{\Delta z} = \frac{\overline{(z + \Delta z)} - \bar{z}}{\Delta z} = \frac{\overline{\Delta z}}{\Delta z} = \frac{\Delta x - i\Delta y}{\Delta x + i\Delta y}. \tag{5}$$

If $\Delta y = 0$, this is $+1$. If $\Delta x = 0$, this is -1. Thus (5) approaches $+1$ along path I in Fig. 331 but -1 along path II. Hence, by definition, the limit of (5) as $\Delta z \to 0$ does not exist at any z. $\blacksquare$

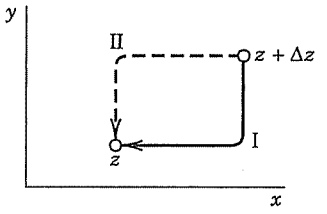


Fig. 331. Paths in (5)

Surprising as Example 4 may be, it merely illustrates that differentiability of a *complex* function is a rather severe requirement.

The idea of proof (approach of z from different directions) is basic and will be used again as the crucial argument in the next section.

Analytic Functions

Complex analysis is concerned with the theory and application of "analytic functions," that is, functions that are differentiable in some domain, so that we can do "calculus in complex." The definition is as follows.

DEFINITION

Analyticity

A function $f(z)$ is said to be *analytic in a domain* D if $f(z)$ is defined and differentiable at all points of D. The function $f(z)$ is said to be *analytic at a point* $z = z_0$ in D if $f(z)$ is analytic in a neighborhood of z_0.

Also, by an **analytic function** we mean a function that is analytic in *some* domain.

Hence analyticity of $f(z)$ at z_0 means that $f(z)$ has a derivative at every point in some neighborhood of z_0 (including z_0 itself since, by definition, z_0 is a point of all its neighborhoods). This concept is *motivated* by the fact that it is of no practical interest if a function is differentiable merely at a single point z_0 but not throughout some neighborhood of z_0. Team Project 26 gives an example.

A more modern term for *analytic in D* is *holomorphic in D.*

EXAMPLE 5 **Polynomials, Rational Functions**

The nonnegative integer powers $1, z, z^2, \cdots$ are analytic in the entire complex plane, and so are **polynomials,** that is, functions of the form

$$f(z) = c_0 + c_1 z + c_2 z^2 + \cdots + c_n z^n$$

where $c_0, \cdots, c_n$ are complex constants.

The quotient of two polynomials $g(z)$ and $h(z)$,

$$f(z) = \frac{g(z)}{h(z)},$$

is called a **rational function.** This f is analytic except at the points where $h(z) = 0$; here we assume that common factors of g and h have been canceled.

Many further analytic functions will be considered in the next sections and chapters. ■

The concepts discussed in this section extend familiar concepts of calculus. Most important is the concept of an analytic function, the exclusive concern of complex analysis. Although many simple functions are not analytic, the large variety of remaining functions will yield a most beautiful branch of mathematics that is very useful in engineering and physics.

PROBLEM SET 13.3

1–10 CURVES AND REGIONS OF PRACTICAL INTEREST

Find and sketch or graph the sets in the complex plane given by

1. $|z - 3 - 2i| = \frac{4}{3}$
2. $1 \leqq |z - 1 + 4i| \leqq 5$
3. $0 < |z - 1| < 1$
4. $-\pi < \operatorname{Re} z < \pi$
5. $\operatorname{Im} z^2 = 2$
6. $\operatorname{Re} z > -1$
7. $|z + 1| = |z - 1|$
8. $|\operatorname{Arg} z| \leqq \frac{1}{4}\pi$
9. $\operatorname{Re} z \leqq \operatorname{Im} z$
10. $\operatorname{Re} (1/z) < 1$

11. WRITING PROJECT. Sets in the Complex Plane. Extend the part of the text on sets in the complex plane by formulating that part in your own words and including examples of your own and comparing with calculus when applicable.

COMPLEX FUNCTIONS AND DERIVATIVES

12–15 Function Values. Find Re f and Im f. Also find their values at the given point z.

12. $f = 3z^2 - 6z + 3i$, $z = 2 + i$

13. $f = z/(z + 1)$, $z = 4 - 5i$

14. $f = 1/(1 - z)$, $z = \frac{1}{2} + \frac{1}{4}i$

15. $f = 1/z^2$, $z = 1 + i$

16–19 Continuity. Find out (and give reason) whether $f(z)$ is continuous at $z = 0$ if $f(0) = 0$ and for $z \neq 0$ the function f is equal to:

16. $[\text{Re}\,(z^2)]/|z|^2$

17. $[\text{Im}\,(z^2)]/|z|$

18. $|z|^2\,\text{Re}\,(1/z)$

19. $(\text{Im}\,z)/(1 - |z|)$

20–24 Derivative. Differentiate

20. $(z^2 - 9)/(z^2 + 1)$

21. $(z^3 + i)^2$

22. $(3z + 4i)/(1.5iz - 2)$

23. $i/(1 - z)^2$

24. $z^2/(z + i)^2$

25. CAS PROJECT. Graphing Functions. Find and graph Re f, Im f, and $|f|$ as surfaces over the z-plane. Also graph the two families of curves Re $f(z) = const$ and Im $f(z) = const$ in the same figure, and the curves $|f(z)| = const$ in another figure, where (a) $f(z) = z^2$, (b) $f(z) = 1/z$, (c) $f(z) = z^4$.

26. TEAM PROJECT. Limit, Continuity, Derivative

(a) Limit. Prove that (1) is equivalent to the pair of relations

$$\lim_{z \to z_0} \text{Re}\, f(z) = \text{Re}\, l, \qquad \lim_{z \to z_0} \text{Im}\, f(z) = \text{Im}\, l.$$

(b) Limit. If $\lim_{z \to z_0} f(z)$ exists, show that this limit is unique.

(c) Continuity. If $z_1, z_2, \cdots$ are complex numbers for which $\lim_{n \to \infty} z_n = a$, and if $f(z)$ is continuous at $z = a$, show that $\lim_{n \to \infty} f(z_n) = f(a)$.

(d) Continuity. If $f(z)$ is differentiable at z_0, show that $f(z)$ is continuous at z_0.

(e) Differentiability. Show that $f(z) = \text{Re}\, z = x$ is not differentiable at any z. Can you find other such functions?

(f) Differentiability. Show that $f(z) = |z|^2$ is differentiable only at $z = 0$; hence it is nowhere analytic.

13.4 Cauchy–Riemann Equations. Laplace's Equation

The Cauchy–Riemann equations are the most important equations in this chapter and one of the pillars on which complex analysis rests. They provide a criterion (a test) for the analyticity of a complex function

$$w = f(z) = u(x, y) + iv(x, y).$$

Roughly, f is analytic in a domain D if and only if the first partial derivatives of u and v satisfy the two **Cauchy–Riemann equations**[4]

$$\textbf{(1)} \qquad u_x = v_y, \qquad u_y = -v_x$$

[4]The French mathematician AUGUSTIN-LOUIS CAUCHY (see Sec. 2.5) and the German mathematicians BERNHARD RIEMANN (1826–1866) and KARL WEIERSTRASS (1815–1897; see also Sec. 15.5) are the founders of complex analysis. Riemann received his Ph.D. (in 1851) under Gauss (Sec. 5.4) at Göttingen, where he also taught until he died, when he was only 39 years old. He introduced the concept of the integral as it is used in basic calculus courses, and made important contributions to differential equations, number theory, and mathematical physics. He also developed the so-called Riemannian geometry, which is the mathematical foundation of Einstein's theory of relativity; see Ref. [GR9] in App. 1.

everywhere in D; here $u_x = \partial u/\partial x$ and $u_y = \partial u/\partial y$ (and similarly for v) are the usual notations for partial derivatives. The precise formulation of this statement is given in Theorems 1 and 2.

Example: $f(z) = z^2 = x^2 - y^2 + 2ixy$ is analytic for all z (see Example 3 in Sec. 13.3), and $u = x^2 - y^2$ and $v = 2xy$ satisfy (1), namely, $u_x = 2x = v_y$ as well as $u_y = -2y = -v_x$. More examples will follow.

THEOREM 1

Cauchy–Riemann Equations

Let $f(z) = u(x, y) + iv(x, y)$ be defined and continuous in some neighborhood of a point $z = x + iy$ and differentiable at z itself. Then at that point, the first-order partial derivatives of u and v exist and satisfy the Cauchy–Riemann equations (1).

Hence if $f(z)$ is analytic in a domain D, those partial derivatives exist and satisfy (1) *at all points of D.*

PROOF By assumption, the derivative $f'(z)$ at z exists. It is given by

$$f'(z) = \lim_{\Delta z \to 0} \frac{f(z + \Delta z) - f(z)}{\Delta z}. \tag{2}$$

The idea of the proof is very simple. By the definition of a limit in complex (Sec. 13.3) we can let Δz approach zero along any path in a neighborhood of z. Thus we may choose the two paths I and II in Fig. 332 and equate the results. By comparing the real parts we shall obtain the first Cauchy–Riemann equation and by comparing the imaginary parts the second. The technical details are as follows.

We write $\Delta z = \Delta x + i\Delta y$. Then $z + \Delta z = x + \Delta x + i(y + \Delta y)$, and in terms of u and v the derivative in (2) becomes

$$f'(z) = \lim_{\Delta z \to 0} \frac{[u(x + \Delta x, y + \Delta y) + iv(x + \Delta x, y + \Delta y)] - [u(x, y) + iv(x, y)]}{\Delta x + i\Delta y}. \tag{3}$$

We first choose path I in Fig. 332. Thus we let $\Delta y \to 0$ first and then $\Delta x \to 0$. After Δy is zero, $\Delta z = \Delta x$. Then (3) becomes, if we first write the two u-terms and then the two v-terms,

$$f'(z) = \lim_{\Delta x \to 0} \frac{u(x + \Delta x, y) - u(x, y)}{\Delta x} + i \lim_{\Delta x \to 0} \frac{v(x + \Delta x, y) - v(x, y)}{\Delta x}.$$

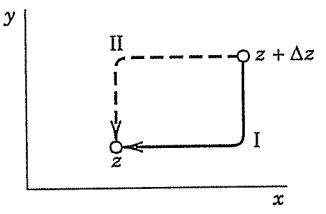


Fig. 332. Paths in (2)

Since $f'(z)$ exists, the two real limits on the right exist. By definition, they are the partial derivatives of u and v with respect to x. Hence the derivative $f'(z)$ of $f(z)$ can be written

$$f'(z) = u_x + iv_x. \tag{4}$$

Similarly, if we choose path II in Fig. 332, we let $\Delta x \to 0$ first and then $\Delta y \to 0$. After Δx is zero, $\Delta z = i\Delta y$, so that from (3) we now obtain

$$f'(z) = \lim_{\Delta y \to 0} \frac{u(x, y + \Delta y) - u(x, y)}{i\,\Delta y} + i \lim_{\Delta y \to 0} \frac{v(x, y + \Delta y) - v(x, y)}{i\,\Delta y}.$$

Since $f'(z)$ exists, the limits on the right exist and give the partial derivatives of u and v with respect to y; noting that $1/i = -i$, we thus obtain

$$f'(z) = -iu_y + v_y. \tag{5}$$

The existence of the derivative $f'(z)$ thus implies the existence of the four partial derivatives in (4) and (5). By equating the real parts u_x and v_y in (4) and (5) we obtain the first Cauchy–Riemann equation (1). Equating the imaginary parts gives the other. This proves the first statement of the theorem and implies the second because of the definition of analyticity. ∎

Formulas (4) and (5) are also quite practical for calculating derivatives $f'(z)$, as we shall see.

EXAMPLE 1 **Cauchy–Riemann Equations**

$f(z) = z^2$ is analytic for all z. It follows that the Cauchy–Riemann equations must be satisfied (as we have verified above).

For $f(z) = \bar{z} = x - iy$ we have $u = x$, $v = -y$ and see that the second Cauchy–Riemann equation is satisfied, $u_y = -v_x = 0$, but the first is not: $u_x = 1 \neq v_y = -1$. We conclude that $f(z) = \bar{z}$ is not analytic, confirming Example 4 of Sec. 13.3. Note the savings in calculation! ∎

The Cauchy–Riemann equations are fundamental because they are not only necessary but also sufficient for a function to be analytic. More precisely, the following theorem holds.

THEOREM 2

Cauchy–Riemann Equations

If two real-valued continuous functions $u(x, y)$ and $v(x, y)$ of two real variables x and y have ***continuous*** *first partial derivatives that satisfy the Cauchy–Riemann equations in some domain D, then the complex function $f(z) = u(x, y) + iv(x, y)$ is analytic in D.*

The proof is more involved than that of Theorem 1 and we leave it optional (see App. 4).

Theorems 1 and 2 are of great practical importance, since by using the Cauchy–Riemann equations we can now easily find out whether or not a given complex function is analytic.

EXAMPLE 2 Cauchy–Riemann Equations. Exponential Function

Is $f(z) = u(x, y) + iv(x, y) = e^x(\cos y + i \sin y)$ analytic?

Solution. We have $u = e^x \cos y$, $v = e^x \sin y$ and by differentiation

$$u_x = e^x \cos y, \qquad v_y = e^x \cos y$$
$$u_y = -e^x \sin y, \qquad v_x = e^x \sin y.$$

We see that the Cauchy–Riemann equations are satisfied and conclude that $f(z)$ is analytic for all z. ($f(z)$ will be the complex analog of e^x known from calculus.) ■

EXAMPLE 3 An Analytic Function of Constant Absolute Value Is Constant

The Cauchy–Riemann equations also help in deriving general properties of analytic functions.

For instance, show that if $f(z)$ is analytic in a domain D and $|f(z)| = k = const$ in D, then $f(z) = const$ in D. (We shall make crucial use of this in Sec. 18.6 in the proof of Theorem 3.)

Solution. By assumption, $|f|^2 = |u + iv|^2 = u^2 + v^2 = k^2$. By differentiation,

$$uu_x + vv_x = 0,$$
$$uu_y + vv_y = 0.$$

Now use $v_x = -u_y$ in the first equation and $v_y = u_x$ in the second, to get

$$\text{(6)} \qquad \begin{aligned} &\text{(a)} \quad uu_x - vu_y = 0, \\ &\text{(b)} \quad uu_y + vu_x = 0. \end{aligned}$$

To get rid of u_y, multiply (6a) by u and (6b) by v and add. Similarly, to eliminate u_x, multiply (6a) by $-v$ and (6b) by u and add. This yields

$$(u^2 + v^2)u_x = 0.$$
$$(u^2 + v^2)u_y = 0.$$

If $k^2 = u^2 + v^2 = 0$, then $u = v = 0$; hence $f = 0$. If $k^2 = u^2 + v^2 \neq 0$, then $u_x = u_y = 0$. Hence, by the Cauchy–Riemann equations, also $v_x = v_y = 0$. Together this implies $u = const$ and $v = const$; hence $f = const$. ■

We mention that if we use the polar form $z = r(\cos \theta + i \sin \theta)$ and set $f(z) = u(r, \theta) + iv(r, \theta)$, then the **Cauchy–Riemann equations** are (Prob. 11)

$$\text{(7)} \qquad u_r = \frac{1}{r} v_\theta, \qquad v_r = -\frac{1}{r} u_\theta \qquad (r > 0).$$

Laplace's Equation. Harmonic Functions

The great importance of complex analysis in engineering mathematics results mainly from the fact that both the real part and the imaginary part of an analytic function satisfy Laplace's equation, the most important PDE of physics, which occurs in gravitation, electrostatics, fluid flow, heat conduction, and so on (see Chaps. 12 and 18).

THEOREM 3

Laplace's Equation

If $f(z) = u(x, y) + iv(x, y)$ is analytic in a domain D, then both u and v satisfy **Laplace's equation**

(8) $$\nabla^2 u = u_{xx} + u_{yy} = 0$$

(∇^2 read "nabla squared") *and*

(9) $$\nabla^2 v = v_{xx} + v_{yy} = 0,$$

in D and have continuous second partial derivatives in D.

PROOF Differentiating $u_x = v_y$ with respect to x and $u_y = -v_x$ with respect to y, we have

(10) $$u_{xx} = v_{yx}, \qquad u_{yy} = -v_{xy}.$$

Now the derivative of an analytic function is itself analytic, as we shall prove later (in Sec. 14.4). This implies that u and v have continuous partial derivatives of all orders; in particular, the mixed second derivatives are equal: $v_{yx} = v_{xy}$. By adding (10) we thus obtain (8). Similarly, (9) is obtained by differentiating $u_x = v_y$ with respect to y and $u_y = -v_x$ with respect to x and subtracting, using $u_{xy} = u_{yx}$. ∎

Solutions of Laplace's equation having ***continuous*** second-order partial derivatives are called **harmonic functions** and their theory is called **potential theory** (see also Sec. 12.10). Hence the real and imaginary parts of an analytic function are harmonic functions.

If two harmonic functions u and v satisfy the Cauchy–Riemann equations in a domain D, they are the real and imaginary parts of an analytic function f in D. Then v is said to be a **harmonic conjugate function** of u in D. (Of course, this has absolutely nothing to do with the use of "conjugate" for $\bar{z}$.)

EXAMPLE 4 **How to Find a Harmonic Conjugate Function by the Cauchy–Riemann Equations**

Verify that $u = x^2 - y^2 - y$ is harmonic in the whole complex plane and find a harmonic conjugate function v of u.

Solution. $\nabla^2 u = 0$ by direct calculation. Now $u_x = 2x$ and $u_y = -2y - 1$. Hence because of the Cauchy–Riemann equations a conjugate v of u must satisfy

$$v_y = u_x = 2x, \qquad v_x = -u_y = 2y + 1.$$

Integrating the first equation with respect to y and differentiating the result with respect to x, we obtain

$$v = 2xy + h(x), \qquad v_x = 2y + \frac{dh}{dx}.$$

A comparison with the second equation shows that $dh/dx = 1$. This gives $h(x) = x + c$. Hence $v = 2xy + x + c$ (c any real constant) is the most general harmonic conjugate of the given u. The corresponding analytic function is

$$f(z) = u + iv = x^2 - y^2 - y + i(2xy + x + c) = z^2 + iz + ic.$$ ∎

Example 4 illustrates that *a conjugate of a given harmonic function is uniquely determined up to an arbitrary real additive constant.*

The Cauchy–Riemann equations are the most important equations in this chapter. Their relation to Laplace's equation opens wide ranges of engineering and physical applications, as we shall show in Chap. 18.

PROBLEM SET 13.4

1–10 CAUCHY–RIEMANN EQUATIONS

Are the following functions analytic? [Use (1) or (7).]

1. $f(z) = z^4$
2. $f(z) = \text{Im}\ (z^2)$
3. $e^{2x}(\cos y + i \sin y)$
4. $f(z) = 1/(1 - z^4)$
5. $e^{-x}(\cos y - i \sin y)$
6. $f(z) = \text{Arg}\ \pi z$
7. $f(z) = \text{Re}\ z + \text{Im}\ z$
8. $f(z) = \ln |z| + i\ \text{Arg}\ z$
9. $f(z) = i/z^8$
10. $f(z) = z^2 + 1/z^2$

11. **(Cauchy–Riemann equations in polar form)** Derive (7) from (1).

12–21 HARMONIC FUNCTIONS

Are the following functions harmonic? If your answer is yes, find a corresponding analytic function $f(z) = u(x, y) + iv(x, y)$.

12. $u = xy$
13. $v = xy$
14. $v = -y/(x^2 + y^2)$
15. $u = \ln |z|$
16. $v = \ln |z|$
17. $u = x^3 - 3xy^2$
18. $u = 1/(x^2 + y^2)$
19. $v = (x^2 - y^2)^2$
20. $u = \cos x \cosh y$
21. $u = e^{-x} \sin 2y$

22–24

Determine a, b, c such that the given functions are harmonic and find a harmonic conjugate.

22. $u = e^{3x} \cos ay$
23. $u = \sin x \cosh cy$
24. $u = ax^3 + by^3$

25. **(Harmonic conjugate)** Show that if u is harmonic and v is a harmonic conjugate of u, then u is a harmonic conjugate of $-v$.

26. **TEAM PROJECT. Conditions for $f(z) = const$.** Let $f(z)$ be analytic. Prove that each of the following conditions is sufficient for $f(z) = const$.
   (a) $\text{Re}\ f(z) = const$
   (b) $\text{Im}\ f(z) = const$
   (c) $f'(z) = 0$
   (d) $|f(z)| = const$ (see Example 3)

27. **(Two further formulas for the derivative).** Formulas (4), (5), and (11) (below) are needed from time to time. Derive

   (11) $f'(z) = u_x - iu_y, \qquad f'(z) = v_y + iv_x.$

28. **CAS PROJECT. Equipotential Lines.** Write a program for graphing equipotential lines $u = const$ of a harmonic function u and of its conjugate v on the same axes. Apply the program to (a) $u = x^2 - y^2$, $v = 2xy$, (b) $u = x^3 - 3xy^2$, $v = 3x^2y - y^3$, (c) $u = e^x \cos y$, $v = e^x \sin y$.

13.5 Exponential Function

In the remaining sections of this chapter we discuss the basic elementary complex functions, the exponential function, trigonometric functions, logarithm, and so on. They will be counterparts to the familiar functions of calculus, to which they reduce when $z = x$ is real. They are indispensable throughout applications, and some of them have interesting properties not shared by their real counterparts.

We begin with one of the most important analytic functions, the complex **exponential function**

$$e^z, \qquad \text{also written} \qquad \exp z.$$

The definition of e^z in terms of the real functions e^x, $\cos y$, and $\sin y$ is

$$e^z = e^x(\cos y + i \sin y). \tag{1}$$

This definition is motivated by the fact the e^z ***extends*** the real exponential function e^x of calculus in a natural fashion. Namely:

(A) $e^z = e^x$ for real $z = x$ because $\cos y = 1$ and $\sin y = 0$ when $y = 0$.

(B) e^z is analytic for all z. (Proved in Example 2 of Sec. 13.4.)

(C) The derivative of e^z is e^z, that is,

$$(e^z)' = e^z. \tag{2}$$

This follows from (4) in Sec. 13.4,

$$(e^z)' = (e^x \cos y)_x + i(e^x \sin y)_x = e^x \cos y + ie^x \sin y = e^z.$$

REMARK. This definition provides for a relatively simple discussion. We could define e^z by the familiar series $1 + x + x^2/2! + x^3/3! + \cdots$ with x replaced by z, but we would then have to discuss complex series at this very early stage. (We will show the connection in Sec. 15.4.)

Further Properties. A function $f(z)$ that is analytic for all z is called an **entire function.** Thus, e^z is entire. Just as in calculus the ***functional relation***

$$e^{z_1+z_2} = e^{z_1}e^{z_2} \tag{3}$$

holds for any $z_1 = x_1 + iy_1$ and $z_2 = x_2 + iy_2$. Indeed, by (1),

$$e^{z_1}e^{z_2} = e^{x_1}(\cos y_1 + i \sin y_1)\, e^{x_2}(\cos y_2 + i \sin y_2).$$

Since $e^{x_1}e^{x_2} = e^{x_1+x_2}$ for these *real* functions, by an application of the addition formulas for the cosine and sine functions (similar to that in Sec. 13.2) we see that

$$e^{z_1}e^{z_2} = e^{x_1+x_2}[\cos (y_1 + y_2) + i \sin (y_1 + y_2)] = e^{z_1+z_2}$$

as asserted. An interesting special case of (3) is $z_1 = x$, $z_2 = iy$; then

$$e^z = e^x e^{iy}. \tag{4}$$

Furthermore, for $z = iy$ we have from (1) the so-called **Euler formula**

$$e^{iy} = \cos y + i \sin y. \tag{5}$$

Hence the **polar form** of a complex number, $z = r(\cos \theta + i \sin \theta)$, may now be written

$$z = re^{i\theta}. \tag{6}$$

From (5) we obtain

$$e^{2\pi i} = 1 \tag{7}$$

as well as the important formulas (verify!)

$$e^{\pi i/2} = i, \qquad e^{\pi i} = -1, \qquad e^{-\pi i/2} = -i, \qquad e^{-\pi i} = -1. \tag{8}$$

Another consequence of (5) is

$$(9) \qquad |e^{iy}| = |\cos y + i \sin y| = \sqrt{\cos^2 y + \sin^2 y} = 1.$$

That is, for pure imaginary exponents the exponential function has absolute value 1, a result you should remember. From (9) and (1),

$$(10) \qquad |e^z| = e^x. \qquad \text{Hence} \qquad \arg e^z = y \pm 2n\pi \ (n = 0, 1, 2, \cdots),$$

since $|e^z| = e^x$ shows that (1) is actually e^z in polar form.

From $|e^z| = e^x \neq 0$ in (10) we see that

$$(11) \qquad e^z \neq 0 \qquad \text{for all } z.$$

So here we have an entire function that never vanishes, in contrast to (nonconstant) polynomials, which are also entire (Example 5 in Sec. 13.3) but always have a zero, as is proved in algebra.

Periodicity of e^z with period $2\pi i$,

$$(12) \qquad e^{z+2\pi i} = e^z \qquad \text{for all } z$$

is a basic property that follows from (1) and the periodicity of $\cos y$ and $\sin y$. Hence all the values that $w = e^z$ can assume are already assumed in the horizontal strip of width 2π

$$(13) \qquad -\pi < y \leqq \pi \qquad \text{(Fig. 333).}$$

This infinite strip is called a **fundamental region** of e^z.

EXAMPLE 1 **Function Values. Solution of Equations.**

Computation of values from (1) provides no problem. For instance, verify that

$$e^{1.4-0.6i} = e^{1.4}(\cos 0.6 - i \sin 0.6) = 4.055(0.8253 - 0.5646i) = 3.347 - 2.289i$$

$$|e^{1.4-0.6i}| = e^{1.4} = 4.055, \qquad \operatorname{Arg} e^{1.4-0.6i} = -0.6.$$

To illustrate (3), take the product of

$$e^{2+i} = e^2(\cos 1 + i \sin 1) \qquad \text{and} \qquad e^{4-i} = e^4(\cos 1 - i \sin 1)$$

and verify that it equals $e^2 e^4(\cos^2 1 + \sin^2 1) = e^6 = e^{(2+i)+(4-i)}$.

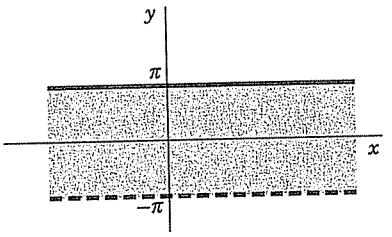


Fig. 333. Fundamental region of the exponential function e^z in the z-plane

To solve the equation $e^z = 3 + 4i$, note first that $|e^z| = e^x = 5$, $x = \ln 5 = 1.609$ is the real part of all solutions. Now, since $e^x = 5$,

$$e^x \cos y = 3, \qquad e^x \sin y = 4, \qquad \cos y = 0.6, \qquad \sin y = 0.8, \qquad y = 0.927.$$

Ans. $z = 1.609 + 0.927i \pm 2n\pi i$ $(n = 0, 1, 2, \cdots)$. These are infinitely many solutions (due to the periodicity of e^z). They lie on the vertical line $x = 1.609$ at a distance 2π from their neighbors. ■

To summarize: many properties of $e^z = \exp z$ parallel those of e^x; an exception is the periodicity of e^z with $2\pi i$, which suggested the concept of a fundamental region. Keep in mind that e^z is an *entire function.* (Do you still remember what that means?)

PROBLEM SET 13.5

1. Using the Cauchy–Riemann equations, show that e^z is entire.

2–8 Values of e^z. Compute e^z in the form $u + iv$ and $|e^z|$, where z equals:

2. $3 + \pi i$
3. $1 + 2i$
4. $\sqrt{2} - \frac{1}{2}\pi i$
5. $7\pi i/2$
6. $(1 + i)\pi$
7. $0.8 - 5i$
8. $9\pi i/2$

9–12 Real and Imaginary Parts. Find Re and Im of:

9. e^{-2z}
10. e^{z^3}
11. e^{z^2}
12. $e^{1/z}$

13–17 Polar Form. Write in polar form:

13. $\sqrt{i}$
14. $1 + i$
15. $\sqrt[n]{z}$
16. $3 + 4i$
17. -9

18–21 Equations. Find all solutions and graph some of them in the complex plane.

18. $e^{3z} = 4$
19. $e^z = -2$
20. $e^z = 0$
21. $e^z = 4 - 3i$

22. TEAM PROJECT. Further Properties of the Exponential Function. (a) Analyticity. Show that e^z is entire. What about $e^{1/z}$? $e^{\bar{z}}$? $e^x(\cos ky + i \sin ky)$? (Use the Cauchy–Riemann equations.)

(b) Special values. Find all z such that (i) e^z is real, (ii) $|e^{-z}| < 1$, (iii) $e^{\bar{z}} = \overline{e^z}$.

(c) Harmonic function. Show that $u = e^{xy} \cos(x^2/2 - y^2/2)$ is harmonic and find a conjugate.

(d) Uniqueness. It is interesting that $f(z) = e^z$ is uniquely determined by the two properties $f(x + i0) = e^x$ and $f'(z) = f(z)$, where f is assumed to be entire. Prove this using the Cauchy–Riemann equations.

13.6 Trigonometric and Hyperbolic Functions

Just as we extended the real e^x to the complex e^z in Sec. 13.5, we now want to extend the familiar *real* trigonometric functions to *complex trigonometric functions.* We can do this by the use of the Euler formulas (Sec. 13.5)

$$e^{ix} = \cos x + i \sin x, \qquad e^{-ix} = \cos x - i \sin x.$$

By addition and subtraction we obtain for the *real* cosine and sine

$$\cos x = \frac{1}{2}(e^{ix} + e^{-ix}), \qquad \sin x = \frac{1}{2i}(e^{ix} - e^{-ix}).$$

This suggests the following definitions for complex values $z = x + iy$:

(1) $$\cos z = \frac{1}{2}(e^{iz} + e^{-iz}), \qquad \sin z = \frac{1}{2i}(e^{iz} - e^{-iz}).$$

It is quite remarkable that here in complex, functions come together that are unrelated in real. This is not an isolated incident but is typical of the general situation and shows the advantage of working in complex.

Furthermore, as in calculus we define

(2) $$\tan z = \frac{\sin z}{\cos z}, \qquad \cot z = \frac{\cos z}{\sin z}$$

and

(3) $$\sec z = \frac{1}{\cos z}, \qquad \csc z = \frac{1}{\sin z}.$$

Since e^z is entire, $\cos z$ and $\sin z$ are entire functions. $\tan z$ and $\sec z$ are not entire; they are analytic except at the points where $\cos z$ is zero; and $\cot z$ and $\csc z$ are analytic except where $\sin z$ is zero. Formulas for the derivatives follow readily from $(e^z)' = e^z$ and (1)–(3); as in calculus,

(4) $$(\cos z)' = -\sin z, \qquad (\sin z)' = \cos z, \qquad (\tan z)' = \sec^2 z,$$

etc. Equation (1) also shows that **Euler's formula** ***is valid in complex:***

(5) $$e^{iz} = \cos z + i \sin z \qquad \text{for all } z.$$

The real and imaginary parts of $\cos z$ and $\sin z$ are needed in computing values, and they also help in displaying properties of our functions. We illustrate this with a typical example.

EXAMPLE 1 **Real and Imaginary Parts. Absolute Value. Periodicity**

Show that

(6) **(a)** $\cos z = \cos x \cosh y - i \sin x \sinh y$

(b) $\sin z = \sin x \cosh y + i \cos x \sinh y$

and

(7) **(a)** $|\cos z|^2 = \cos^2 x + \sinh^2 y$

(b) $|\sin z|^2 = \sin^2 x + \sinh^2 y$

and give some applications of these formulas.

Solution. From (1),

$$\begin{aligned}\cos z &= \tfrac{1}{2}(e^{i(x+iy)} + e^{-i(x+iy)}) \\ &= \tfrac{1}{2}e^{-y}(\cos x + i \sin x) + \tfrac{1}{2}e^{y}(\cos x - i \sin x) \\ &= \tfrac{1}{2}(e^{y} + e^{-y}) \cos x - \tfrac{1}{2}i(e^{y} - e^{-y}) \sin x.\end{aligned}$$

This yields (6a) since, as is known form calculus,

(8) $$\cosh y = \tfrac{1}{2}(e^{y} + e^{-y}), \qquad \sinh y = \tfrac{1}{2}(e^{y} - e^{-y});$$

(6b) is obtained similarly. From (6a) and $\cosh^2 y = 1 + \sinh^2 y$ we obtain

$$|\cos z|^2 = (\cos^2 x)\,(1 + \sinh^2 y) + \sin^2 x \sinh^2 y.$$

Since $\sin^2 x + \cos^2 x = 1$, this gives (7a), and (7b) is obtained similarly.

For instance, $\cos(2 + 3i) = \cos 2 \cosh 3 - i \sin 2 \sinh 3 = -4.190 - 9.109i$.

From (6) we see that $\cos z$ and $\sin z$ are ***periodic with period 2π,*** just as in real. Periodicity of $\tan z$ and $\cot z$ with period π now follows.

Formula (7) points to an essential difference between the real and the complex cosine and sine; whereas $|\cos x| \leqq 1$ and $|\sin x| \leqq 1$, the complex cosine and sine functions are ***no longer bounded*** but approach infinity in absolute value as $y \to \infty$, since then $\sinh y \to \infty$ in (7). ■

EXAMPLE 2 **Solutions of Equations. Zeros of $\cos z$ and $\sin z$**

Solve (a) $\cos z = 5$ (which has no real solution!), (b) $\cos z = 0$, (c) $\sin z = 0$.

Solution. (a) $e^{2iz} - 10e^{iz} + 1 = 0$ from (1) by multiplication by e^{iz}. This is a quadratic equation in e^{iz}, with solutions (rounded off to 3 decimals)

$$e^{iz} = e^{-y+ix} = 5 \pm \sqrt{25 - 1} = 9.899 \quad \text{and} \quad 0.101.$$

Thus $e^{-y} = 9.899$ or 0.101, $e^{ix} = 1$, $y = \pm 2.292$, $x = 2n\pi$. *Ans.* $z = \pm 2n\pi \pm 2.292i$ $(n = 0, 1, 2, \cdots)$.

Can you obtain this from (6a)?

(b) $\cos x = 0$, $\sinh y = 0$ by (7a), $y = 0$. *Ans.* $z = \pm\frac{1}{2}(2n + 1)\pi$ $(n = 0, 1, 2, \cdots)$.

(c) $\sin x = 0$, $\sinh y = 0$ by (7b). *Ans.* $z = \pm n\pi$ $(n = 0, 1, 2, \cdots)$. Hence the only zeros of $\cos z$ and $\sin z$ are those of the real cosine and sine functions. ■

General formulas *for the real trigonometric functions continue to hold for complex values.* This follows immediately from the definitions. We mention in particular the addition rules

$$\text{(9)} \qquad \begin{aligned} \cos(z_1 \pm z_2) &= \cos z_1 \cos z_2 \mp \sin z_1 \sin z_2 \\ \sin(z_1 \pm z_2) &= \sin z_1 \cos z_2 \pm \sin z_2 \cos z_1 \end{aligned}$$

and the formula

$$\text{(10)} \qquad \cos^2 z + \sin^2 z = 1.$$

Some further useful formulas are included in the problem set.

Hyperbolic Functions

The complex **hyperbolic cosine** and **sine** are defined by the formulas

$$\text{(11)} \qquad \cosh z = \tfrac{1}{2}(e^z + e^{-z}), \qquad \sinh z = \tfrac{1}{2}(e^z - e^{-z}).$$

This is suggested by the familiar definitions for a real variable [see (8)]. These functions are entire, with derivatives

$$\text{(12)} \qquad (\cosh z)' = \sinh z, \qquad (\sinh z)' = \cosh z,$$

as in calculus. The other hyperbolic functions are defined by

(13)
$$\tanh z = \frac{\sinh z}{\cosh z}, \qquad \coth z = \frac{\cosh z}{\sinh z},$$
$$\operatorname{sech} z = \frac{1}{\cosh z}, \qquad \operatorname{csch} z = \frac{1}{\sinh z}.$$

Complex Trigonometric and Hyperbolic Functions Are Related. If in (11), we replace z by iz and then use (1), we obtain

(14) $$\cosh iz = \cos z, \qquad \sinh iz = i \sin z.$$

Similarly, if in (1) we replace z by iz and then use (11), we obtain conversely

(15) $$\cos iz = \cosh z, \qquad \sin iz = i \sinh z.$$

Here we have another case of *unrelated* real functions that have *related* complex analogs, pointing again to the advantage of working in complex in order to get both a more unified formalism and a deeper understanding of special functions. This is one of the main reasons for the importance of complex analysis to the engineer and physicist.

PROBLEM SET 13.6

1. Prove that $\cos z$, $\sin z$, $\cosh z$, $\sinh z$ are entire functions.
2. Verify by differentiation that Re $\cos z$ and Im $\sin z$ are harmonic.

3–6 FORMULAS FOR HYPERBOLIC FUNCTIONS

Show that

3. $\cosh z = \cosh x \cos y + i \sinh x \sin y$
   $\sinh z = \sinh x \cos y + i \cosh x \sin y.$

4. $\cosh (z_1 + z_2) = \cosh z_1 \cosh z_2 + \sinh z_1 \sinh z_2$
   $\sinh (z_1 + z_2) = \sinh z_1 \cosh z_2 + \cosh z_1 \sinh z_2.$

5. $\cosh^2 z - \sinh^2 z = 1$
6. $\cosh^2 z + \sinh^2 z = \cosh 2z$

7–15 Function Values. Compute (in the form $u + iv$)

7. $\cos (1 + i)$
8. $\sin (1 + i)$
9. $\sin 5i$, $\cos 5i$
10. $\cos 3\pi i$
11. $\cosh (-2 + 3i)$, $\cos (-3 - 2i)$
12. $-i \sinh (-\pi + 2i)$, $\sin (2 + \pi i)$
13. $\cosh (2n + 1)\pi i$, $n = 1, 2, \cdots$
14. $\sinh (4 - 3i)$
15. $\cosh (4 - 6\pi i)$

16. **(Real and imaginary parts)** Show that
$$\operatorname{Re} \tan z = \frac{\sin x \cos x}{\cos^2 x + \sinh^2 y},$$
$$\operatorname{Im} \tan z = \frac{\sinh y \cosh y}{\cos^2 x + \sinh^2 y}.$$

17–21 Equations. Find all solutions of the following equations.

17. $\cosh z = 0$
18. $\sin z = 100$
19. $\cos z = 2i$
20. $\cosh z = -1$
21. $\sinh z = 0$

22. Find all z for which (a) $\cos z$, (b) $\sin z$ has real values.

23–25 Equations and Inequalities. Using the definitions, prove:

23. $\cos z$ is even, $\cos (-z) = \cos z$, and $\sin z$ is odd, $\sin (-z) = -\sin z$.
24. $|\sinh y| \leqq |\cos z| \leqq \cosh y$, $|\sinh y| \leqq |\sin z| \leqq \cosh y$. Conclude that the complex cosine and sine are not bounded in the whole complex plane.
25. $\sin z_1 \cos z_2 = \frac{1}{2}[\sin (z_1 + z_2) + \sin (z_1 - z_2)]$

13.7 Logarithm. General Power

We finally introduce the *complex logarithm,* which is more complicated than the real logarithm (which it includes as a special case) and historically puzzled mathematicians for some time (so if you first get puzzled—which need not happen!—be patient and work through this section with extra care).

The **natural logarithm** of $z = x + iy$ is denoted by $\ln z$ (sometimes also by $\log z$) and is defined as the inverse of the exponential function; that is, $w = \ln z$ is defined for $z \neq 0$ by the relation

$$e^w = z.$$

(Note that $z = 0$ is impossible, since $e^w \neq 0$ for all w; see Sec. 13.5.) If we set $w = u + iv$ and $z = re^{i\theta}$, this becomes

$$e^w = e^{u+iv} = re^{i\theta}.$$

Now from Sec. 13.5 we know that e^{u+iv} has the absolute value e^u and the argument v. These must be equal to the absolute value and argument on the right:

$$e^u = r, \qquad v = \theta.$$

$e^u = r$ gives $u = \ln r$, where $\ln r$ is the familiar *real* natural logarithm of the positive number $r = |z|$. Hence $w = u + iv = \ln z$ is given by

(1) $$\ln z = \ln r + i\theta \qquad (r = |z| > 0, \quad \theta = \arg z).$$

Now comes an important point (without analog in real calculus). Since the argument of z is determined only up to integer multiples of 2π, ***the complex natural logarithm*** $\ln z$ ***($z \neq 0$) is infinitely many-valued.***

The value of $\ln z$ corresponding to the principal value $\operatorname{Arg} z$ (see Sec. 13.2) is denoted by $\operatorname{Ln} z$ (Ln with capital L) and is called the **principal value** of $\ln z$. Thus

(2) $$\operatorname{Ln} z = \ln |z| + i \operatorname{Arg} z \qquad (z \neq 0).$$

The uniqueness of $\operatorname{Arg} z$ for given z ($\neq 0$) implies that $\operatorname{Ln} z$ is single-valued, that is, a function in the usual sense. Since the other values of $\arg z$ differ by integer multiples of 2π, the other values of $\ln z$ are given by

(3) $$\ln z = \operatorname{Ln} z \pm 2n\pi i \qquad (n = 1, 2, \cdots).$$

They all have the same real part, and their imaginary parts differ by integer multiples of 2π.

If z is positive real, then $\operatorname{Arg} z = 0$, and $\operatorname{Ln} z$ becomes identical with the real natural logarithm known from calculus. If z is negative real (so that the natural logarithm of calculus is not defined!), then $\operatorname{Arg} z = \pi$ and

$$\operatorname{Ln} z = \ln |z| + \pi i \qquad (z \text{ negative real}).$$

From (1) and $e^{\ln r} = r$ for positive real r we obtain

$$e^{\ln z} = z \tag{4a}$$

as expected, but since $\arg(e^z) = y \pm 2n\pi$ is multivalued, so is

$$\ln(e^z) = z \pm 2n\pi i, \qquad n = 0, 1, \cdots. \tag{4b}$$

EXAMPLE 1 **Natural Logarithm. Principal Value**

$$\ln 1 = 0, \pm 2\pi i, \pm 4\pi i, \cdots \qquad \operatorname{Ln} 1 = 0$$

$$\ln 4 = 1.386\,294 \pm 2n\pi i \qquad \operatorname{Ln} 4 = 1.386\,294$$

$$\ln(-1) = \pm\pi i, \pm 3\pi i, \pm 5\pi i, \cdots \qquad \operatorname{Ln}(-1) = \pi i$$

$$\ln(-4) = 1.386\,294 \pm (2n+1)\pi i \qquad \operatorname{Ln}(-4) = 1.386\,294 + \pi i$$

$$\ln i = \pi i/2, -3\pi/2, 5\pi i/2, \cdots \qquad \operatorname{Ln} i = \pi i/2$$

$$\ln 4i = 1.386\,294 + \pi i/2 \pm 2n\pi i \qquad \operatorname{Ln} 4i = 1.386\,294 + \pi i/2$$

$$\ln(-4i) = 1.386\,294 - \pi i/2 \pm 2n\pi i \qquad \operatorname{Ln}(-4i) = 1.386\,294 - \pi i/2$$

$$\ln(3-4i) = \ln 5 + i \arg(3-4i) \qquad \operatorname{Ln}(3-4i) = 1.609\,438 - 0.927\,295i$$

$$= 1.609\,438 - 0.927\,295i \pm 2n\pi i \qquad \text{(Fig. 334)}$$

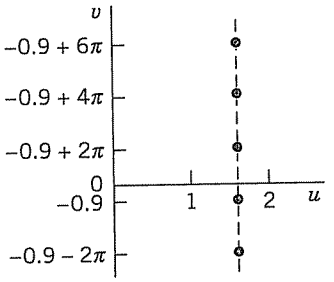


Fig. 334. Some values of $\ln(3-4i)$ in Example 1

The familiar relations for the natural logarithm continue to hold for complex values, that is,

$$\text{(5)} \qquad \text{(a)} \quad \ln(z_1 z_2) = \ln z_1 + \ln z_2, \qquad \text{(b)} \quad \ln(z_1/z_2) = \ln z_1 - \ln z_2$$

but these relations are to be understood in the sense that each value of one side is also contained among the values of the other side; see the next example.

EXAMPLE 2 **Illustration of the Functional Relation (5) in Complex**

Let

$$z_1 = z_2 = e^{\pi i} = -1.$$

If we take the principal values

$$\operatorname{Ln} z_1 = \operatorname{Ln} z_2 = \pi i,$$

then (5a) holds provided we write $\ln(z_1 z_2) = \ln 1 = 2\pi i$; however, it is not true for the principal value, $\operatorname{Ln}(z_1 z_2) = \operatorname{Ln} 1 = 0$.

THEOREM 1

Analyticity of the Logarithm

For every $n = 0, \pm 1, \pm 2, \cdots$ formula (3) defines a function, which is analytic, except at 0 and on the negative real axis, and has the derivative

(6) $$(\ln z)' = \frac{1}{z} \qquad (z \text{ not } 0 \text{ or negative real}).$$

PROOF We show that the Cauchy–Riemann equations are satisfied. From (1)–(3) we have

$$\ln z = \ln r + i(\theta + c) = \frac{1}{2} \ln (x^2 + y^2) + i\left(\arctan \frac{y}{x} + c\right)$$

where the constant c is a multiple of 2π. By differentiation,

$$u_x = \frac{x}{x^2 + y^2} = v_y = \frac{1}{1 + (y/x)^2} \cdot \frac{1}{x}$$

$$u_y = \frac{y}{x^2 + y^2} = -v_x = -\frac{1}{1 + (y/x)^2} \left(-\frac{y}{x^2}\right).$$

Hence the Cauchy–Riemann equations hold. [Confirm this by using these equations in polar form, which we did not use since we proved them only in the problems (to Sec. 13.4).] Formula (4) in Sec. 13.4 now gives (6),

$$(\ln z)' = u_x + iv_x = \frac{x}{x^2 + y^2} + i \frac{1}{1 + (y/x)^2} \left(-\frac{y}{x^2}\right) = \frac{x - iy}{x^2 + y^2} = \frac{1}{z}. \quad \blacksquare$$

Each of the infinitely many functions in (3) is called a **branch** of the logarithm. The negative real axis is known as a **branch cut** and is usually graphed as shown in Fig. 335. The branch for $n = 0$ is called the **principal branch** of $\ln z$.

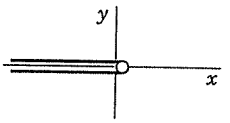


Fig. 335. Branch cut for $\ln z$

General Powers

General powers of a complex number $z = x + iy$ are defined by the formula

(7) $$z^c = e^{c \ln z} \qquad (c \text{ complex}, z \neq 0).$$

Since $\ln z$ is infinitely many-valued, z^c will, in general, be multivalued. The particular value

$$z^c = e^{c \operatorname{Ln} z}$$

is called the **principal value** *of* z^c.

If $c = n = 1, 2, \cdots$, then z^n is single-valued and identical with the usual nth power of z. If $c = -1, -2, \cdots$, the situation is similar.

If $c = 1/n$, where $n = 2, 3, \cdots$, then

$$z^c = \sqrt[n]{z} = e^{(1/n) \ln z} \qquad (z \neq 0),$$

the exponent is determined up to multiples of $2\pi i/n$ and we obtain the n distinct values of the nth root, in agreement with the result in Sec. 13.2. If $c = p/q$, the quotient of two positive integers, the situation is similar, and z^c has only finitely many distinct values. However, if c is real irrational or genuinely complex, then z^c is infinitely many-valued.

EXAMPLE 3 **General Power**

$$i^i = e^{i \ln i} = \exp(i \ln i) = \exp\left[i\left(\frac{\pi}{2} i \pm 2n\pi i\right)\right] = e^{-(\pi/2) \mp 2n\pi}.$$

All these values are real, and the principal value ($n = 0$) is $e^{-\pi/2}$.

Similarly, by direct calculation and multiplying out in the exponent,

$$(1 + i)^{2-i} = \exp\left[(2 - i) \ln (1 + i)\right] = \exp\left[(2 - i)\left\{\ln \sqrt{2} + \tfrac{1}{4}\pi i \pm 2n\pi i\right\}\right]$$
$$= 2e^{\pi/4 \pm 2n\pi}\left[\sin\left(\tfrac{1}{2} \ln 2\right) + i \cos\left(\tfrac{1}{2} \ln 2\right)\right].$$

It is a ***convention*** that for real positive $z = x$ the expression z^c means $e^{c \ln x}$ where $\ln x$ is the elementary real natural logarithm (that is, the principal value $\operatorname{Ln} z$ ($z = x > 0$) in the sense of our definition). Also, if $z = e$, the base of the natural logarithm, $z^c = e^c$ is *conventionally* regarded as the unique value obtained from (1) in Sec. 13.5.

From (7) we see that for any complex number a,

$$(8) \qquad a^z = e^{z \ln a}.$$

We have now introduced the complex functions needed in practical work, some of them (e^z, $\cos z$, $\sin z$, $\cosh z$, $\sinh z$) entire (Sec. 13.5), some of them ($\tan z$, $\cot z$, $\tanh z$, $\coth z$) analytic except at certain points, and one of them ($\ln z$) splitting up into infinitely many functions, each analytic except at 0 and on the negative real axis.

For the **inverse trigonometric** and **hyperbolic functions** see the problem set.

PROBLEM SET 13.7

1–9 **Principal Value Ln z.** Find Ln z when z equals:

1. -10
2. $2 + 2i$
3. $2 - 2i$
4. $-5 \pm 0.1i$
5. $-3 - 4i$
6. -100
7. $0.6 + 0.8i$
8. $-ei$
9. $1 - i$

10–16 **All Values of ln z.** Find all values and graph some of them in the complex plane.

10. $\ln 1$
11. $\ln (-1)$
12. $\ln e$
13. $\ln (-6)$
14. $\ln (4 + 3i)$
15. $\ln (-e^{-i})$
16. $\ln (e^{3i})$
17. Show that the set of values of $\ln (i^2)$ differs from the set of values of $2 \ln i$.

18–21 **Equations.** Solve for z:

18. $\ln z = (2 - \frac{1}{2}i)\pi$
19. $\ln z = 0.3 + 0.7i$
20. $\ln z = e - \pi i$
21. $\ln z = 2 + \frac{1}{4}\pi i$

22–28 **General Powers.** Showing the details of your work, find the principal value of:

22. i^{2i}, $(2i)^i$
23. 4^{3+i}
24. $(1 - i)^{1+i}$
25. $(1 + i)^{1-i}$
26. $(-1)^{1-2i}$
27. $i^{1/2}$
28. $(3 - 4i)^{1/3}$

29. How can you find the answer to Prob. 24 from the answer to Prob. 25?

30. TEAM PROJECT. Inverse Trigonometric and Hyperbolic Functions. By definition, the **inverse sine** $w = \arcsin z$ is the relation such that $\sin w = z$. The **inverse cosine** $w = \arccos z$ is the relation such that $\cos w = z$. The **inverse tangent, inverse cotangent, inverse hyperbolic sine,** etc., are defined and denoted in a similar fashion. (Note that all these relations are ***multivalued.***) Using $\sin w = (e^{iw} - e^{-iw})/(2i)$ and similar representations of $\cos w$, etc., show that

(a) $\arccos z = -i \ln (z + \sqrt{z^2 - 1})$

(b) $\arcsin z = -i \ln (iz + \sqrt{1 - z^2})$

(c) $\operatorname{arccosh} z = \ln (z + \sqrt{z^2 - 1})$

(d) $\operatorname{arcsinh} z = \ln (z + \sqrt{z^2 + 1})$

(e) $\arctan z = \dfrac{i}{2} \ln \dfrac{i + z}{i - z}$

(f) $\operatorname{arctanh} z = \dfrac{1}{2} \ln \dfrac{1 + z}{1 - z}$

(g) Show that $w = \arcsin z$ is infinitely many-valued, and if w_1 is one of these values, the others are of the form $w_1 \pm 2n\pi$ and $\pi - w_1 \pm 2n\pi$, $n = 0, 1, \cdots$. (The *principal value of* $w = u + iv = \arcsin z$ is defined to be the value for which $-\pi/2 \leqq u \leqq \pi/2$ if $v \geqq 0$ and $-\pi/2 < u < \pi/2$ if $v < 0$.)

CHAPTER 13 REVIEW QUESTIONS AND PROBLEMS

1. Add, subtract, multiply, and divide $26 - 7i$ and $3 + 4i$ as well as their complex conjugates.
2. Write the two given numbers in Prob. 1 in polar form. Find the principal value of their arguments.
3. What is the triangle inequality? Its geometric meaning? Its significance?
4. If you know the values of $\sqrt[6]{1}$, how do you get from them the values of $\sqrt[6]{z}$ for any z?
5. State the definition of the derivative from memory. It looks similar to that in calculus. But what is the big difference?
6. What is an analytic function? How would you test for analyticity?
7. Can a function be differentiable at a point without being analytic there? If yes, give an example.
8. Are $|z|$, $\bar{z}$, $\operatorname{Re} z$, $\operatorname{Im} z$ analytic? Give reason.
9. State the definitions of e^z, $\cos z$, $\sin z$, $\cosh z$, $\sinh z$ and the relations between these functions. Do these relations have analogs in real?
10. What properties of e^z are similar to those of e^x ? Which one is different?
11. What is the fundamental region of e^z ? Its significance?
12. What is an entire function? Give examples.
13. Why is $\ln z$ much more complicated than $\ln x$? Explain from memory.
14. What is the principal value of $\ln z$?
15. How is the general power z^c defined? Give examples.

16–21 **Complex Numbers.** Find, in the form $x + iy$, showing the details:

16. $(1 + i)^{12}$
17. $(-2 + 6i)^2$
18. $1/(3 - 7i)$
19. $(1 - i)/(1 + i)^2$
20. $\sqrt{-5 - 12i}$
21. $(43 - 19i)/(8 + i)$

22–26 **Polar Form.** Represent in polar form, with the principal argument:

22. $1 - 3i$
23. $-6 + 6i$
24. $\sqrt{20}/(4 + 2i)$
25. $-12i$
26. $2 + 2i$

27–30 **Roots.** Find and graph all values of

27. $\sqrt{8i}$
28. $\sqrt[4]{256}$
29. $\sqrt[4]{-1}$
30. $\sqrt{32 - 24i}$

31–35 **Analytic Functions.** Find $f(z) = u(x, y) + iv(x, y)$ with u or v as given. Check for analyticity.

31. $u = x/(x^2 + y^2)$
32. $v = e^{-3x} \sin 3y$
33. $u = x^2 - 2xy - y^2$
34. $u = \cos 2x \cosh 2y$
35. $v = e^{x^2 - y^2} \sin 2xy$

36–39 **Harmonic Functions.** Are the following functions harmonic? If so, find a harmonic conjugate.

36. x^2y^2
37. xy
38. $e^{-x/2} \cos \frac{1}{2}y$
39. $x^2 + y^2$

40–45 **Special Function Values.** Find the values of

40. $\sin (3 + 4\pi i)$
41. $\sinh 4\pi i$
42. $\cos (5\pi + 2i)$
43. $\operatorname{Ln} (0.8 + 0.6i)$
44. $\tan (1 + i)$
45. $\cosh (1 + \pi i)$

SUMMARY OF CHAPTER 13
Complex Numbers and Functions

For arithmetic operations with **complex numbers**

(1) $$z = x + iy = re^{i\theta} = r(\cos\theta + i\sin\theta),$$

$r = |z| = \sqrt{x^2 + y^2}$, $\theta = \arctan(y/x)$, and for their representation in the complex plane, see Secs. 13.1 and 13.2.

A complex function $f(z) = u(x, y) + iv(x, y)$ is **analytic** in a domain D if it has a **derivative** (Sec. 13.3)

(2) $$f'(z) = \lim_{\Delta z \to 0} \frac{f(z + \Delta z) - f(z)}{\Delta z}$$

everywhere in D. Also, $f(z)$ is *analytic at a point* $z = z_0$ if it has a derivative in a neighborhood of z_0 (not merely at z_0 itself).

If $f(z)$ is analytic in D, then $u(x, y)$ and $v(x, y)$ satisfy the (very important!) **Cauchy–Riemann equations** (Sec. 13.4)

(3) $$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \qquad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}$$

everywhere in D. Then u and v also satisfy **Laplace's equation**

(4) $$u_{xx} + u_{yy} = 0, \qquad v_{xx} + v_{yy} = 0$$

everywhere in D. If $u(x, y)$ and $v(x, y)$ are continuous and have *continuous* partial derivatives in D that satisfy (3) in D, then $f(z) = u(x, y) + iv(x, y)$ is analytic in D. See Sec. 13.4. (More on Laplace's equation and complex analysis follows in Chap. 18.)

The complex **exponential function** (Sec. 13.5)

(5) $$e^z = \exp z = e^x(\cos y + i\sin y)$$

reduces to e^x if $z = x$ ($y = 0$). It is periodic with $2\pi i$ and has the derivative e^z.

The **trigonometric functions** are (Sec. 13.6)

(6) $$\cos z = \frac{1}{2}(e^{iz} + e^{-iz}) = \cos x \cosh y - i \sin x \sinh y$$

$$\sin z = \frac{1}{2i}(e^{iz} - e^{-iz}) = \sin x \cosh y + i \cos x \sinh y$$

and, furthermore,

$$\tan z = (\sin z)/\cos z, \qquad \cot z = 1/\tan z, \quad \text{etc.}$$

The **hyperbolic functions** are (Sec. 13.6)

(7) $\cosh z = \frac{1}{2}(e^z + e^{-z}) = \cos iz, \qquad \sinh z = \frac{1}{2}(e^z - e^{-z}) = -i \sin iz$

etc. The functions (5)–(7) are **entire,** that is, analytic everywhere in the complex plane.

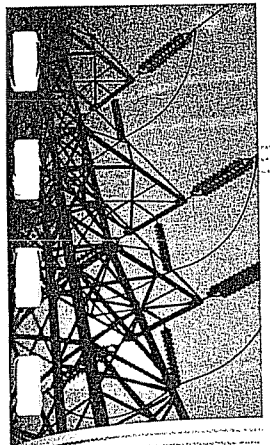
The **natural logarithm** is (Sec. 13.7)

(8) $$\ln z = \ln |z| + i \arg z = \ln |z| + i \operatorname{Arg} z \pm 2n\pi i$$

where $z \neq 0$ and $n = 0, 1, \cdots$. $\operatorname{Arg} z$ is the **principal value** of $\arg z$, that is, $-\pi < \operatorname{Arg} z \leqq \pi$. We see that $\ln z$ is infinitely many-valued. Taking $n = 0$ gives the **principal value** $\operatorname{Ln} z$ of $\ln z$; thus $\operatorname{Ln} z = \ln |z| + i \operatorname{Arg} z$.

General powers are defined by (Sec. 13.7)

(9) $$z^c = e^{c \ln z} \qquad (c \text{ complex}, z \neq 0).$$



CHAPTER 14

Complex Integration

Two main reasons account for the importance of integration in the complex plane. The practical reason is that complex integration can evaluate certain real integrals appearing in applications that are not accessible by real integral calculus. The theoretical reason is that some basic properties of analytic functions are difficult to prove by other methods. A striking property of this type is the existence of higher derivatives of an analytic function.

Complex integration also plays a role in connection with special functions, such as the gamma function (see [GR1], p. 255), the error function, various polynomials (see [GR10]) and others, and the application of these functions in physics.

In this chapter we define and explain complex integrals. The most important result in the chapter is **Cauchy's integral theorem** or the *Cauchy–Goursat theorem,* as it is also called (Sec. 14.2). It implies **Cauchy's integral formula** (Sec. 14.3), which in turn implies the existence of all higher derivatives of an analytic function. Hence in this respect, complex analytic functions behave much more simply than real-valued functions of real variables, which may have derivatives only up to a certain order.

A further method of complex integration, known as integration by residues, and its application to real integrals will need complex series and follows in Chap. 16.

Prerequisite: Chap. 13
References and Answers to Problems: App. 1 Part D, App. 2.

14.1 Line Integral in the Complex Plane

As in calculus we distinguish between definite integrals and indefinite integrals or antiderivatives. An **indefinite integral** is a function whose derivative equals a given analytic function in a region. By inverting known differentiation formulas we may find many types of indefinite integrals.

Complex definite integrals are called (complex) **line integrals.** They are written

$$\int_C f(z)\,dz.$$

Here the **integrand** $f(z)$ is integrated over a given curve C or a portion of it (an *arc,* but we shall say ***"curve"*** in either case, for simplicity). This curve C in the complex plane is called the **path of integration.** We may represent C by a parametric representation

$$z(t) = x(t) + iy(t) \qquad (a \leqq t \leqq b). \tag{1}$$

The sense of increasing t is called the **positive sense** on C, and we say that C is **oriented** by (1).

For instance, $z(t) = t + 3it$ $(0 \leqq t \leqq 2)$ gives a portion (a segment) of the line $y = 3x$. The function $z(t) = 4\cos t + 4i \sin t$ $(-\pi \leqq t \leqq \pi)$ represents the circle $|z| = 4$, and so on. More examples follow below.

We assume C to be a **smooth curve,** that is, C has a continuous and nonzero derivative

$$\dot{z}(t) = \frac{dz}{dt} = \dot{x}(t) + i\dot{y}(t)$$

at each point. Geometrically this means that C has everywhere a continuously turning tangent, as follows directly from the definition

$$\dot{z}(t) = \lim_{\Delta t \to 0} \frac{z(t + \Delta t) - z(t)}{\Delta t} \qquad \text{(Fig. 336).}$$

Here we use a dot since a prime $'$ denotes the derivative with respect to z.

Definition of the Complex Line Integral

This is similar to the method in calculus. Let C be a smooth curve in the complex plane given by (1), and let $f(z)$ be a continuous function given (at least) at each point of C. We now subdivide (we ***"partition"***) the interval $a \leqq t \leqq b$ in (1) by points

$$t_0 (= a), \quad t_1, \quad \cdots, \quad t_{n-1}, \quad t_n (= b)$$

where $t_0 < t_1 < \cdots < t_n$. To this subdivision there corresponds a subdivision of C by points

$$z_0, \quad z_1, \quad \cdots, \quad z_{n-1}, \quad z_n (= Z) \qquad \text{(Fig. 337),}$$

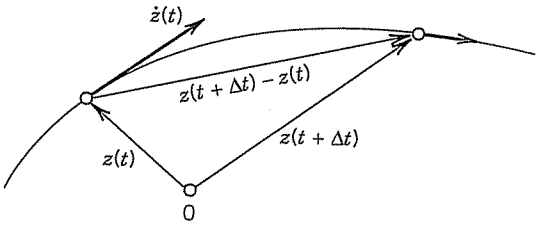


Fig. 336. Tangent vector $\dot{z}(t)$ of a curve C in the complex plane given by $z(t)$. The arrowhead on the curve indicates the *positive sense* (sense of increasing t).

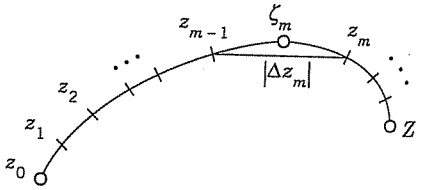


Fig. 337. Complex line integral

where $z_j = z(t_j)$. On each portion of subdivision of C we choose an arbitrary point, say, a point ζ_1 between z_0 and z_1 (that is, $\zeta_1 = z(t)$ where t satisfies $t_0 \leqq t \leqq t_1$), a point ζ_2 between z_1 and z_2, etc. Then we form the sum

$$(2) \qquad S_n = \sum_{m=1}^{n} f(\zeta_m)\, \Delta z_m \qquad \text{where} \qquad \Delta z_m = z_m - z_{m-1}.$$

We do this for each $n = 2, 3, \cdots$ in a completely independent manner, but so that the greatest $|\Delta t_m| = |t_m - t_{m-1}|$ approaches zero as $n \to \infty$. This implies that the greatest

$|\Delta z_m|$ also approaches zero. Indeed, it cannot exceed the length of the arc of C from z_{m-1} to z_m and the latter goes to zero since the arc length of the smooth curve C is a continuous function of t. The limit of the sequence of complex numbers $S_2, S_3, \cdots$ thus obtained is called the **line integral** (or simply the *integral*) of $f(z)$ over the path of integration C with the orientation given by (1). This line integral is denoted by

$$\text{(3)} \qquad \int_C f(z)\,dz, \qquad \text{or by} \qquad \oint_C f(z)\,dz$$

if C is a **closed path** (one whose terminal point Z coincides with its initial point z_0, as for a circle or for a curve shaped like an 8).

General Assumption. *All paths of integration for complex line integrals are assumed to be* **piecewise smooth,** *that is, they consist of finitely many smooth curves joined end to end.*

Basic Properties Directly Implied by the Definition

1. **Linearity.** Integration is a **linear operation,** that is, we can integrate sums term by term and can take out constant factors from under the integral sign. This means that if the integrals of f_1 and f_2 over a path C exist, so does the integral of $k_1f_1 + k_2f_2$ over the same path and

$$\text{(4)} \qquad \int_C [k_1f_1(z) + k_2f_2(z)]\,dz = k_1 \int_C f_1(z)\,dz + k_2 \int_C f_2(z)\,dz.$$

2. **Sense reversal** in integrating over the ***same*** path, from z_0 to Z (left) and from Z to z_0 (right), introduces a minus sign as shown,

$$\text{(5)} \qquad \int_{z_0}^{Z} f(z)\,dz = -\int_{Z}^{z_0} f(z)\,dz.$$

3. **Partitioning of path** (see Fig. 338)

$$\text{(6)} \qquad \int_C f(z)\,dz = \int_{C_1} f(z)\,dz + \int_{C_2} f(z)\,dz.$$

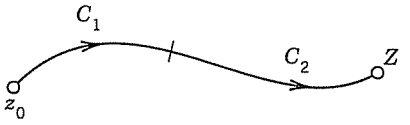


Fig. 338. Partitioning of path [formula (6)]

Existence of the Complex Line Integral

Our assumptions that $f(z)$ is continuous and C is piecewise smooth imply the existence of the line integral (3). This can be seen as follows.

As in the preceding chapter let us write $f(z) = u(x, y) + iv(x, y)$. We also set

$$\zeta_m = \xi_m + i\eta_m \qquad \text{and} \qquad \Delta z_m = \Delta x_m + i\Delta y_m.$$

Then (2) may be written

$$S_n = \sum (u + iv)(\Delta x_m + i\Delta y_m) \tag{7}$$

where $u = u(\zeta_m, \eta_m)$, $v = v(\zeta_m, \eta_m)$ and we sum over m from 1 to n. Performing the multiplication, we may now split up S_n into four sums:

$$S_n = \sum u\,\Delta x_m - \sum v\,\Delta y_m + i\left[\sum u\,\Delta y_m + \sum v\,\Delta x_m\right].$$

These sums are real. Since f is continuous, u and v are continuous. Hence, if we let n approach infinity in the aforementioned way, then the greatest Δx_m and Δy_m will approach zero and each sum on the right becomes a real line integral:

$$\lim_{n\to\infty} S_n = \int_C f(z)\,dz = \int_C u\,dx - \int_C v\,dy + i\left[\int_C u\,dy + \int_C v\,dx\right]. \tag{8}$$

This shows that under our assumptions on f and C the line integral (3) exists and its value is independent of the choice of subdivisions and intermediate points ζ_m. ∎

First Evaluation Method: Indefinite Integration and Substitution of Limits

This method is the analog of the evaluation of definite integrals in calculus by the well-known formula

$$\int_a^b f(x)\,dx = F(b) - F(a) \qquad [F'(x) = f(x)].$$

It is simpler than the next method, but it is suitable for analytic functions only. To formulate it, we need the following concept of general interest.

A domain D is called **simply connected** if every **simple closed curve** (closed curve without self-intersections) encloses only points of D.

For instance, a circular disk is simply connected, whereas an annulus (Sec. 13.3) is not simply connected. (Explain!)

THEOREM 1

Indefinite Integration of Analytic Functions

Let $f(z)$ be analytic in a simply connected domain D. Then there exists an indefinite integral of $f(z)$ in the domain D, that is, an analytic function $F(z)$ such that $F'(z) = f(z)$ in D, and for all paths in D joining two points z_0 and z_1 in D we have

$$\int_{z_0}^{z_1} f(z)\,dz = F(z_1) - F(z_0) \qquad [F'(z) = f(z)]. \tag{9}$$

(Note that we can write z_0 and z_1 instead of C, since we get the same value for all those C from z_0 to z_1.)

This theorem will be proved in the next section.

Simple connectedness is quite essential in Theorem 1, as we shall see in Example 5.

Since analytic functions are our main concern, and since differentiation formulas will often help in finding $F(z)$ for a given $f(z) = F'(z)$, the present method is of great practical interest.

If $f(z)$ is entire (Sec. 13.5), we can take for D the complex plane (which is certainly simply connected).

EXAMPLE 1 $$\int_0^{1+i} z^2\,dz = \frac{1}{3}\,z^3\Big|_0^{1+i} = \frac{1}{3}\,(1+i)^3 = -\frac{2}{3} + \frac{2}{3}\,i$$

EXAMPLE 2 $$\int_{-\pi i}^{\pi i} \cos z\,dz = \sin z\Big|_{-\pi i}^{\pi i} = 2\sin \pi i = 2i \sinh \pi = 23.097i$$

EXAMPLE 3 $$\int_{8+\pi i}^{8-3\pi i} e^{z/2}\,dz = 2e^{z/2}\Big|_{8+\pi i}^{8-3\pi i} = 2(e^{4-3\pi i/2} - e^{4+\pi i/2}) = 0$$

since e^z is periodic with period $2\pi i$.

EXAMPLE 4 $\displaystyle\int_{-i}^{i} \frac{dz}{z} = \operatorname{Ln} i - \operatorname{Ln}(-i) = \frac{i\pi}{2} - \left(-\frac{i\pi}{2}\right) = i\pi$. Here D is the complex plane without 0 and the negative real axis (where Ln z is not analytic). Obviously, D is a simply connected domain.

Second Evaluation Method: Use of a Representation of a Path

This method is not restricted to analytic functions but applies to any continuous complex function.

THEOREM 2

Integration by the Use of the Path

Let C be a piecewise smooth path, represented by $z = z(t)$, where $a \leqq t \leqq b$. Let $f(z)$ be a continuous function on C. Then

(10) $$\int_C f(z)\,dz = \int_a^b f[z(t)]\dot{z}(t)\,dt \qquad \left(\dot{z} = \frac{dz}{dt}\right).$$

PROOF The left side of (10) is given by (8) in terms of real line integrals, and we show that the right side of (10) also equals (8). We have $z = x + iy$, hence $\dot{z} = \dot{x} + i\dot{y}$. We simply write u for $u[x(t), y(t)]$ and v for $v[x(t), y(t)]$. We also have $dx = \dot{x}\,dt$ *and* $dy = \dot{y}\,dt$. Consequently, in (10)

$$\int_a^b f[z(t)]\dot{z}(t)\,dt = \int_a^b (u + iv)(\dot{x} + i\dot{y})\,dt$$

$$= \int_C [u\,dx - v\,dy + i(u\,dy + v\,dx)]$$

$$= \int_C (u\,dx - v\,dy) + i\int_C (u\,dy + v\,dx).$$

COMMENT. In (7) and (8) of the existence proof of the complex line integral we referred to real line integrals. If one wants to avoid this, one can take (10) as a *definition* of the complex line integral.

Steps in Applying Theorem 2

(A) Represent the path C in the form $z(t)$ $(a \leqq t \leqq b)$.

(B) Calculate the derivative $\dot{z}(t) = dz/dt$.

(C) Substitute $z(t)$ for every z in $f(z)$ (hence $x(t)$ for x and $y(t)$ for y).

(D) Integrate $f[z(t)]\dot{z}(t)$ over t from a to b.

EXAMPLE 5 **A Basic Result: Integral of $1/z$ Around the Unit Circle**

We show that by integrating $1/z$ counterclockwise around the unit circle (the circle of radius 1 and center 0; see Sec. 13.3) we obtain

$$\text{(11)} \qquad \oint_C \frac{dz}{z} = 2\pi i \qquad (C \text{ the unit circle, counterclockwise}).$$

This is a very important result that we shall need quite often.

Solution. **(A)** We may represent the unit circle C in Fig. 327 of Sec. 13.3 by

$$z(t) = \cos t + i \sin t = e^{it} \qquad (0 \leqq t \leqq 2\pi),$$

so that counterclockwise integration corresponds to an increase of t from 0 to 2π.

(B) Differentiation gives $\dot{z}(t) = ie^{it}$ (chain rule!).

(C) By substitution, $f(z(t)) = 1/z(t) = e^{-it}$.

(D) From (10) we thus obtain the result

$$\oint_C \frac{dz}{z} = \int_0^{2\pi} e^{-it} ie^{it}\, dt = i \int_0^{2\pi} dt = 2\pi i.$$

Check this result by using $z(t) = \cos t + i \sin t$.

Simple connectedness is essential in Theorem 1. Equation (9) in Theorem 1 gives 0 for any closed path because then $z_1 = z_0$, so that $F(z_1) - F(z_0) = 0$. Now $1/z$ is not analytic at $z = 0$. But any *simply connected* domain containing the unit circle must contain $z = 0$, so that Theorem 1 does not apply—it is not enough that $1/z$ is analytic in an annulus, say, $\frac{1}{2} < |z| < \frac{3}{2}$, because an annulus is not simply connected! ∎

EXAMPLE 6 **Integral of $1/z^m$ with Integer Power m**

Let $f(z) = (z - z_0)^m$ where m is the integer and z_0 a constant. Integrate counterclockwise around the circle C of radius ρ with center at z_0 (Fig. 339).

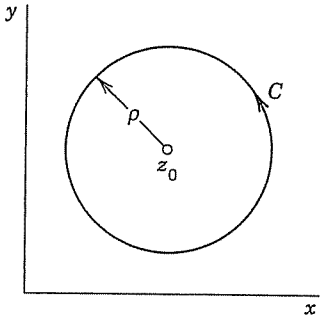


Fig. 339. Path in Example 6

Solution. We may represent C in the form

$$z(t) = z_0 + \rho(\cos t + i \sin t) = z_0 + \rho e^{it} \qquad (0 \leqq t \leqq 2\pi).$$

Then we have

$$(z - z_0)^m = \rho^m e^{imt}, \qquad dz = i\rho e^{it}\, dt$$

and obtain

$$\oint_C (z - z_0)^m\, dz = \int_0^{2\pi} \rho^m e^{imt}\, i\rho e^{it}\, dt = i\rho^{m+1} \int_0^{2\pi} e^{i(m+1)t}\, dt.$$

By the Euler formula (5) in Sec. 13.6 the right side equals

$$i\rho^{m+1} \left[\int_0^{2\pi} \cos (m + 1)t\, dt + i \int_0^{2\pi} \sin (m + 1)t\, dt \right].$$

If $m = -1$, we have $\rho^{m+1} = 1$, $\cos 0 = 1$, $\sin 0 = 0$. We thus obtain $2\pi i$. For integer $m \neq 1$ each of the two integrals is zero because we integrate over an interval of length 2π, equal to a period of sine and cosine. Hence the result is

(12)
$$\oint_C (z - z_0)^m\, dz = \begin{cases} 2\pi i & (m = -1), \\ 0 & (m \neq -1 \text{ and integer}). \end{cases}$$

Dependence on path. Now comes a very important fact. If we integrate a given function $f(z)$ from a point z_0 to a point z_1 along different paths, the integrals will in general have different values. In other words, ***a complex line integral depends not only on the endpoints of the path but in general also on the path itself.*** The next example gives a first impression of this, and a systematic discussion follows in the next section.

EXAMPLE 7 **Integral of a Nonanalytic Function. Dependence on Path**

Integrate $f(z) = \operatorname{Re} z = x$ from 0 to $1 + 2i$ (a) along C^* in Fig. 340, (b) along C consisting of C_1 and C_2.

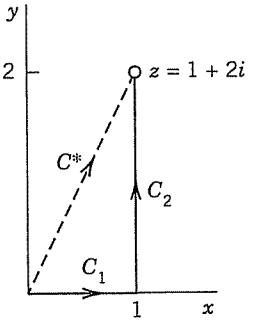


Fig. 340. Paths in Example 7

Solution. (a) C^* can be represented by $z(t) = t + 2it$ $(0 \leqq t \leqq 1)$. Hence $\dot{z}(t) = 1 + 2i$ and $f[z(t)] = x(t) = t$ on C^*. We now calculate

$$\int_{C^*} \operatorname{Re} z\, dz = \int_0^1 t(1 + 2i)\, dt = \tfrac{1}{2}(1 + 2i) = \tfrac{1}{2} + i.$$

(b) We now have

$$C_1\text{: } z(t) = t, \qquad \dot{z}(t) = 1, \qquad f(z(t)) = x(t) = t \qquad (0 \leqq t \leqq 1)$$

$$C_2\text{: } z(t) = 1 + it, \qquad \dot{z}(t) = i, \qquad f(z(t)) = x(t) = 1 \qquad (0 \leqq t \leqq 2).$$

Using (6) we calculate

$$\int_C \operatorname{Re} z\, dz = \int_{C_1} \operatorname{Re} z\, dz + \int_{C_2} \operatorname{Re} z\, dz = \int_0^1 t\, dt + \int_0^2 1 \cdot i\, dt = \tfrac{1}{2} + 2i.$$

Note that this result differs from the result in (a). ■

Bounds for Integrals. *ML*-Inequality

There will be a frequent need for estimating the absolute value of complex line integrals. The basic formula is

(13) $$\left|\int_C f(z)\, dz\right| \leqq ML \qquad \textbf{(}\textit{ML}\textbf{-inequality);}$$

L is the length of C and M a constant such that $|f(z)| \leqq M$ everywhere on C.

PROOF Taking the absolute value in (2) and applying the generalized inequality (6*) in Sec. 13.2, we obtain

$$|S_n| = \left|\sum_{m=1}^{n} f(\zeta_m)\, \Delta z_m\right| \leqq \sum_{m=1}^{n} |f(\zeta_m)|\, |\Delta z_m| \leqq M \sum_{m=1}^{n} |\Delta z_m|.$$

Now $|\Delta z_m|$ is the length of the chord whose endpoints are z_{m-1} and z_m (see Fig. 337 on p. 638). Hence the sum on the right represents the length L^* of the broken line of chords whose endpoints are $z_0, z_1, \cdots, z_n\ (= Z)$. If n approaches infinity in such a way that the greatest $|\Delta t_m|$ and thus $|\Delta z_m|$ approach zero, then L^* approaches the length L of the curve C, by the definition of the length of a curve. From this the inequality (13) follows. ■

We cannot see from (13) how close to the bound ML the actual absolute value of the integral is, but this will be no handicap in applying (13). For the time being we explain the practical use of (13) by a simple example.

EXAMPLE 8 **Estimation of an Integral**

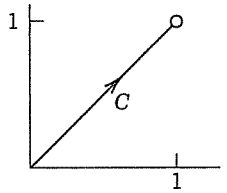


Fig. 341. Path in Example 8

Find an upper bound for the absolute value of the integral

$$\int_C z^2\, dz, \qquad C \text{ the straight-line segment from 0 to } 1 + i, \text{ Fig. 341.}$$

Solution. $L = \sqrt{2}$ and $|f(z)| = |z^2| \leqq 2$ on C gives by (13)

$$\left|\int_C z^2\, dz\right| \leqq 2\sqrt{2} = 2.8284.$$

The absolute value of the integral is $\left|-\frac{2}{3} + \frac{2}{3}\, i\right| = \frac{2}{3}\sqrt{2} = 0.9428$ (see Example 1). ■

Summary on Integration. Line integrals of $f(z)$ can always be evaluated by (10), using a representation (1) of the path of integration. If $f(z)$ is analytic, indefinite integration by (9) as in calculus will be simpler.

PROBLEM SET 14.1

1–9 PARAMETRIC REPRESENTATIONS

Find and sketch the path and its orientation given by:

1. $z(t) = (1 + 3i)t$ $(1 \leqq t \leqq 4)$
2. $z(t) = 5 - 2it$ $(-3 \leqq t \leqq 3)$
3. $z(t) = 4 + i + 3e^{it}$ $(0 \leqq t \leqq 2\pi)$
4. $z(t) = 1 + i + e^{-\pi it}$ $(0 \leqq t \leqq 2)$
5. $z(t) = e^{it}$ $(0 \leqq t \leqq \pi)$
6. $z(t) = 3 + 4i + 5e^{it}$ $(\pi \leqq t \leqq 2\pi)$
7. $z(t) = 6\cos 2t + 5i \sin 2t$ $(0 \leqq t \leqq \pi)$
8. $z(t) = 1 + 2t + 8it^2$ $(-1 \leqq t \leqq 1)$
9. $z(t) = t + \frac{1}{2}it^3$ $(-1 \leqq t \leqq 2)$

10–18 PARAMETRIC REPRESENTATIONS

Sketch and represent parametrically:

10. Segment from $1 + i$ to $4 - 2i$
11. Unit circle (clockwise)
12. Segment from $a + ib$ to $c + id$
13. Hyperbola $xy = 1$ from $1 + i$ to $4 + \frac{1}{4}i$
14. Semi-ellipse $x^2/a^2 + y^2/b^2 = 1$, $y \geqq 0$
15. Parabola $y = 4 - 4x^2$ $(-1 \leqq x \leqq 1)$
16. $|z - 2 + 3i| = 4$ (counterclockwise)
17. $|z + a + ib| = r$ (clockwise)
18. Ellipse $4(x - 1)^2 + 9(y + 2)^2 = 36$

19–29 INTEGRATION

Integrate by the first method or state why it does not apply and then use the second method. (Show the details of your work.)

19. $\int_C \operatorname{Re} z \, dz$, C the shortest path from 0 to $1 + i$
20. $\int_C \operatorname{Re} z \, dz$, C the parabola $y = x^2$ from 0 to $1 + i$
21. $\int_C e^{2z} \, dz$, C the shortest path from πi to $2\pi i$
22. $\int_C \sin z \, dz$, C any path from 0 to $2i$
23. $\int_C \cos^2 z \, dz$ from $-\pi i$ along $|z| = \pi$ to πi in the right half-plane
24. $\int_C (z + z^{-1}) \, dz$, C the unit circle (counterclockwise)
25. $\int_C \cosh 4z \, dz$, C any path from $-\pi i/8$ to $\pi i/8$
26. $\int_C \bar{z} \, dz$, C from $-1 + i$ along the parabola $y = x^2$ to $1 + i$
27. $\int_C \sec^2 z \, dz$, C any path from $\pi/4$ to $\pi i/4$
28. $\int_C \operatorname{Im} z^2 \, dz$ counterclockwise around the triangle with vertices $z = 0, 1, i$
29. $\int_C z e^{z^2/2} \, dz$, C from i along the axes to 1

30. **(Sense reversal)** Verify (5) for $f(z) = z^2$, where C is the segment from $-1 - i$ to $1 + i$.
31. **(Path partitioning)** Verify (6) for $f(z) = 1/z$ and C_1 and C_2 the upper and lower halfs of the unit circle
32. **(*ML*-inequality)** Find an upper bound of the absolute value of the integral in Prob. 19.
33. **(Linearity)** Illustrate (4) with an example of your own. Prove (4).
34. **TEAM PROJECT. Integration. (a) Comparison.** Write a short report comparing the essential points of the two integration methods.

    (b) Comparison. Evaluate $\int_C f(z) \, dz$ by Theorem 1 and check the result by Theorem 2, where:

    (i) $f(z) = z^4$ and C is the semicircle $|z| = 2$ from $-2i$ to $2i$ in the right half-plane,

    (ii) $f(z) = e^{2z}$ and C is the shortest path from 0 to $1 + 2i$.

    (c) Continuous deformation of path. Experiment with a family of paths with common endpoints, say, $z(t) = t + ia \sin t$, $0 \leqq t \leqq \pi$, with real parameter a. Integrate nonanalytic functions ($\operatorname{Re} z$, $\operatorname{Re}(z^2)$, etc.) and explore how the result depends on a. Then take analytic functions of your choice. (Show the details of your work.) Compare and comment.

    (d) Continuous deformation of path. Choose another family, for example, semi-ellipses $z(t) = a\cos t + i \sin t$, $-\pi/2 \leqq t \leqq \pi/2$, and experiment as in (c).
35. **CAS PROJECT. Integration.** Write programs for the two integration methods. Apply them to problems of your choice. Could you make them into a joint program that also decides which of the two methods to use in a given case?

14.2 Cauchy's Integral Theorem

We have just seen in Sec. 14.1 that a line integral of a function $f(z)$ generally depends not merely on the endpoints of the path, but also on the choice of the path itself. This dependence often complicates situations. Hence conditions under which this does ***not*** occur are of considerable importance. Namely, if $f(z)$ is analytic in a domain D and D is simply connected (see Sec. 14.1 and also below), then the integral will not depend on the choice of a path between given points. This result (Theorem 2) follows from Cauchy's integral theorem, along with other basic consequences that make ***Cauchy's integral theorem the most important theorem in this chapter*** and fundamental throughout complex analysis.

Let us begin by repeating and illustrating the definition of simple connectedness (Sec. 14.1) and adding some more details.

1. A **simple closed path** is a closed path (Sec. 14.1) that does not intersect or touch itself (Fig. 342). For example, a circle is simple, but a curve shaped like an 8 is not simple.

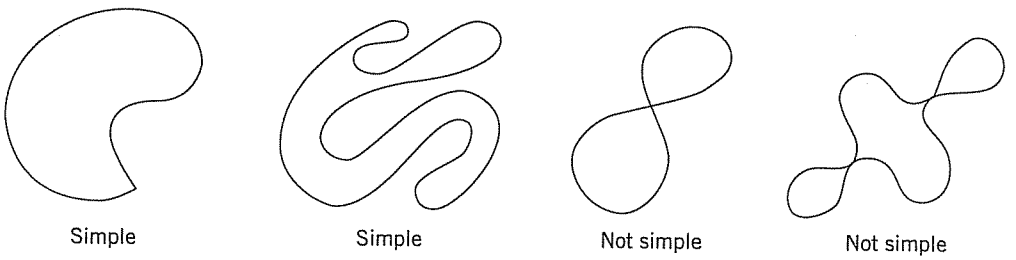


Fig. 342. Closed paths

2. A **simply connected domain** D in the complex plane is a domain (Sec. 13.3) such that every simple closed path in D encloses only points of D. *Examples:* The interior of a circle ("open disk"), ellipse, or any simple closed curve. A domain that is not simply connected is called **multiply connected.** *Examples:* An annulus (Sec. 13.3), a disk without the center, for example, $0 < |z| < 1$. See also Fig. 343.

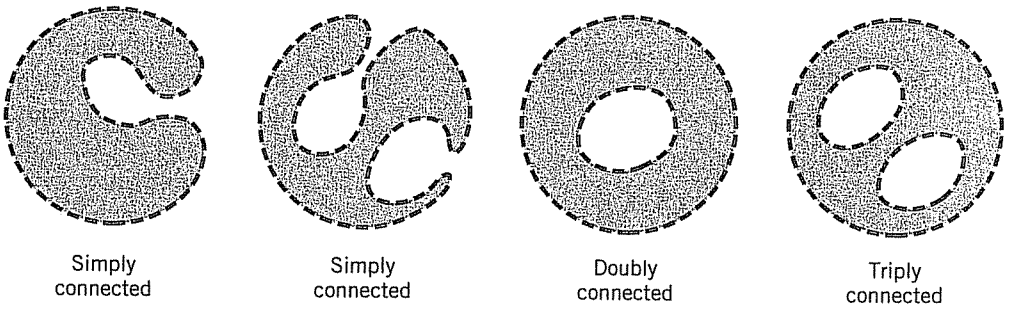


Fig. 343. Simply and multiply connected domains

More precisely, a **bounded domain** D (that is, a domain that lies entirely in some circle about the origin) is called ***p*-fold connected** if its boundary consists of p closed connected sets without common points. These sets can be curves, segments, or single points (such as $z = 0$ for $0 < |z| < 1$, for which $p = 2$). Thus, D has $p - 1$ **"holes",** where "hole" may also mean a segment or even a single point. Hence an annulus is doubly connected ($p = 2$).

THEOREM 1

Cauchy's Integral Theorem

If $f(z)$ is analytic in a simply connected domain D, then for every simple closed path C in D,

$$\oint_C f(z)\,dz = 0. \tag{1}$$

See Fig. 344.

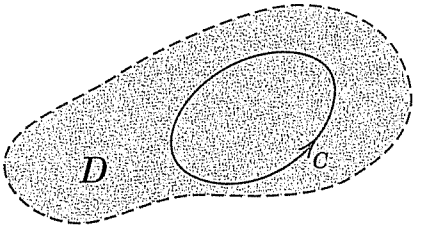


Fig. 344. Cauchy's integral theorem

Before we prove the theorem, let us consider some examples in order to really understand what is going on. A simple closed path is sometimes called a *contour* and an integral over such a path a **contour integral.** Thus, (1) and our examples involve contour integrals.

EXAMPLE 1 **No Singularities (Entire Functions)**

$$\oint_C e^z\,dz = 0, \qquad \oint_C \cos z\,dz = 0, \qquad \oint_C z^n\,dz = 0 \qquad (n = 0, 1, \cdots)$$

for any closed path, since these functions are entire (analytic for all z). ■

EXAMPLE 2 **Singularities Outside the Contour**

$$\oint_C \sec z\,dz = 0, \qquad \oint_C \frac{dz}{z^2 + 4} = 0$$

where C is the unit circle, $\sec z = 1/\cos z$ is not analytic at $z = \pm\pi/2, \pm 3\pi/2, \cdots$, but all these points lie outside C; none lies on C or inside C. Similarly for the second integral, whose integrand is not analytic at $z = \pm 2i$ outside C. ■

EXAMPLE 3 **Nonanalytic Function**

$$\oint_C \bar{z}\,dz = \int_0^{2\pi} e^{-it} ie^{it}\,dt = 2\pi i$$

where C: $z(t) = e^{it}$ is the unit circle. This does not contradict Cauchy's theorem because $f(z) = \bar{z}$ is not analytic. ■

EXAMPLE 4 **Analyticity Sufficient, Not Necessary**

$$\oint_C \frac{dz}{z^2} = 0$$

where C is the unit circle. This result does *not* follow from Cauchy's theorem, because $f(z) = 1/z^2$ is not analytic at $z = 0$. Hence *the condition that f be analytic in D is sufficient rather than necessary for* (1) *to be true.* ■

EXAMPLE 5 Simple Connectedness Essential

$$\oint_C \frac{dz}{z} = 2\pi i$$

for counterclockwise integration around the unit circle (see Sec. 14.1). C lies in the annulus $\frac{1}{2} < |z| < \frac{3}{2}$ where $1/z$ is analytic, but this domain is not simply connected, so that Cauchy's theorem cannot be applied. Hence ***the condition that the domain D be simply connected is essential.***

In other words, by Cauchy's theorem, if $f(z)$ is analytic on a simple closed path C and everywhere inside C, with no exception, not even a single point, then (1) holds. The point that causes trouble here is $z = 0$ where $1/z$ is not analytic. ■

PROOF Cauchy proved his integral theorem under the additional assumption that the derivative $f'(z)$ is continuous (which is true, but would need an extra proof). His proof proceeds as follows. From (8) in Sec. 14.1 we have

$$\oint_C f(z)\,dz = \oint_C (u\,dx - v\,dy) + i \oint_C (u\,dy + v\,dx).$$

Since $f(z)$ is analytic in D, its derivative $f'(z)$ exists in D. Since $f'(z)$ is assumed to be continuous, (4) and (5) in Sec. 13.4 imply that u and v have *continuous* partial derivatives in D. Hence Green's theorem (Sec. 10.4) (with u and $-v$ instead of F_1 and F_2) is applicable and gives

$$\oint_C (u\,dx - v\,dy) = \iint_R \left(-\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} \right) dx\,dy$$

where R is the region bounded by C. The second Cauchy–Riemann equation (Sec. 13.4) shows that the integrand on the right is identically zero. Hence the integral on the left is zero. In the same fashion it follows by the use of the first Cauchy–Riemann equation that the last integral in the above formula is zero. This completes Cauchy's proof. ■

Goursat's proof *without the condition that $f'(z)$ is continuous*[1] is much more complicated. We leave it optional and include it in App. 4.

Independence of Path

We know from the preceding section that the value of a line integral of a given function $f(z)$ from a point z_1 to a point z_2 will in general depend on the path C over which we integrate, not merely on z_1 and z_2. It is important to characterize situations in which this difficulty of path dependence does not occur. This task suggests the following concept. We call an integral of $f(z)$ **independent of path in a domain D** if for every z_1, z_2 in D its value depends (besides on $f(z)$, of course) only on the initial point z_1 and the terminal point z_2, but not on the choice of the path C in D [so that every path in D from z_1 to z_2 gives the same value of the integral of $f(z)$].

[1] ÉDOUARD GOURSAT (1858–1936), French mathematician. Cauchy published the theorem in 1825. The removal of that condition by Goursat (see *Transactions Amer. Math. Soc.*, vol. 1, 1900) is quite important, for instance, in connection with the fact that derivatives of analytic functions are also analytic, as we shall prove soon. Goursat also made important contributions to PDEs.

THEOREM 2

Independence of Path

If $f(z)$ is analytic in a simply connected domain D, then the integral of $f(z)$ is independent of path in D.

PROOF Let z_1 and z_2 be any points in D. Consider two paths C_1 and C_2 in D from z_1 to z_2 without further common points, as in Fig. 345. Denote by C_2^* the path C_2 with the orientation reserved (Fig. 346). Integrate from z_1 over C_1 to z_2 and over C_2^* back to z_1. This is a simple closed path, and Cauchy's theorem applies under our assumptions of the present theorem and gives zero:

$$(2') \qquad \int_{C_1} f\,dz + \int_{C_2^*} f\,dz = 0, \qquad \text{thus} \qquad \int_{C_1} f\,dz = -\int_{C_2^*} f\,dz.$$

But the minus sign on the right disappears if we integrate in the reverse direction, from z_1 to z_2, which shows that the integrals of $f(z)$ over C_1 and C_2 are equal,

$$(2) \qquad \int_{C_1} f(z)\,dz = \int_{C_2} f(z)\,dz \qquad \text{(Fig. 345).}$$

This proves the theorem for paths that have only the endpoints in common. For paths that have finitely many further common points, apply the present argument to each "loop" (portions of C_1 and C_2 between consecutive common points; four loops in Fig. 347). For paths with infinitely many common points we would need additional argumentation not to be presented here. ∎

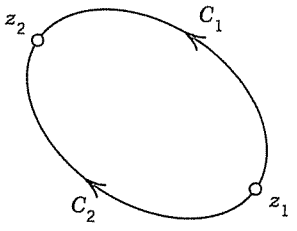


Fig. 345. Formula (2)

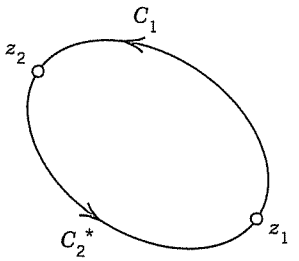


Fig. 346. Formula (2′)

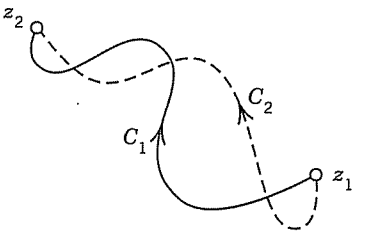


Fig. 347. Paths with more common points

Principle of Deformation of Path

This idea is related to path independence. We may imagine that the path C_2 in (2) was obtained from C_1 by continuously moving C_1 (with ends fixed!) until it coincides with C_2. Figure 348 shows two of the infinitely many intermediate paths for which the integral always retains its value (because of Theorem 2). Hence we may impose a continuous deformation of the path of an integral, keeping the ends fixed. As long as our deforming path always contains only points at which $f(z)$ is analytic, the integral retains the same value. This is called the **principle of deformation of path.**

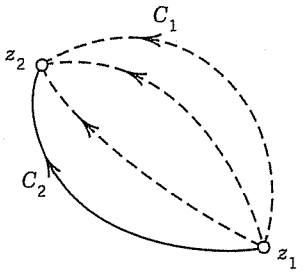


Fig. 348. Continuous deformation of path

EXAMPLE 6 **A Basic Result: Integral of Integer Powers**

From Example 6 in Sec. 14.1 and the principle of deformation of path it follows that

$$\oint (z - z_0)^m \, dz = \begin{cases} 2\pi i & (m = -1) \\ 0 & (m \neq -1 \text{ and integer}) \end{cases} \tag{3}$$

for counterclockwise integration around ***any simple closed path containing z_0 in its interior.***

Indeed, the circle $|z - z_0| = \rho$ in Example 6 of Sec. 14.1 can be continuously deformed in two steps into a path as just indicated, namely, by first deforming, say, one semicircle and then the other one. (Make a sketch). ∎

Existence of Indefinite Integral

We shall now justify our indefinite integration method in the preceding section [formula (9) in Sec. 14.1]. The proof will need Cauchy's integral theorem.

THEOREM 3

Existence of Indefinite Integral

If $f(z)$ is analytic in a simply connected domain D, then there exists an indefinite integral $F(z)$ of $f(z)$ in D—thus, $F'(z) = f(z)$—which is analytic in D, and for all paths in D joining any two points z_0 and z_1 in D, the integral of $f(z)$ from z_0 to z_1 can be evaluated by formula (9) *in* Sec. 14.1.

PROOF The conditions of Cauchy's integral theorem are satisfied. Hence the line integral of $f(z)$ from any z_0 in D to any z in D is independent of path in D. We keep z_0 fixed. Then this integral becomes a function of z, call if $F(z)$,

$$F(z) = \int_{z_0}^{z} f(z^*) \, dz^* \tag{4}$$

which is uniquely determined. We show that this $F(z)$ is analytic in D and $F'(z) = f(z)$. The idea of doing this is as follows. Using (4) we form the difference quotient

$$\frac{F(z + \Delta z) - F(z)}{\Delta z} = \frac{1}{\Delta z} \left[\int_{z_0}^{z+\Delta z} f(z^*) \, dz^* - \int_{z_0}^{z} f(z^*) \, dz^* \right] = \frac{1}{\Delta z} \int_{z}^{z+\Delta z} f(z^*) \, dz^*. \tag{5}$$

We now subtract $f(z)$ from (5) and show that the resulting expression approaches zero as $\Delta z \to 0$. The details are as follows.

We keep z fixed. Then we choose $z + \Delta z$ in D so that the whole segment with endpoints z and $z + \Delta z$ is in D (Fig. 349). This can be done because D is a domain, hence it contains a neighborhood of z. We use this segment as the path of integration in (5). Now we subtract $f(z)$. This is a constant because z is kept fixed. Hence we can write

$$\int_z^{z+\Delta z} f(z)\, dz^* = f(z) \int_z^{z+\Delta z} dz^* = f(z)\, \Delta z. \qquad \text{Thus} \qquad f(z) = \frac{1}{\Delta z} \int_z^{z+\Delta z} f(z)\, dz^*.$$

By this trick and from (5) we get a single integral:

$$\frac{F(z + \Delta z) - F(z)}{\Delta z} - f(z) = \frac{1}{\Delta z} \int_z^{z+\Delta z} [f(z^*) - f(z)]\, dz^*.$$

Since $f(z)$ is analytic, it is continuous. An $\epsilon > 0$ being given, we can thus find a $\delta > 0$ such that $|f(z^*) - f(z)| < \epsilon$ when $|z^* - z| < \delta$. Hence, letting $|\Delta z| < \delta$, we see that the *ML*-inequality (Sec. 14.1) yields

$$\left| \frac{F(z + \Delta z) - F(z)}{\Delta z} - f(z) \right| = \frac{1}{|\Delta z|} \left| \int_z^{z+\Delta z} [f(z^*) - f(z)]\, dz^* \right| \leqq \frac{1}{|\Delta z|}\, \epsilon |\Delta z| = \epsilon.$$

By the definition of limit and derivative, this proves that

$$F'(z) = \lim_{\Delta z \to 0} \frac{F(z + \Delta z) - F(z)}{\Delta z} = f(z).$$

Since z is any point in D, this implies that $F(z)$ is analytic in D and is an indefinite integral or antiderivative of $f(z)$ in D, written

$$F(z) = \int f(z)\, dz.$$

Also, if $G'(z) = f(z)$, then $F'(z) - G'(z) \equiv 0$ in D; hence $F(z) - G(z)$ is constant in D (see Team Project 26 in Problem Set 13.4). That is, two indefinite integrals of $f(z)$ can differ only by a constant. The latter drops out in (9) of Sec. 14.1, so that we can use any indefinite integral of $f(z)$. This proves Theorem 3. ∎

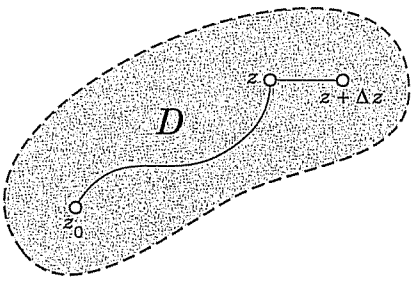


Fig. 349. Path of integration

Cauchy's Integral Theorem for Multiply Connected Domains

Cauchy's theorem applies to multiply connected domains. We first explain this for a **doubly connected domain** D with outer boundary curve C_1 and inner C_2 (Fig. 350). If a function $f(z)$ is analytic in any domain D^* that contains D and its boundary curves, we claim that

(6)
$$\oint_{C_1} f(z)\,dz = \oint_{C_2} f(z)\,dz \qquad \text{(Fig. 350)}$$

both integrals being taken counterclockwise (or both clockwise, and regardless of whether or not the full interior of C_2 belongs to D^*).

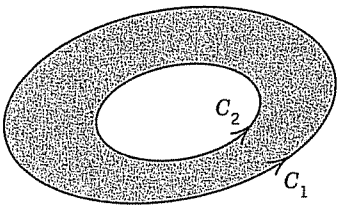


Fig. 350. Paths in (5)

PROOF By two cuts $\tilde{C}_1$ and $\tilde{C}_2$ (Fig. 351) we cut D into two simply connected domains D_1 and D_2 in which and on whose boundaries $f(z)$ is analytic. By Cauchy's integral theorem the integral over the entire boundary of D_1 (taken in the sense of the arrows in Fig. 351) is zero, and so is the integral over the boundary of D_2, and thus their sum. In this sum the integrals over the cuts $\tilde{C}_1$ and $\tilde{C}_2$ cancel because we integrate over them in both directions—this is the key—and we are left with the integrals over C_1 (counterclockwise) and C_2 (clockwise; see Fig. 351); hence by reversing the integration over C_2 (to counterclockwise) we have

$$\oint_{C_1} f\,dz - \oint_{C_2} f\,dz = 0$$

and (6) follows. ∎

For domains of higher connectivity the idea remains the same. Thus, for a **triply connected domain** we use three cuts $\tilde{C}_1$, $\tilde{C}_2$, $\tilde{C}_3$ (Fig. 352). Adding integrals as before, the integrals over the cuts cancel and the sum of the integrals over C_1 (counterclockwise) and C_2, C_3 (clockwise) is zero. Hence the integral over C_1 equals the sum of the integrals over C_2 and C_3, all three now taken counterclockwise. Similarly for quadruply connected domains, and so on.

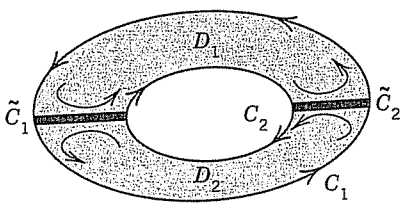


Fig. 351. Doubly connected domain

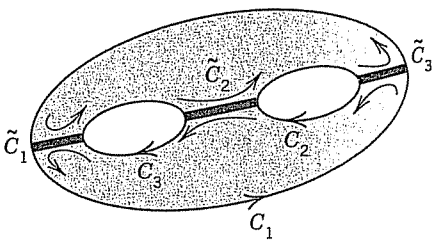


Fig. 352. Triply connected domain

PROBLEM SET 14.2

1–11 CAUCHY'S INTEGRAL THEOREM APPLICABLE?

Integrate $f(z)$ counterclockwise around the unit circle, indicating whether Cauchy's integral theorem applies. (Show the details of your work.)

1. $f(z) = \operatorname{Re} z$
2. $f(z) = 1/(3z - \pi i)$
3. $f(z) = e^{z^2/2}$
4. $f(z) = 1/\bar{z}$
5. $f(z) = \tan z^2$
6. $f(z) = \sec(z/2)$
7. $f(z) = 1/(z^8 - 1.2)$
8. $f(z) = 1/(4z - 3)$
9. $f(z) = 1/(2|z|^3)$
10. $f(z) = \bar{z}^2$
11. $f(z) = z^2 \cot z$

12–17 COMMENTS ON TEXT AND EXAMPLES

12. **(Singularities)** Can we conclude in Example 2 that the integral of $1/(z^2 + 4)$ taken over (a) $|z - 2| = 2$, (b) $|z - 2| = 3$ is zero? Give reasons.

13. **(Cauchy's integral theorem)** Verify Theorem 1 for the integral of z^2 over the boundary of the square with vertices $1 + i$, $-1 + i$, $-1 - i$, and $1 - i$ (counterclockwise).

14. **(Cauchy's integral theorem)** For what contours C will it follow from Theorem 1 that

$$\text{(a)} \quad \oint_C \frac{dz}{z} = 0, \qquad \text{(b)} \quad \oint_C \frac{\cos z}{z^6 - z^2}\, dz = 0,$$

$$\text{(c)} \quad \oint_C \frac{e^{1/z}}{z^2 + 9}\, dz = 0?$$

15. **(Deformation principle)** Can we conclude from Example 4 that the integral is also zero over the contour in Problem 13?

16. **(Deformation principle)** If the integral of a function $f(z)$ over the unit circle equals 3 and over the circle $|z| = 2$ equals 5, can we conclude that $f(z)$ is analytic everywhere in the annulus $1 < |z| < 2$?

17. **(Path independence)** Verify Theorem 2 for the integral of $\cos z$ from 0 to $(1 + i)\pi$ (a) over the shortest path, (b) over the x-axis to π and then straight up to $(1 + i)\pi$.

18. **TEAM PROJECT. Cauchy's Integral Theorem.**
(a) Main Aspects. Each of the problems in Examples 1–5 explains a basic fact in connection with Cauchy's theorem. Find five examples of your own, more complicated ones if possible, each illustrating one of those facts.

(b) Partial fractions. Write $f(z)$ in terms of partial fractions and integrate it counterclockwise over the unit circle, where

$$\text{(i)} \quad f(z) = \frac{2z + 3i}{z^2 + \frac{1}{4}} \qquad \text{(ii)} \quad f(z) = \frac{z + 1}{z^2 + 2z}.$$

(c) Deformation of path. Review (c) and (d) of Team Project 34, Sec. 14.1, in the light of the principle of deformation of path. Then consider another family of paths with common endpoints, say, $z(t) = t + ia(t - t^2)$, $0 \leqq t \leqq 1$, and experiment with the integration of analytic and nonanalytic functions of your choice over these paths (e.g., z, $\operatorname{Im} z$, z^2, $\operatorname{Re} z^2$, $\operatorname{Im} z^2$, etc).

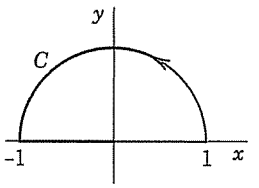
19–30 FURTHER CONTOUR INTEGRALS

Evaluate (showing the details and using partial fractions if necessary)

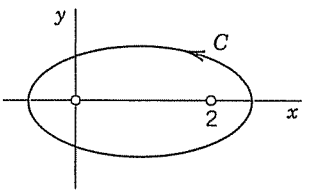
19. $\displaystyle\oint_C \frac{dz}{2z - i}$, C the circle $|z| = 3$ (counterclockwise)

20. $\displaystyle\oint_C \tanh z\, dz$, C the circle $|z - \tfrac{1}{4}\pi i| = \tfrac{1}{2}$ (clockwise)

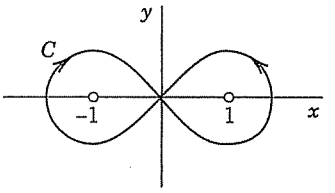
21. $\displaystyle\oint_C \operatorname{Re} 2z\, dz$, C as shown



22. $\displaystyle\oint_C \frac{7z - 6}{z^2 - 2z}\, dz$, C as shown



23. $\displaystyle\oint_C \frac{dz}{z^2 - 1}$, C as shown



24. $\displaystyle\oint_C \frac{e^{2z}}{4z}\, dz$, C consists of $|z| = 2$ (clockwise) and $|z| = \frac{1}{2}$ (counterclockwise)

25. $\oint_C \frac{\cos z}{z}\,dz$, C consists of $|z| = 1$ (counterclockwise) and $|z| = 3$ (clockwise)

26. $\oint_C \operatorname{Ln}(2 + z)\,dz$, C the boundary of the square with vertices $\pm 1, \pm i$

27. $\oint_C \frac{dz}{z^2 + 1}$, C: (a) $|z| = \frac{1}{2}$, (b) $|z - i| = \frac{3}{2}$ (counterclockwise)

28. $\oint_C \frac{dz}{z^2 + 1}$, C: (a) $|z + i| = 1$, (b) $|z - i| = 1$ (counterclockwise)

29. $\oint_C \frac{\sin z}{z + 2i}\,dz$, C: $|z - 4 - 2i| = 5.5$ (clockwise)

30. $\oint_C \frac{\tan(z/2)}{z^4 - 16}\,dz$, C the boundary of the square with vertices $\pm 1, \pm i$ (clockwise)

14.3 Cauchy's Integral Formula

The most important consequence of Cauchy's integral theorem is Cauchy's integral formula. This formula is useful for evaluating integrals, as we show below. Even more important is its key role in proving the surprising fact that analytic functions have derivatives of all orders (Sec. 14.4), in establishing Taylor series representations (Sec. 15.4), and so on. Cauchy's integral formula and its conditions of validity may be stated as follows.

THEOREM 1

Cauchy's Integral Formula

Let $f(z)$ be analytic in a simply connected domain D. Then for any point z_0 in D and any simple closed path C in D that encloses z_0 (Fig. 353),

$$\text{(1)} \qquad \oint_C \frac{f(z)}{z - z_0}\,dz = 2\pi i f(z_0) \qquad \textbf{(Cauchy's integral formula)}$$

the integration being taken counterclockwise. Alternatively (for representing $f(z_0)$ by a contour integral, divide (1) by $2\pi i$),

$$\text{(1*)} \qquad f(z_0) = \frac{1}{2\pi i}\oint_C \frac{f(z)}{z - z_0}\,dz \qquad \textbf{(Cauchy's integral formula).}$$

PROOF By addition and subtraction, $f(z) = f(z_0) + [f(z) - f(z_0)]$. Inserting this into (1) on the left and taking the constant factor $f(z_0)$ out from under the integral sign, we have

$$\text{(2)} \qquad \oint_C \frac{f(z)}{z - z_0}\,dz = f(z_0)\oint_C \frac{dz}{z - z_0} + \oint_C \frac{f(z) - f(z_0)}{z - z_0}\,dz.$$

The first term on the right equals $f(z_0) \cdot 2\pi i$ (see Example 6 in Sec. 14.2 with $m = -1$). This proves the theorem, provided the second integral on the right is zero. This is what we are now going to show. Its integrand is analytic, except at z_0. Hence by (6) in Sec. 14.2 we can replace C by a small circle K of radius ρ and center z_0 (Fig. 354), without

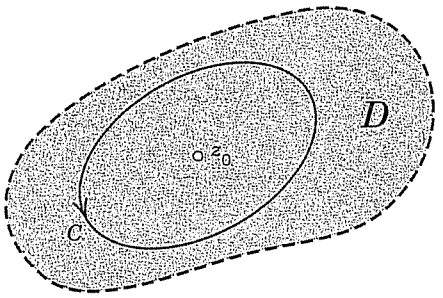


Fig. 353. Cauchy's integral formula

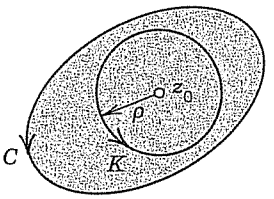


Fig. 354. Proof of Cauchy's integral formula

altering the value of the integral. Since $f(z)$ is analytic, it is continuous (Team Project 26, Sec. 13.3). Hence an $\epsilon > 0$ being given, we can find a $\delta > 0$ such that $|f(z) - f(z_0)| < \epsilon$ for all z in the disk $|z - z_0| < \delta$. Choosing the radius ρ of K smaller than δ, we thus have the inequality

$$\left| \frac{f(z) - f(z_0)}{z - z_0} \right| < \frac{\epsilon}{\rho}$$

at each point of K. The length of K is $2\pi\rho$. Hence, by the *ML*-inequality in Sec. 14.1,

$$\left| \oint_K \frac{f(z) - f(z_0)}{z - z_0} \, dz \right| < \frac{\epsilon}{\rho} \, 2\pi\rho = 2\pi\epsilon.$$

Since ϵ (> 0) can be chosen arbitrarily small, it follows that the last integral in (2) must have the value zero, and the theorem is proved. ∎

EXAMPLE 1 **Cauchy's Integral Formula**

$$\oint_C \frac{e^z}{z - 2} \, dz = 2\pi i e^z \Big|_{z=2} = 2\pi i e^2 = 46.4268i$$

for any contour enclosing $z_0 = 2$ (since e^z is entire), and zero for any contour for which $z_0 = 2$ lies outside (by Cauchy's integral theorem). ∎

EXAMPLE 2 **Cauchy's Integral Formula**

$$\oint_C \frac{z^3 - 6}{2z - i} \, dz = \oint_C \frac{\frac{1}{2}z^3 - 3}{z - \frac{1}{2}i} \, dz$$

$$= 2\pi i \left[\tfrac{1}{2}z^3 - 3\right]\Big|_{z=i/2}$$

$$= \frac{\pi}{8} - 6\pi i \qquad (z_0 = \tfrac{1}{2}i \text{ inside } C).$$ ∎

EXAMPLE 3 **Integration Around Different Contours**

Integrate

$$g(z) = \frac{z^2 + 1}{z^2 - 1} = \frac{z^2 + 1}{(z + 1)(z - 1)}$$

counterclockwise around each of the four circles in Fig. 355.

Solution. $g(z)$ is not analytic at -1 and 1. These are the points we have to watch for. We consider each circle separately.

(a) The circle $|z - 1| = 1$ encloses the point $z_0 = 1$ where $g(z)$ is not analytic. Hence in (1) we have to write

$$g(z) = \frac{z^2 + 1}{z^2 - 1} = \frac{z^2 + 1}{z + 1} \frac{1}{z - 1};$$

thus

$$f(z) = \frac{z^2 + 1}{z + 1}$$

and (1) gives

$$\oint_C \frac{z^2 + 1}{z^2 - 1} dz = 2\pi i f(1) = 2\pi i \left[\frac{z^2 + 1}{z + 1} \right]_{z=1} = 2\pi i.$$

(b) gives the same as (a) by the principle of deformation of path.

(c) The function $g(z)$ is as before, but $f(z)$ changes because we must take $z_0 = -1$ (instead of 1). This gives a factor $z - z_0 = z + 1$ in (1). Hence we must write

$$g(z) = \frac{z^2 + 1}{z - 1} \frac{1}{z + 1};$$

thus

$$f(z) = \frac{z^2 + 1}{z - 1}.$$

Compare this for a minute with the previous expression and then go on:

$$\oint_C \frac{z^2 + 1}{z^2 - 1} dz = 2\pi i f(-1) = 2\pi i \left[\frac{z^2 + 1}{z - 1} \right]_{z=-1} = -2\pi i.$$

(d) gives 0. Why?

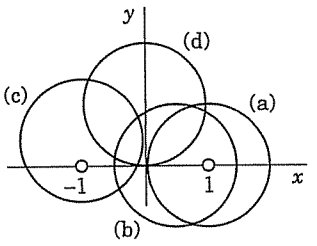


Fig. 355. Example 3

Multiply connected domains may be handled as in Sec. 14.2. For instance, if $f(z)$ is analytic on C_1 and C_2 and in the ring-shaped domain bounded by C_1 and C_2 (Fig. 356) and z_0 is any point in that domain, then

$$\textbf{(3)} \qquad f(z_0) = \frac{1}{2\pi i} \oint_{C_1} \frac{f(z)}{z - z_0} dz + \frac{1}{2\pi i} \oint_{C_2} \frac{f(z)}{z - z_0} dz,$$

where the outer integral (over C_1) is taken counterclockwise and the inner clockwise, as indicated in Fig. 356.

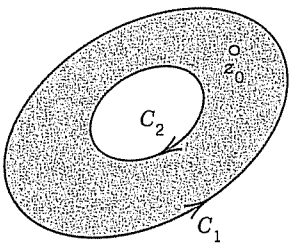


Fig. 356. Formula (3)

Our discussion in this section has illustrated the use of Cauchy's integral formula in integration. In the next section we show that the formula plays the key role in proving the surprising fact that an analytic function has derivatives of all orders, which are thus analytic functions themselves.

PROBLEM SET 14.3

1–4 CONTOUR INTEGRATION

Integrate $(z^2 - 4)/(z^2 + 4)$ counterclockwise around the circle:

1. $|z - i| = 2$
2. $|z - 1| = 2$
3. $|z + 3i| = 2$
4. $|z| = \pi/2$

5–17 CONTOUR INTEGRATION

Using Cauchy's integral formula (and showing the details), integrate counterclockwise (or as indicated)

5. $\oint_C \dfrac{z+2}{z-2}\,dz$, C: $|z - 1| = 2$
6. $\oint_C \dfrac{e^{3z}}{3z - i}\,dz$, C: $|z| = 1$
7. $\oint_C \dfrac{\sinh \pi z}{z^2 - 3z}\,dz$, C: $|z| = 1$
8. $\oint_C \dfrac{dz}{z^2 - 1}$, C: $|z - 1| = \pi/2$
9. $\oint_C \dfrac{dz}{z^2 - 1}$, C: $|z + 1| = 1$
10. $\oint_C \dfrac{e^z}{z - 2i}\,dz$, C: $|z - 2i| = 4$
11. $\oint_C \dfrac{\cos z}{2z}\,dz$, C: $|z| = \frac{1}{2}$
12. $\oint_C \dfrac{\tan z}{z - i}\,dz$, C the boundary of the triangle with vertices 0 and $\pm 1 + 2i$
13. $\oint_C \dfrac{e^{-3\pi z}}{2z + i}\,dz$, C the boundary of the square with vertices ± 1, $\pm i$
14. $\oint_C \dfrac{\operatorname{Ln}(z+1)}{z^2 + 1}\,dz$, C consists of $|z - 2i| = 2$ (counterclockwise) and $|z - 2i| = \frac{1}{2}$ (clockwise)
15. $\oint_C \dfrac{\operatorname{Ln}(z-1)}{z - 5}\,dz$, C: $|z - 4| = 2$
16. $\oint_C \dfrac{\sin z}{z^2 - 2iz}\,dz$, C consists of $|z| = 3$ (counterclockwise) and $|z| = 1$ (clockwise)
17. $\oint_C \dfrac{\cosh^2 z}{(z - 1 - i)z^2}\,dz$, C as in Prob. 16
18. Show that $\oint_C (z - z_1)^{-1}(z - z_2)^{-1}\,dz = 0$ for a simple closed path C enclosing z_1 and z_2, which are arbitrary.
19. **CAS PROJECT. Contour Integration.** Experiment to find out to what extent your CAS can do contour integration (a) by using the second method in Sec. 14.1, (b) by Cauchy's integral formula.
20. **TEAM PROJECT. Cauchy's Integral Theorem.** Gain additional insight into the proof of Cauchy's integral theorem by producing (2) with a contour enclosing z_0 (as in Fig. 353) and taking the limit as in the text. Choose

(a) $\oint_C \dfrac{z^3 - 6}{z - \frac{1}{2}i}\,dz$, (b) $\oint_C \dfrac{\sin z}{z - \frac{1}{2}\pi}\,dz$,

and (c) two other examples of your choice.

14.4 Derivatives of Analytic Functions

In this section we use Cauchy's integral formula to show the basic fact that ***complex analytic functions have derivatives of all orders.*** This is very surprising because it differs strikingly from the situation in real calculus. Indeed, if a *real* function is once differentiable, nothing follows about the existence of second or higher derivatives. Thus, in this respect, complex analytic functions behave much more simply than real functions that are once differentiable.

The existence of those derivatives will result from a general integral formula, as follows.

THEOREM 1

Derivatives of an Analytic Function

If $f(z)$ is analytic in a domain D, then it has derivatives of all orders in D, which are then also analytic functions in D. The values of these derivatives at a point z_0 in D are given by the formulas

$$(1') \qquad f'(z_0) = \frac{1}{2\pi i} \oint_C \frac{f(z)}{(z - z_0)^2}\, dz$$

$$(1'') \qquad f''(z_0) = \frac{2!}{2\pi i} \oint_C \frac{f(z)}{(z - z_0)^3}\, dz$$

and in general

$$(1) \qquad f^{(n)}(z_0) = \frac{n!}{2\pi i} \oint_C \frac{f(z)}{(z - z_0)^{n+1}}\, dz \qquad (n = 1, 2, \cdots);$$

here C is any simple closed path in D that encloses z_0 and whose full interior belongs to D; and we integrate counterclockwise around C (Fig. 357).

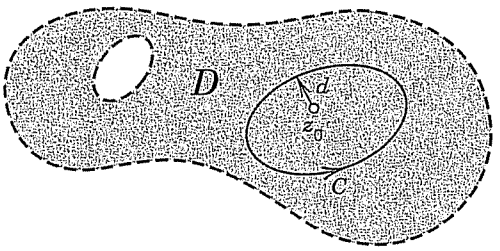


Fig. 357. Theorem 1 and its proof

COMMENT. For memorizing (1), it is useful to observe that these formulas are obtained formally by differentiating the Cauchy formula (1*), Sec. 14.3, under the integral sign *with respect to* z_0.

PROOF We prove (1′), starting from the definition of the derivative

$$f'(z_0) = \lim_{\Delta z \to 0} \frac{f(z_0 + \Delta z) - f(z_0)}{\Delta z}.$$

On the right we represent $f(z_0 + \Delta z)$ and $f(z_0)$ by Cauchy's integral formula:

$$\frac{f(z_0 + \Delta z) - f(z_0)}{\Delta z} = \frac{1}{2\pi i \Delta z}\left[\oint_C \frac{f(z)}{z - (z_0 + \Delta z)}\,dz - \oint_C \frac{f(z)}{z - z_0}\,dz\right].$$

We now write the two integrals as a single integral. Taking the common denominator gives the numerator $f(z)\{z - z_0 - [z - (z_0 + \Delta z)]\} = f(z)\,\Delta z$, so that a factor Δz drops out and we get

$$\frac{f(z_0 + \Delta z) - f(z_0)}{\Delta z} = \frac{1}{2\pi i}\oint_C \frac{f(z)}{(z - z_0 - \Delta z)(z - z_0)}\,dz.$$

Clearly, we can now establish (1′) by showing that, as $\Delta z \to 0$, the integral on the right approaches the integral in (1′). To do this, we consider the difference between these two integrals. We can write this difference as a single integral by taking the common denominator and simplifying the numerator (as just before). This gives

$$\oint_C \frac{f(z)}{(z - z_0 - \Delta z)(z - z_0)}\,dz - \oint_C \frac{f(z)}{(z - z_0)^2}\,dz = \oint_C \frac{f(z)\,\Delta z}{(z - z_0 - \Delta z)(z - z_0)^2}\,dz.$$

We show by the *ML*-inequality (Sec. 14.1) that the integral on the right approaches zero as $\Delta z \to 0$.

Being analytic, the function $f(z)$ is continuous on C, hence bounded in absolute value, say, $|f(z)| \leqq K$. Let d be the smallest distance from z_0 to the points of C (see Fig. 357). Then for all z on C,

$$|z - z_0|^2 \geqq d^2, \qquad \text{hence} \qquad \frac{1}{|z - z_0|^2} \leqq \frac{1}{d^2}.$$

Furthermore, by the triangle inequality for all z on C we then also have

$$d \leqq |z - z_0| = |z - z_0 - \Delta z + \Delta z| \leqq |z - z_0 - \Delta z| + |\Delta z|.$$

We now subtract $|\Delta z|$ on both sides and let $|\Delta z| \leqq d/2$, so that $-|\Delta z| \geqq -d/2$. Then

$$\tfrac{1}{2}d \leqq d - |\Delta z| \leqq |z - z_0 - \Delta z|. \qquad \text{Hence} \qquad \frac{1}{|z - z_0 - \Delta z|} \leqq \frac{2}{d}.$$

Let L be the length of C. If $|\Delta z| \leqq d/2$, then by the *ML*-inequality

$$\left|\oint_C \frac{f(z)\,\Delta z}{(z - z_0 - \Delta z)(z - z_0)^2}\,dz\right| \leqq KL\,|\Delta z|\,\frac{2}{d}\cdot\frac{1}{d^2}.$$

This approaches zero as $\Delta z \to 0$. Formula (1′) is proved.

Note that we used Cauchy's integral formula (1*), Sec. 14.3, but if all we had known about $f(z_0)$ is the fact that it can be represented by (1*), Sec. 14.3, our argument would have established the existence of the derivative $f'(z_0)$ of $f(z)$. This is essential to the continuation and completion of this proof, because it implies that (1″) can be proved by a similar argument, with f replaced by f', and that the general formula (1) follows by induction. ∎

EXAMPLE 1 **Evaluation of Line Integrals**

From (1′), for any contour enclosing the point πi (counterclockwise)

$$\oint_C \frac{\cos z}{(z - \pi i)^2}\, dz = 2\pi i(\cos z)'\Big|_{z=\pi i} = -2\pi i \sin \pi i = 2\pi \sinh \pi.$$ ∎

EXAMPLE 2 From (1″), for any contour enclosing the point $-i$ we obtain by counterclockwise integration

$$\oint_C \frac{z^4 - 3z^2 + 6}{(z + i)^3}\, dz = \pi i(z^4 - 3z^2 + 6)''\Big|_{z=-i} = \pi i\big[12z^2 - 6\big]_{z=-i} = -18\pi i.$$ ∎

EXAMPLE 3 By (1′), for any contour for which 1 lies inside and $\pm 2i$ lie outside (counterclockwise),

$$\oint_C \frac{e^z}{(z - 1)^2(z^2 + 4)}\, dz = 2\pi i\left(\frac{e^z}{z^2 + 4}\right)'\Big|_{z=1}$$

$$= 2\pi i \frac{e^z(z^2 + 4) - e^z 2z}{(z^2 + 4)^2}\Big|_{z=1} = \frac{6e\pi}{25}\, i \approx 2.050i.$$ ∎

Cauchy's Inequality. Liouville's and Morera's Theorems

As a new aspect, let us now show that Cauchy's integral theorem is also fundamental in deriving general results on analytic functions.

Cauchy's Inequality. Theorem 1 yields a basic inequality that has many applications. To get it, all we have to do is to choose for C in (1) a circle of radius r and center z_0 and apply the ML-inequality (Sec. 14.1); with $|f(z)| \leqq M$ on C we obtain from (1)

$$|f^{(n)}(z_0)| = \frac{n!}{2\pi}\left|\oint_C \frac{f(z)}{(z - z_0)^{n+1}}\, dz\right| \leqq \frac{n!}{2\pi}\, M\, \frac{1}{r^{n+1}}\, 2\pi r.$$

This gives **Cauchy's inequality**

(2)
$$|f^{(n)}(z_0)| \leqq \frac{n!M}{r^n}.$$

To gain a first impression of the importance of this inequality, let us prove a famous theorem on entire functions (definition in Sec. 13.5). (For Liouville, see Sec. 5.7.)

THEOREM 2

Liouville's Theorem

If an entire function is bounded in absolute value in the whole complex plane, then this function must be a constant.

PROOF By assumption, $|f(z)|$ is bounded, say, $|f(z)| < K$ for all z. Using (2), we see that $|f'(z_0)| < K/r$. Since $f(z)$ is entire, this holds for every r, so that we can take r as large as we please and conclude that $f'(z_0) = 0$. Since z_0 is arbitrary, $f'(z) = u_x + iv_x = 0$ for all z (see (4) in Sec. 13.4), hence $u_x = v_x = 0$, and $u_y = v_y = 0$ by the Cauchy–Riemann equations. Thus $u = const$, $v = const$, and $f = u + iv = const$ for all z. This completes the proof. ∎

Another very interesting consequence of Theorem 1 is

THEOREM 3

Morera's[2] Theorem (Converse of Cauchy's Integral Theorem)

If $f(z)$ is continuous in a simply connected domain D and if

(3) $$\oint_C f(z)\,dz = 0$$

for every closed path in D, then $f(z)$ is analytic in D.

PROOF In Sec. 14.2 we showed that if $f(z)$ is analytic in a simply connected domain D, then

$$F(z) = \int_{z_0}^{z} f(z^*)\,dz^*$$

is analytic in D and $F'(z) = f(z)$. In the proof we used only the continuity of $f(z)$ and the property that its integral around every closed path in D is zero; from these assumptions we concluded that $F(z)$ is analytic. By Theorem 1, the derivative of $F(z)$ is analytic, that is, $f(z)$ is analytic in D, and Morera's theorem is proved. ∎

PROBLEM SET 14.4

1–8 CONTOUR INTEGRATION

Integrate counterclockwise around the circle $|z| = 2$. (n is a positive integer, a is arbitrary.) Show the details of your work.

1. $\dfrac{\cosh 3z}{z^5}$

2. $\dfrac{\sin z}{(z - \pi i/2)^4}$

3. $\dfrac{e^z \cos z}{(z - \pi/2)^2}$

4. $\dfrac{\cos z}{z^{2n+1}}$

5. $\dfrac{\sinh az}{z^4}$

6. $\dfrac{\operatorname{Ln}(z + 3) + \cos z}{(z + 1)^2}$

7. $\dfrac{z^n}{(z - a)^{n+1}}$

8. $\dfrac{e^z}{(z - a)^n}$

[2]GIACINTO MORERA (1856–1909), Italian mathematician who worked in Genoa and Turin.

9–13 INTEGRATION AROUND DIFFERENT CONTOURS

Integrate around C. Show the details.

9. $\dfrac{(1 + 2z)\cos z}{(2z - 1)^2}$, C the unit circle, counterclockwise

10. $\dfrac{\sin 4z}{(z - 4)^3}$, C consists of $|z| = 5$ (counterclockwise) and $|z - 3| = \frac{3}{2}$ (clockwise)

11. $\dfrac{\tan \pi z}{z^2}$, C the ellipse $16x^2 + y^2 = 1$, counterclockwise

12. $\dfrac{e^{2z}}{z(z - 2i)^2}$, C consists of $|z - i| = 3$ (counterclockwise) and $|z| = 1$ (clockwise)

13. $\dfrac{e^{z/2}}{(z - a)^4}$, C the circle $|z - 2 - i| = 3$, counterclockwise

14. **TEAM PROJECT. Theory on Growth**

(a) Growth of entire functions. If $f(z)$ is not a constant and is analytic for all (finite) z, and R and M are any positive real numbers (no matter how large), show that there exist values of z for which $|z| > R$ and $|f(z)| > M$.

(b) Growth of polynomials. If $f(z)$ is a polynomial of degree $n > 0$ and M is an arbitrary positive real number (no matter how large), show that there exists a positive real number R such that $|f(z)| > M$ for all $|z| > R$.

(c) Exponential function. Show that $f(z) = e^z$ has the property characterized in (a) but does not have that characterized in (b).

(d) Fundamental theorem of algebra. *If $f(z)$ is a polynomial in z, not a constant, then $f(z) = 0$ for at least one value of z.* Prove this, using (a).

15. **(Proof of Theorem 1)** Complete the proof of Theorem 1 by performing the induction mentioned at the end.

CHAPTER 14 REVIEW QUESTIONS AND PROBLEMS

1. What is a path of integration? What did we assume about paths?
2. State the definition of a complex line integral from memory.
3. What do we mean by saying that complex integration is a linear operation?
4. Make a list of integration methods discussed. Illustrate each with a simple example.
5. Which integration methods apply to analytic functions only?
6. What value do you get if you integrate $1/z$ counterclockwise around the unit circle? (You should memorize this basic result.) If you integrate $1/z^2$, $1/z^3, \cdots$?
7. Which theorem in this chapter do you regard as most important? State it from memory.
8. What is independence of path? What is the principle of deformation of path? Why is this important?
9. Do not confuse Cauchy's integral theorem and Cauchy's integral formula. State both. How are they related?
10. How can you extend Cauchy's integral theorem to doubly and triply connected domains?
11. If integrating $f(z)$ over the boundary circles of an annulus D gives different values, can $f(z)$ be analytic in D? (Give reason.)
12. Is $\left|\int_C f(z)\,dz\right| = \int_C |f(z)|\,dz$? How would you find a bound for the integral on the left?
13. Is $\operatorname{Re}\int_C f(z)\,dz = \int_C \operatorname{Re} f(z)\,dz$? Give examples.
14. How did we use integral formulas for derivatives in integration?
15. What is Liouville's theorem? Give examples. State consequences.

16–30 INTEGRATION

Integrate by a suitable method:

16. $4z^3 + 2z$ from $-i$ to $2 + i$ along any path
17. $5z - 3/z$ counterclockwise around the unit circle
18. $z + 1/z$ counterclockwise around $|z + 3i| = 2$
19. e^{2z} from $-2 + 3\pi i$ along the straight segment to $-2 + 5\pi i$
20. $e^{z^2}/(z - 1)^2$ counterclockwise around $|z| = 2$
21. $z/(z^2 + 1)$ clockwise around $|z + i| = 1$
22. Re z from 0 to 4 and then vertically up to $4 + 3i$
23. $\cosh 4z$ from 0 to $2i$ along the imaginary axis
24. e^z/z over C consisting of $|z| = 1$ (counterclockwise) and $|z| = \frac{1}{2}$ (clockwise)
25. $(\sin z)/z$ clockwise around a circle containing $z = 0$ in its interior
26. Im z counterclockwise around $|z| = r$
27. $(\operatorname{Ln} z)/(z - 2i)^2$ counterclockwise around $|z - 2i| = 1$
28. $(\tan \pi z)/(z - 1)^2$ counterclockwise around $|z - 1| = 0.2$
29. $|z| + z$ clockwise around the unit circle
30. $(z - i)^{-3}(z^3 + \sin z)$ counterclockwise around any circle with center i

SUMMARY OF CHAPTER 14
Complex Integration

The **complex line integral** of a function $f(z)$ taken over a path C is denoted by

(1) $$\int_C f(z)\,dz \qquad \text{or, if } C \text{ is closed, also by} \qquad \oint_C f(z) \qquad \text{(Sec. 14.1).}$$

If $f(z)$ is analytic in a simply connected domain D, then we can evaluate (1) as in calculus by indefinite integration and substitution of limits, that is,

(2) $$\int_C f(z)\,dz = F(z_1) - F(z_0) \qquad [F'(z) = f(z)]$$

for every path C in D from a point z_0 to a point z_1 (see Sec. 14.1). These assumptions imply **independence of path,** that is, (2) depends only on z_0 and z_1 (and on $f(z)$, of course) but not on the choice of C (Sec. 14.2). The existence of an $F(z)$ such that $F'(z) = f(z)$ is proved in Sec. 14.2 by Cauchy's integral theorem (see below).

A general method of integration, not restricted to analytic functions, uses the equation $z = z(t)$ of C, where $a \leqq t \leqq b$,

(3) $$\int_C f(z)\,dz = \int_a^b f(z(t))\dot{z}(t)\,dt \qquad \left(\dot{z} = \frac{dz}{dt}\right).$$

Cauchy's integral theorem is the most important theorem in this chapter. It states that if $f(z)$ is analytic in a simply connected domain D, then for every closed path C in D (Sec. 14.2),

(4) $$\oint_C f(z)\,dz = 0.$$

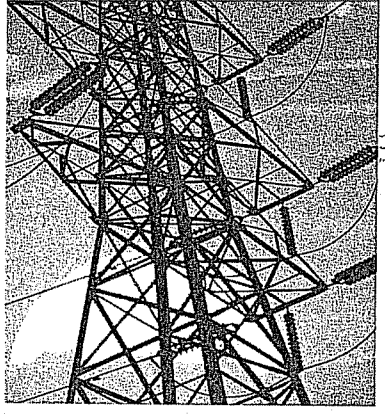
Under the same assumptions and for any z_0 in D and closed path C in D containing z_0 in its interior we also have **Cauchy's integral formula**

(5) $$f(z_0) = \frac{1}{2\pi i}\oint_C \frac{f(z)}{z - z_0}\,dz.$$

Furthermore, under these assumptions $f(z)$ has derivatives of all orders in D that are themselves analytic functions in D and (Sec. 14.4)

(6) $$f^{(n)}(z_0) = \frac{n!}{2\pi i}\oint_C \frac{f(z)}{(z - z_0)^{n+1}}\,dz \qquad (n = 1, 2, \cdots).$$

This implies *Morera's theorem* (the converse of Cauchy's integral theorem) and *Cauchy's inequality* (Sec. 14.4), which in turn implies *Liouville's theorem* that an entire function that is bounded in the whole complex plane must be constant.



CHAPTER 15

Power Series, Taylor Series

Complex power series, in particular, Taylor series, are analogs of real power and Taylor series in calculus. However, they are much more fundamental in complex analysis than their real counterparts in calculus. The reason is that power series represent analytic functions (Sec. 15.3) and, conversely, every analytic function can be represented by power series, called Taylor series (Sec. 15.4).

Use Sec. 15.1 for reference if you are familiar with convergence tests for *real* series—in complex this is quite similar. The last section (15.5) on uniform convergence is *optional.*

Prerequisite: Chaps. 13, 14.
Sections that may be omitted in a shorter course: 14.1, 14.5.
References and Answers to Problems: App. 1 Part D, App. 2.

15.1 Sequences, Series, Convergence Tests

In this section we define the basic concepts for ***complex*** sequences and series and discuss tests for convergence and divergence. This is very similar to *real* sequences and series in calculus. ***If you feel at home with the latter and want to take for granted that the ratio test also holds in complex, skip this section and go to Sec.* 15.2.**

Sequences

The basic definitions are as in calculus. An *infinite sequence* or, briefly, a **sequence,** is obtained by assigning to each positive integer n a number z_n, called a **term** of the sequence, and is written

$$z_1, z_2, \cdots \qquad \text{or} \qquad \{z_1, z_2, \cdots\} \qquad \text{or briefly} \qquad \{z_n\}.$$

We may also write $z_0, z_1, \cdots$ or $z_2, z_3, \cdots$ or start with some other integer if convenient.

A **real sequence** is one whose terms are real.

Convergence. A **convergent sequence** $z_1, z_2, \cdots$ is one that has a limit c, written

$$\lim_{n\to\infty} z_n = c \qquad \text{or simply} \qquad z_n \to c.$$

By definition of **limit** this means that for every $\epsilon > 0$ we can find an N such that

$$|z_n - c| < \epsilon \qquad \text{for all } n > N; \tag{1}$$

geometrically, all terms z_n with $n > N$ lie in the open disk of radius ϵ and center c (Fig. 358) and only finitely many terms do not lie in that disk. [For a *real* sequence, (1) gives an open interval of length 2ϵ and real midpoint c on the real line; see Fig. 359.]

A **divergent sequence** is one that does not converge.

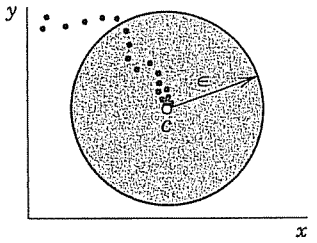


Fig. 358. Convergent complex sequence

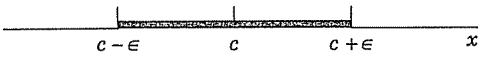


Fig. 359. Convergent real sequence

EXAMPLE 1 **Convergent and Divergent Sequences**

The sequence $\{i^n/n\} = \{i, -1/2, -i/3, 1/4, \cdots\}$ is convergent with limit 0.

The sequence $\{i^n\} = \{i, -1, -i, 1, \cdots\}$ is divergent, and so is $\{z_n\}$ with $z_n = (1 + i)^n$. ■

EXAMPLE 2 **Sequences of the Real and the Imaginary Parts**

The sequence $\{z_n\}$ with $z_n = x_n + iy_n = 1 - 1/n^2 + i(2 + 4/n)$ is $6i$, $3/4 + 4i$, $8/9 + 10i/3$, $15/16 + 3i$, $\cdots$. (Sketch it.) It converges with the limit $c = 1 + 2i$. Observe that $\{x_n\}$ has the limit $1 = \operatorname{Re} c$ and $\{y_n\}$ has the limit $2 = \operatorname{Im} c$. This is typical. It illustrates the following theorem by which the convergence of a *complex* sequence can be referred back to that of the two *real* sequences of the real parts and the imaginary parts. ■

THEOREM 1

Sequences of the Real and the Imaginary Parts

A sequence $z_1, z_2, \cdots, z_n, \cdots$ of complex numbers $z_n = x_n + iy_n$ (where $n = 1, 2, \cdots$) converges to $c = a + ib$ if and only if the sequence of the real parts $x_1, x_2, \cdots$ converges to a and the sequence of the imaginary parts $y_1, y_2, \cdots$ converges to b.

PROOF Convergence $z_n \to c = a + ib$ implies convergence $x_n \to a$ and $y_n \to b$ because if $|z_n - c| < \epsilon$, then z_n lies within the circle of radius ϵ about $c = a + ib$, so that (Fig. 360a)

$$|x_n - a| < \epsilon, \qquad |y_n - b| < \epsilon.$$

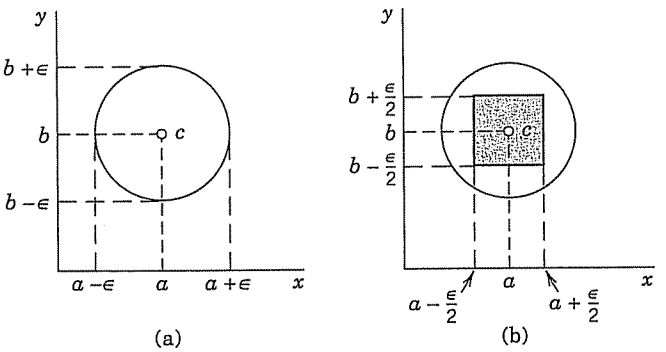


Fig. 360. Proof of Theorem 1

Conversely, if $x_n \to a$ and $y_n \to b$ as $n \to \infty$, then for a given $\epsilon > 0$ we can choose N so large that, for every $n > N$,

$$|x_n - a| < \frac{\epsilon}{2}, \qquad |y_n - b| < \frac{\epsilon}{2}.$$

These two inequalities imply that $z_n = x_n + iy_n$ lies in a square with center c and side ϵ. Hence, z_n must lie within a circle of radius ϵ with center c (Fig. 360b). ∎

Series

Given a sequence $z_1, z_2, \cdots, z_m, \cdots$, we may form the sequence of the sums

$$s_1 = z_1, \qquad s_2 = z_1 + z_2, \qquad s_3 = z_1 + z_2 + z_3, \quad \cdots$$

and in general

$$s_n = z_1 + z_2 + \cdots + z_n \qquad (n = 1, 2, \cdots). \tag{2}$$

s_n is called the ***n*th partial sum** of the *infinite series* or **series**

$$\sum_{m=1}^{\infty} z_m = z_1 + z_2 + \cdots. \tag{3}$$

The $z_1, z_2, \cdots$ are called the **terms** of the series. (Our usual ***summation letter*** is n, unless we need n for another purpose, as here, and we then use m as the summation letter.)

A **convergent series** is one whose sequence of partial sums converges, say,

$$\lim_{n\to\infty} s_n = s. \qquad \text{Then we write} \qquad s = \sum_{m=1}^{\infty} z_m = z_1 + z_2 + \cdots$$

and call s the **sum** or *value* of the series. A series that is not convergent is called a **divergent series.**

If we omit the terms of s_n from (3), there remains

$$R_n = z_{n+1} + z_{n+2} + z_{n+3} + \cdots. \tag{4}$$

This is called the **remainder** *of the series* (3) *after the term* z_n. Clearly, if (3) converges and has the sum s, then

$$s = s_n + R_n, \qquad \text{thus} \qquad R_n = s - s_n.$$

Now $s_n \to s$ by the definition of convergence; hence $R_n \to 0$. In applications, when s is unknown and we compute an approximation s_n of s, then $|R_n|$ is the error, and $R_n \to 0$ means that we can make $|R_n|$ as small as we please, by choosing n large enough.

An application of Theorem 1 to the partial sums immediately relates the convergence of a complex series to that of the two series of its real parts and of its imaginary parts:

THEOREM 2

Real and Imaginary Parts

A series (3) *with* $z_m = x_m + iy_m$ *converges and has the sum* $s = u + iv$ *if and only if* $x_1 + x_2 + \cdots$ *converges and has the sum* u *and* $y_1 + y_2 + \cdots$ *converges and has the sum* v.

Tests for Convergence and Divergence of Series

Convergence tests in complex are practically the same as in calculus. We apply them before we use a series, to make sure that the series converges.

Divergence can often be shown very simply as follows.

THEOREM 3

Divergence

If a series $z_1 + z_2 + \cdots$ *converges, then* $\lim_{m\to\infty} z_m = 0$. *Hence if this does not hold, the series diverges.*

PROOF If $z_1 + z_2 + \cdots$ converges, with the sum s, then, since $z_m = s_m - s_{m-1}$,

$$\lim_{m\to\infty} z_m = \lim_{m\to\infty} (s_m - s_{m-1}) = \lim_{m\to\infty} s_m - \lim_{m\to\infty} s_{m-1} = s - s = 0. \qquad \blacksquare$$

CAUTION! $z_m \to 0$ is *necessary* for convergence but *not sufficient,* as we see from the harmonic series $1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots$, which satisfies this condition but diverges, as is shown in calculus (see, for example, Ref. [GR11] in App. 1).

The practical difficulty in proving convergence is that in most cases the sum of a series is unknown. Cauchy overcame this by showing that a series converges if and only if its partial sums eventually get close to each other:

THEOREM 4

Cauchy's Convergence Principle for Series

A series $z_1 + z_2 + \cdots$ *is convergent if and only if for every given* $\epsilon > 0$ *(no matter how small) we can find an* N *(which depends on* ϵ, *in general) such that*

(5) $$|z_{n+1} + z_{n+2} + \cdots + z_{n+p}| < \epsilon \qquad \text{for every } n > N \text{ and } p = 1, 2, \cdots$$

The somewhat involved proof is left optional (see App. 4).

Absolute Convergence. A series $z_1 + z_2 + \cdots$ is called **absolutely convergent** if the series of the absolute values of the terms

$$\sum_{m=1}^{\infty} |z_m| = |z_1| + |z_2| + \cdots$$

is convergent.

If $z_1 + z_2 + \cdots$ converges but $|z_1| + |z_2| + \cdots$ diverges, then the series $z_1 + z_2 + \cdots$ is called, more precisely, **conditionally convergent.**

EXAMPLE 3 **A Conditionally Convergent Series**

The series $1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + - \cdots$ converges, but only conditionally since the harmonic series diverges, as mentioned above (after Theorem 3).

If a series is absolutely convergent, it is convergent.

This follows readily from Cauchy's principle (see Team Project 30). This principle also yields the following general convergence test.

THEOREM 5 **Comparison Test**

If a series $z_1 + z_2 + \cdots$ is given and we can find a convergent series $b_1 + b_2 + \cdots$ with nonnegative real terms such that $|z_1| \leqq b_1$, $|z_2| \leqq b_2$, $\cdots$, then the given series converges, even absolutely.

PROOF By Cauchy's principle, since $b_1 + b_2 + \cdots$ converges, for any given $\epsilon > 0$ we can find an N such that

$$b_{n+1} + \cdots + b_{n+p} < \epsilon \qquad \text{for every } n > N \text{ and } p = 1, 2, \cdots.$$

From this and $|z_1| \leqq b_1$, $|z_2| \leqq b_2$, $\cdots$ we conclude that for those n and p,

$$|z_{n+1}| + \cdots + |z_{n+p}| \leqq b_{n+1} + \cdots + b_{n+p} < \epsilon.$$

Hence, again by Cauchy's principle, $|z_1| + |z_2| + \cdots$ converges, so that $z_1 + z_2 + \cdots$ is absolutely convergent.

A good comparison series is the geometric series, which behaves as follows.

THEOREM 6 **Geometric Series**

The geometric series

(6*)
$$\sum_{m=0}^{\infty} q^m = 1 + q + q^2 + \cdots$$

converges with the sum $1/(1 - q)$ if $|q| < 1$ and diverges if $|q| \geqq 1$.

PROOF If $|q| \geqq 1$, then $|q^m| \geqq 1$ and Theorem 3 implies divergence.

Now let $|q| < 1$. The nth partial sum is

$$s_n = 1 + q + \cdots + q^n.$$

From this,

$$qs_n = \qquad q + \cdots + q^n + q^{n+1}.$$

On subtraction, most terms on the right cancel in pairs, and we are left with

$$s_n - qs_n = (1 - q)s_n = 1 - q^{n+1}.$$

Now $1 - q \neq 0$ since $q \neq 1$, and we may solve for s_n, finding

$$(6) \qquad s_n = \frac{1 - q^{n+1}}{1 - q} = \frac{1}{1 - q} - \frac{q^{n+1}}{1 - q}.$$

Since $|q| < 1$, the last term approaches zero as $n \to \infty$. Hence if $|q| < 1$, the series is convergent and has the sum $1/(1 - q)$. This completes the proof. ∎

Ratio Test

This is the most important test in our further work. We get it by taking the geometric series as comparison series $b_1 + b_2 + \cdots$ in Theorem 5:

THEOREM 7

Ratio Test

If a series $z_1 + z_2 + \cdots$ with $z_n \neq 0$ $(n = 1, 2, \cdots)$ has the property that for every n greater than some N,

$$(7) \qquad \left| \frac{z_{n+1}}{z_n} \right| \leqq q < 1 \qquad (n > N)$$

(where $q < 1$ is fixed), this series converges absolutely. If for every $n > N$,

$$(8) \qquad \left| \frac{z_{n+1}}{z_n} \right| \geqq 1 \qquad (n > N),$$

the series diverges.

PROOF If (8) holds, then $|z_{n+1}| \geqq |z_n|$ for $n > N$, so that divergence of the series follows from Theorem 3.

If (7) holds, then $|z_{n+1}| \leqq |z_n|\, q$ for $n > N$, in particular,

$$|z_{N+2}| \leqq |z_{N+1}|q, \qquad |z_{N+3}| \leqq |z_{N+2}|q \leqq |z_{N+1}|q^2, \qquad \text{etc.},$$

and in general, $|z_{N+p}| \leqq |z_{N+1}|q^{p-1}$. Since $q < 1$, we obtain from this and Theorem 6

$$|z_{N+1}| + |z_{N+2}| + |z_{N+3}| + \cdots \leqq |z_{N+1}|\,(1 + q + q^2 + \cdots) \leqq |z_{N+1}|\,\frac{1}{1 - q}.$$

Absolute convergence of $z_1 + z_2 + \cdots$ now follows from Theorem 5. ∎

CAUTION! The inequality (7) implies $|z_{n+1}/z_n| < 1$, but this does ***not*** imply convergence, as we see from the harmonic series, which satisfies $z_{n+1}/z_n = n/(n + 1) < 1$ for all n but diverges.

If the sequence of the ratios in (7) and (8) converges, we get the more convenient

THEOREM 8

Ratio Test

If a series $z_1 + z_2 + \cdots$ *with* $z_n \neq 0$ $(n = 1, 2, \cdots)$ *is such that* $\lim_{n\to\infty} \left|\frac{z_{n+1}}{z_n}\right| = L$, *then:*

(a) *If* $L < 1$, *the series converges absolutely.*

(b) *If* $L > 1$, *the series diverges.*

(c) *If* $L = 1$, *the series may converge or diverge, so that the test fails and permits no conclusion.*

PROOF **(a)** We write $k_n = |z_{n+1}/z_n|$ and let $L = 1 - b < 1$. Then by the definition of limit, the k_n must eventually get close to $1 - b$, say, $k_n \leqq q = 1 - \frac{1}{2}b < 1$ for all n greater than some N. Convergence of $z_1 + z_2 + \cdots$ now follows from Theorem 7.

(b) Similarly, for $L = 1 + c > 1$ we have $k_n \geqq 1 + \frac{1}{2}c > 1$ for all $n > N^*$ (sufficiently large), which implies divergence of $z_1 + z_2 + \cdots$ by Theorem 7.

(c) The harmonic series $1 + \frac{1}{2} + \frac{1}{3} + \cdots$ has $z_{n+1}/z_n = n/(n+1)$, hence $L = 1$, and diverges. The series

$$1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \frac{1}{25} + \cdots \qquad \text{has} \qquad \frac{z_{n+1}}{z_n} = \frac{n^2}{(n+1)^2},$$

hence also $L = 1$, but it converges. Convergence follows from (Fig. 361)

$$s_n = 1 + \frac{1}{4} + \cdots + \frac{1}{n^2} \leqq 1 + \int_1^n \frac{dx}{x^2} = 2 - \frac{1}{n},$$

so that $s_1, s_2, \cdots$ is a bounded sequence and is monotone increasing (since the terms of the series are all positive); both properties together are sufficient for the convergence of the real sequence $s_1, s_2, \cdots$. (In calculus this is proved by the so-called *integral test*, whose idea we have used.) ∎

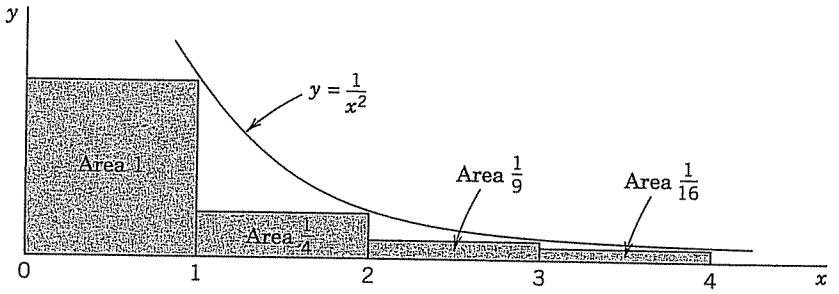


Fig. 361. Convergence of the series $1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \cdots$

EXAMPLE 4 **Ratio Test**

Is the following series convergent or divergent? (First guess, then calculate.)

$$\sum_{n=0}^{\infty} \frac{(100 + 75i)^n}{n!} = 1 + (100 + 75i) + \frac{1}{2!}(100 + 75i)^2 + \cdots$$

Solution. By Theorem 8, the series is convergent, since

$$\left|\frac{z_{n+1}}{z_n}\right| = \frac{|100 + 75i|^{n+1}/(n+1)!}{|100 + 75i|^n/n!} = \frac{|100 + 75i|}{n+1} = \frac{125}{n+1} \quad \to \quad L = 0. \qquad \blacksquare$$

EXAMPLE 5 **Theorem 7 More General than Theorem 8**

Let $a_n = i/2^{3n}$ and $b_n = 1/2^{3n+1}$. Is the following series convergent or divergent?

$$a_0 + b_0 + a_1 + b_1 + \cdots = i + \frac{1}{2} + \frac{i}{8} + \frac{1}{16} + \frac{i}{64} + \frac{1}{128} + \cdots$$

Solution. The ratios of the absolute values of successive terms are $\frac{1}{2}, \frac{1}{4}, \frac{1}{2}, \frac{1}{4}, \cdots$. Hence convergence follows from Theorem 7. Since the sequence of these ratios has no limit, Theorem 8 is not applicable. $\blacksquare$

Root Test

The ratio test and the root test are the two practically most important tests. The ratio test is usually simpler, but the root test is somewhat more general.

THEOREM 9

Root Test

If a series $z_1 + z_2 + \cdots$ is such that for every n greater than some N,

$$\sqrt[n]{|z_n|} \leqq q < 1 \qquad (n > N) \tag{9}$$

(where $q < 1$ is fixed), this series converges absolutely. If for infinitely many n,

$$\sqrt[n]{|z_n|} \geqq 1, \tag{10}$$

the series diverges.

PROOF If (9) holds, then $|z_n| \leqq q^n < 1$ for all $n > N$. Hence the series $|z_1| + |z_2| + \cdots$ converges by comparison with the geometric series, so that the series $z_1 + z_2 + \cdots$ converges absolutely. If (10) holds, then $|z_n| \geqq 1$ for infinitely many n. Divergence of $z_1 + z_2 + \cdots$ now follows from Theorem 3. $\blacksquare$

CAUTION! Equation (9) implies $\sqrt[n]{|z_n|} < 1$, but this does not imply convergence, as we see from the harmonic series, which satisfies $\sqrt[n]{1/n} < 1$ (for $n > 1$) but diverges.

If the sequence of the roots in (9) and (10) converges, we more conveniently have

THEOREM 10

Root Test

If a series $z_1 + z_2 + \cdots$ is such that $\lim_{n\to\infty} \sqrt[n]{|z_n|} = L$, then:

(a) *The series converges absolutely if $L < 1$.*

(b) *The series diverges if $L > 1$.*

(c) *If $L = 1$, the test fails; that is, no conclusion is possible.*

PROOF The proof parallels that of Theorem 8.

(a) Let $L = 1 - a^* < 1$. Then by the definition of a limit we have $\sqrt[n]{|z_n|} < q = 1 - \frac{1}{2}a^* < 1$ for all n greater than some (sufficiently large) N^*. Hence $|z_n| < q^n < 1$ for all $n > N^*$. Absolute convergence of the series $z_1 + z_2 + \cdots$ now follows by the comparison with the geometric series.

(b) If $L > 1$, then we also have $\sqrt[n]{|z_n|} > 1$ for all sufficiently large n. Hence $|z_n| > 1$ for those n. Theorem 3 now implies that $z_1 + z_2 + \cdots$ diverges.

(c) Both the *divergent* harmonic series and the *convergent* series $1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \frac{1}{25} + \cdots$ give $L = 1$. This can be seen from $(\ln n)/n \to 0$ and

$$\sqrt[n]{\frac{1}{n}} = \frac{1}{n^{1/n}} = \frac{1}{e^{(1/n)\ln n}} \to \frac{1}{e^0}, \qquad \sqrt[n]{\frac{1}{n^2}} = \frac{1}{n^{2/n}} = \frac{1}{e^{(2/n)\ln n}} \to \frac{1}{e^0}. \quad \blacksquare$$

PROBLEM SET 15.1

1–10 SEQUENCES

Are the following sequences $z_1, z_2, \cdots, z_n, \cdots$ bounded? Convergent? Find their limit points. (Show the details of your work.)

1. $z_n = (-1)^n + i/2^n$
2. $z_n = e^{-n\pi i/4}$
3. $z_n = (-1)^n/(n + i)$
4. $z_n = (1 + i)^n$
5. $z_n = \operatorname{Ln}((2 + i)^n)$
6. $z_n = (3 + 4i)^n/n!$
7. $z_n = \sin(n\pi/4) + i^n$
8. $z_n = \left[(1 + 3i)/\sqrt{10}\right]^n$
9. $z_n = (0.9 + 0.1i)^{2n}$
10. $z_n = (5 + 5i)^{-n}$

11. Illustrate Theorem 1 by an example of your own.
12. **(Uniqueness of limit)** Show that if a sequence converges, its limit is unique.
13. **(Addition)** If $z_1, z_2, \cdots$ converges with the limit l and $z_1^*, z_2^*, \cdots$ converges with the limit l^*, show that $z_1 + z_1^*, z_2 + z_2^*, \cdots$ converges with the limit $l + l^*$.
14. **(Multiplication)** Show that under the assumptions of Prob. 13 the sequence $z_1 z_1^*, z_2 z_2^*, \cdots$ converges with the limit ll^*.
15. **(Boundedness)** Show that a complex sequence is bounded if and only if the two corresponding sequences of the real parts and of the imaginary parts are bounded.

16–24 SERIES

Are the following series convergent or divergent? (Give a reason.)

16. $\sum_{n=0}^{\infty} \frac{(10 - 15i)^n}{n!}$
17. $\sum_{n=0}^{\infty} \frac{(-1)^n(1 + 2i)^{2n+1}}{(2n + 1)!}$
18. $\sum_{n=0}^{\infty} \frac{i^n}{n^2 - 2i}$
19. $\sum_{n=1}^{\infty} \frac{1}{\sqrt{n}}$
20. $\sum_{n=2}^{\infty} \frac{1}{\ln n}$
21. $\sum_{n=1}^{\infty} \frac{i^n}{n}$
22. $\sum_{n=0}^{\infty} \frac{(n!)^3}{(3n)!}(1 + i)^n$
23. $\sum_{n=0}^{\infty} \frac{n - i}{3n + 2i}$
24. $\sum_{n=1}^{\infty} n^2\left(\frac{i}{3}\right)^n$

25. What is the difference between (7) and just stating $|z_{n+1}/z_n| < 1$?
26. Illustrate Theorem 2 by an example of your choice.
27. For what n do we obtain the term of greatest absolute value of the series in Example 4? About how big is it? First guess, then calculate it by the Stirling formula in Sec. 24.4.
28. Give another example showing that Theorem 7 is more general than Theorem 8.
29. **CAS PROJECT. Sequences and Series. (a)** Write a program for graphing complex **sequences.** Apply it to sequences of your choice that have interesting "geometrical" properties (e.g., lying on an ellipse, spiraling toward its limit, etc.).

    (b) Write a program for computing and graphing numeric values of the first n partial sums of a **series** of complex numbers. Use the program to experiment with the rapidity of convergence of series of your choice.
30. **TEAM PROJECT. Series. (a) Absolute convergence.** Show that if a series converges absolutely, it is convergent.

    (b) Write a short report on the basic concepts and properties of series of numbers, explaining in each case whether or not they carry over from real series (discussed in calculus) to complex series, with reasons given.

(c) **Estimate of the remainder.** Let $|z_{n+1}/z_n| \leqq q < 1$, so that the series $z_1 + z_2 + \cdots$ converges by the ratio test. Show that the remainder $R_n = z_{n+1} + z_{n+2} + \cdots$ satisfies the inequality $|R_n| \leqq |z_{n+1}|/(1 - q)$.

(d) Using (c), find how many terms suffice for computing the sum s of the series

$$\sum_{n=1}^{\infty} \frac{n + i}{2^n n}$$

with an error not exceeding 0.05 and compute s to this accuracy.

(e) Find other applications of the estimate in (c).

15.2 Power Series

Power series are the most important series in complex analysis because we shall see that their sums are analytic functions, and every analytic function can be represented by power series (Theorem 5 in Sec. 15.3 and Theorem 1 in Sec. 15.4).

A **power series** *in powers of* $z - z_0$ is a series of the form

$$\sum_{n=0}^{\infty} a_n(z - z_0)^n = a_0 + a_1(z - z_0) + a_2(z - z_0)^2 + \cdots \tag{1}$$

where z is a complex variable, a_0, a_1, $\cdots$ are complex (or real) constants, called the **coefficients** of the series, and z_0 is a complex (or real) constant, called the **center** of the series. This generalizes real power series of calculus.

If $z_0 = 0$, we obtain as a particular case a *power series in powers of* z*:*

$$\sum_{n=0}^{\infty} a_n z^n = a_0 + a_1 z + a_2 z^2 + \cdots. \tag{2}$$

Convergence Behavior of Power Series

Power series have variable terms (functions of z), but ***if we fix z, then all the concepts for series with constant terms in the last section apply.*** Usually a series with variable terms will converge for some z and diverge for others. For a power series the situation is simple. The series (1) may converge in a disk with center z_0 or in the whole z-plane or only at z_0. We illustrate this with typical examples and then prove it.

EXAMPLE 1 **Convergence in a Disk. Geometric Series**

The *geometric series*

$$\sum_{n=0}^{\infty} z^n = 1 + z + z^2 + \cdots$$

converges absolutely if $|z| < 1$ and diverges if $|z| \geqq 1$ (see Theorem 6 in Sec. 15.1). ▩

EXAMPLE 2 **Convergence for Every z**

The power series (which will be the Maclaurin series of e^z in Sec. 15.4)

$$\sum_{n=0}^{\infty} \frac{z^n}{n!} = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \cdots$$

is absolutely convergent for every z, In fact, by the ratio test, for any fixed z,

$$\left|\frac{z^{n+1}/(n+1)!}{z^n/n!}\right| = \frac{|z|}{n+1} \to 0 \quad \text{as} \quad n \to \infty.$$

EXAMPLE 3 **Convergence Only at the Center. (Useless Series)**

The following power series converges only at $z = 0$, but diverges for every $z \neq 0$, as we shall show.

$$\sum_{n=0}^{\infty} n! z^n = 1 + z + 2z^2 + 6z^3 + \cdots$$

In fact, from the ratio test we have

$$\left|\frac{(n+1)!z^{n+1}}{n!z^n}\right| = (n+1)|z| \to \infty \quad \text{as} \quad n \to \infty \quad (z \text{ fixed and } \neq 0).$$

THEOREM 1

Convergence of a Power Series

(a) *Every power series* (1) *converges at the center* z_0.

(b) *If* (1) *converges at a point* $z = z_1 \neq z_0$, *it converges absolutely for every* z *closer to* z_0 *than* z_1, *that is,* $|z - z_0| < |z_1 - z_0|$. See Fig. 362.

(c) *If* (1) *diverges at a* $z = z_2$, *it diverges for every* z *farther away from* z_0 *than* z_2. See Fig. 362.

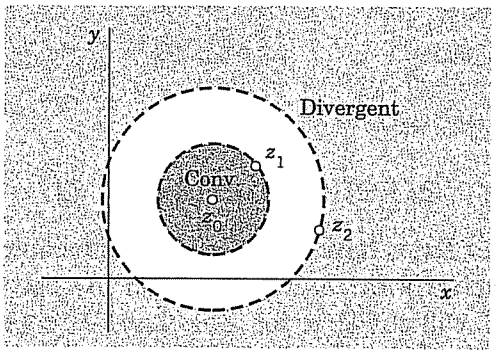


Fig. 362. Theroem 1

PROOF **(a)** For $z = z_0$ the series reduces to the single term a_0.

(b) Convergence at $z = z_1$ gives by Theorem 3 in Sec. 15.1 $a_n(z_1 - z_0)^n \to 0$ as $n \to \infty$. This implies boundedness in absolute value,

$$|a_n(z_1 - z_0)^n| < M \qquad \text{for every } n = 0, 1, \cdots.$$

Multiplying and dividing $a_n(z - z_0)^n$ by $(z_1 - z_0)^n$ we obtain from this

$$|a_n(z - z_0)^n| = \left|a_n(z_1 - z_0)^n \left(\frac{z - z_0}{z_1 - z_0}\right)^n\right| \leqq M \left|\frac{z - z_0}{z_1 - z_0}\right|^n.$$

Summation over n gives

$$\sum_{n=1}^{\infty} |a_n(z - z_0)^n| \leqq M \sum_{n=1}^{\infty} \left| \frac{z - z_0}{z_1 - z_0} \right|^n. \tag{3}$$

Now our assumption $|z - z_0| < |z_1 - z_0|$ implies that $|(z - z_0)/(z_1 - z_0)| < 1$. Hence the series on the right side of (3) is a converging geometric series (see Theorem 6 in Sec. 15.1). Absolute convergence of (1) as stated in (b) now follows by the comparison test in Sec. 15.1.

(c) If this were false, we would have convergence at a z_3 farther away from z_0 than z_2. This would imply convergence at z_2, by (b), a contradiction to our assumption of divergence at z_2. ∎

Radius of Convergence of a Power Series

Convergence for every z (the nicest case, Example 2) or for no $z \neq z_0$ (the useless case, Example 3) needs no further discussion, and we put these cases aside for a moment. We consider the ***smallest*** circle with center z_0 that includes all the points at which a given power series (1) converges. Let R denote its radius. The circle

$$|z - z_0| = R \qquad \text{(Fig. 363)}$$

is called the **circle of convergence** and its radius R the **radius of convergence** of (1). Theorem 1 then implies convergence everywhere within that circle, that is, for all z for which

$$|z - z_0| < R \tag{4}$$

(the open disk with center z_0 and radius R). Also, since R is as *small* as possible, the series (1) diverges for all z for which

$$|z - z_0| > R. \tag{5}$$

No general statements can be made about the convergence of a power series (1) ***on the circle of convergence*** itself. The series (1) may converge at some or all or none of these points. Details will not be essential to us. Hence a simple example may just give us the idea.

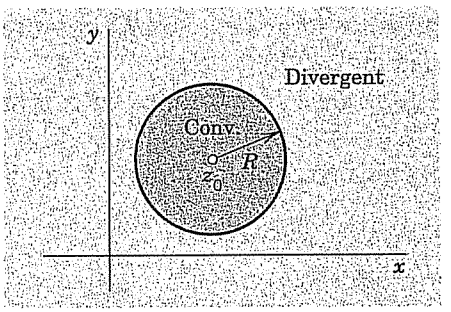


Fig. 363. Circle of convergence

EXAMPLE 4 **Behavior on the Circle of Convergence**

On the circle of convergence (radius $R = 1$ in all three series),

$\Sigma z^n/n^2$ converges everywhere since $\Sigma 1/n^2$ converges,

$\Sigma z^n/n$ converges at -1 (by Leibniz's test) but diverges at 1,

Σz^n diverges everywhere. ∎

Notations $R = \infty$ and $R = 0$. To incorporate these two excluded cases in the present notation, we write

$R = \infty$ if the series (1) converges for all z (as in Example 2),

$R = 0$ if (1) converges only at the center $z = z_0$ (as in Example 3).

These are convenient notations, but nothing else.

Real Power Series. In this case in which powers, coefficients, and center are real, formula (4) gives the **convergence interval** $|x - x_0| < R$ of length $2R$ on the real line.

Determination of the Radius of Convergence from the Coefficients. For this important practical task we can use

THEOREM 2

Radius of Convergence R

Suppose that the sequence $|a_{n+1}/a_n|$, $n = 1, 2, \cdots$, converges with limit L^*. If $L^* = 0$, then $R = \infty$; that is, the power series (1) converges for all z. If $L^* \neq 0$ (hence $L^* > 0$), then

$$R = \frac{1}{L^*} = \lim_{n\to\infty} \left| \frac{a_n}{a_{n+1}} \right| \tag{6}$$

(Cauchy–Hadamard formula[1]).

If $|a_{n+1}/a_n| \to \infty$, then $R = 0$ (convergence only at the center z_0).

PROOF For (1) the ratio of the terms in the ratio test (Sec. 15.1) is

$$\left| \frac{a_{n+1}(z - z_0)^{n+1}}{a_n(z - z_0)^n} \right| = \left| \frac{a_{n+1}}{a_n} \right| |z - z_0|. \qquad \text{The limit is} \qquad L = L^*|z - z_0|.$$

Let $L^* \neq 0$, thus $L^* > 0$. We have convergence if $L = L^*|z - z_0| < 1$, thus $|z - z_0| < 1/L^*$, and divergence if $|z - z_0| > 1/L^*$. By (4) and (5) this shows that $1/L^*$ is the convergence radius and proves (6).

If $L^* = 0$, then $L = 0$ for every z, which gives convergence for all z by the ratio test. If $|a_{n+1}/a_n| \to \infty$, then $|a_{n+1}/a_n||z - z_0| > 1$ for any $z \neq z_0$ and all sufficiently large n. This implies divergence for all $z \neq z_0$ by the ratio test (Theorem 7, Sec. 15.1). ∎

[1]Named after the French mathematicians A. L. CAUCHY (see Sec. 2.5) and JACQUES HADAMARD (1865–1963). Hadamard made basic contributions to the theory of power series and devoted his lifework to partial differential equations.

Formula (6) will not help if L^* does not exist, but extensions of Theorem 2 are still possible, as we discuss in Example 6 below.

EXAMPLE 5 Radius of Convergence

By (6) the radius of convergence of the power series $\sum_{n=0}^{\infty} \frac{(2n)!}{(n!)^2}(z-3i)^n$ is

$$R = \lim_{n\to\infty}\left[\frac{(2n)!}{(n!)^2} \Big/ \frac{(2n+2)!}{((n+1)!)^2}\right] = \lim_{n\to\infty}\left[\frac{(2n)!}{(2n+2)!}\cdot\frac{((n+1)!)^2}{(n!)^2}\right] = \lim_{n\to\infty}\frac{(n+1)^2}{(2n+2)(2n+1)} = \frac{1}{4}.$$

The series converges in the open disk $|z-3i| < \frac{1}{4}$ of radius $\frac{1}{4}$ and center $3i$. ■

EXAMPLE 6 Extension of Theorem 2

Find the radius of convergence R of the power series

$$\sum_{n=0}^{\infty}\left[1+(-1)^n+\frac{1}{2^n}\right]z^n = 3+\frac{1}{2}z+\left(2+\frac{1}{4}\right)z^2+\frac{1}{8}z^3+\left(2+\frac{1}{16}\right)z^4+\cdots.$$

Solution. The sequence of the ratios $1/6, 2(2+\frac{1}{4}), 1/(8(2+\frac{1}{4})), \cdots$ does not converge, so that Theorem 2 is of no help. It can be shown that

$$(6^*) \qquad R = 1/\tilde{L}, \qquad \tilde{L} = \lim_{n\to\infty}\sqrt[n]{|a_n|}.$$

This still does not help here, since $\{\sqrt[n]{|a_n|}\}$ does not converge because $\sqrt[n]{|a_n|} = \sqrt[n]{1/2^n} = 1/2$ for odd n, whereas for even n we have

$$\sqrt[n]{|a_n|} = \sqrt[n]{2+1/2^n} \to 1 \qquad \text{as} \qquad n\to\infty,$$

so that $\sqrt[n]{|a_n|}$ has the two limit points $1/2$ and 1. It can further be shown that

$$(6^{**}) \qquad R = 1/\tilde{l}, \qquad \tilde{l} \text{ the greatest limit point of the sequence } \left\{\sqrt[n]{|a_n|}\right\}.$$

Here $\tilde{l} = 1$, so that $R = 1$. *Answer.* The series converges for $|z| < 1$. ■

Summary. Power series converge in an open circular disk or some even for every z (or some only at the center, but they are useless); for the radius of convergence, see (6) or Example 6.

Except for the useless ones, power series have sums that are analytic functions (as we show in the next section); this accounts for their importance in complex analysis.

PROBLEM SET 15.2

1. **(Powers missing)** Show that if $\Sigma a_n z^n$ has radius of convergence R (assumed finite), then $\Sigma a_n z^{2n}$ has radius of convergence $\sqrt{R}$. Give examples.
2. **(Convergence behavior)** Illustrate the facts shown by Examples 1–3 by further examples of your own.

3–18 RADIUS OF CONVERGENCE

Find the center and the radius of convergence of the following power series. (Show the details.)

3. $\sum_{n=1}^{\infty}\frac{(z+i)^n}{n^2}$
4. $\sum_{n=0}^{\infty}\frac{n^n}{n!}(z+2i)^n$
5. $\sum_{n=0}^{\infty}\frac{n!}{n^n}(z+1)^n$
6. $\sum_{n=0}^{\infty}\frac{2^{100n}}{n!}z^n$
7. $\sum_{n=0}^{\infty}\left(\frac{a}{b}\right)^n z^n$
8. $\sum_{n=0}^{\infty}\frac{(-1)^n}{2^{2n}(n!)^2}z^{2n}$
9. $\sum_{n=0}^{\infty}(n-i)^n z^n$
10. $\sum_{n=0}^{\infty}\frac{(2z)^{2n}}{(2n)!}$
11. $\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n}z^n$
12. $\sum_{n=0}^{\infty}\frac{4^n}{(1+i)^n}(z-5)^n$

13. $\sum_{n=2}^{\infty} n(n-1)(z-3+2i)^n$

14. $\sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} z^{2n}$

15. $\sum_{n=0}^{\infty} 2^n(z-i)^{4n}$

16. $\sum_{n=0}^{\infty} \left(\frac{2+3i}{5-i}\right)^n (z-\pi)^n$

17. $\sum_{n=0}^{\infty} \frac{n^4}{2^n} z^{2n}$

18. $\sum_{n=0}^{\infty} \frac{(4n)!}{2^n(n!)^4} (z+\pi i)^n$

19. **CAS PROJECT. Radius of Convergence.** Write a program for computing R from (6), (6*), or (6**), in this order, depending on the existence of the limits needed. Test the program on series of your choice and such that all three formulas (6), (6*), and (6**) will come up.

20. **TEAM PROJECT. Radius of Convergence.** (a) Formula (6) for R contains $|a_n/a_{n+1}|$, not $|a_{n+1}/a_n|$. How could you memorize this by using a qualitative argument?

(b) Change of coefficients. What happens to R $(0 < R < \infty)$ if you (i) multiply all a_n by $k \neq 0$, (ii) multiply a_n by $k^n \neq 0$, (iii) replace a_n by $1/a_n$?

(c) Example 6 extends Theorem 2 to nonconvergent cases of a_n/a_{n+1}. Do you understand the principle of "mixing" by which Example 6 was obtained? Use this principle for making up further examples.

(d) Does there exist a power series in powers of z that converges at $z = 30 + 10i$ and diverges at $z = 31 - 6i$? (Give reason.)

15.3 Functions Given by Power Series

The main goal of this section is to show that power series represent analytic functions (Theorem 5). Along our way we shall see that power series behave nicely under addition, multiplication, differentiation, and integration, which makes these series very useful in complex analysis.

To simplify the formulas in this section, we take $z_0 = 0$ and write

$$\sum_{n=0}^{\infty} a_n z^n. \tag{1}$$

This is no restriction because a series in powers of $\hat{z} - z_0$ with any z_0 can always be reduced to the form (1) if we set $\hat{z} - z_0 = z$.

Terminology and Notation. If any given power series (1) has a nonzero radius of convergence R (thus $R > 0$), its sum is a function of z, say $f(z)$. Then we write

$$f(z) = \sum_{n=0}^{\infty} a_n z^n = a_0 + a_1 z + a_2 z^2 + \cdots \qquad (|z| < R). \tag{2}$$

We say that $f(z)$ *is* **represented** *by the power series* or that *it is* **developed** *in the power series.* For instance, the geometric series *represents* the function $f(z) = 1/(1-z)$ in the interior of the unit circle $|z| = 1$. (See Theorem 6 in Sec. 15.1.)

Uniqueness of a Power Series Representation. This is our next goal. It means that *a function $f(z)$ cannot be represented by two different power series with the same center.* We claim that if $f(z)$ can at all be developed in a power series with center z_0, the development is unique. This important fact is frequently used in complex analysis (as well as in calculus). We shall prove it in Theorem 2. The proof will follow from

THEOREM 1

Continuity of the Sum of a Power Series

If a function $f(z)$ can be represented by a power series (2) with radius of convergence $R > 0$, then $f(z)$ is continuous at $z = 0$.

PROOF From (2) with $z = 0$ we have $f(0) = a_0$. Hence by the definition of continuity we must show that $\lim_{z \to 0} f(z) = f(0) = a_0$. That is, we must show that for a given $\epsilon > 0$ there is a $\delta > 0$ such that $|z| < \delta$ implies $|f(z) - a_0| < \epsilon$. Now (2) converges absolutely for $|z| \leqq r$ with any r such that $0 < r < R$, by Theorem 1 in Sec. 15.2. Hence the series

$$\sum_{n=1}^{\infty} |a_n| r^{n-1} = \frac{1}{r} \sum_{n=1}^{\infty} |a_n| r^n$$

converges. Let $S \neq 0$ be its sum. ($S = 0$ is trivial.) Then for $0 < |z| \leqq r$,

$$|f(z) - a_0| = \left| \sum_{n=1}^{\infty} a_n z^n \right| \leqq |z| \sum_{n=1}^{\infty} |a_n| \, |z|^{n-1} \leqq |z| \sum_{n=1}^{\infty} |a_n| r^{n-1} = |z| S$$

and $|z|S < \epsilon$ when $|z| < \delta$, where $\delta > 0$ is less than r and less than ϵ/S. Hence $|z|S < \delta S < (\epsilon/S)S = \epsilon$. This proves the theorem. ∎

From this theorem we can now readily obtain the desired uniqueness theorem (again assuming $z_0 = 0$ without loss of generality):

THEOREM 2

Identity Theorem for Power Series. Uniqueness

Let the power series $a_0 + a_1 z + a_2 z^2 + \cdots$ and $b_0 + b_1 z + b_2 z^2 + \cdots$ both be convergent for $|z| < R$, where R is positive, and let them both have the same sum for all these z. Then the series are identical, that is, $a_0 = b_0$, $a_1 = b_1$, $a_2 = b_2$, $\cdots$.

Hence if a function $f(z)$ can be represented by a power series with any center z_0, this representation is ***unique.***

PROOF We proceed by induction. By assumption,

$$a_0 + a_1 z + a_2 z^2 + \cdots = b_0 + b_1 z + b_2 z^2 + \cdots \qquad (|z| < R).$$

The sums of these two power series are continuous at $z = 0$, by Theorem 1. Hence if we consider $|z| > 0$ and let $z \to 0$ on both sides, we see that $a_0 = b_0$: the assertion is true for $n = 0$. Now assume that $a_n = b_n$ for $n = 0, 1, \cdots, m$. Then on both sides we may omit the terms that are equal and divide the result by z^{m+1} ($\neq 0$); this gives

$$a_{m+1} + a_{m+2} z + a_{m+3} z^2 + \cdots = b_{m+1} + b_{m+2} z + b_{m+3} z^2 + \cdots.$$

Similarly as before by letting $z \to 0$ we conclude from this that $a_{m+1} = b_{m+1}$. This completes the proof. ∎

Operations on Power Series

Interesting in itself, this discussion will serve as a preparation for our main goal, namely, to show that functions represented by power series are analytic.

Termwise addition or subtraction of two power series with radii of convergence R_1 and R_2 yields a power series with radius of convergence at least equal to the smaller of R_1 and R_2. *Proof.* Add (or subtract) the partial sums s_n and s_n^* term by term and use $\lim (s_n \pm s_n^*) = \lim s_n \pm \lim s_n^*$.

Termwise multiplication of two power series

$$f(z) = \sum_{k=0}^{\infty} a_k z^k = a_0 + a_1 z + \cdots$$

and

$$g(z) = \sum_{m=0}^{\infty} b_m z^m = b_0 + b_1 z + \cdots$$

means the multiplication of each term of the first series by each term of the second series and the collection of like powers of z. This gives a power series, which is called the **Cauchy product** of the two series and is given by

$$a_0 b_0 + (a_0 b_1 + a_1 b_0) z + (a_0 b_2 + a_1 b_1 + a_2 b_0) z^2 + \cdots = \sum_{n=0}^{\infty} (a_0 b_n + a_1 b_{n-1} + \cdots + a_n b_0) z^n.$$

We mention without proof that this power series converges absolutely for each z within the circle of convergence of each of the two given series and has the sum $s(z) = f(z)g(z)$. For a proof, see [D5] listed in App. 1.

Termwise differentiation and integration of power series is permissible, as we show next. We call **derived series** *of the power series* (1) the power series obtained from (1) by termwise differentiation, that is,

$$\text{(3)} \qquad \sum_{n=1}^{\infty} n a_n z^{n-1} = a_1 + 2a_2 z + 3a_3 z^2 + \cdots.$$

THEOREM 3

Termwise Differentiation of a Power Series

The derived series of a power series has the same radius of convergence as the original series.

PROOF This follows from (6) in Sec. 15.2 because

$$\lim_{n\to\infty} \frac{n|a_n|}{(n+1)\,|a_{n+1}|} = \lim_{n\to\infty} \frac{n}{n+1} \lim_{n\to\infty} \left|\frac{a_n}{a_{n+1}}\right| = \lim_{n\to\infty} \left|\frac{a_n}{a_{n+1}}\right|$$

or, if the limit does not exist, from (6**) in Sec. 15.2 by noting that $\sqrt[n]{n} \to 1$ as $n \to \infty$. ∎

EXAMPLE 1 **Application of Theorem 3**

Find the radius of convergence R of the following series by applying Theorem 3.

$$\sum_{n=2}^{\infty} \binom{n}{2} z^n = z^2 + 3z^3 + 6z^4 + 10z^5 + \cdots.$$

Solution. Differentiate the geometric series twice term by term and multiply the result by $z^2/2$. This yields the given series. Hence $R = 1$ by Theorem 3. ∎

THEOREM 4

Termwise Integration of Power Series

The power series

$$\sum_{n=0}^{\infty} \frac{a_n}{n+1} z^{n+1} = a_0 z + \frac{a_1}{2} z^2 + \frac{a_2}{3} z^3 + \cdots$$

obtained by integrating the series $a_0 + a_1 z + a_2 z^2 + \cdots$ term by term has the same radius of convergence as the original series.

The proof is similar to that of Theorem 3.

With Theorem 3 as a tool, we are now ready to establish our main result in this section.

Power Series Represent Analytic Functions

THEOREM 5

Analytic Functions. Their Derivatives

A power series with a nonzero radius of convergence R represents an analytic function at every point interior to its circle of convergence. The derivatives of this function are obtained by differentiating the original series term by term. All the series thus obtained have the same radius of convergence as the original series. Hence, by the first statement, each of them represents an analytic function.

PROOF **(a)** We consider any power series (1) with positive radius of convergence R. Let $f(z)$ be its sum and $f_1(z)$ the sum of its derived series; thus

$$(4) \qquad f(z) = \sum_{n=0}^{\infty} a_n z^n \qquad \text{and} \qquad f_1(z) = \sum_{n=1}^{\infty} n a_n z^{n-1}.$$

We show that $f(z)$ is analytic and has the derivative $f_1(z)$ in the interior of the circle of convergence. We do this by proving that for any fixed z with $|z| < R$ and $\Delta z \to 0$ the difference quotient $[f(z + \Delta z) - f(z)]/\Delta z$ approaches $f_1(z)$. By termwise addition we first have from (4)

$$(5) \qquad \frac{f(z+\Delta z) - f(z)}{\Delta z} - f_1(z) = \sum_{n=2}^{\infty} a_n \left[\frac{(z+\Delta z)^n - z^n}{\Delta z} - n z^{n-1} \right].$$

Note that the summation starts with 2, since the constant term drops out in taking the difference $f(z + \Delta z) - f(z)$, and so does the linear term when we subtract $f_1(z)$ from the difference quotient.

(b) We claim that the series in (5) can be written

$$(6) \qquad \sum_{n=2}^{\infty} a_n \Delta z[(z+\Delta z)^{n-2} + 2z(z+\Delta z)^{n-3} + \cdots + (n-2)z^{n-3}(z+\Delta z) + (n-1)z^{n-2}].$$

The somewhat technical proof of this is given in App. 4.

(c) We consider (6). The brackets contain $n-1$ terms, and the largest coefficient is $n-1$. Since $(n-1)^2 \leqq n(n-1)$, we see that for $|z| \leqq R_0$ and $|z+\Delta z| \leqq R_0$, $R_0 < R$, the absolute value of this series (6) cannot exceed

$$(7) \qquad |\Delta z| \sum_{n=2}^{\infty} |a_n| n(n-1) R_0^{n-2}.$$

This series with a_n instead of $|a_n|$ is the second derived series of (2) at $z = R_0$ and converges absolutely by Theorem 3 of this section and Theorem 1 of Sec. 15.2. Hence our present series (7) converges. Let the sum of (7) (without the factor $|\Delta z|$) be $K(R_0)$. Since (6) is the right side of (5), our present result is

$$\left| \frac{f(z+\Delta z) - f(z)}{\Delta z} - f_1(z) \right| \leqq |\Delta z| \, K(R_0).$$

Letting $\Delta z \to 0$ and noting that R_0 ($< R$) is arbitrary, we conclude that $f(z)$ is analytic at any point interior to the circle of convergence and its derivative is represented by the derived series. From this the statements about the higher derivatives follow by induction. ∎

Summary. The results in this section show that power series are about as nice as we could hope for: we can differentiate and integrate them term by term (Theorems 3 and 4). Theorem 5 accounts for the great importance of power series in complex analysis: the sum of such a series (with a positive radius of convergence) is an analytic function and has derivatives of all orders, which thus in turn are analytic functions. But this is only part of the story. In the next section we show that, conversely, *every* given analytic function $f(z)$ can be represented by power series, called **Taylor series** and being the complex analog of the real Taylor series of calculus.

PROBLEM SET 15.3

1–10 RADIUS OF CONVERGENCE BY DIFFERENTIATION OR INTEGRATION

Find the radius of convergence in two ways: (a) directly by the Cauchy–Hadamard formula in Sec. 15.2, (b) from a series of simpler terms by using Theorem 3 or Theorem 4.

1. $\displaystyle\sum_{n=2}^{\infty} \frac{n(n-1)}{3^n} (z-2i)^n$

2. $\displaystyle\sum_{n=1}^{\infty} \frac{4^n}{n(n+1)} z^n$

3. $\displaystyle\sum_{n=1}^{\infty} \frac{n}{2^n} (z+i)^{2n}$

4. $\displaystyle\sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1} \left(\frac{z}{\pi}\right)^{2n+1}$

5. $\displaystyle\sum_{n=1}^{\infty} \frac{3^n n(n+1)}{5^n} (z-1)^{2n}$

6. $\displaystyle\sum_{n=k}^{\infty} \binom{n}{k} \left(\frac{z}{4}\right)^n$

7. $\displaystyle\sum_{n=1}^{\infty} \frac{(-7)^n}{n(n+1)(n+2)} z^{2n}$

8. $\displaystyle\sum_{n=1}^{\infty} \frac{2n(2n-1)}{n^n} z^{2n-2}$

9. $\sum_{n=0}^{\infty} \left[\binom{n+k}{k}\right]^{-1} z^{n+k}$

10. $\sum_{n=0}^{\infty} \binom{n+m}{m} z^n$

11. **(Addition and subtraction)** Write out the details of the proof on termwise addition and subtraction of power series.

12. **(Cauchy product)** Show that $(1-z)^{-2} = \Sigma_{n=0}^{\infty}(n+1)z^n$ (a) by using the Cauchy product, (b) by differentiating a suitable series.

13. **(Cauchy product)** Show that the Cauchy product of $\Sigma_{n=0}^{\infty} z^n/n!$ multiplied by itself gives $\Sigma_{n=0}^{\infty}(2z)^n/n!$.

14. **(On Theorem 3)** Prove that $\sqrt[n]{n} \to 1$ as $n \to \infty$ (as claimed in the proof of Theorem 3).

15. **(On Theorems 3 and 4)** Find further examples of your own.

16–20 APPLICATIONS OF THE IDENTITY THEOREM

State clearly and explicitly where and how you are using Theorem 2.

16. **(Binomial coefficients)** Using $(1+z)^p(1+z)^q = (1+z)^{p+q}$, obtain the basic relation

$$\sum_{n=0}^{r} \binom{p}{n}\binom{q}{r-n} = \binom{p+q}{r}.$$

17. **(Odd function)** If $f(z)$ in (1) is *odd* (i.e., $f(-z) = -f(z)$), show that $a_n = 0$ for even n. Give examples.

18. **(Even functions)** If $f(z)$ in (1) is *even* (i.e., $f(-z) = f(z)$), show that $a_n = 0$ for odd n. Give examples.

19. Find applications of Theorem 2 in differential equations and elsewhere

20. **TEAM PROJECT. Fibonacci numbers.**[2] **(a)** The Fibonacci numbers are recursively defined by $a_0 = a_1 = 1$, $a_{n+1} = a_n + a_{n-1}$ if $n = 1, 2, \cdots$. Find the limit of the sequence (a_{n+1}/a_n).

(b) Fibonacci's rabbit problem. Compute a list of $a_1, \cdots, a_{12}$. Show that $a_{12} = 233$ is the number of pairs of rabbits after 12 months if initially there is 1 pair and each pair generates 1 pair per month, beginning in the second month of existence (no deaths occurring).

(c) Generating function. Show that the *generating function* of the **Fibonacci numbers** is $f(z) = 1/(1-z-z^2)$; that is, if a power series (1) represents this $f(z)$, its coefficients must be the Fibonacci numbers and conversely. *Hint.* Start from $f(z)(1-z-z^2) = 1$ and use Theorem 2.

15.4 Taylor and Maclaurin Series

The **Taylor series**[3] of a function $f(z)$, the complex analog of the real Taylor series is

$$\text{(1)} \qquad f(z) = \sum_{n=1}^{\infty} a_n(z-z_0)^n \qquad \text{where} \qquad a_n = \frac{1}{n!}f^{(n)}(z_0)$$

or, by (1), Sec. 14.4,

$$\text{(2)} \qquad a_n = \frac{1}{2\pi i}\oint_C \frac{f(z^*)}{(z^*-z_0)^{n+1}}\,dz^*.$$

In (2) we integrate counterclockwise around a simple closed path C that contains z_0 in its interior and is such that $f(z)$ is analytic in a domain containing C and every point inside C.

A **Maclaurin series**[3] is a Taylor series with center $z_0 = 0$.

---

[2]LEONARDO OF PISA, called FIBONACCI (= son of Bonaccio), about 1180–1250, Italian mathematician, credited with the first renaissance of mathematics on Christian soil.

[3]BROOK TAYLOR (1685–1731), English mathematician who introduced real Taylor series. COLIN MACLAURIN (1698–1746), Scots mathematician, professor at Edinburgh.

The **remainder** of the Taylor series (1) after the term $a_n(z - z_0)^n$ is

$$R_n(z) = \frac{(z - z_0)^{n+1}}{2\pi i} \oint_C \frac{f(z^*)}{(z^* - z_0)^{n+1}(z^* - z)}\, dz^* \tag{3}$$

(proof below). Writing out the corresponding partial sum of (1), we thus have

$$f(z) = f(z_0) + \frac{z - z_0}{1!} f'(z_0) + \frac{(z - z_0)^2}{2!} f''(z_0) + \cdots + \frac{(z - z_0)^n}{n!} f^{(n)}(z_0) + R_n(z). \tag{4}$$

This is called **Taylor's formula** *with remainder.*

We see that ***Taylor series are power series.*** From the last section we know that power series represent analytic functions. And we now show that *every* analytic function can be represented by power series, namely, by Taylor series (with various centers). This makes Taylor series very important in complex analysis. Indeed, they are more fundamental in complex analysis than their real counterparts are in calculus.

THEOREM 1

Taylor's Theorem

Let $f(z)$ be analytic in a domain D, and let $z = z_0$ be any point in D. Then there exists precisely one Taylor series (1) *with center z_0 that represents $f(z)$. This representation is valid in the largest open disk with center z_0 in which $f(z)$ is analytic. The remainders $R_n(z)$ of* (1) *can be represented in the form* (3). *The coefficients satisfy the inequality*

$$|a_n| \leqq \frac{M}{r^n} \tag{5}$$

where M is the maximum of $|f(z)|$ on a circle $|z - z_0| = r$ in D whose interior is also in D.

PROOF The key tool is Cauchy's integral formula in Sec. 14.3; writing z and z^* instead of z_0 and z (so that z^* is the variable of integration), we have

$$f(z) = \frac{1}{2\pi i} \oint_C \frac{f(z^*)}{z^* - z}\, dz^*. \tag{6}$$

z lies inside C, for which we take a circle of radius r with center z_0 and interior in D (Fig. 364). We develop $1/(z^* - z)$ in (6) in powers of $z - z_0$. By a ***standard algebraic manipulation*** (worth remembering!) we first have

$$\frac{1}{z^* - z} = \frac{1}{z^* - z_0 - (z - z_0)} = \frac{1}{(z^* - z_0)\left(1 - \dfrac{z - z_0}{z^* - z_0}\right)}. \tag{7}$$

For later use we note that since z^* is on C while z is inside C, we have

(7*) $$\left|\frac{z - z_0}{z^* - z_0}\right| < 1 \qquad \text{(Fig. 364).}$$

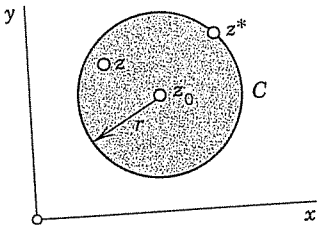


Fig. 364. Cauchy formula (6)

To (7) we now apply the sum formula for a finite geometric sum

(8*) $$1 + q + \cdots + q^n = \frac{1 - q^{n+1}}{1 - q} = \frac{1}{1 - q} - \frac{q^{n+1}}{1 - q} \qquad (q \neq 1),$$

which we use in the form (take the last term to the other side and interchange sides)

(8) $$\frac{1}{1 - q} = 1 + q + \cdots + q^n + \frac{q^{n+1}}{1 - q}.$$

Applying this with $q = (z - z_0)/(z^* - z_0)$ to the right side of (7), we get

$$\frac{1}{z^* - z} = \frac{1}{z^* - z_0}\left[1 + \frac{z - z_0}{z^* - z_0} + \left(\frac{z - z_0}{z^* - z_0}\right)^2 + \cdots + \left(\frac{z - z_0}{z^* - z_0}\right)^n\right] + \frac{1}{z^* - z}\left(\frac{z - z_0}{z^* - z_0}\right)^{n+1}.$$

We insert this into (6). Powers of $z - z_0$ do not depend on the variable of integration z^*, so that we may take them out from under the integral sign. This yields

$$f(z) = \frac{1}{2\pi i}\oint_C \frac{f(z^*)}{z^* - z_0}\,dz^* + \frac{z - z_0}{2\pi i}\oint_C \frac{f(z^*)}{(z^* - z_0)^2}\,dz^* + \cdots + \frac{(z - z_0)^n}{2\pi i}\oint_C \frac{f(z^*)}{(z^* - z_0)^{n+1}}\,dz^* + R_n(z)$$

with $R_n(z)$ given by (3). The integrals are those in (2) related to the derivatives, so that we have proved the Taylor formula (4).

Since analytic functions have derivatives of all orders, we can take n in (4) as large as we please. If we let n approach infinity, we obtain (1). Clearly, (1) will converge and represent $f(z)$ if and only if

(9) $$\lim_{n\to\infty} R_n(z) = 0.$$

We prove (9) as follows. Since z^* lies on C, whereas z lies inside C (Fig. 364), we have $|z^* - z| > 0$. Since $f(z)$ is analytic inside and on C, it is bounded, and so is the function $f(z^*)/(z^* - z)$, say,

$$\left| \frac{f(z^*)}{z^* - z} \right| \leqq \tilde{M}$$

for all z^* on C. Also, C has the radius $r = |z^* - z_0|$ and the length $2\pi r$. Hence by the *ML*-inequality (Sec. 14.1) we obtain from (3)

$$(10) \qquad \begin{aligned} |R_n| &= \frac{|z - z_0|^{n+1}}{2\pi} \left| \oint_C \frac{f(z^*)}{(z^* - z_0)^{n+1}(z^* - z)} \, dz^* \right| \\ &\leqq \frac{|z - z_0|^{n+1}}{2\pi} \tilde{M} \frac{1}{r^{n+1}} 2\pi r = \tilde{M} \left| \frac{z - z_0}{r} \right|^{n+1}. \end{aligned}$$

Now $|z - z_0| < r$ because z lies *inside* C. Thus $|z - z_0|/r < 1$, so that the right side approaches 0 as $n \to \infty$. This proves the convergence of the Taylor series. Uniqueness follows from Theorem 2 in the last section. Finally, (5) follows from (1) and the Cauchy inequality in Sec. 14.4. This proves Taylor's theorem. ∎

Accuracy of Approximation. We can achieve any preassinged accuracy in approximating $f(z)$ by a partial sum of (1) by choosing n large enough. This is the practical aspect of formula (9).

Singularity, Radius of Convergence. On the circle of convergence of (1) there is at least one **singular point** of $f(z)$, that is, a point $z = c$ at which $f(z)$ is not analytic (but such that every disk with center c contains points at which $f(z)$ *is* analytic). We also say that $f(z)$ **is singular** at c or **has a singularity** at c. Hence the radius of convergence R of (1) is usually equal to the distance from z_0 to the nearest singular point of $f(z)$.

(Sometimes R can be greater than that distance: Ln z is singular on the negative real axis, whose distance from $z_0 = -1 + i$ is 1, but the Taylor series of Ln z with center $z_0 = -1 + i$ has radius of convergence $\sqrt{2}$.)

Power Series as Taylor Series

Taylor series are power series—of course! Conversely, we have

THEOREM 2

Relation to the Last Section

A power series with a nonzero radius of convergence is the Taylor series of its sum.

PROOF Given the power series

$$f(z) = a_0 + a_1(z - z_0) + a_2(z - z_0)^2 + a_3(z - z_0)^3 + \cdots.$$

Then $f(z_0) = a_0$. By Theorem 5 in Sec. 15.3 we obtain

$$f'(z) = a_1 + 2a_2(z - z_0) + 3a_3(z - z_0)^2 + \cdots, \qquad \text{thus} \quad f'(z_0) = a_1$$

$$f''(z) = 2a_2 + 3 \cdot 2(z - z_0) + \cdots, \qquad \text{thus} \quad f''(z_0) = 2!\,a_2$$

and in general $f^{(n)}(z_0) = n!\,a_n$. With these coefficients the given series becomes the Taylor series of $f(z)$ with center z_0. ■

Comparison with Real Functions. One surprising property of complex analytic functions is that they have derivatives of all orders, and now we have discovered the other surprising property that they can always be represented by power series of the form (1). This is not true in general for ***real functions;*** there are real functions that have derivatives of all orders but cannot be represented by a power series. (Example: $f(x) = \exp(-1/x^2)$ if $x \neq 0$ and $f(0) = 0$; this function cannot be represented by a Maclaurin series in an open disk with center 0 because all its derivatives at 0 are zero.)

Important Special Taylor Series

These are as in calculus, with x replaced by complex z. Can you see why? (*Answer.* The coefficient formulas are the same.)

EXAMPLE 1 **Geometric Series**

Let $f(z) = 1/(1 - z)$. Then we have $f^{(n)}(z) = n!/(1 - z)^{n+1}$, $f^{(n)}(0) = n!$. Hence the Maclaurin expansion of $1/(1 - z)$ is the geometric series

$$\frac{1}{1-z} = \sum_{n=0}^{\infty} z^n = 1 + z + z^2 + \cdots \qquad (|z| < 1). \tag{11}$$

$f(z)$ is singular at $z = 1$; this point lies on the circle of convergence. ■

EXAMPLE 2 **Exponential Function**

We know that the exponential function e^z (Sec. 13.5) is analytic for all z, and $(e^z)' = e^z$. Hence from (1) with $z_0 = 0$ we obtain the Maclaurin series

$$e^z = \sum_{n=0}^{\infty} \frac{z^n}{n!} = 1 + z + \frac{z^2}{2!} + \cdots. \tag{12}$$

This series is also obtained if we replace x in the familiar Maclaurin series of e^x by z.

Furthermore, by setting $z = iy$ in (12) and separating the series into the real and imaginary parts (see Theorem 2, Sec. 15.1) we obtain

$$e^{iy} = \sum_{n=0}^{\infty} \frac{(iy)^n}{n!} = \sum_{k=0}^{\infty} (-1)^k \frac{y^{2k}}{(2k)!} + i \sum_{k=0}^{\infty} (-1)^k \frac{y^{2k+1}}{(2k+1)!}.$$

Since the series on the right are the familiar Maclaurin series of the real functions $\cos y$ and $\sin y$, this shows that we have rediscovered the **Euler formula**

$$e^{iy} = \cos y + i \sin y. \tag{13}$$

Indeed, one may use (12) for ***defining*** e^z and derive from (12) the basic properties of e^z. For instance, the differentiation formula $(e^z)' = e^z$ follows readily from (12) by termwise differentiation. ■

EXAMPLE 3 **Trigonometric and Hyperbolic Functions**

By substituting (12) into (1) of Sec. 13.6 we obtain

$$\cos z = \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n}}{(2n)!} = 1 - \frac{z^2}{2!} + \frac{z^4}{4!} - + \cdots$$

(14)

$$\sin z = \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n+1}}{(2n+1)!} = z - \frac{z^3}{3!} + \frac{z^5}{5!} - + \cdots.$$

When $z = x$ these are the familiar Maclaurin series of the real functions $\cos x$ and $\sin x$. Similarly, by substituting (12) into (11), Sec. 13.6, we obtain

$$\cosh z = \sum_{n=0}^{\infty} \frac{z^{2n}}{(2n)!} = 1 + \frac{z^2}{2!} + \frac{z^4}{4!} + \cdots$$

(15)

$$\sinh z = \sum_{n=0}^{\infty} \frac{z^{2n+1}}{(2n+1)!} = z + \frac{z^3}{3!} + \frac{z^5}{5!} \cdots. \qquad \blacksquare$$

EXAMPLE 4 **Logarithm**

From (1) it follows that

$$\text{(16)} \qquad \operatorname{Ln}(1+z) = z - \frac{z^2}{2} + \frac{z^3}{3} - + \cdots \qquad (|z| < 1).$$

Replacing z by $-z$ and multiplying both sides by -1, we get

$$\text{(17)} \qquad -\operatorname{Ln}(1-z) = \operatorname{Ln}\frac{1}{1-z} = z + \frac{z^2}{2} + \frac{z^3}{3} + \cdots \qquad (|z| < 1).$$

By adding both series we obtain

$$\text{(18)} \qquad \operatorname{Ln}\frac{1+z}{1-z} = 2\left(z + \frac{z^3}{3} + \frac{z^5}{5} + \cdots\right) \qquad (|z| < 1). \qquad \blacksquare$$

Practical Methods

The following examples show ways of obtaining Taylor series more quickly than by the use of the coefficient formulas. Regardless of the method used, the result will be the same. This follows from the uniqueness (see Theorem 1).

EXAMPLE 5 **Substitution**

Find the Maclaurin series of $f(z) = 1/(1+z^2)$.

Solution. By substituting $-z^2$ for z in (11) we obtain

$$\text{(19)} \qquad \frac{1}{1+z^2} = \frac{1}{1-(-z^2)} = \sum_{n=0}^{\infty} (-z^2)^n = \sum_{n=0}^{\infty} (-1)^n z^{2n} = 1 - z^2 + z^4 - z^6 + \cdots \qquad (|z| < 1). \qquad \blacksquare$$

EXAMPLE 6 Integration

Find the Maclaurin series of $f(z) = \arctan z$.

Solution. We have $f'(z) = 1/(1 + z^2)$. Integrating (19) term by term and using $f(0) = 0$ we get

$$\arctan z = \sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1} z^{2n+1} = z - \frac{z^3}{3} + \frac{z^5}{5} - + \cdots \qquad (|z| < 1);$$

this series represents the principal value of $w = u + iv = \arctan z$ defined as that value for which $|u| < \pi/2$. ∎

EXAMPLE 7 Development by Using the Geometric Series

Develop $1/(c - z)$ in powers of $z - z_0$, where $c - z_0 \neq 0$.

Solution. This was done in the proof of Theorem 1, where $c = z^*$. The beginning was simple algebra and then the use of (11) with z replaced by $(z - z_0)/(c - z_0)$:

$$\frac{1}{c-z} = \frac{1}{c - z_0 - (z - z_0)} = \frac{1}{(c - z_0)\left(1 - \dfrac{z - z_0}{c - z_0}\right)} = \frac{1}{c - z_0} \sum_{n=0}^{\infty} \left(\frac{z - z_0}{c - z_0}\right)^n$$

$$= \frac{1}{c - z_0}\left(1 + \frac{z - z_0}{c - z_0} + \left(\frac{z - z_0}{c - z_0}\right)^2 + \cdots\right).$$

This series converges for

$$\left|\frac{z - z_0}{c - z_0}\right| < 1, \qquad \text{that is,} \qquad |z - z_0| < |c - z_0|.$$ ∎

EXAMPLE 8 Binomial Series, Reduction by Partial Fractions

Find the Taylor series of the following function with center $z_0 = 1$.

$$f(z) = \frac{2z^2 + 9z + 5}{z^3 + z^2 - 8z - 12}$$

Solution. We develop $f(z)$ in partial fractions and the first fraction in a **binomial series**

(20)
$$\frac{1}{(1+z)^m} = (1+z)^{-m} = \sum_{n=0}^{\infty} \binom{-m}{n} z^n$$
$$= 1 - mz + \frac{m(m+1)}{2!} z^2 - \frac{m(m+1)(m+2)}{3!} z^3 + \cdots$$

with $m = 2$ and the second fraction in a geometric series, and then add the two series term by term. This gives

$$f(z) = \frac{1}{(z+2)^2} + \frac{2}{z-3} = \frac{1}{[3 + (z-1)]^2} - \frac{2}{2 - (z-1)} = \frac{1}{9}\left(\frac{1}{[1 + \frac{1}{3}(z-1)]^2}\right) - \frac{1}{1 - \frac{1}{2}(z-1)}$$

$$= \frac{1}{9}\sum_{n=0}^{\infty} \binom{-2}{n} \left(\frac{z-1}{3}\right)^n - \sum_{n=0}^{\infty} \left(\frac{z-1}{2}\right)^n = \sum_{n=0}^{\infty} \left[\frac{(-1)^n(n+1)}{3^{n+2}} - \frac{1}{2^n}\right](z-1)^n$$

$$= -\frac{8}{9} - \frac{31}{54}(z-1) - \frac{23}{108}(z-1)^2 - \frac{275}{1944}(z-1)^3 - \cdots.$$

We see that the first series converges for $|z - 1| < 3$ and the second for $|z - 1| < 2$. This had to be expected because $1/(z + 2)^2$ is singular at -2 and $2/(z - 3)$ at 3, and these points have distance 3 and 2, respectively, from the center $z_0 = 1$. Hence the whole series converges for $|z - 1| < 2$. ∎

PROBLEM SET 15.4

1–12 TAYLOR AND MACLAURIN SERIES

Find the Taylor or Maclaurin series of the given function with the given point as center and determine the radius of convergence.

1. e^{-2z}, 0
2. $1/(1 - z^3)$, 0
3. e^z, $-2i$
4. $\cos^2 z$, 0
5. $\sin z$, $\pi/2$
6. $1/z$, 1
7. $1/(1 - z)$, i
8. $\operatorname{Ln}(1 - z)$, i
9. $e^{-z^2/2}$, 0
10. $e^{z^2}\int_0^z e^{-t^2}\,dt$, 0
11. $z^6 - z^4 + z^2 - 1$, 1
12. $\sinh(z - 2i)$, $2i$

13–16 HIGHER TRANSCENDENTAL FUNCTIONS

Find the Maclaurin series by termwise integrating the integrand. (The integrals cannot be evaluated by the usual methods of calculus. They define the **error function** erf z, **sine integral** Si(z), and **Fresnel integrals**[4] S(z) and C(z), which occur in statistics, heat conduction, optics, and other applications. These are special so-called higher transcendental functions.)

13. $\operatorname{erf} z = \dfrac{2}{\sqrt{\pi}}\displaystyle\int_0^z e^{-t^2}\,dt$
14. $\operatorname{Si}(z) = \displaystyle\int_0^z \frac{\sin t}{t}\,dt$
15. $\operatorname{S}(z) = \displaystyle\int_0^z \sin t^2\,dt$
16. $\operatorname{C}(z) = \displaystyle\int_0^z \cos t^2\,dt$

17. **CAS PROJECT. sec, tan, arcsin. (a) Euler numbers.** The Maclaurin series

$$\sec z = E_0 - \frac{E_2}{2!}z^2 + \frac{E_4}{4!}z^4 - + \cdots \tag{21}$$

defines the *Euler numbers* E_{2n}. Show that $E_0 = 1$, $E_2 = -1$, $E_4 = 5$, $E_6 = -61$. Write a program that computes the E_{2n} from the coefficient formula in (1) or extracts them as a list from the series. (For tables see Ref. [GR1], p. 810, listed in App. 1.)

(b) Bernoulli numbers. The Maclaurin series

$$\frac{z}{e^z - 1} = 1 + B_1 z + \frac{B_2}{2!}z^2 + \frac{B_3}{3!}z^3 + \cdots \tag{22}$$

defines the *Bernoulli numbers* B_n. Using undetermined coefficients, show that

$$B_1 = -\frac{1}{2}, \quad B_2 = \frac{1}{6}, \quad B_3 = 0, \quad B_4 = -\frac{1}{30}, \quad B_5 = 0, \quad B_6 = \frac{1}{42}, \cdots. \tag{23}$$

Write a program for computing B_n.

(c) Tangent. Using (1), (2), Sec. 13.6, and (22), show that $\tan z$ has the following Maclaurin series and calculate from it a table of $B_0, \cdots, B_{20}$:

$$\tan z = \frac{2i}{e^{2iz} - 1} - \frac{4i}{e^{4iz} - 1} - i = \sum_{n=1}^{\infty} (-1)^{n-1}\frac{2^{2n}(2^{2n} - 1)}{(2n)!}B_{2n}z^{2n-1}. \tag{24}$$

18. **(Inverse sine)** Developing $1/\sqrt{1 - z^2}$ and integrating, show that

$$\arcsin z = z + \left(\frac{1}{2}\right)\frac{z^3}{3} + \left(\frac{1 \cdot 3}{2 \cdot 4}\right)\frac{z^5}{5} + \left(\frac{1 \cdot 3 \cdot 5}{2 \cdot 4 \cdot 6}\right)\frac{z^7}{7} + \cdots \quad (|z| < 1).$$

Show that this series represents the principal value of arcsin z (defined in Team Project 30, Sec. 13.7).

19. **(Undetermined coefficients)** Using the relation $\sin z = \tan z \cos z$ and the Maclaurin series of $\sin z$ and $\cos z$, find the first four nonzero terms of the Maclaurin series of $\tan z$. (Show the details.)

20. **TEAM PROJECT. Properties from Maclaurin Series.** Clearly, from series we can compute function values. In this project we show that properties of functions can often be discovered from their Taylor or Maclaurin series. Using suitable series, prove the following.

(a) The formulas for the derivatives of e^z, $\cos z$, $\sin z$, $\cosh z$, $\sinh z$, and $\operatorname{Ln}(1 + z)$

(b) $\frac{1}{2}(e^{iz} + e^{-iz}) = \cos z$

(c) $\sin z \neq 0$ for all pure imaginary $z = iy \neq 0$

[4] AUGUSTIN FRESNEL (1788–1827), French physicist and engineer, known for his work in optics.

15.5 Uniform Convergence. *Optional*

We know that power series are *absolutely convergent* (Sec. 15.2, Theorem 1) and, as another basic property, we now show that they are *uniformly convergent.* Since uniform convergence is of general importance, for instance, in connection with termwise integration of series, we shall discuss it quite thoroughly.

To define uniform convergence, we consider a series whose terms are any complex functions $f_0(z), f_1(z), \cdots$:

$$\sum_{m=0}^{\infty} f_m(z) = f_0(z) + f_1(z) + f_2(z) + \cdots. \tag{1}$$

(This includes power series as a special case in which $f_m(z) = a_m (z - z_0)^m$.) We assume that the series (1) converges for all z in some region G. We call its sum $s(z)$ and its nth partial sum $s_n(z)$; thus

$$s_n(z) = f_0(z) + f_1(z) + \cdots + f_n(z).$$

Convergence in G means the following. If we pick a $z = z_1$ in G, then, by the definition of convergence at z_1, for given $\epsilon > 0$ we can find an $N_1(\epsilon)$ such that

$$|s(z_1) - s_n(z_1)| < \epsilon \qquad \text{for all } n > N_1(\epsilon).$$

If we pick a z_2 in G, keeping ϵ as before, we can find an $N_2(\epsilon)$ such that

$$|s(z_2) - s_n(z_2)| < \epsilon \qquad \text{for all } n > N_2(\epsilon),$$

and so on. Hence, given an $\epsilon > 0$, to each z in G there corresponds a number $N_z(\epsilon)$. This number tells us how many terms we need (what s_n we need) at a z to make $|s(z) - s_n(z)|$ smaller than ϵ. Thus this number $N_z(\epsilon)$ measures the speed of convergence.

Small $N_z(\epsilon)$ means rapid convergence, large $N_z(\epsilon)$ means slow convergence at the point z considered. Now, if we can find an $N(\epsilon)$ larger than all these $N_z(\epsilon)$ for all z in G, we say that the convergence of the series (1) in G is *uniform.* Hence this basic concept is defined as follows.

DEFINITION

Uniform Convergence

A series (1) with sum $s(z)$ is called **uniformly convergent** in a region G if for every $\epsilon > 0$ we can find an $N = N(\epsilon)$, ***not depending on z,*** such that

$$|s(z) - s_n(z)| < \epsilon \qquad \text{for all } n > N(\epsilon) \textbf{ and all } z \textbf{ in } G.$$

Uniformity of convergence is thus a property that always refers to an *infinite set* in the z-plane, that is, a set consisting of infinitely many points.

15.5 Uniform Convergence. *Optional*

We know that power series are *absolutely convergent* (Sec. 15.2, Theorem 1) and, as another basic property, we now show that they are *uniformly convergent.* Since uniform convergence is of general importance, for instance, in connection with termwise integration of series, we shall discuss it quite thoroughly.

To define uniform convergence, we consider a series whose terms are any complex functions $f_0(z), f_1(z), \cdots$:

$$\sum_{m=0}^{\infty} f_m(z) = f_0(z) + f_1(z) + f_2(z) + \cdots. \tag{1}$$

(This includes power series as a special case in which $f_m(z) = a_m (z - z_0)^m$.) We assume that the series (1) converges for all z in some region G. We call its sum $s(z)$ and its nth partial sum $s_n(z)$; thus

$$s_n(z) = f_0(z) + f_1(z) + \cdots + f_n(z).$$

Convergence in G means the following. If we pick a $z = z_1$ in G, then, by the definition of convergence at z_1, for given $\epsilon > 0$ we can find an $N_1(\epsilon)$ such that

$$|s(z_1) - s_n(z_1)| < \epsilon \qquad \text{for all } n > N_1(\epsilon).$$

If we pick a z_2 in G, keeping ϵ as before, we can find an $N_2(\epsilon)$ such that

$$|s(z_2) - s_n(z_2)| < \epsilon \qquad \text{for all } n > N_2(\epsilon),$$

and so on. Hence, given an $\epsilon > 0$, to each z in G there corresponds a number $N_z(\epsilon)$. This number tells us how many terms we need (what s_n we need) at a z to make $|s(z) - s_n(z)|$ smaller than ϵ. Thus this number $N_z(\epsilon)$ measures the speed of convergence.

Small $N_z(\epsilon)$ means rapid convergence, large $N_z(\epsilon)$ means slow convergence at the point z considered. Now, if we can find an $N(\epsilon)$ larger than all these $N_z(\epsilon)$ for all z in G, we say that the convergence of the series (1) in G is *uniform.* Hence this basic concept is defined as follows.

DEFINITION

Uniform Convergence

A series (1) with sum $s(z)$ is called **uniformly convergent** in a region G if for every $\epsilon > 0$ we can find an $N = N(\epsilon)$, ***not depending on z,*** such that

$$|s(z) - s_n(z)| < \epsilon \qquad \text{for all } n > N(\epsilon) \textbf{ and all } z \textbf{ in } G.$$

Uniformity of convergence is thus a property that always refers to an *infinite set* in the z-plane, that is, a set consisting of infinitely many points.

EXAMPLE 1 **Geometric Series**

Show that the geometric series $1 + z + z^2 + \cdots$ is (a) uniformly convergent in any closed disk $|z| \leqq r < 1$, (b) not uniformly convergent in its whole disk of convergence $|z| < 1$.

Solution. (a) For z in that closed disk we have $|1 - z| \geqq 1 - r$ (sketch it). This implies that $1/|1 - z| \leqq 1/(1 - r)$. Hence (remember (8) in Sec. 15.4 with $q = z$)

$$|s(z) - s_n(z)| = \left| \sum_{m=n+1}^{\infty} z^m \right| = \left| \frac{z^{n+1}}{1 - z} \right| \leqq \frac{r^{n+1}}{1 - r}.$$

Since $r < 1$, we can make the right side as small as we want by choosing n large enough, and since the right side does not depend on z (in the closed disk considered), this means that the convergence is uniform.

(b) For given real K (no matter how large) and n we can always find a z in the disk $|z| < 1$ such that

$$\left| \frac{z^{n+1}}{1 - z} \right| = \frac{|z|^{n+1}}{|1 - z|} > K,$$

simply by taking z close enough to 1. Hence no single $N(\epsilon)$ will suffice to make $|s(z) - s_n(z)|$ smaller than a given $\epsilon > 0$ *throughout the whole disk.* By definition, this shows that the convergence of the geometric series in $|z| < 1$ is not uniform. ∎

This example suggests that *for a power series, the uniformity of convergence may at most be disturbed near the circle of convergence.* This is true:

THEOREM 1

Uniform Convergence of Power Series

A power series

$$(2) \qquad \sum_{m=0}^{\infty} a_m(z - z_0)^m$$

with a nonzero radius of convergence R is uniformly convergent in every circular disk $|z - z_0| \leqq r$ of radius $r < R$.

PROOF For $|z - z_0| \leqq r$ and any positive integers n and p we have

$$(3) \quad |a_{n+1}(z - z_0)^{n+1} + \cdots + a_{n+p}(z - z_0)^{n+p}| \leqq |a_{n+1}|r^{n+1} + \cdots + |a_{n+p}|r^{n+p}.$$

Now (2) converges absolutely if $|z - z_0| = r < R$ (by Theorem 1 in Sec. 15.2). Hence it follows from the Cauchy convergence principle (Sec. 15.1) that, an $\epsilon > 0$ being given, we can find an $N(\epsilon)$ such that

$$|a_{n+1}|r^{n+1} + \cdots + |a_{n+p}|r^{n+p} < \epsilon \qquad \text{for } n > N(\epsilon) \quad \text{and} \quad p = 1, 2, \cdots.$$

From this and (3) we obtain

$$|a_{n+1}(z - z_0)^{n+1} + \cdots + a_{n+p}(z - z_0)^{n+p}| < \epsilon$$

for all z in the disk $|z - z_0| \leqq r$, every $n > N(\epsilon)$, and every $p = 1, 2, \cdots$. Since $N(\epsilon)$ is independent of z, this shows uniform convergence, and the theorem is proved. ∎

Theorem 1 meets with our immediate need and concern, which is power series. The remainder of this section should provide a deeper understanding of the concept of uniform convergence in connection with arbitrary series of variable terms.

Properties of Uniformly Convergent Series

Uniform convergence derives its main importance from two facts:

1. If a series of *continuous* terms is uniformly convergent, its sum is also continuous (Theorem 2, below).

2. Under the same assumptions, termwise integration is permissible (Theorem 3).

This raises two questions:

1. How can a converging series of continuous terms manage to have a discontinuous sum? (Example 2)

2. How can something go wrong in termwise integration? (Example 3)

Another natural question is:

3. What is the relation between absolute convergence and uniform convergence? The surprising answer: none. (Example 5)

These are the ideas we shall discuss.

If we add *finitely many* continuous functions, we get a continuous function as their sum. Example 2 will show that this is no longer true for an infinite series, even if it converges absolutely. However, if it converges *uniformly,* this cannot happen, as follows.

THEOREM 2

Continuity of the Sum

Let the series

$$\sum_{m=0}^{\infty} f_m(z) = f_0(z) + f_1(z) + \cdots$$

be uniformly convergent in a region G. Let F(z) be its sum. Then if each term $f_m(z)$ is continuous at a point z_1 in G, the function F(z) is continuous at z_1.

PROOF Let $s_n(z)$ be the nth partial sum of the series and $R_n(z)$ the corresponding remainder:

$$s_n = f_0 + f_1 + \cdots + f_n, \qquad R_n = f_{n+1} + f_{n+2} + \cdots.$$

Since the series converges uniformly, for a given $\epsilon > 0$ we can find an $N = N(\epsilon)$ such that

$$|R_N(z)| < \frac{\epsilon}{3} \qquad \text{for all } z \text{ in } G.$$

Since $s_N(z)$ is a sum of finitely many functions that are continuous at z_1, this sum is continuous at z_1. Therefore, we can find a $\delta > 0$ such that

$$|s_N(z) - s_N(z_1)| < \frac{\epsilon}{3} \quad \text{for all } z \text{ in } G \text{ for which } |z - z_1| < \delta.$$

Using $F = s_N + R_N$ and the triangle inequality (Sec. 13.2), for these z we thus obtain

$$\begin{aligned} |F(z) - F(z_1)| &= |s_N(z) + R_N(z) - [s_N(z_1) + R_N(z_1)]| \\ &\leqq |s_N(z) - s_N(z_1)| + |R_N(z)| + |R_N(z_1)| < \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon. \end{aligned}$$

This implies that $F(z)$ is continuous at z_1, and the theorem is proved. ∎

EXAMPLE 2 Series of Continuous Terms with a Discontinuous Sum

Consider the series

$$x^2 + \frac{x^2}{1+x^2} + \frac{x^2}{(1+x^2)^2} + \frac{x^2}{(1+x^2)^3} + \cdots \qquad (x \text{ real}).$$

This is a geometric series with $q = 1/(1+x^2)$ times a factor x^2. Its nth partial sum is

$$s_n(x) = x^2\left[1 + \frac{1}{1+x^2} + \frac{1}{(1+x^2)^2} + \cdots + \frac{1}{(1+x^2)^n}\right].$$

We now use the trick by which one finds the sum of a geometric series, namely, we multiply $s_n(x)$ by $-q = -1/(1+x^2)$,

$$-\frac{1}{1+x^2}\,s_n(x) = -x^2\left[\frac{1}{1+x^2} + \cdots + \frac{1}{(1+x^2)^n} + \frac{1}{(1+x^2)^{n+1}}\right].$$

Adding this to the previous formula, simplifying on the left, and canceling most terms on the right, we obtain

$$\frac{x^2}{1+x^2}\,s_n(x) = x^2\left[1 - \frac{1}{(1+x^2)^{n+1}}\right],$$

thus

$$s_n(x) = 1 + x^2 - \frac{1}{(1+x^2)^n}.$$

The exciting Fig. 365 "explains" what is going on. We see that if $x \neq 0$, the sum is

$$s(x) = \lim_{n\to\infty} s_n(x) = 1 + x^2,$$

but for $x = 0$ we have $s_n(0) = 1 - 1 = 0$ for all n, hence $s(0) = 0$. So we have the surprising fact that the sum is discontinuous (at $x = 0$), although all the terms are continuous and the series converges even absolutely (its terms are nonnegative, thus equal to their absolute value!).

Theorem 2 now tells us that the convergence cannot be uniform in an interval containing $x = 0$. We can also verify this directly. Indeed, for $x \neq 0$ the remainder has the absolute value

$$|R_n(x)| = |s(x) - s_n(x)| = \frac{1}{(1+x^2)^n}$$

and we see that for a given ϵ (< 1) we cannot find an N depending only on ϵ such that $|R_n| < \epsilon$ for all $n > N(\epsilon)$ and all x, say, in the interval $0 \leqq x \leqq 1$. ∎

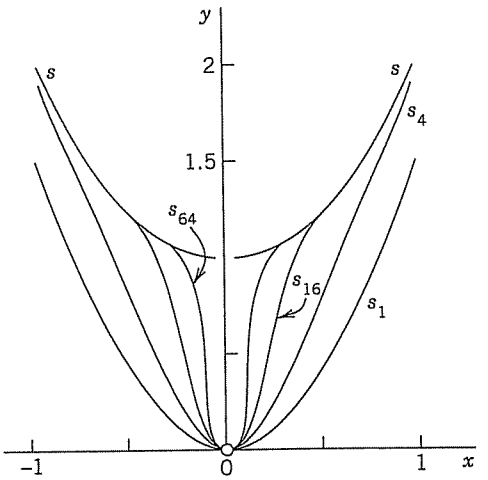


Fig. 365. Partial sums in Example 2

Termwise Integration

This is our second topic in connection with uniform convergence, and we begin with an example to become aware of the danger of just blindly integrating term-by-term.

EXAMPLE 3 **Series for which Termwise Integration is Not Permissible**

Let $u_m(x) = mxe^{-mx^2}$ and consider the series

$$\sum_{m=0}^{\infty} f_m(x) \qquad \text{where} \qquad f_m(x) = u_m(x) - u_{m-1}(x)$$

in the interval $0 \leqq x \leqq 1$. The nth partial sum is

$$s_n = u_1 - u_0 + u_2 - u_1 + \cdots + u_n - u_{n-1} = u_n - u_0 = u_n.$$

Hence the series has the sum $F(x) = \lim_{n\to\infty} s_n(x) = \lim_{n\to\infty} u_n(x) = 0 \quad (0 \leqq x \leqq 1)$. From this we obtain

$$\int_0^1 F(x)\,dx = 0.$$

On the other hand, by integrating term by term and using $f_1 + f_2 + \cdots + f_n = s_n$, we have

$$\sum_{m=1}^{\infty} \int_0^1 f_m(x)\,dx = \lim_{n\to\infty} \sum_{m=1}^{n} \int_0^1 f_m(x)\,dx = \lim_{n\to\infty} \int_0^1 s_n(x)\,dx.$$

Now $s_n = u_n$ and the expression on the right becomes

$$\lim_{n\to\infty} \int_0^1 u_n(x)\,dx = \lim_{n\to\infty} \int_0^1 nxe^{-nx^2}\,dx = \lim_{n\to\infty} \frac{1}{2}\,(1 - e^{-n}) = \frac{1}{2},$$

but not 0. This shows that the series under consideration cannot be integrated term by term from $x = 0$ to $x = 1$. ▩

The series in Example 3 is not uniformly convergent in the interval of integration, and we shall now prove that in the case of a uniformly convergent series of continuous functions we may integrate term by term.

THEOREM 3 **Termwise Integration**

Let

$$F(z) = \sum_{m=0}^{\infty} f_m(z) = f_0(z) + f_1(z) + \cdots$$

be a uniformly convergent series of continuous functions in a region G. Let C be any path in G. Then the series

$$\sum_{m=0}^{\infty} \int_C f_m(z)\,dz = \int_C f_0(z)\,dz + \int_C f_1(z)\,dz + \cdots \tag{4}$$

is convergent and has the sum $\int_C F(z)\,dz$.

PROOF From Theorem 2 it follows that $F(z)$ is continuous. Let $s_n(z)$ be the nth partial sum of the given series and $R_n(z)$ the corresponding remainder. Then $F = s_n + R_n$ and by integration,

$$\int_C F(z)\,dz = \int_C s_n(z)\,dz + \int_C R_n(z)\,dz.$$

Let L be the length of C. Since the given series converges uniformly, for every given $\epsilon > 0$ we can find a number N such that $|R_n(z)| < \epsilon/L$ for all $n > N$ and all z in G. By applying the ML-inequality (Sec. 14.1) we thus obtain

$$\left|\int_C R_n(z)\,dz\right| < \frac{\epsilon}{L}\,L = \epsilon \qquad \text{for all } n > N.$$

Since $R_n = F - s_n$, this means that

$$\left|\int_C F(z)\,dz - \int_C s_n(z)\,dz\right| < \epsilon \qquad \text{for all } n > N.$$

Hence, the series (4) converges and has the sum indicated in the theorem. ∎

Theorems 2 and 3 characterize the two most important properties of uniformly convergent series. Also, since differentiation and integration are inverse processes, Theorem 3 implies

THEOREM 4

Termwise Differentiation

Let the series $f_0(z) + f_1(z) + f_2(z) + \cdots$ be convergent in a region G and let $F(z)$ be its sum. Suppose that the series $f_0'(z) + f_1'(z) + f_2'(z) + \cdots$ converges uniformly in G and its terms are continuous in G. Then

$$F'(z) = f_0'(z) + f_1'(z) + f_2'(z) + \cdots \qquad \textit{for all } z \textit{ in } G.$$

Test for Uniform Convergence

Uniform convergence is usually proved by the following comparison test.

THEOREM 5

Weierstrass[5] M-Test for Uniform Convergence

Consider a series of the form (1) *in a region G of the z-plane. Suppose that one can find a convergent series of constant terms,*

$$M_0 + M_1 + M_2 + \cdots, \tag{5}$$

such that $|f_m(z)| \leqq M_m$ for all z in G and every $m = 0, 1, \cdots$. Then (1) *is uniformly convergent in G.*

[5]KARL WEIERSTRASS (1815–1897), great German mathematician, whose lifework was the development of complex analysis based on the concept of power series (see the footnote in Sec. 13.4). He also made basic contributions to the calculus, the calculus of variations, approximation theory, and differential geometry. He obtained the concept of uniform convergence in 1841 (published 1894, *sic!*); the first publication on the concept was by G. G. STOKES (see Sec 10.9) in 1847.

The simple proof is left to the student (Team Project 18).

EXAMPLE 4 **Weierstrass *M*-Test**

Does the following series converge uniformly in the disk $|z| \leqq 1$?

$$\sum_{m=1}^{\infty} \frac{z^m + 1}{m^2 + \cosh m|z|}$$

Solution. Uniform convergence follows by the Weierstrass M-test and the convergence of $\Sigma 1/m^2$ (see Sec. 15.1, in the proof of Theorem 8) because

$$\left| \frac{z^m + 1}{m^2 + \cosh m|z|} \right| \leqq \frac{|z|^m + 1}{m^2}$$

$$\leqq \frac{2}{m^2}.$$

■

No Relation Between Absolute and Uniform Convergence

We finally show the surprising fact that there are series that converge absolutely but not uniformly, and others that converge uniformly but not absolutely, so that there is no relation between the two concepts.

EXAMPLE 5 **No Relation Between Absolute and Uniform Convergence**

The series in Example 2 converges absolutely but not uniformly, as we have shown. On the other hand, the series

$$\sum_{m=1}^{\infty} \frac{(-1)^{m-1}}{x^2 + m} = \frac{1}{x^2 + 1} - \frac{1}{x^2 + 2} + \frac{1}{x^2 + 3} - + \cdots \qquad (x \text{ real})$$

converges uniformly on the whole real line but not absolutely.

Proof. By the familiar Leibniz test of calculus (see App. A3.3) the remainder R_n does not exceed its first term in absolute value, since we have a series of alternating terms whose absolute values form a monotone decreasing sequence with limit zero. Hence given $\epsilon > 0$, for all x we have

$$|R_n(x)| \leqq \frac{1}{x^2 + n + 1} < \frac{1}{n} < \epsilon \qquad \text{if } n > N(\epsilon) \geqq \frac{1}{\epsilon}.$$

This proves uniform convergence, since $N(\epsilon)$ does not depend on x.

The convergence is not absolute because for any *fixed* x we have

$$\left| \frac{(-1)^{m-1}}{x^2 + m} \right| = \frac{1}{x^2 + m}$$

$$> \frac{k}{m}$$

where k is a suitable constant, and $k\Sigma 1/m$ diverges. ■

PROBLEM SET 15.5

1–8 UNIFORM CONVERGENCE

Prove that the given series converges uniformly in the indicated region.

1. $\sum_{n=0}^{\infty} (z - 2i)^{2n}, \quad |z - 2i| \leqq 0.999$

2. $\sum_{n=0}^{\infty} \frac{z^{2n+1}}{(2n+1)!}, \quad |z| \leqq 10^{10}$

3. $\sum_{n=0}^{\infty} \frac{\pi^n}{n^4} z^{2n}, \quad |z| \leqq 0.56$

4. $\sum_{n=1}^{\infty} \frac{\sin^n |\pi z|}{n(n+1)}$, $|z| \leqq 10^5$

5. $\sum_{n=1}^{\infty} \frac{z^n}{n^2}$, $|z| \leqq 1$

6. $\sum_{n=1}^{\infty} \frac{z^n}{n^2 \cosh n|z|}$, $|z| \leqq 1$

7. $\sum_{n=0}^{\infty} \frac{\tanh^n |z|}{n^2+1}$ $|z| \leqq 10^{10}$

8. $\sum_{n=1}^{\infty} \frac{\cos n|z|}{n^2}$, $|z| \leqq 10^{20}$

9–16 POWER SERIES

Find the region of uniform convergence. (Give reason.)

9. $\sum_{n=0}^{\infty} \frac{(z+1-2i)^n}{4^n}$

10. $\sum_{n=0}^{\infty} \frac{(z-i)^{2n}}{(2n)!}$

11. $\sum_{n=1}^{\infty} \frac{(-1)^n}{2^n n} z^n$

12. $\sum_{n=2}^{\infty} \binom{n}{2} (2z-i)^n$

13. $\sum_{n=1}^{\infty} \frac{n!}{n^2} z^n$

14. $\sum_{n=1}^{\infty} (3^n \tanh n) z^{2n}$

15. $\sum_{n=1}^{\infty} \frac{z^{2n}}{5^n n^2}$

16. $\sum_{n=0}^{\infty} \frac{(-1)^n z^{2n}}{(2n)!}$

17. **CAS PROJECT. Graphs of Partial Sums. (a) Figure 365.** Produce this exciting figure using your software and adding further curves, say, those of s_{256}, s_{1024}, etc.

(b) Power series. Study the nonuniformity of convergence experimentally by plotting partial sums near the endpoints of the convergence interval for real $z = x$.

18. **TEAM PROJECT. Uniform Convergence. (a) Weierstrass *M*-test.** Give a proof.

(b) Termwise differentiation. Derive Theorem 4 from Theorem 3.

(c) Subregions. Prove that uniform convergence of a series in a region G implies uniform convergence in any portion of G. Is the converse true?

(d) Example 2. Find the precise region of convergence of the series in Example 2 with x replaced by a complex variable z.

(e) Figure 366. Show that $x^2 \sum_{m=1}^{\infty} (1+x^2)^{-m} = 1$ if $x \neq 0$ and 0 if $x = 0$. Verify by computation that the partial sums s_1, s_2, s_3, look as shown in Fig. 366.

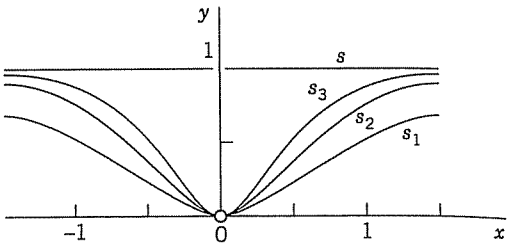


Fig. 366. Sum s and partial sums in Team Project 18(e)

19–20 HEAT EQUATION

Show that (9) in Sec. 12.5 with coefficients (10) is a solution of the heat equation for $t > 0$, assuming that $f(x)$ is continuous on the interval $0 \leqq x \leqq L$ and has one-sided derivatives at all interior points of that interval. Proceed as follows.

19. Show that $|B_n|$ is bounded, say $|B_n| < K$ for all n. Conclude that

$$|u_n| < Ke^{-\lambda_n^2 t_0} \qquad \text{if} \qquad t \geqq t_0 > 0$$

and, by the Weierstrass test, the series (9) converges uniformly with respect to x and t for $t \geqq t_0$, $0 \leqq x \leqq L$. Using Theorem 2, show that $u(x, t)$ is continuous for $t \geqq t_0$ and thus satisfies the boundary conditions (2) for $t \geqq t_0$.

20. Show that $|\partial u_n/\partial t| < \lambda_n^2 Ke^{-\lambda_n^2 t_0}$ if $t \geqq t_0$ and the series of the expressions on the right converges, by the ratio test. Conclude from this, the Weierstrass test, and Theorem 4 that the series (9) can be differentiated term by term with respect to t and the resulting series has the sum $\partial u/\partial t$. Show that (9) can be differentiated twice with respect to x and the resulting series has the sum $\partial^2 u/\partial x^2$. Conclude from this and the result to Prob. 19 that (9) is a solution of the heat equation for all $t \geqq t_0$. (The proof that (9) satisfies the given initial condition can be found in Ref. [C10] listed in App. 1.)

CHAPTER 15 REVIEW QUESTIONS AND PROBLEMS

1. What are power series? Why are these series very important in complex analysis?
2. State from memory the ratio test, the root test, and the Cauchy–Hadamard formula for the radius of convergence.
3. What is absolute convergence? Conditional convergence? Uniform convergence?
4. What do you know about the convergence of power series?
5. What is a Taylor series? What was the idea of obtaining it from Cauchy's integral formula?
6. Give examples of practical methods for obtaining Taylor series.
7. What have power series to do with analytic functions?

8. Can properties of functions be discovered from their Maclaurin series? If so, give examples.
9. Make a list of Maclaurin series of e^z, $\cos z$, $\sin z$, $\cosh z$, $\sinh z$, $\operatorname{Ln}(1 - z)$ from memory.
10. What do you know about adding and multiplying power series?

11–20 RADIUS OF CONVERGENCE

Find the radius of convergence. Can you identify the sum as a familiar function in some of the problems? (Show the details of your work.)

11. $\sum_{n=0}^{\infty} \frac{(3z)^n}{n!}$
12. $\sum_{n=1}^{\infty} \frac{(-2)^{n+1}}{2n} z^n$
13. $\sum_{n=0}^{\infty} \frac{z^{2n+1}}{2n+1}$
14. $\sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} z^n$
15. $\sum_{n=1}^{\infty} \frac{n^5}{n!} (z - 3i)^{2n}$
16. $\sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} (z - 2)^{2n+1}$
17. $\sum_{n=0}^{\infty} \pi^n (z - 2i)^{2n}$
18. $\sum_{n=0}^{\infty} \frac{(2z)^{2n}}{(2n)!}$
19. $\sum_{n=1}^{\infty} \frac{3^n}{n^{10}} z^n$
20. $\sum_{n=0}^{\infty} \frac{(z - i)^n}{(3 + 4i)^n}$

21–30 TAYLOR AND MACLAURIN SERIES

Find the Taylor or Maclaurin series with the given point as center and determine the radius of convergence. (Show details.)

21. e^z, πi
22. $\operatorname{Ln} z$, 2
23. $1/(1 - z)$, -1
24. $1/(4 - 3z)$, $1 + i$
25. $1/(1 - z)^3$, 0
26. $1/z^2$, i
27. $1/z$, $-i$
28. $\int_0^z t^{-1}(e^t - 1)\, dt$, 0
29. $\cos z$, $\frac{1}{2}\pi$
30. $\sin^2 z$, 0
31. Does every function $f(z)$ have a Taylor series?
32. Does there exist a Taylor series in powers of $z - 1 - i$ that diverges at $5 + 5i$ but converges at $4 + 6i$?
33. Do we obtain an analytic function if we replace x by z in the Maclaurin series of a real function $f(x)$?
34. Using Maclaurin series, show that if $f(z)$ is even, its integral (with a suitable constant of integration) is odd.
35. Obtain the first few terms of the Maclaurin series of $\tan z$ by using the Cauchy product and

$$\sin z = \cos z \tan z.$$

SUMMARY OF CHAPTER 15

Power Series, Taylor Series

Sequences, series, and convergence tests are discussed in Sec. 15.1. A **power series** is of the form (Sec. 15.2)

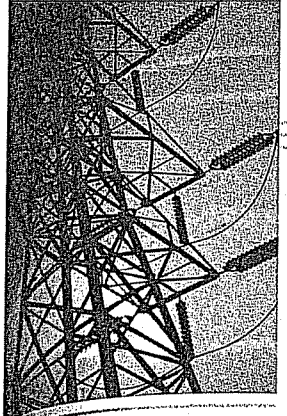
$$(1) \qquad \sum_{n=0}^{\infty} a_n(z - z_0)^n = a_0 + a_1(z - z_0) + a_2(z - z_0)^2 + \cdots;$$

z_0 is its *center*. The series (1) converges for $|z - z_0| < R$ and diverges for $|z - z_0| > R$, where R is the **radius of convergence.** Some power series converge for all z (then we write $R = \infty$). In exceptional cases a power series may converge only at the center; such a series is practically useless. Also, $R = \lim |a_n/a_{n+1}|$ if this limit exists. The series (1) converges absolutely (Sec. 15.2) and **uniformly** (Sec. 15.5) in every closed disk $|z - z_0| \leqq r < R$ $(R > 0)$. It represents an analytic function $f(z)$ for $|z - z_0| < R$. The derivatives $f'(z)$, $f''(z)$, $\cdots$ are obtained by termwise differentiation of (1), and these series have the same radius of convergence R as (1). See Sec. 15.3.

Conversely, *every* analytic function $f(z)$ can be represented by power series. These **Taylor series** of $f(z)$ are of the form (Sec. 15.4)

$$(2) \qquad f(z) = \sum_{n=0}^{\infty} \frac{1}{n!} f^{(n)}(z_0)(z - z_0)^n \qquad (|z - z_0| < R),$$

as in calculus. They converge for all z in the open disk with center z_0 and radius generally equal to the distance from z_0 to the nearest **singularity** of $f(z)$ (point at which $f(z)$ ceases to be analytic as defined in Sec. 15.4). If $f(z)$ is **entire** (analytic for all z; see Sec. 13.5), then (2) converges for all z. The functions e^z, $\cos z$, $\sin z$, etc. have Maclaurin series, that is, Taylor series with center 0, similar to those in calculus (Sec. 15.4).



CHAPTER 16

Laurent Series. Residue Integration

Laurent series generalize Taylor series. Indeed, whereas a Taylor series has positive integer powers (and a constant term) and converges in a disk, a **Laurent series** (Sec. 16.1) is a series of positive ***and negative*** integer powers of $z - z_0$ and converges in an annulus (a circular ring) with center z_0. Hence by a Laurent series we can represent a given function $f(z)$ that is analytic in an annulus and may have singularities outside the ring as well as in the "hole" of the annulus.

We know that for a given function the Taylor series with a given center z_0 is unique. We shall see that, in contrast, a function $f(z)$ can have several Laurent series with the same center z_0 and valid in several concentric annuli. The most important of these series is that which converges for $0 < |z - z_0| < R$, that is, everywhere near the center z_0 except at z_0 itself, where z_0 is a singular point of $f(z)$. The series (or finite sum) of the negative powers of *this* Laurent series is called the **principal part** of the singularity of $f(z)$ at z_0, and is used to classify this singularity (Sec. 16.2). The coefficient of the power $1/(z - z_0)$ of *this* series is called the **residue** of $f(z)$ at z_0. Residues are used in an elegant and powerful integration method, called **residue integration,** for complex contour integrals (Sec. 16.3) as well as for certain complicated real integrals (Sec. 16.4).

Prerequisite: Chaps. 13, 14, Sec. 15.2.
Sections that may be omitted in a shorter course: 16.2, 16.4.
References and Answers to Problems: App. 1 Part D, App. 2.

16.1 Laurent Series

Laurent series generalize Taylor series. If in an application we want to develop a function $f(z)$ in powers of $z - z_0$ when $f(z)$ is singular at z_0 (as defined in Sec. 15.4), we cannot use a Taylor series. Instead we may use a new kind of series, called **Laurent series,**[1] consisting of positive integer powers of $z - z_0$ (and a constant) as well as ***negative integer powers*** of $z - z_0$; this is the new feature.

Laurent series are also used for classifying singularities (Sec. 16.2) and in a powerful integration method ("residue integration", Sec. 16.3).

A Laurent series of $f(z)$ converges in an annulus (in the "hole" of which $f(z)$ may have singularities), as follows.

[1]PIERRE ALPHONSE LAURENT (1813–1854), French military engineer and mathematician, published the theorem in 1843.

THEOREM 1

Laurent's Theorem

Let $f(z)$ be analytic in a domain containing two concentric circles C_1 and C_2 with center z_0 and the annulus between them (blue in Fig. 367). *Then $f(z)$ can be represented by the Laurent series*

$$\text{(1)} \qquad \begin{aligned} f(z) &= \sum_{n=0}^{\infty} a_n(z-z_0)^n + \sum_{n=1}^{\infty} \frac{b_n}{(z-z_0)^n} \\ &= a_0 + a_1(z-z_0) + a_2(z-z_0)^2 + \cdots \\ &\quad \cdots + \frac{b_1}{z-z_0} + \frac{b_2}{(z-z_0)^2} + \cdots \end{aligned}$$

consisting of nonnegative and negative powers. The coefficients of this Laurent series are given by the integrals

$$\text{(2)} \qquad a_n = \frac{1}{2\pi i} \oint_C \frac{f(z^*)}{(z^*-z_0)^{n+1}}\, dz^*, \quad b_n = \frac{1}{2\pi i} \oint_C (z^*-z_0)^{n-1} f(z^*)\, dz^*,$$

taken counterclockwise around any simple closed path C that lies in the annulus and encircles the inner circle, as in Fig. 367. [The variable of integration is denoted by z^* since z is used in (1).]

This series converges and represents $f(z)$ in the enlarged open annulus obtained from the given annulus by continuously increasing the outer circle C_1 and decreasing C_2 until each of the two circles reaches a point where $f(z)$ is singular.

In the important special case that z_0 is the only singular point of $f(z)$ inside C_2, this circle can be shrunk to the point z_0, giving convergence in a disk except at the center. In this case the series (or finite sum) of the negative powers of (1) *is called the* **principal part** *of the singularity of $f(z)$ at z_0.*

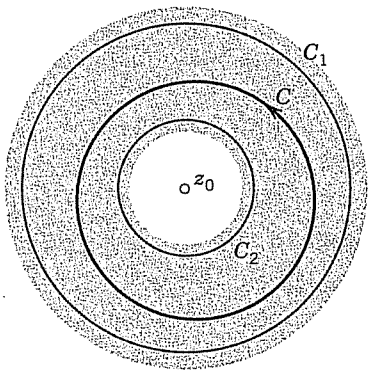


Fig. 367. Laurent's theorem

COMMENT. Obviously, instead of (1), (2) we may write (denoting b_n by a_{-n})

$$\text{(1')} \qquad f(z) = \sum_{n=-\infty}^{\infty} a_n(z-z_0)^n$$

where all the coefficients are now given by a single integral formula, namely,

$$(2') \qquad a_n = \frac{1}{2\pi i} \oint_C \frac{f(z^*)}{(z^* - z_0)^{n+1}}\, dz^* \qquad (n = 0, \pm 1, \pm 2, \cdots).$$

PROOF We prove Laurent's theorem. **(a)** ***The nonnegative powers*** are those of a Taylor series. To see this, we use Cauchy's integral formula (3) in Sec. 14.3 with z^* (instead of z) as the variable of integration and z instead of z_0. Let $g(z)$ and $h(z)$ denote the functions represented by the two terms in (3), Sec. 14.3. Then

$$(3) \qquad f(z) = g(z) + h(z) = \frac{1}{2\pi i} \oint_{C_1} \frac{f(z^*)}{z^* - z}\, dz^* - \frac{1}{2\pi i} \oint_{C_2} \frac{f(z^*)}{z^* - z}\, dz^*.$$

Here z is any point in the given annulus and we integrate counterclockwise over both C_1 and C_2, so that the minus sign appears since in (3) of Sec. 14.3 the integration over C_2 is taken clockwise. We transform each of these two integrals as in Sec. 15.4. The first integral is precisely as in Sec. 15.4. Hence we get precisely the same result, namely, the Taylor series of $g(z)$,

$$(4) \qquad g(z) = \frac{1}{2\pi i} \oint_{C_1} \frac{f(z^*)}{z^* - z}\, dz^* = \sum_{n=0}^{\infty} a_n (z - z_0)^n$$

with coefficients [see (2), Sec. 15.4, counterclockwise integration]

$$(5) \qquad a_n = \frac{1}{2\pi i} \oint_{C_1} \frac{f(z^*)}{(z^* - z_0)^{n+1}}\, dz^*.$$

Here we can replace C_1 by C (see Fig. 367), by the principle of deformation of path, since z_0, the point where the integrand in (5) is not analytic, is not a point of the annulus. This proves the formula for the a_n in (2).

(b) ***The negative powers*** in (1) and the formula for b_n in (2) are obtained if we consider $h(z)$ (the second integral times $-1/(2\pi i)$ in (3). Since z lies in the annulus, it lies in the exterior of the path C_2. Hence the situation differs from that for the first integral. The essential point is that instead of [see (7*) in Sec. 15.4]

$$(6) \quad (a) \qquad \left| \frac{z - z_0}{z^* - z_0} \right| < 1 \qquad \text{we now have} \qquad (b) \qquad \left| \frac{z^* - z_0}{z - z_0} \right| < 1.$$

Consequently, we must develop the expression $1/(z^* - z)$ in the integrand of the second integral in (3) in powers of $(z^* - z_0)/(z - z_0)$ (instead of the reciprocal of this) to get a *convergent* series. We find

$$\frac{1}{z^* - z} = \frac{1}{z^* - z_0 - (z - z_0)} = \frac{-1}{(z - z_0)\left(1 - \dfrac{z^* - z_0}{z - z_0}\right)}.$$

Compare this for a moment with (7) in Sec. 15.4, to really understand the difference. Then go on and apply formula (8), Sec. 15.4, for a finite geometric sum, obtaining

$$\frac{1}{z^* - z} = -\frac{1}{z - z_0}\left\{1 + \frac{z^* - z_0}{z - z_0} + \left(\frac{z^* - z_0}{z - z_0}\right)^2 + \cdots + \left(\frac{z^* - z_0}{z - z_0}\right)^n\right\} - \frac{1}{z - z^*}\left(\frac{z^* - z_0}{z - z_0}\right)^{n+1}.$$

Multiplication by $-f(z^*)/2\pi i$ and integration over C_2 on both sides now yield

$$\begin{aligned} h(z) &= -\frac{1}{2\pi i}\oint_{C_2}\frac{f(z^*)}{z^* - z}\,dz^* \\ &= \frac{1}{2\pi i}\left\{\frac{1}{z - z_0}\oint_{C_2} f(z^*)\,dz^* + \frac{1}{(z - z_0)^2}\oint_{C_2}(z^* - z_0)f(z^*)\,dz^* + \cdots \right. \\ &\qquad + \frac{1}{(z - z_0)^n}\oint_{C_2}(z^* - z_0)^{n-1}f(z^*)\,dz^* \\ &\qquad \left. + \frac{1}{(z - z_0)^{n+1}}\oint_{C_2}(z^* - z_0)^n f(z^*)\,dz^*\right\} + R_n^*(z) \end{aligned}$$

with the last term on the right given by

$$R_n^*(z) = \frac{1}{2\pi i(z - z_0)^{n+1}}\oint_{C_2}\frac{(z^* - z_0)^{n+1}}{z - z^*}f(z^*)\,dz^*. \tag{7}$$

As before, we can integrate over C instead of C_2 in the integrals on the right. We see that on the right, the power $1/(z - z_0)^n$ is multiplied by b_n as given in (2). This establishes Laurent's theorem, provided

$$\lim_{n\to\infty} R_n^*(z) = 0. \tag{8}$$

(c) ***Convergence proof of* (8).** Very often (1) will have only finitely many negative powers. Then there is nothing to be proved. Otherwise, we begin by noting that $f(z^*)/(z - z^*)$ in (7) is bounded in absolute value, say,

$$\left|\frac{f(z^*)}{z - z^*}\right| < \tilde{M} \qquad \text{for all } z^* \text{ on } C_2$$

because $f(z^*)$ is analytic in the annulus and on C_2, and z^* lies on C_2 and z outside, so that $z - z^* \neq 0$. From this and the *ML*-inequality (Sec. 14.1) applied to (7) we get the inequality (L = length of C_2, $|z^* - z_0|$ = radius of C_2 = *const*)

$$|R_n^*(z)| \leqq \frac{1}{2\pi|z - z_0|^{n+1}}\,|z^* - z_0|^{n+1}\,\tilde{M}L = \frac{\tilde{M}L}{2\pi}\left|\frac{z^* - z_0}{z - z_0}\right|^{n+1}.$$

From (6b) we see that the expression on the right approaches zero as n approaches infinity. This proves (8). The representation (1) with coefficients (2) is now established in the given annulus.

(d) *Convergence of* (1) *in the enlarged annulus.* The first series in (1) is a Taylor series [representing $g(z)$]; hence it converges in the disk D with center z_0 whose radius equals the distance of the singularity (or singularities) closest to z_0. Also, $g(z)$ must be singular at all points outside C_1 where $f(z)$ is singular.

The second series in (1), representing $h(z)$, is a power series in $Z = 1/(z - z_0)$. Let the given annulus be $r_2 < |z - z_0| < r_1$, where r_1 and r_2 are the radii of C_1 and C_2, respectively (Fig. 367). This corresponds to $1/r_2 > |Z| > 1/r_1$. Hence this power series in Z must converge at least in the disk $|Z| < 1/r_2$. This corresponds to the exterior $|z - z_0| > r_2$ of C_2, so that $h(z)$ is analytic for all z outside C_2. Also, $h(z)$ must be singular inside C_2 where $f(z)$ is singular, and the series of the negative powers of (1) converges for all z in the exterior E of the circle with center z_0 and radius equal to the maximum distance from z_0 to the singularities of $f(z)$ inside C_2. The domain common to D and E is the enlarged open annulus characterized near the end of Laurent's theorem, whose proof is now complete. ■

Uniqueness. *The Laurent series of a given analytic function $f(z)$ in its annulus of convergence is unique* (see Team Project 24). *However, $f(z)$ may have different Laurent series in two annuli with the same center;* see the examples below. The uniqueness is essential. As for a Taylor series, to obtain the coefficients of Laurent series, we do not generally use the integral formulas (2); instead, we use various other methods, some of which we shall illustrate in our examples. If a Laurent series has been found by any such process, the uniqueness guarantees that it must be ***the*** Laurent series of the given function in the given annulus.

EXAMPLE 1 **Use of Maclaurin Series**

Find the Laurent series of $z^{-5} \sin z$ with center 0.

Solution. By (14), Sec. 15.4, we obtain

$$z^{-5} \sin z = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} z^{2n-4} = \frac{1}{z^4} - \frac{1}{6z^2} + \frac{1}{120} - \frac{1}{5040} z^2 + - \cdots \qquad (|z| > 0).$$

Here the "annulus" of convergence is the whole complex plane without the origin and the principal part of the series at 0 is $z^{-4} - \frac{1}{6} z^{-2}$. ■

EXAMPLE 2 **Substitution**

Find the Laurent series of $z^2 e^{1/z}$ with center 0.

Solution. From (12) in Sec. 15.4 with z replaced by $1/z$ we obtain a Laurent series whose principal part is an infinite series,

$$z^2 e^{1/z} = z^2 \left(1 + \frac{1}{1!z} + \frac{1}{2!z^2} + \cdots\right) = z^2 + z + \frac{1}{2} + \frac{1}{3!z} + \frac{1}{4!z^2} + \cdots \qquad (|z| > 0). \blacksquare$$

EXAMPLE 3 **Development of $1/(1 - z)$**

Develop $1/(1 - z)$ (a) in nonnegative powers of z, (b) in negative powers of z.

Solution.

(a) $$\frac{1}{1 - z} = \sum_{n=0}^{\infty} z^n \qquad \text{(valid if } |z| < 1).$$

(b) $$\frac{1}{1 - z} = \frac{-1}{z(1 - z^{-1})} = -\sum_{n=0}^{\infty} \frac{1}{z^{n+1}} = -\frac{1}{z} - \frac{1}{z^2} - \cdots \qquad \text{(valid if } |z| > 1). \blacksquare$$

EXAMPLE 4 **Laurent Expansions in Different Concentric Annuli**

Find all Laurent series of $1/(z^3 - z^4)$ with center 0.

Solution. Multiplying by $1/z^3$, we get from Example 3

(I) $$\frac{1}{z^3 - z^4} = \sum_{n=0}^{\infty} z^{n-3} = \frac{1}{z^3} + \frac{1}{z^2} + \frac{1}{z} + 1 + z + \cdots \qquad (0 < |z| < 1),$$

(II) $$\frac{1}{z^3 - z^4} = -\sum_{n=0}^{\infty} \frac{1}{z^{n+4}} = -\frac{1}{z^4} - \frac{1}{z^5} - \cdots \qquad (|z| > 1).$$ ■

EXAMPLE 5 **Use of Partial Fractions**

Find all Taylor and Laurent series of $f(z) = \dfrac{-2z + 3}{z^2 - 3z + 2}$ with center 0.

Solution. In terms of partial fractions,

$$f(z) = -\frac{1}{z - 1} - \frac{1}{z - 2}.$$

(a) and (b) in Example 3 take care of the first fraction. For the second fraction,

(c) $$-\frac{1}{z - 2} = \frac{1}{2\left(1 - \frac{1}{2}z\right)} = \sum_{n=0}^{\infty} \frac{1}{2^{n+1}} z^n \qquad (|z| < 2),$$

(d) $$-\frac{1}{z - 2} = -\frac{1}{z\left(1 - \frac{2}{z}\right)} = -\sum_{n=0}^{\infty} \frac{2^n}{z^{n+1}} \qquad (|z| > 2).$$

(I) From (a) and (c), valid for $|z| < 1$ (see Fig. 368),

$$f(z) = \sum_{n=0}^{\infty} \left(1 + \frac{1}{2^{n+1}}\right) z^n = \frac{3}{2} + \frac{5}{4} z + \frac{9}{8} z^2 + \cdots.$$

(II) From (c) and (b), valid for $1 < |z| < 2$,

$$f(z) = \sum_{n=0}^{\infty} \frac{1}{2^{n+1}} z^n - \sum_{n=0}^{\infty} \frac{1}{z^{n+1}} = \frac{1}{2} + \frac{1}{4} z + \frac{1}{8} z^2 + \cdots - \frac{1}{z} - \frac{1}{z^2} - \cdots.$$

(III) From (d) and (b), valid for $|z| > 2$,

$$f(z) = -\sum_{n=0}^{\infty} (2^n + 1) \frac{1}{z^{n+1}} = -\frac{2}{z} - \frac{3}{z^2} - \frac{5}{z^3} - \frac{9}{z^4} - \cdots.$$ ■

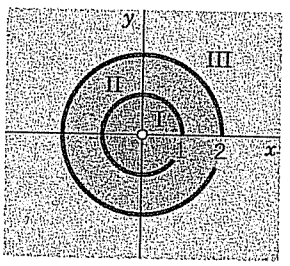


Fig. 368. Regions of convergence in Example 5

If $f(z)$ in Laurent's theorem is analytic inside C_2, the coefficients b_n in (2) are zero by Cauchy's integral theorem, so that the Laurent series reduces to a Taylor series. Examples 3(a) and 5(I) illustrate this.

PROBLEM SET 16.1

1–6 LAURENT SERIES NEAR A SINGULARITY AT 0

Expand the given function in a Laurent series that converges for $0 < |z| < R$ and determine the precise region of convergence. (Show the details of your work.)

1. $\dfrac{1}{z^4 - z^5}$
2. $z \cos \dfrac{1}{z}$
3. $\dfrac{e^{-z}}{z^3}$
4. $\dfrac{\cosh 2z}{z^2}$
5. $z^{-3}e^{1/z^2}$
6. $\dfrac{e^z}{z^2 - z^3}$

7–14 LAURENT SERIES NEAR A SINGULARITY AT z_0

Expand the given function in a Laurent series that converges for $0 < |z - z_0| < R$ and determine the precise region of convergence. (Show details.)

7. $\dfrac{e^z}{z - 1}$, $z_0 = 1$
8. $\dfrac{\sin z}{(z - \frac{1}{4}\pi)^3}$, $z_0 = \frac{1}{4}\pi$
9. $\dfrac{1}{z^2 + 1}$, $z_0 = i$
10. $\dfrac{\cos z}{(z - \pi)^4}$, $z_0 = \pi$
11. $\dfrac{1}{(z + i)^2 - (z + i)}$, $z_0 = -i$
12. $\dfrac{z^3}{(z + i)^2}$, $z_0 = -i$
13. $\dfrac{z^2 - 4}{z - 1}$, $z_0 = 1$
14. $z^2 \sinh \dfrac{1}{z}$, $z_0 = 0$

15–23 TAYLOR AND LAURENT SERIES

Find all Taylor and Laurent series with center $z = z_0$ and determine the precise regions of convergence.

15. $\dfrac{1}{1 - z^3}$, $z_0 = 0$
16. $\dfrac{1}{1 - z^2}$, $z_0 = 1$
17. $\dfrac{z^2}{1 - z^4}$, $z_0 = 0$
18. $\dfrac{1}{z}$, $z_0 = 1$
19. $\dfrac{z^3 - 2iz^2}{(z - i)^2}$, $z_0 = i$
20. $\dfrac{\sinh z}{(z - 1)^4}$, $z_0 = 1$
21. $\dfrac{4z - 1}{z^4 - 1}$, $z_0 = 0$
22. $\dfrac{1}{z^2}$, $z_0 = i$
23. $\dfrac{\sin z}{z + \frac{1}{2}\pi}$, $z_0 = -\frac{1}{2}\pi$

24. **TEAM PROJECT. Laurent Series. (a) Uniqueness.** Prove that the Laurent expansion of a given analytic function in a given annulus is unique.

(b) Accumulation of singularities. Does $\tan(1/z)$ have a Laurent series that converges in a region $0 < |z| < R$? (Give a reason.)

(c) Integrals. Expand the following functions in a Laurent series that converges for $|z| > 0$:

$$\frac{1}{z^2}\int_0^z \frac{e^t - 1}{t}\,dt, \qquad \frac{1}{z^3}\int_0^z \frac{\sin t}{t}\,dt.$$

25. **CAS PROJECT. Partial Fractions.** Write a program for obtaining Laurent series by the use of partial fractions. Using the program, verify the calculations in Example 5 of the text. Apply the program to two other functions of your choice.

16.2 Singularities and Zeros. Infinity

Roughly, a *singular point* of an analytic function $f(z)$ is a z_0 at which $f(z)$ ceases to be analytic, and a *zero* is a z at which $f(z) = 0$. Precise definitions follow below. In this section we show that Laurent series can be used for classifying singularities and Taylor series for discussing zeros.

Singularities were defined in Sec. 15.4, as we shall now recall and extend. We also remember that, by definition, a function is a ***single-valued*** relation, as was emphasized in Sec. 13.3.

We say that a function $f(z)$ **is singular** or **has a singularity** at a point $z = z_0$ if $f(z)$ is not analytic (perhaps not even defined) at $z = z_0$, but every neighborhood of $z = z_0$ contains points at which $f(z)$ is analytic. We also say that $z = z_0$ is a **singular point** of $f(z)$.

We call $z = z_0$ an **isolated singularity** of $f(z)$ if $z = z_0$ has a neighborhood without further singularities of $f(z)$. *Example:* $\tan z$ has isolated singularities at $\pm\pi/2$, $\pm 3\pi/2$, etc.; $\tan(1/z)$ has a nonisolated singularity at 0. (Explain!)

Isolated singularities of $f(z)$ at $z = z_0$ can be classified by the Laurent series

$$(1) \qquad f(z) = \sum_{n=0}^{\infty} a_n(z - z_0)^n + \sum_{n=1}^{\infty} \frac{b_n}{(z - z_0)^n} \qquad \text{(Sec. 16.1)}$$

valid ***in the immediate neighborhood*** of the singular point $z = z_0$, except at z_0 itself, that is, in a region of the form

$$0 < |z - z_0| < R.$$

The sum of the first series is analytic at $z = z_0$, as we know from the last section. The second series, containing the negative powers, is called the **principal part** of (1), as we remember from the last section. If it has only finitely many terms, it is of the form

$$(2) \qquad \frac{b_1}{z - z_0} + \cdots + \frac{b_m}{(z - z_0)^m} \qquad (b_m \neq 0).$$

Then the singularity of $f(z)$ at $z = z_0$ is called a **pole,** and m is called its **order.** Poles of the first order are also known as **simple poles.**

If the principal part of (1) has infinitely many terms, we say that $f(z)$ has at $z = z_0$ an **isolated essential singularity.**

We leave aside nonisolated singularities.

EXAMPLE 1 **Poles. Essential Singularities**

The function

$$f(z) = \frac{1}{z(z - 2)^5} + \frac{3}{(z - 2)^2}$$

has a simple pole at $z = 0$ and a pole of fifth order at $z = 2$. Examples of functions having an isolated essential singularity at $z = 0$ are

$$e^{1/z} = \sum_{n=0}^{\infty} \frac{1}{n!z^n} = 1 + \frac{1}{z} + \frac{1}{2!z^2} + \cdots$$

and

$$\sin \frac{1}{z} = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n + 1)!z^{2n+1}} = \frac{1}{z} - \frac{1}{3!z^3} + \frac{1}{5!z^5} - + \cdots.$$

Section 16.1 provides further examples. For instance, Example 1 shows that $z^{-5} \sin z$ has a fourth-order pole at 0. Example 4 shows that $1/(z^3 - z^4)$ has a third-order pole at 0 and a Laurent series with infinitely many negative powers. This is no contradiction, since this series is valid for $|z| > 1$; it merely tells us that in classifying singularities it is quite important to consider the Laurent series valid ***in the immediate neighborhood*** of a singular point. In Example 4 this is the series (I), which has three negative powers. ■

The classification of singularities into poles and essential singularities is not merely a formal matter, because the behavior of an analytic function in a neighborhood of an essential singularity is entirely different from that in the neighborhood of a pole.

EXAMPLE 2 **Behavior Near a Pole**

$f(z) = 1/z^2$ has a pole at $z = 0$, and $|f(z)| \to \infty$ as $z \to 0$ in any manner. This illustrates the following theorem. ■

THEOREM 1

Poles

If $f(z)$ is analytic and has a pole at $z = z_0$, then $|f(z)| \to \infty$ as $z \to z_0$ in any manner.

The proof is left to the student (see Prob. 12).

EXAMPLE 3 **Behavior Near an Essential Singularity**

The function $f(z) = e^{1/z}$ has an essential singularity at $z = 0$. It has no limit for approach along the imaginary axis; it becomes infinite if $z \to 0$ through positive real values, but it approaches zero if $z \to 0$ through negative real values. It takes on any given value $c = c_0 e^{i\alpha} \neq 0$ in an arbitrarily small ϵ-neighborhood of $z = 0$. To see the letter, we set $z = re^{i\theta}$, and then obtain the following complex equation for r and θ, which we must solve:

$$e^{1/z} = e^{(\cos\theta - i\sin\theta)/r} = c_0 e^{i\alpha}.$$

Equating the absolute values and the arguments, we have $e^{(\cos\theta)/r} = c_0$, that is

$$\cos\theta = r\ln c_0, \qquad \text{and} \qquad -\sin\theta = \alpha r$$

respectively. From these two equations and $\cos^2\theta + \sin^2\theta = r^2(\ln c_0)^2 + \alpha^2 r^2 = 1$ we obtain the formulas

$$r^2 = \frac{1}{(\ln c_0)^2 + \alpha^2} \qquad \text{and} \qquad \tan\theta = -\frac{\alpha}{\ln c_0}.$$

Hence r can be made arbitrarily small by adding multiples of 2π to α, leaving c unaltered. This illustrates the very famous *Picard's theorem* (with $z = 0$ as the exceptional value). For the rather complicated proof, see Ref. [D4], vol. 2, p. 258. For Picard, see Sec. 1.7. ■

THEOREM 2

Picard's Theorem

If $f(z)$ is analytic and has an isolated essential singularity at a point z_0, it takes on every value, with at most one exceptional value, in an arbitrarily small ϵ-neighborhood of z_0.

Removable Singularities. We say that a function $f(z)$ has a *removable singularity* at $z = z_0$ if $f(z)$ is not analytic at $z = z_0$, but can be made analytic there by assigning a suitable value $f(z_0)$. Such singularities are of no interest since they can be removed as just indicated. *Example:* $f(z) = (\sin z)/z$ becomes analytic at $z = 0$ if we define $f(0) = 1$.

Zeros of Analytic Functions

A **zero** of an analytic function $f(z)$ in a domain D is a $z = z_0$ in D such that $f(z_0) = 0$. A zero has **order** n if not only f but also the derivatives $f', f'', \cdots, f^{(n-1)}$ are all 0 at $z = z_0$ but $f^{(n)}(z_0) \neq 0$. A first-order zero is also called a **simple zero.** For a second-order zero, $f(z_0) = f'(z_0) = 0$ but $f''(z_0) \neq 0$. And so on.

EXAMPLE 4 **Zeros**

The function $1 + z^2$ has simple zeros at $\pm i$. The function $(1 - z^4)^2$ has second-order zeros at ± 1 and $\pm i$. The function $(z - a)^3$ has a third-order zero at $z = a$. The function e^z has no zeros (see Sec. 13.5). The function $\sin z$ has simple zeros at $0, \pm\pi, \pm 2\pi, \cdots$, and $\sin^2 z$ has second-order zeros at these points. The function $1 - \cos z$ has second-order zeros at $0, \pm 2\pi, \pm 4\pi, \cdots$, and the function $(1 - \cos z)^2$ has fourth-order zeros at these points. ■

Taylor Series at a Zero. At an nth-order zero $z = z_0$ of $f(z)$, the derivatives $f'(z_0), \cdots, f^{(n-1)}(z_0)$ are zero, by definition. Hence the first few coefficients $a_0, \cdots, a_{n-1}$ of the Taylor series (1), Sec. 15.4, are zero, too, whereas $a_n \neq 0$, so that this series takes the form

$$\begin{aligned} f(z) &= a_n(z - z_0)^n + a_{n+1}(z - z_0)^{n+1} + \cdots \\ &= (z - z_0)^n [a_n + a_{n+1}(z - z_0) + a_{n+2}(z - z_0)^2 + \cdots] \qquad (a_n \neq 0). \end{aligned} \tag{3}$$

This is characteristic of such a zero, because if $f(z)$ has such a Taylor series, it has an nth-order zero at $z = z_0$, as follows by differentiation.

Whereas nonisolated singularities may occur, for zeros we have

THEOREM 3

Zeros

The zeros of an analytic function $f(z)$ ($\not\equiv 0$) are isolated; that is, each of them has a neighborhood that contains no further zeros of $f(z)$.

PROOF The factor $(z - z_0)^n$ in (3) is zero only at $z = z_0$. The power series in the brackets $[\cdots]$ represents an analytic function (by Theorem 5 in Sec. 15.3), call it $g(z)$. Now $g(z_0) = a_n \neq 0$, since an analytic function is continuous, and because of this continuity, also $g(z) \neq 0$ in some neighborhood of $z = z_0$. Hence the same holds of $f(z)$. ∎

This theorem is illustrated by the functions in Example 4.

Poles are often caused by zeros in the denominator. (*Example:* $\tan z$ has poles where $\cos z$ is zero.) This is a major reason for the importance of zeros. The key to the connection is the following theorem, whose proof follows from (3) (see Team Project 24).

THEOREM 4

Poles and Zeros

Let $f(z)$ be analytic at $z = z_0$ and have a zero of nth order at $z = z_0$. Then $1/f(z)$ has a pole of nth order at $z = z_0$; and so does $h(z)/f(z)$, provided $h(z)$ is analytic at $z = z_0$ and $h(z_0) \neq 0$.

Riemann Sphere. Point at Infinity

When we want to study complex functions for large $|z|$, the complex plane will generally become rather inconvenient. Then it may be better to use a representation of complex numbers on the so-called **Riemann sphere.** This is a sphere S of diameter 1 touching the complex z-plane at $z = 0$ (Fig. 369), and we let the image of a point P (a number z in the plane) be the intersection P^* of the segment PN with S, where N is the "North Pole" diametrically opposite to the origin in the plane. Then to each z there corresponds a point on S.

Conversely, each point on S represents a complex number z, except for N, which does not correspond to any point in the complex plane. This suggests that we introduce an additional point, called the **point at infinity** and denoted ∞ ("infinity") and let its image be N. The complex plane together with ∞ is called the **extended complex plane.** The complex plane is often called the *finite complex plane,* for distinction, or simply the

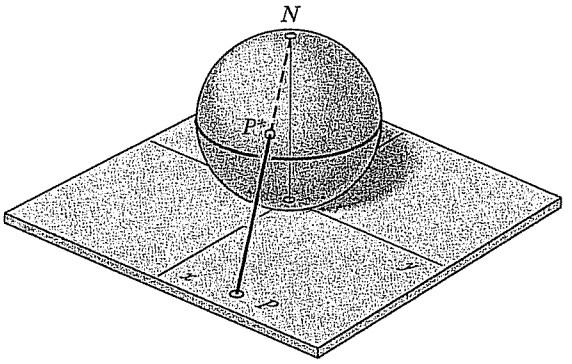


Fig. 369. Riemann sphere

complex plane as before. The sphere S is called the **Riemann sphere.** The mapping of the extended complex plane onto the sphere is known as a **stereographic projection.** (What is the image of the Northern Hemisphere? Of the Western Hemisphere? Of a straight line through the origin?)

Analytic or Singular at Infinity

If we want to investigate a function $f(z)$ for large $|z|$, we may now set $z = 1/w$ and investigate $f(z) = f(1/w) \equiv g(w)$ in a neighborhood of $w = 0$. We define $f(z)$ to be **analytic** or **singular at infinity** if $g(w)$ is analytic or singular, respectively, at $w = 0$. We also define

$$g(0) = \lim_{w \to 0} g(w) \tag{4}$$

if this limit exists.

Furthermore, we say that $f(z)$ has an *nth-order zero at infinity* if $f(1/w)$ has such a zero at $w = 0$. Similarly for poles and essential singularities.

EXAMPLE 5 **Functions Analytic or Singular at Infinity. Entire and Meromorphic Functions**

The function $f(z) = 1/z^2$ is analytic at ∞ since $g(w) = f(1/w) = w^2$ is analytic at $w = 0$, and $f(z)$ has a second-order zero at ∞. The function $f(z) = z^3$ is singular at ∞ and has a third-order pole there since the function $g(w) = f(1/w) = 1/w^3$ has such a pole at $w = 0$. The function e^z has an essential singularity at ∞ since $e^{1/w}$ has such a singularity at $w = 0$. Similarly, $\cos z$ and $\sin z$ have an essential singularity at ∞.

Recall that an **entire function** is one that is analytic everywhere in the (finite) complex plane. Liouville's theorem (Sec. 14.4) tells us that the only *bounded* entire functions are the constants, hence any nonconstant entire function must be unbounded. Hence it has a singularity at ∞, a pole if it is a polynomial or an essential singularity if it is not. The functions just considered are typical in this respect.

An analytic function whose only singularities in the finite plane are poles is called a **meromorphic function.** Examples are rational functions with nonconstant denominator, $\tan z$, $\cot z$, $\sec z$, and $\csc z$. ■

In this section we used Laurent series for investigating singularities. In the next section we shall use these series for an elegant integration method.

PROBLEM SET 16.2

1–10 SINGULARITIES

Determine the location and kind of the singularities of the following functions in the finite plane and at infinity. In the case of poles also state the order.

1. $\tan^2 \pi z$
2. $z + \dfrac{2}{z} - \dfrac{3}{z^2}$
3. $\cot z^2$
4. $z^3 e^{1/(z-1)}$
5. $\cos z - \sin z$
6. $1/(\cos z - \sin z)$

7. $\dfrac{\sin 3z}{(z^4-1)^4}$

8. $\dfrac{4}{z-1}+\dfrac{2}{(z-1)^2}-\dfrac{8}{(z-1)^3}$

9. $\cosh\left[1/(z^2+1)\right]$

10. $e^{1/(z-1)}/(e^z-1)$

11. **(Essential singularity)** Discuss e^{1/z^2} in a similar way as $e^{1/z}$ is discussed in Example 3.

12. **(Poles)** Verify Theorem 1 for $f(z)=z^{-3}-z^{-1}$. Prove Theorem 1.

13–22 ZEROS

Determine the location and order of the zeros.

13. $(z+16i)^4$

14. $(z^4-16)^4$

15. $z^{-3}\sin^3 \pi z$

16. $\cosh^2 z$

17. $(3z^2+1)e^{-z}$

18. $(z^2-1)^2(e^{z^2}-1)$

19. $(z^2+4)(e^z-1)^2$

20. $(\sin z-1)^3$

21. $(1-\cos z)^2$

22. e^z-e^{2z}

23. **(Zeros)** If $f(z)$ is analytic and has a zero of order n at $z=z_0$, show that $f^2(z)$ has a zero of order $2n$.

24. **TEAM PROJECT. Zeros. (a) Derivative.** Show that if $f(z)$ has a zero of order $n>1$ at $z=z_0$, then $f'(z)$ has a zero of order $n-1$ at z_0.

(b) Poles and zeros. Prove Theorem 4.

(c) Isolated k-points. Show that the points at which a nonconstant analytic function $f(z)$ has a given value k are isolated.

(d) Identical functions. If $f_1(z)$ are analytic in a domain D and equal at a sequence of points z_n in D that converges in D, show that $f_1(z)\equiv f_2(z)$ in D.

25. **(Riemann sphere)** Assuming that we let the image of the x-axis be meridians 0° and 180°, describe and sketch (or graph) the images of the following regions on the Riemann sphere: (a) $|z|>100$, (b) the lower half-plane, (c) $\frac{1}{2}\leqq|z|\leqq 2$.

16.3 Residue Integration Method

The purpose of Cauchy's residue integration method is the evaluation of integrals

$$\oint_C f(z)\,dz$$

taken around a simple close path C. The idea is as follows.

If $f(z)$ is analytic everywhere on C and inside C, such an integral is zero by Cauchy's integral theorem (Sec. 14.2), and we are done.

If $f(z)$ has a singularity at a point $z=z_0$ inside C, but is otherwise analytic on C and inside C, then $f(z)$ has a Laurent series

$$f(z)=\sum_{n=0}^{\infty}a_n(z-z_0)^n+\frac{b_1}{z-z_0}+\frac{b_2}{(z-z_0)^2}+\cdots$$

that converges for all points near $z=z_0$ (except at $z=z_0$ itself), in some domain of the form $0<|z-z_0|<R$ (sometimes called a **deleted neighborhood,** an old-fashioned term that we shall not use). Now comes the key idea. The coefficient b_1 of the first negative power $1/(z-z_0)$ of this Laurent series is given by the integral formula (2) in Sec. 16.1 with $n=1$, namely,

$$b_1=\frac{1}{2\pi i}\oint_C f(z)\,dz.$$

Now, since we can obtain Laurent series by various methods, without using the integral formulas for the coefficients (see the examples in Sec. 16.1), we can find b_1 by one of those methods and then use the formula for b_1 for evaluating the integral, that is,

(1) $$\oint_C f(z)\,dz=2\pi i b_1.$$

Here we integrate counterclockwise around a simple closed path C that contains $z = z_0$ in its interior (but no other singular points of $f(z)$ on or inside C!).

The coefficient b_1 is called the **residue** of $f(z)$ at $z = z_0$ and we denote it by

$$b_1 = \operatorname*{Res}_{z=z_0} f(z). \tag{2}$$

EXAMPLE 1 Evaluation of an Integral by Means of a Residue

Integrate the function $f(z) = z^{-4} \sin z$ counterclockwise around the unit circle C.

Solution. From (14) in Sec. 15.4 we obtain the Laurent series

$$f(z) = \frac{\sin z}{z^4} = \frac{1}{z^3} - \frac{1}{3!z} + \frac{z}{5!} - \frac{z^3}{7!} + - \cdots$$

which converges for $|z| > 0$ (that is, for all $z \neq 0$). This series shows that $f(z)$ has a pole of third order at $z = 0$ and the residue $b_1 = -1/3!$. From (1) we thus obtain the answer

$$\oint_C \frac{\sin z}{z^4}\, dz = 2\pi i b_1 = -\frac{\pi i}{3}. \qquad \blacksquare$$

EXAMPLE 2 CAUTION! Use the Right Laurent Series!

Integrate $f(z) = 1/(z^3 - z^4)$ clockwise around the circle C: $|z| = 1/2$.

Solution. $z^3 - z^4 = z^3(1 - z)$ shows that $f(z)$ is singular at $z = 0$ and $z = 1$. Now $z = 1$ lies outside C. Hence it is of no interest here. So we need the residue of $f(z)$ at 0. We find it from the Laurent series that converges for $0 < |z| < 1$. This is series (I) in Example 4, Sec. 16.1,

$$\frac{1}{z^3 - z^4} = \frac{1}{z^3} + \frac{1}{z^2} + \frac{1}{z} + 1 + z + \cdots \qquad (0 < |z| < 1).$$

We see from it that this residue is 1. Clockwise integration thus yields

$$\oint_C \frac{dz}{z^3 - z^4} = -2\pi i \operatorname*{Res}_{z=0} f(z) = -2\pi i.$$

CAUTION! Had we used the wrong series (II) in Example 4, Sec. 16.1,

$$\frac{1}{z^3 - z^4} = -\frac{1}{z^4} - \frac{1}{z^5} - \frac{1}{z^6} - \cdots \qquad (|z| > 1),$$

we would have obtained the wrong answer, 0, because this series has no power $1/z$. $\blacksquare$

Formulas for Residues

To calculate a residue at a pole, we need not produce a whole Laurent series, but, more economically, we can derive formulas for residues once and for all.

Simple Poles. Two formulas for the residue of $f(z)$ at a simple pole at z_0 are

$$\operatorname*{Res}_{z=z_0} f(z) = b_1 = \lim_{z \to z_0} (z - z_0) f(z) \tag{3}$$

and, assuming that $f(z) = p(z)/q(z)$, $p(z_0) \neq 0$, and $q(z)$ has a simple zero at z_0 (so that $f(z)$ has at z_0 a simple pole, by Theorem 4 in Sec. 16.2),

$$\operatorname*{Res}_{z=z_0} f(z) = \operatorname*{Res}_{z=z_0} \frac{p(z)}{q(z)} = \frac{p(z_0)}{q'(z_0)}. \tag{4}$$

PROOF For a simple pole at $z = z_0$ the Laurent series (1), Sec. 16.1, is

$$f(z) = \frac{b_1}{z - z_0} + a_0 + a_1(z - z_0) + a_2(z - z_0)^2 + \cdots \qquad (0 < |z - z_0| < R).$$

Here $b_1 \neq 0$. (Why?) Multiplying both sides by $z - z_0$ and then letting $z \to z_0$, we obtain the formula (3):

$$\lim_{z \to z_0} (z - z_0) f(z) = b_1 + \lim_{z \to z_0} (z - z_0)[a_0 + a_1(z - z_0) + \cdots] = b_1$$

where the last equality follows from continuity (Theorem 1, Sec. 15.3).

We prove (4). The Taylor series of $q(z)$ at a simple zero z_0 is

$$q(z) = (z - z_0)q'(z_0) + \frac{(z - z_0)^2}{2!} q''(z_0) + \cdots.$$

Substituting this into $f = p/q$ and then f into (3) gives

$$\operatorname*{Res}_{z=z_0} f(z) = \lim_{z \to z_0} (z - z_0) \frac{p(z)}{q(z)} = \lim_{z \to z_0} \frac{(z - z_0)p(z)}{(z - z_0)[q'(z_0) + (z - z_0)q''(z_0)/2 + \cdots]}.$$

$z - z_0$ cancels. By continuity, the limit of the denominator is $q'(z_0)$ and (4) follows. ∎

EXAMPLE 3 **Residue at a Simple Pole**

$f(z) = (9z + i)/(z^3 + z)$ has a simple pole at i because $z^2 + 1 = (z + i)(z - i)$, and (3) gives the residue

$$\operatorname*{Res}_{z=i} \frac{9z + i}{z(z^2 + 1)} = \lim_{z \to i} (z - i) \frac{9z + i}{z(z + i)(z - i)} = \left[\frac{9z + i}{z(z + i)}\right]_{z=i} = \frac{10i}{-2} = -5i.$$

By (4) with $p(i) = 9i + i$ and $q'(z) = 3z^2 + 1$ we confirm the result,

$$\operatorname*{Res}_{z=i} \frac{9z + i}{z(z^2 + 1)} = \left[\frac{9z + i}{3z^2 + 1}\right]_{z=i} = \frac{10i}{-2} = -5i.$$ ∎

Poles of Any Order. The residue of $f(z)$ at an mth-order pole at z_0 is

(5) $$\operatorname*{Res}_{z=z_0} f(z) = \frac{1}{(m-1)!} \lim_{z \to z_0} \left\{ \frac{d^{m-1}}{dz^{m-1}} \left[(z - z_0)^m f(z) \right] \right\}.$$

In particular, for a second-order pole ($m = 2$),

(5*) $$\operatorname*{Res}_{z=z_0} f(z) = \lim_{z \to z_0} \left\{ [(z - z_0)^2 f(z)]' \right\}.$$

PROOF The Laurent series of $f(z)$ converging near z_0 (except at z_0 itself) is (Sec. 16.2)

$$f(z) = \frac{b_m}{(z - z_0)^m} + \frac{b_{m-1}}{(z - z_0)^{m-1}} + \cdots + \frac{b_1}{z - z_0} + a_0 + a_1(z - z_0) + \cdots$$

where $b_m \neq 0$. The residue wanted is b_1. Multiplying both sides by $(z - z_0)^m$ gives

$$(z - z_0)^m f(z) = b_m + b_{m-1}(z - z_0) + \cdots + b_1(z - z_0)^{m-1} + a_0(z - z_0)^m + \cdots.$$

We see that b_1 is now the coefficient of the power $(z - z_0)^{m-1}$ of the power series of $g(z) = (z - z_0)^m f(z)$. Hence Taylor's theorem (Sec. 15.4) gives (5):

$$b_1 = \frac{1}{(m-1)!} g^{(m-1)}(z_0)$$

$$= \frac{1}{(m-1)!} \frac{d^{m-1}}{dz^{m-1}} [(z - z_0)^m f(z)]. \quad \blacksquare$$

EXAMPLE 4 Residue at a Pole of Higher Order

$f(z) = 50z/(z^3 + 2z^2 - 7z + 4)$ has a pole of second order at $z = 1$ because the denominator equals $(z + 4)(z - 1)^2$ (verify!). From (5*) we obtain the residue

$$\operatorname*{Res}_{z=1} f(z) = \lim_{z \to 1} \frac{d}{dz} [(z-1)^2 f(z)]$$

$$= \lim_{z \to 1} \frac{d}{dz} \left(\frac{50z}{z+4} \right)$$

$$= \frac{200}{5^2} = 8. \quad \blacksquare$$

Several Singularities Inside the Contour. Residue Theorem

Residue integration can be extended from the case of a single singularity to the case of several singularities within the contour C. This is the purpose of the residue theorem. The extension is surprisingly simple.

THEOREM 1

Residue Theorem

Let $f(z)$ be analytic inside a simple closed path C and on C, except for finitely many singular points $z_1, z_2, \cdots, z_k$ inside C. Then the integral of $f(z)$ taken counterclockwise around C equals $2\pi i$ times the sum of the residues of $f(z)$ at $z_1, \cdots, z_k$:

(6)
$$\oint_C f(z)\, dz = 2\pi i \sum_{j=1}^{k} \operatorname*{Res}_{z=z_j} f(z).$$

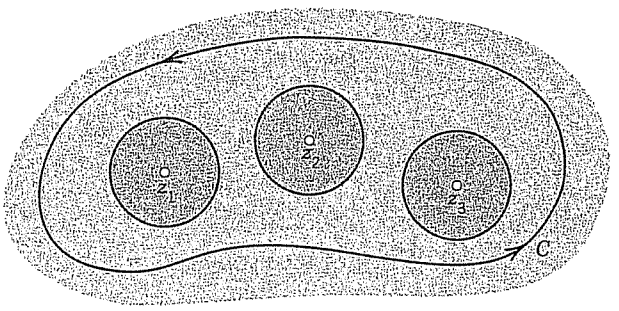


Fig. 370. Residue theorem

PROOF We enclose each of the singular points z_j in a circle C_j with radius small enough that those k circles and C are all separated (Fig. 370). Then $f(z)$ is analytic in the multiply connected domain D bounded by C and $C_1, \cdots, C_k$ and on the entire boundary of D. From Cauchy's integral theorem we thus have

$$\oint_C f(z)\,dz + \oint_{C_1} f(z)\,dz + \oint_{C_2} f(z)\,dz + \cdots + \oint_{C_k} f(z)\,dz = 0, \tag{7}$$

the integral along C being taken *counterclockwise* and the other integrals *clockwise* (as in Figs. 351 and 352, Sec. 14.2). We take the integrals over $C_1, \cdots, C_k$ to the right and compensate the resulting minus sign by reversing the sense of integration. Thus,

$$\oint_C f(z)\,dz = \oint_{C_1} f(z)\,dz + \oint_{C_2} f(z)\,dz + \cdots + \oint_{C_k} f(z)\,dz \tag{8}$$

where all the integrals are now taken counterclockwise. By (1) and (2),

$$\oint_{C_j} f(z)\,dz = 2\pi i \operatorname*{Res}_{z=z_j} f(z), \qquad j = 1, \cdots, k,$$

so that (8) gives (6) and the residue theorem is proved. ∎

This important theorem has various applications in connection with complex and real integrals. Let us first consider some complex integrals. (Real integrals follow in the next section.)

EXAMPLE 5 **Integration by the Residue Theorem. Several Contours**

Evaluate the following integral counterclockwise around any simple closed path such that (a) 0 and 1 are inside C, (b) 0 is inside, 1 outside, (c) 1 is inside, 0 outside, (d) 0 and 1 are outside.

$$\oint_C \frac{4 - 3z}{z^2 - z}\,dz$$

Solution. The integrand has simple poles at 0 and 1, with residues [by (3)]

$$\operatorname*{Res}_{z=0} \frac{4 - 3z}{z(z-1)} = \left[\frac{4 - 3z}{z - 1}\right]_{z=0} = -4, \qquad \operatorname*{Res}_{z=1} \frac{4 - 3z}{z(z-1)} = \left[\frac{4 - 3z}{z}\right]_{z=1} = 1.$$

[Confirm this by (4).] *Ans.* (a) $2\pi i(-4 + 1) = -6\pi i$, (b) $-8\pi i$, (c) $2\pi i$, (d) 0. ∎

EXAMPLE 6 **Another Application of the Residue Theorem**

Integrate $(\tan z)/(z^2 - 1)$ counterclockwise around the circle C: $|z| = 3/2$.

Solution. $\tan z$ is not analytic at $\pm\pi/2, \pm 3\pi/2, \cdots$, but all these points lie outside the contour C. Because of the denominator $z^2 - 1 = (z - 1)(z + 1)$ the given function has simple poles at ± 1. We thus obtain from (4) and the residue theorem

$$\oint_C \frac{\tan z}{z^2 - 1}\,dz = 2\pi i\left(\operatorname*{Res}_{z=1} \frac{\tan z}{z^2 - 1} + \operatorname*{Res}_{z=-1} \frac{\tan z}{z^2 - 1}\right)$$

$$= 2\pi i\left(\left.\frac{\tan z}{2z}\right|_{z=1} + \left.\frac{\tan z}{2z}\right|_{z=-1}\right)$$

$$= 2\pi i \tan 1 = 9.7855i.$$

∎

EXAMPLE 7 Poles and Essential Singularities

Evaluate the following integral, where C is the ellipse $9x^2 + y^2 = 9$ (counterclockwise, sketch it).

$$\oint_C \left(\frac{ze^{\pi z}}{z^4 - 16} + ze^{\pi/z} \right) dz$$

Solution. Since $z^4 - 16 = 0$ at $\pm 2i$ and ± 2, the first term of the integrand has simple poles at $\pm 2i$ inside C, with residues [by (4); note that $e^{2\pi i} = 1$]

$$\operatorname*{Res}_{z=2i} \frac{ze^{\pi z}}{z^4 - 16} = \left[\frac{ze^{\pi z}}{4z^3} \right]_{z=2i} = -\frac{1}{16},$$

$$\operatorname*{Res}_{z=-2i} \frac{ze^{\pi z}}{z^4 - 16} = \left[\frac{ze^{\pi z}}{4z^3} \right]_{z=-2i} = -\frac{1}{16}$$

and simple poles at ± 2, which lie outside C, so that they are of no interest here. The second term of the integrand has an essential singularity at 0, with residue $\pi^2/2$ as obtained from

$$ze^{\pi/z} = z\left(1 + \frac{\pi}{z} + \frac{\pi^2}{2!z^2} + \frac{\pi^3}{3!z^3} + \cdots\right) = z + \pi + \frac{\pi^2}{2} \cdot \frac{1}{z} + \cdots \qquad (|z| > 0).$$

Ans. $2\pi i(-\frac{1}{16} - \frac{1}{16} + \frac{1}{2}\pi^2) = \pi(\pi^2 - \frac{1}{4})i = 30.221i$ by the residue theorem. ■

PROBLEM SET 16.3

1. Verify the calculations in Example 3 and find the other residues.
2. Verify the calculations in Example 4 and find the other residue.

3–12 RESIDUES

Find all the singular points and the corresponding residues. (Show the details of your work.)

3. $\dfrac{1}{4 + z^2}$
4. $\dfrac{\cos z}{z^6}$
5. $\dfrac{\sin z}{z^6}$
6. $\dfrac{z^2 + 1}{z^2 - z}$
7. $\cot z$
8. $\sec z$
9. $\dfrac{1}{(z^2 - 1)^2}$
10. $\dfrac{1/3}{z^4 - 1}$
11. $\tan z$
12. $\dfrac{z^2}{z^4 - 1}$
13. **CAS PROJECT. Residue at a Pole.** Write a program for calculating the residue at a pole of any order. Use it for solving Probs. 3–8.

14–25 RESIDUE INTEGRATION

Evaluate (counterclockwise). (Show the details.)

14. $\displaystyle\oint_C \frac{\sin \pi z}{z^4}\, dz, \quad C: |z - i| = 2$
15. $\displaystyle\oint_C e^{1/z}\, dz, \quad C: |z| = 1$
16. $\displaystyle\oint_C \frac{dz}{\sinh \frac{1}{2}\pi z}, \quad C: |z - 1| = 1.4$
17. $\displaystyle\oint_C \tan \pi z\, dz, \quad C: |z| = 1$
18. $\displaystyle\oint_C \tan \pi z\, dz, \quad C: |z| = 2$
19. $\displaystyle\oint_C \frac{e^z}{\cos z}\, dz, \quad C: |z| = 4.5$
20. $\displaystyle\oint_C \coth z\, dz, \quad C: |z| = 1$
21. $\displaystyle\oint_C \frac{e^z}{\cos \pi z}\, dz, \quad C: |z - i| = 1.5$
22. $\displaystyle\oint_C \frac{\cosh z}{z^2 - 3iz}\, dz, \quad C: |z| = 1$
23. $\displaystyle\oint_C \frac{\tan \pi z}{z^3}\, dz, \quad C: |z + \frac{1}{2}i| = 1$
24. $\displaystyle\oint_C \frac{1 - 4z + 6z^2}{(z^2 + \frac{1}{4})(2 - z)}\, dz, \quad C: |z| = 1$
25. $\displaystyle\oint_C \frac{30z^2 - 23z + 5}{(2z - 1)^2(3z - 1)}\, dz, \quad C: |z| = 1$

16.4 Residue Integration of Real Integrals

It is quite surprising that certain classes of complicated ***real*** integrals can be integrated by the residue theorem, as we shall see.

Integrals of Rational Functions of cos θ and sin θ

We first consider integrals of the type

$$J = \int_0^{2\pi} F(\cos\theta, \sin\theta)\, d\theta \tag{1}$$

where $F(\cos\theta, \sin\theta)$ is a real rational function of $\cos\theta$ and $\sin\theta$ [for example, $(\sin^2\theta)/(5 - 4\cos\theta)$] and is finite (does not become infinite) on the interval of integration. Setting $e^{i\theta} = z$, we obtain

$$\cos\theta = \frac{1}{2}(e^{i\theta} + e^{-i\theta}) = \frac{1}{2}\left(z + \frac{1}{z}\right)$$
$$\sin\theta = \frac{1}{2i}(e^{i\theta} - e^{-i\theta}) = \frac{1}{2i}\left(z - \frac{1}{z}\right). \tag{2}$$

Since F is rational in $\cos\theta$ and $\sin\theta$, Eq. (2) shows that F is now a rational function of z, say, $f(z)$. Since $dz/d\theta = ie^{i\theta}$, we have $d\theta = dz/iz$ and the given integral takes the form

$$J = \oint_C f(z)\,\frac{dz}{iz} \tag{3}$$

and, as θ ranges from 0 to 2π in (1), the variable $z = e^{i\theta}$ ranges counterclockwise once around the unit circle $|z| = 1$. (Review Sec. 13.5 if necessary.)

EXAMPLE 1 **An Integral of the Type (1)**

Show by the present method that $\displaystyle\int_0^{2\pi} \frac{d\theta}{\sqrt{2} - \cos\theta} = 2\pi.$

Solution. We use $\cos\theta = \frac{1}{2}(z + 1/z)$ and $d\theta = dz/iz$. Then the integral becomes

$$\oint_C \frac{dz/iz}{\sqrt{2} - \frac{1}{2}\left(z + \frac{1}{z}\right)} = \oint_C \frac{dz}{-\frac{i}{2}(z^2 - 2\sqrt{2}z + 1)}$$
$$= -\frac{2}{i}\oint_C \frac{dz}{(z - \sqrt{2} - 1)(z - \sqrt{2} + 1)}.$$

We see that the integrand has a simple pole at $z_1 = \sqrt{2} + 1$ outside the unit circle C, so that it is of no interest here, and another simple pole at $z_2 = \sqrt{2} - 1$ (where $z - \sqrt{2} + 1 = 0$) inside C with residue [by (3), Sec. 16.3]

$$\operatorname*{Res}_{z=z_2} \frac{1}{(z - \sqrt{2} - 1)(z - \sqrt{2} + 1)} = \left[\frac{1}{z - \sqrt{2} - 1}\right]_{z=\sqrt{2}-1}$$
$$= -\frac{1}{2}.$$

Answer: $2\pi i(-2/i)(-1/2) = 2\pi$. (Here $-2/i$ is the factor in front of the last integral.)

As another large class, let us consider real integrals of the form

$$\int_{-\infty}^{\infty} f(x)\,dx. \tag{4}$$

Such an integral, whose interval of integration is not finite is called an **improper integral,** and it has the meaning

$$\int_{-\infty}^{\infty} f(x)\,dx = \lim_{a\to-\infty} \int_{a}^{0} f(x)\,dx + \lim_{b\to\infty} \int_{0}^{b} f(x)\,dx. \tag{5'}$$

If both limits exist, we may couple the two independent passages to $-\infty$ and ∞, and write

$$\int_{-\infty}^{\infty} f(x)\,dx = \lim_{R\to\infty} \int_{-R}^{R} f(x)\,dx. \tag{5}$$

The limit in (5) is called the **Cauchy principal value** of the integral. It is written

$$\text{pr. v.} \int_{-\infty}^{\infty} f(x)\,dx.$$

It may exist even if the limits in (5′) do not. *Example:*

$$\lim_{R\to\infty} \int_{-R}^{R} x\,dx = \lim_{R\to\infty} \left(\frac{R^2}{2} - \frac{R^2}{2}\right) = 0, \qquad \text{but} \qquad \lim_{b\to\infty} \int_{0}^{b} x\,dx = \infty.$$

We assume that the function $f(x)$ in (4) is a real rational function whose denominator is different from zero for all real x and is of degree at least two units higher than the degree of the numerator. Then the limits in (5′) exist, and we may start from (5). We consider the corresponding contour integral

$$\oint_C f(z)\,dz \tag{5*}$$

around a path C in Fig. 371. Since $f(x)$ is rational, $f(z)$ has finitely many poles in the upper half-plane, and if we choose R large enough, then C encloses all these poles. By the residue theorem we then obtain

$$\oint_C f(z)\,dz = \int_S f(z)\,dz + \int_{-R}^{R} f(x)\,dx = 2\pi i \sum \text{Res } f(z)$$

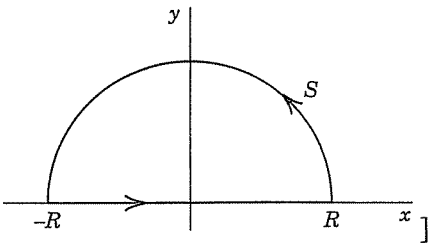


Fig. 371. Path C of the contour integral in (5*)

where the sum consists of all the residues of $f(z)$ at the points in the upper half-plane at which $f(z)$ has a pole. From this we have

$$\int_{-R}^{R} f(x)\,dx = 2\pi i \sum \operatorname{Res} f(z) - \int_S f(z)\,dz. \tag{6}$$

We prove that, if $R \to \infty$, the value of the integral over the semicircle S approaches zero. If we set $z = Re^{i\theta}$, then S is represented by $R = const$, and as z ranges along S, the variable θ ranges from 0 to π. Since, by assumption, the degree of the denominator of $f(z)$ is at least two units higher than the degree of the numerator, we have

$$|f(z)| < \frac{k}{|z|^2} \qquad (|z| = R > R_0)$$

for sufficiently large constants k and R_0. By the *ML*-inequality in Sec. 14.1,

$$\left|\int_S f(z)\,dz\right| < \frac{k}{R^2}\,\pi R = \frac{k\pi}{R} \qquad (R > R_0).$$

Hence, as R approaches infinity, the value of the integral over S approaches zero, and (5) and (6) yield the result

$$\int_{-\infty}^{\infty} f(x)\,dx = 2\pi i \sum \operatorname{Res} f(z) \tag{7}$$

where we sum over all the residues of $f(z)$ at the poles of $f(z)$ in the upper half-plane.

EXAMPLE 2 An Improper Integral from 0 to ∞

Using (7), show that

$$\int_0^{\infty} \frac{dx}{1+x^4} = \frac{\pi}{2\sqrt{2}}.$$

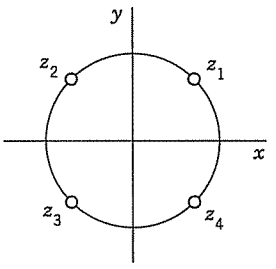


Fig. 372. Example 2

Solution. Indeed, $f(z) = 1/(1+z^4)$ has four simple poles at the points (make a sketch)

$$z_1 = e^{\pi i/4}, \qquad z_2 = e^{3\pi i/4}, \qquad z_3 = e^{-3\pi i/4}, \qquad z_4 = e^{-\pi i/4}.$$

The first two of these poles lie in the upper half-plane (Fig. 372). From (4) in the last section we find the residues

$$\operatorname*{Res}_{z=z_1} f(z) = \left[\frac{1}{(1+z^4)'}\right]_{z=z_1} = \left[\frac{1}{4z^3}\right]_{z=z_1} = \frac{1}{4}\,e^{-3\pi i/4} = -\frac{1}{4}\,e^{\pi i/4}.$$

$$\operatorname*{Res}_{z=z_2} f(z) = \left[\frac{1}{(1+z^4)'}\right]_{z=z_2} = \left[\frac{1}{4z^3}\right]_{z=z_2} = \frac{1}{4}\,e^{-9\pi i/4} = \frac{1}{4}\,e^{-\pi i/4}.$$

(Here we used $e^{\pi i} = -1$ and $e^{-2\pi i} = 1$.) By (1) in Sec. 13.6 and (7) in this section,

$$\int_{-\infty}^{\infty} \frac{dx}{1+x^4} = -\frac{2\pi i}{4}\,(e^{\pi i/4} - e^{-\pi i/4}) = -\frac{2\pi i}{4}\cdot 2i\cdot \sin\frac{\pi}{4} = \pi \sin\frac{\pi}{4} = \frac{\pi}{\sqrt{2}}.$$

Since $1/(1+x^4)$ is an even function, we thus obtain, as asserted,

$$\int_0^{\infty} \frac{dx}{1+x^4} = \frac{1}{2}\int_{-\infty}^{\infty} \frac{dx}{1+x^4} = \frac{\pi}{2\sqrt{2}}. \qquad \blacksquare$$

Fourier Integrals

The method of evaluating (4) by creating a closed contour (Fig. 371) and "blowing it up" extends to integrals

$$(8) \qquad \int_{-\infty}^{\infty} f(x)\cos sx\,dx \qquad \text{and} \qquad \int_{-\infty}^{\infty} f(x)\sin sx\,dx \qquad (s \text{ real})$$

as they occur in connection with the Fourier integral (Sec. 11.7).

If $f(x)$ is a rational function satisfying the assumption on the degree as for (4), we may consider the corresponding integral

$$\oint_C f(z)e^{isz}\,dz \qquad (s \text{ real and positive})$$

over the contour C in Fig. 371 on p. 719. Instead of (7) we now get

$$(9) \qquad \int_{-\infty}^{\infty} f(x)e^{isx}\,dx = 2\pi i \sum \operatorname{Res}\left[f(z)e^{isz}\right] \qquad (s > 0)$$

where we sum the residues of $f(z)e^{isz}$ at its poles in the upper half-plane. Equating the real and the imaginary parts on both sides of (9), we have

$$\textbf{(10)} \qquad \begin{aligned} \int_{-\infty}^{\infty} f(x)\cos sx\,dx &= -2\pi \sum \operatorname{Im}\operatorname{Res}\left[f(z)e^{isz}\right], \\ \int_{-\infty}^{\infty} f(x)\sin sx\,dx &= 2\pi \sum \operatorname{Re}\operatorname{Res}\left[f(z)e^{isz}\right]. \end{aligned} \qquad (s > 0)$$

To establish (9), we must show [as for (4)] that the value of the integral over the semicircle S in Fig. 371 approaches 0 as $R \to \infty$. Now $s > 0$ and S lies in the upper half-plane $y \geqq 0$. Hence

$$|e^{isz}| = |e^{is(x+iy)}| = |e^{isx}|\,|e^{-sy}| = 1\cdot e^{-sy} \leqq 1 \qquad (s > 0, \quad y \geqq 0).$$

From this we obtain the inequality $|f(z)\,e^{isz}| = |f(z)|\,|e^{isz}| \leqq |f(z)|$ $(s > 0, \quad y \geqq 0)$. This reduces our present problem to that for (4). Continuing as before gives (9) and (10). $\blacksquare$

EXAMPLE 3 An Application of (10)

Show that $$\int_{-\infty}^{\infty} \frac{\cos sx}{k^2 + x^2}\, dx = \frac{\pi}{k}\, e^{-ks}, \qquad \int_{-\infty}^{\infty} \frac{\sin sx}{k^2 + x^2}\, dx = 0 \qquad (s > 0,\ k > 0).$$

Solution. In fact, $e^{isz}/(k^2 + z^2)$ has only one pole in the upper half-plane, namely, a simple pole at $z = ik$, and from (4) in Sec. 16.3 we obtain

$$\operatorname*{Res}_{z=ik} \frac{e^{isz}}{k^2 + z^2} = \left[\frac{e^{isz}}{2z}\right]_{z=ik} = \frac{e^{-ks}}{2ik}.$$

Thus

$$\int_{-\infty}^{\infty} \frac{e^{isx}}{k^2 + x^2}\, dx = 2\pi i\, \frac{e^{-ks}}{2ik} = \frac{\pi}{k}\, e^{-ks}.$$

Since $e^{isx} = \cos sx + i \sin sx$, this yields the above results [see also (15) in Sec. 11.7.] ∎

Another Kind of Improper Integral

We consider an improper integral

$$\int_A^B f(x)\, dx \tag{11}$$

whose integrand becomes infinite at a point a in the interval of integration,

$$\lim_{x \to a} |f(x)| = \infty.$$

By definition, this integral (11) means

$$\int_A^B f(x)\, dx = \lim_{\epsilon \to 0} \int_A^{a-\epsilon} f(x)\, dx + \lim_{\eta \to 0} \int_{a+\eta}^B f(x)\, dx \tag{12}$$

where both ϵ and η approach zero independently and through positive values. It may happen that neither of these two limits exists if ϵ and η go to 0 independently, but the limit

$$\lim_{\epsilon \to 0} \left[\int_A^{a-\epsilon} f(x)\, dx + \int_{a+\epsilon}^B f(x)\, dx\right] \tag{13}$$

exists. This is called the **Cauchy principal value** of the integral. It is written

$$\text{pr. v.} \int_A^B f(x)\, dx.$$

For example,

$$\text{pr. v.} \int_{-1}^1 \frac{dx}{x^3} = \lim_{\epsilon \to 0} \left[\int_{-1}^{-\epsilon} \frac{dx}{x^3} + \int_{\epsilon}^1 \frac{dx}{x^3}\right] = 0;$$

the principal value exists, although the integral itself has no meaning.

In the case of simple poles on the real axis we shall obtain a formula for the principal value of an integral from $-\infty$ to ∞. This formula will result from the following theorem.

THEOREM 1

Simple Poles on the Real Axis

If $f(z)$ has a simple pole at $z = a$ on the real axis, then (Fig. 373)

$$\lim_{r \to 0} \int_{C_2} f(z)\, dz = \pi i \operatorname*{Res}_{z=a} f(z).$$

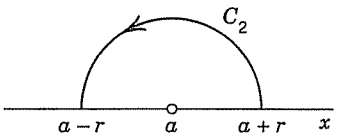


Fig. 373. Theorem 1

PROOF By the definition of a simple pole (Sec. 16.2) the integrand $f(z)$ has for $0 < |z - a| < R$ the Laurent series

$$f(z) = \frac{b_1}{z - a} + g(z), \qquad b_1 = \operatorname*{Res}_{z=a} f(z).$$

Here $g(z)$ is analytic on the semicircle of integration (Fig. 373)

$$C_2\colon \quad z = a + re^{i\theta}, \qquad 0 \leqq \theta \leqq \pi$$

and for all z between C_2 and the x-axis, and thus bounded on C_2, say, $|g(z)| \leqq M$. By integration,

$$\int_{C_2} f(z)\, dz = \int_0^{\pi} \frac{b_1}{re^{i\theta}}\, ire^{i\theta}\, d\theta + \int_{C_2} g(z)\, dz = b_1 \pi i + \int_{C_2} g(z)\, dz.$$

The second integral on the right cannot exceed $M\pi r$ in absolute value, by the ML-inequality (Sec. 14.1), and $ML = M\pi r \to 0$ as $r \to 0$. ∎

Figure 374 shows the idea of applying Theorem 1 to obtain the principal value of the integral of a rational function $f(x)$ from $-\infty$ to ∞. For sufficiently large R the integral over the entire contour in Fig. 374 has the value J given by $2\pi i$ times the sum of the residues of $f(z)$ at the singularities in the upper half-plane. We assume that $f(x)$ satisfies the degree

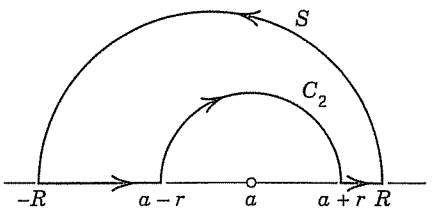


Fig. 374. Application of Theorem 1

condition imposed in connection with (4). Then the value of the integral over the large semicircle S approaches 0 as $R \to \infty$. For $r \to 0$ the integral over C_2 (clockwise!) approaches the value

$$K = -\pi i \operatorname*{Res}_{z=a} f(z)$$

by Theorem 1. Together this shows that the principal value P of the integral from $-\infty$ to ∞ plus K equals J; hence $P = J - K = J + \pi i \operatorname{Res}_{z=a} f(z)$. If $f(z)$ has several simple poles on the real axis, then K will be $-\pi i$ times the sum of the corresponding residues. Hence the desired formula is

$$\text{(14)} \qquad \text{pr. v.} \int_{-\infty}^{\infty} f(x)\,dx = 2\pi i \sum \operatorname{Res} f(z) + \pi i \sum \operatorname{Res} f(z)$$

where the first sum extends over all poles in the upper half-plane and the second over all poles on the real axis, the latter being simple by assumption.

EXAMPLE 4 Poles on the Real Axis

Find the principal value

$$\text{pr. v.} \int_{-\infty}^{\infty} \frac{dx}{(x^2 - 3x + 2)(x^2 + 1)}.$$

Solution. Since

$$x^2 - 3x + 2 = (x - 1)(x - 2),$$

the integrand $f(x)$, considered for complex z, has simple poles at

$$z = 1, \qquad \operatorname*{Res}_{z=1} f(z) = \left[\frac{1}{(z - 2)(z^2 + 1)}\right]_{z=1} = -\frac{1}{2},$$

$$z = 2, \qquad \operatorname*{Res}_{z=2} f(z) = \left[\frac{1}{(z - 1)(z^2 + 1)}\right]_{z=2} = \frac{1}{5},$$

$$z = i, \qquad \operatorname*{Res}_{z=i} f(z) = \left[\frac{1}{(z^2 - 3z + 2)(z + i)}\right]_{z=i} = \frac{1}{6 + 2i} = \frac{3 - i}{20},$$

and at $z = -i$ in the lower half-plane, which is of no interest here. From (14) we get the answer

$$\text{pr. v.} \int_{-\infty}^{\infty} \frac{dx}{(x^2 - 3x + 2)(x^2 + 1)} = 2\pi i \left(\frac{3 - i}{20}\right) + \pi i \left(-\frac{1}{2} + \frac{1}{5}\right) = \frac{\pi}{10}.$$

More integrals of the kind considered in this section are included in the problem set. Try also your CAS, which may sometimes give you false results on complex integrals.

PROBLEM SET 16.4

1–8 INTEGRALS INVOLVING COSINE AND SINE

Evaluate the following integrals. (Show the details of your work.)

1. $\int_0^{2\pi} \frac{d\theta}{7 + 6\cos\theta}$

2. $\int_0^{\pi} \frac{d\theta}{2 + \cos\theta}$

3. $\int_0^{2\pi} \frac{d\theta}{37 - 12\cos\theta}$

4. $\int_0^{2\pi} \frac{d\theta}{8 - 2\sin\theta}$

5. $\int_0^{2\pi} \frac{d\theta}{5 - 4\sin\theta}$

6. $\int_0^{2\pi} \frac{\sin^2\theta}{5 - 4\cos\theta}\, d\theta$

7. $\int_0^{2\pi} \frac{\cos\theta}{13 - 12\cos 2\theta}\, d\theta.$

Hint. $\cos 2\theta = \frac{1}{2}\left(z^2 + \frac{1}{z^2}\right)$

8. $\int_0^{2\pi} \frac{1 + 4\cos\theta}{17 - 8\cos\theta}\, d\theta$

9–22 IMPROPER INTEGRALS: INFINITE INTERVAL OF INTEGRATION

Evaluate (showing the details):

9. $\int_{-\infty}^{\infty} \frac{dx}{x^2 + 1}$

10. $\int_{-\infty}^{\infty} \frac{x}{x^4 + 1}\, dx$

11. $\int_{-\infty}^{\infty} \frac{dx}{x^6 + 1}$

12. $\int_{-\infty}^{\infty} \frac{dx}{(x^2 - 2x + 5)^2}$

13. $\int_{-\infty}^{\infty} \frac{dx}{(x^2 + 4)^2}$

14. $\int_{-\infty}^{\infty} \frac{dx}{x^4 + 16}$

15. $\int_{-\infty}^{\infty} \frac{x^3}{1 + x^8}\, dx$

16. $\int_{-\infty}^{\infty} \frac{dx}{(x^2 + 1)(x^2 + 9)}$

17. $\int_{-\infty}^{\infty} \frac{x}{(x^2 - 2x + 2)^2}\, dx$

18. $\int_{-\infty}^{\infty} \frac{x^2 + 1}{x^4 + 1}\, dx$

19. $\int_{-\infty}^{\infty} \frac{\sin x}{x^4 + 1}\, dx$

20. $\int_{-\infty}^{\infty} \frac{\cos x}{x^4 + 1}\, dx$

21. $\int_{-\infty}^{\infty} \frac{\sin 3x}{x^4 + 1}\, dx$

22. $\int_{-\infty}^{\infty} \frac{\cos 4x}{x^4 + 5x^2 + 4}\, dx$

23–27 IMPROPER INTEGRALS: POLES ON THE REAL AXIS

Find the Cauchy principal value (showing details):

23. $\int_{-\infty}^{\infty} \frac{x + 2}{x^3 + x}\, dx$

24. $\int_{-\infty}^{\infty} \frac{x^2}{x^4 - 1}\, dx$

25. $\int_{-\infty}^{\infty} \frac{x + 5}{x^3 - x}\, dx$

26. $\int_{-\infty}^{\infty} \frac{dx}{x^4 + 3x^2 - 4}$

27. $\int_{-\infty}^{\infty} \frac{dx}{x^4 - 1}$

28. **TEAM PROJECT. Comments on Real Integrals.**
(a) Formula (10) follows from (9). Give the details.

(b) Use of auxiliary results. Integrating e^{-z^2} around the boundary C of the rectangle with vertices $-a$, a, $a + ib$, $-a + ib$, letting $a \to \infty$, and using

$$\int_0^{\infty} e^{-x^2}\, dx = \frac{\sqrt{\pi}}{2},$$

show that

$$\int_0^{\infty} e^{-x^2} \cos 2bx\, dx = \frac{\sqrt{\pi}}{2}\, e^{-b^2}.$$

(This integral is needed in heat conduction in Sec. 12.6.)

(c) Inspection. Solve Probs. 15 and 21 without calculation.

29. **CAS EXPERIMENT. Check your CAS.** Find out to what extent your CAS can evaluate integrals of the form (1), (4), and (8) correctly. Do this by comparing the results of direct integration (which may come out false) with those of using residues.

30. **CAS EXPERIMENT. Simple Poles on the Real Axis.** Experiment with integrals $\int_{-\infty}^{\infty} f(x)\, dx$, $f(x) = [(x - a_1)(x - a_2) \cdots (x - a_k)]^{-1}$, a_j real and all different, $k > 1$. Conjecture that the principal value of these integrals is 0. Try to prove this for a special k, say, $k = 3$. For general k.

CHAPTER 16 REVIEW QUESTIONS AND PROBLEMS

1. Laurent series generalize Taylor series. Explain the details.
2. Can a function have several Laurent series with the same center? Explain. If your answer is yes, give examples.
3. What is the principal part of a Laurent series? Its significance?
4. What is a pole? An essential singularity? Give examples.
5. What is Picard's theorem? Why did it occur in this chapter?
6. What is the Riemann sphere? The extended complex plane? Its significance?
7. Is e^{1/z^2} analytic or singular at infinity? $\cosh z$? $(z-4)^3$? Explain.
8. What is the residue? Why is it important?
9. State formulas for residues from memory.
10. State some further methods for calculating residues.
11. What is residue integration? To what kind of complex integrals does it apply?
12. By what idea can we apply residue integration to real integrals from $-\infty$ to ∞? Give simple examples.
13. What is a zero of an analytic function? How are zeros classified?
14. What are improper integrals? Cauchy principal values? Give examples.
15. Can the residue at a singular point be 0? At a simple pole?
16. What is a meromorphic function? An entire function? Give examples.

17–28 COMPLEX INTEGRALS

Integrate counterclockwise around C. (Show the details.)

17. $\dfrac{\tan z}{z^4}$, C: $|z| = 1$
18. $\dfrac{\sin 2z}{z^4}$, C: $|z| = 1$
19. $\dfrac{10z}{2z+i}$, C: $|z-2i| = 3$
20. $\dfrac{iz+1}{z^2-iz+2}$, C: $|z-1| = 3$
21. $\dfrac{\cosh 5z}{z^2+4}$, C: $|z-i| = 2$
22. $\dfrac{4z^3+7z}{\cos z}$, C: $|z+1| = 1$
23. $\cot 8z$, C: $|z| = 0.2$
24. $\dfrac{z^2 \sin z}{4z^2-1}$, C: $|z-1| = 2$
25. $\dfrac{\cos z}{z^n}$, $n = 1, 2, \cdots$, C: $|z| = 1$
26. $\dfrac{z^2+1}{z^2-2z}$, C: $\dfrac{1}{2}x^2 + y^2 = 1$
27. $\dfrac{15z+9}{z^3-9z}$, C: $|z-3| = 2$
28. $\dfrac{15z+9}{z^3-9z}$, C: $|z| = 4$

29–35 REAL INTEGRALS

Evaluate by the methods of this chapter (showing the details):

29. $\displaystyle\int_0^{2\pi} \frac{d\theta}{25 - 24\cos\theta}$
30. $\displaystyle\int_0^{\pi} \frac{d\theta}{k + \cos\theta}$, $k > 1$
31. $\displaystyle\int_0^{2\pi} \frac{d\theta}{1 - \frac{1}{2}\sin\theta}$
32. $\displaystyle\int_0^{2\pi} \frac{\sin\theta}{3 + \cos\theta}\, d\theta$
33. $\displaystyle\int_{-\infty}^{\infty} \frac{x}{(1+x^2)^2}\, dx$
34. $\displaystyle\int_0^{\infty} \frac{dx}{(1+x^2)^2}$
35. $\displaystyle\int_0^{\infty} \frac{1+2x^2}{1+4x^4}\, dx$
36. Obtain the answer to Prob. 18 in Sec. 16.4 from the present Prob. 35.

SUMMARY OF CHAPTER 16
Laurent Series. Residue Integration

A **Laurent series** is a series of the form

$$(1) \qquad f(z) = \sum_{n=0}^{\infty} a_n(z - z_0)^n + \sum_{n=1}^{\infty} \frac{b_n}{(z - z_0)^n} \qquad \text{(Sec. 16.1)}$$

or, more briefly written [but this means the same as (1)!]

$$(1^*) \qquad f(z) = \sum_{n=-\infty}^{\infty} a_n(z - z_0)^n, \qquad a_n = \frac{1}{2\pi i} \oint_C \frac{f(z^*)}{(z^* - z_0)^{n+1}}\, dz^*$$

where $n = 0, \pm 1, \pm 2, \cdots$. This series converges in an open annulus (ring) A with center z_0. In A the function $f(z)$ is analytic. At points not in A it may have singularities. The first series in (1) is a power series. In a given annulus, a Laurent series of $f(z)$ is unique, but $f(z)$ may have different Laurent series in different annuli with the same center.

Of particular importance is the Laurent series (1) that converges in a neighborhood of z_0 except at z_0 itself, say, for $0 < |z - z_0| < R$ ($R > 0$, suitable). The series (or finite sum) of the negative powers in this Laurent series is called the **principal part** of $f(z)$ at z_0. The coefficient b_1 of $1/(z - z_0)$ in this series is called the **residue** of $f(z)$ at z_0 and is given by [see (1) and (1*)]

$$(2) \quad b_1 = \operatorname*{Res}_{z \to z_0} f(z) = \frac{1}{2\pi i} \oint_C f(z^*)\, dz^*. \qquad \text{Thus} \qquad \oint_C f(z^*)\, dz^* = 2\pi i \operatorname*{Res}_{z = z_0} f(z).$$

b_1 can be used for **integration** as shown in (2) because it can be found from

$$(3) \qquad \operatorname*{Res}_{z = z_0} f(z) = \frac{1}{(m-1)!} \lim_{z \to z_0} \left(\frac{d^{m-1}}{dz^{m-1}} \left[(z - z_0)^m f(z) \right] \right), \qquad \text{(Sec. 16.3),}$$

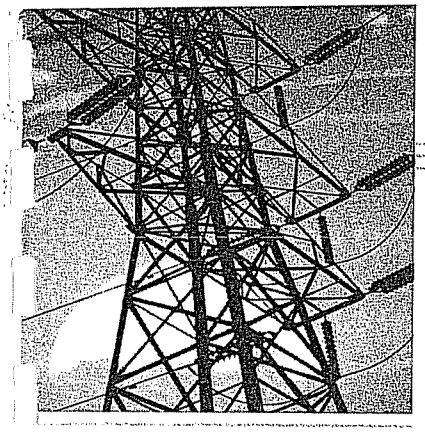
provided $f(z)$ has at z_0 a **pole of order *m*;** by definition this means that that principal part has $1/(z - z_0)^m$ as its highest negative power. Thus for a simple pole ($m = 1$),

$$\operatorname*{Res}_{z = z_0} f(z) = \lim_{z \to z_0} (z - z_0) f(z); \qquad \text{also,} \qquad \operatorname*{Res}_{z = z_0} \frac{p(z)}{q(z)} = \frac{p(z_0)}{q'(z_0)}.$$

If the principal part is an infinite series, the singularity of $f(z)$ at z_0 is called an **essential singularity** (Sec. 16.2).

Section 16.2 also discusses the *extended complex plane,* that is, the complex plane with an improper point ∞ ("infinity") attached.

Residue integration may also be used to evaluate certain classes of complicated real integrals (Sec. 16.4).



CHAPTER 17

Conformal Mapping

If a complex function $w = f(z)$ is defined in a domain D of the z-plane, then to each point in D there corresponds a point in the w-plane. In this way we obtain a **mapping** of D onto the range of values of $f(z)$ in the w-plane. We shall see that if $f(z)$ is an *analytic* function, then the mapping given by $w = f(z)$ is **conformal** (angle-preserving), except at points where the derivative $f'(z)$ is zero.

Conformality appeared early in history in connection with constructing maps of the globe, which can be conformal (can give directions correctly) or "equiareal" (give areas correctly, except for a scale factor), but cannot have both properties, as can be proved (see [GR8] in App. 1).

Conformality is the most important geometric property of analytic functions and gives the possibility of a geometric approach to complex analysis. Indeed, just as in calculus we use curves of real functions $y = f(x)$ for studying "geometric" properties of functions, in complex analysis we can use conformal mappings for obtaining a deeper understanding of properties of functions, notably of those discussed in Chap. 13.

Indeed, we shall first define the concepts of conformal mapping and then consider mappings by those elementary analytic functions in Chap. 13.

This is one purpose of this chapter. A second purpose, more important to the engineer and physicist, is the use of conformal mapping in connection with potential problems. In fact, in this chapter and in the next one we shall see that conformal mapping yields a standard method for solving **boundary value problems** in (two-dimensional) potential theory by transforming a complicated region into a simpler one. Corresponding applications will concern problems from electrostatics, heat flow, and fluid flow.

In the last section (17.5) we explain the concept of a **Riemann surface,** which fits well into the present discussion of "geometric" ideas.

Prerequisite: Chap. 13.
Sections that may be omitted in a shorter course: 17.3 and 17.5
References and Answers to Problems: App. 1 Part D, App. 2.

17.1 Geometry of Analytic Functions: Conformal Mapping

A complex function

$$w = f(z) = u(x, y) + iv(x, y) \qquad (z = x + iy) \tag{1}$$

of a complex variable z gives a **mapping** of its domain of definition D in the complex z-plane ***into*** the complex w-plane or ***onto*** its range of values in that plane.[1] For any point z_0 in D the point $w_0 = f(z_0)$ is called the **image** of z_0 with respect to f. More generally, for the points of a curve C in D the image points form the **image** of C; similarly for other point sets in D. Also, instead of *the mapping by a function* $w = f(z)$ we shall say more briefly ***the mapping*** $w = f(z)$.

EXAMPLE 1 **Mapping $w = f(z) = z^2$**

Using polar forms $z = re^{i\theta}$ and $w = Re^{i\phi}$, we have $w = z^2 = r^2e^{2i\theta}$. Comparing moduli and arguments gives $R = r^2$ and $\phi = 2\theta$. Hence circles $r = r_0$ are mapped onto circles $R = r_0^2$ and rays $\theta = \theta_0$ onto rays $\phi = 2\theta_0$. Figure 375 shows this for the region $1 \leqq |z| \leqq 3/2$, $\pi/6 \leqq \theta \leqq \pi/3$, which is mapped onto the region $1 \leqq |w| \leqq 9/4$, $\pi/3 \leqq \theta \leqq 2\pi/3$.

In Cartesian coordinates we have $z = x + iy$ and

$$u = \operatorname{Re}(z^2) = x^2 - y^2, \qquad v = \operatorname{Im}(z^2) = 2xy.$$

Hence vertical lines $x = c = const$ are mapped onto $u = c^2 - y^2$, $v = 2cy$. From this we can eliminate y. We obtain $y^2 = c^2 - u$ and $v^2 = 4c^2y^2$. Together,

$$v^2 = 4c^2(c^2 - u) \qquad \text{(Fig. 376).}$$

These parabolas open to the left. Similarly, horizontal lines $y = k = const$ are mapped onto parabolas opening to the right,

$$v^2 = 4k^2(k^2 + u) \qquad \text{(Fig. 376).}$$

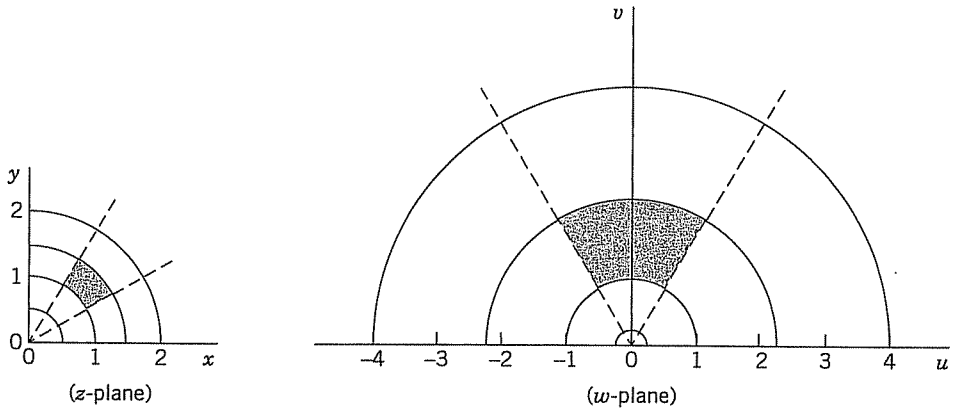


Fig. 375. Mapping $w = z^2$. Lines $|z| = const$, $\arg z = const$ and their images in the w-plane

[1]The general terminology is as follows. A mapping of a set A into a set B is called **surjective** or a mapping of A **onto** B if every element of B is the image of at least one element of A. It is called **injective** or **one-to-one** if different elements of A have different images in B. Finally, it is called **bijective** if it is both surjective and injective.

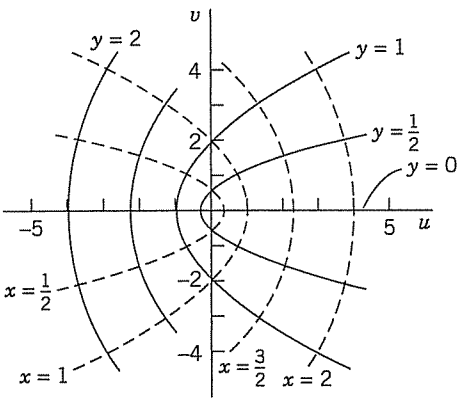


Fig. 376. Images of $x = const$, $y = const$ under $w = z^2$

Conformal Mapping

A mapping $w = f(z)$ is called **conformal** if it preserves angles between oriented curves in magnitude as well as in sense. Figure 377 shows what this means. The **angle** α $(0 \leqq \alpha \leqq \pi)$ between two intersecting curves C_1 and C_2 is defined to be the angle between their oriented tangents at the intersection point z_0. And *conformality* means that the images C_1^* and C_2^* of C_1 and C_2 make the same angle as the curves themselves in both magnitude and direction.

THEOREM 1

Conformality of Mapping by Analytic Functions

The mapping $w = f(z)$ by an analytic function f is conformal, except at **critical points,** *that is, points at which the derivative f' is zero.*

PROOF $w = z^2$ has a critical point at $z = 0$, where $f'(z) = 2z = 0$ and the angles are doubled (see Fig. 375), so that conformality fails.

The idea of proof is to consider a curve

$$C: z(t) = x(t) + iy(t) \tag{2}$$

in the domain of $f(z)$ and to show that $w = f(z)$ rotates all tangents at a point z_0 (where $f'(z_0) \neq 0$) through the same angle. Now $\dot{z}(t) = dz/dt = \dot{x}(t) + i\dot{y}(t)$ is tangent to C in (2) because this is the limit of $(z_1 - z_0)/\Delta t$ (which has the direction of the secant $z_1 - z_0$

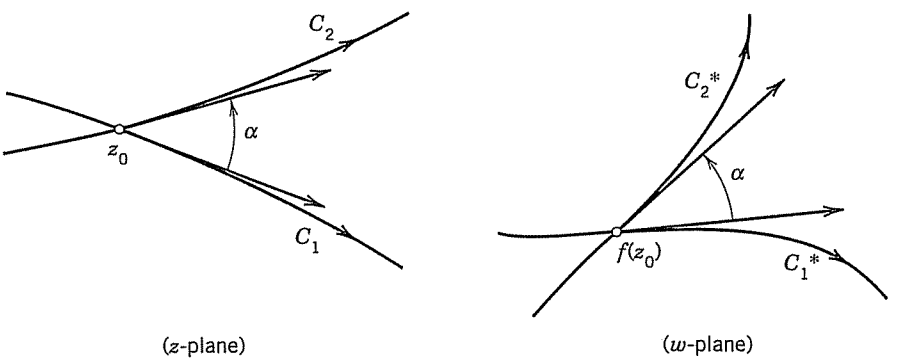


Fig. 377. Curves C_1 and C_2 and their respective images C_1^* and C_2^* under a conformal mapping $w = f(z)$

in Fig. 378) as z_1 approaches z_0 along C. The image C^* of C is $w = f(z(t))$. By the chain rule, $\dot{w} = f'(z(t))\dot{z}(t)$. Hence the tangent direction of C^* is given by the argument (use (9) in Sec. 13.2)

$$\arg \dot{w} = \arg f' + \arg \dot{z} \tag{3}$$

where $\arg \dot{z}$ gives the tangent direction of C. This shows that the mapping rotates *all* directions at a point z_0 in the domain of analyticity of f through the same angle $\arg f'(z_0)$, which exists as long as $f'(z_0) \neq 0$. But this means conformality, as Fig. 377 illustrates for an angle α between two curves, whose images C_1^* and C_2^* make the same angle (because of the rotation). ∎

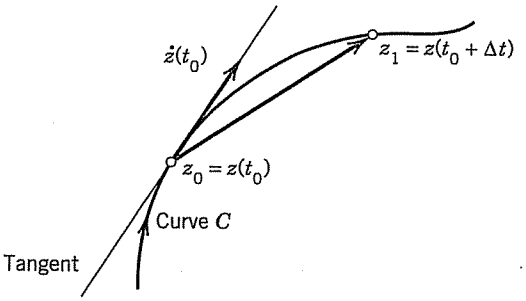


Fig. 378. Secant and tangent of the curve C

In the remainder of this section and in the next ones we shall consider various conformal mappings that are of practical interest, for instance, in modeling potential problems.

EXAMPLE 2 **Conformality of $w = z^n$**

The mapping $w = z^n$, $n = 2, 3, \cdots$, is conformal, except at $z = 0$, where $w' = nz^{n-1} = 0$. For $n = 2$ this is shown in Fig. 375; we see that at 0 the angles are doubled. For general n the angles at 0 are multiplied by a factor n under the mapping. Hence the sector $0 \leqq \theta \leqq \pi/n$ is mapped by z^n onto the upper half-plane $v \geqq 0$ (Fig. 379). ∎

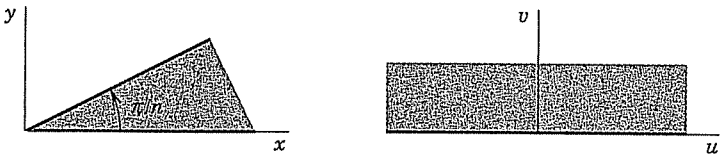


Fig. 379. Mapping by $w = z^n$

EXAMPLE 3 **Mapping $w = z + 1/z$. Joukowski Airfoil**

In terms of polar coordinates this mapping is

$$w = u + iv = r(\cos\theta + i\sin\theta) + \frac{1}{r}(\cos\theta - i\sin\theta).$$

By separating the real and imaginary parts we thus obtain

$$u = a\cos\theta, \qquad v = b\sin\theta \qquad \text{where} \qquad a = r + \frac{1}{r}, \quad b = r - \frac{1}{r}.$$

Hence circles $|z| = r = const \neq 1$ are mapped onto ellipses $x^2/a^2 + y^2/b^2 = 1$. The circle $r = 1$ is mapped onto the segment $-2 \leqq u \leqq 2$ of the u-axis. See Fig. 380.

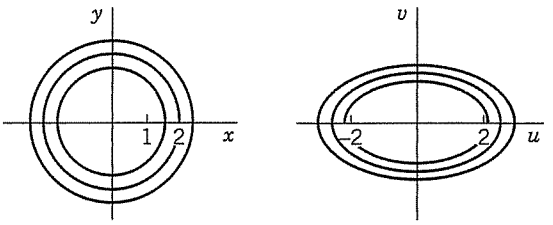


Fig. 380. Example 3

Now the derivative of w is

$$w' = 1 - \frac{1}{z^2} = \frac{(z+1)(z-1)}{z^2}$$

which is 0 at $z = \pm 1$. These are the points at which the mapping is not conformal. The two circles in Fig. 381 pass through $z = -1$. The larger is mapped onto a *Joukowski airfoil.* The dashed circle passes through both -1 and 1 and is mapped onto a curved segment.

Another interesting application of $w = z + 1/z$ (the flow around a cylinder) will be considered in Sec. 18.4. ■

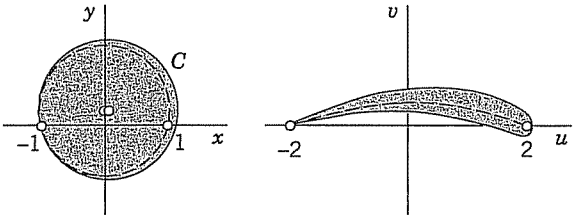


Fig. 381. Joukowski airfoil

EXAMPLE 4 **Conformality of $w = e^z$**

From (10) in Sec. 13.5 we have $|e^z| = e^x$ and $\operatorname{Arg} z = y$. Hence e^z maps a vertical straight line $x = x_0 = const$ onto the circle $|w| = e^{x_0}$ and a horizontal straight line $y = y_0 = const$ onto the ray $\arg w = y_0$. The rectangle in Fig. 382 is mapped onto a region bounded by circles and rays as shown.

The fundamental region $-\pi < \operatorname{Arg} z \leqq \pi$ of e^z in the z-plane is mapped bijectively and conformally onto the entire w-plane without the origin $w = 0$ (because $e^z = 0$ for no z). Figure 383 shows that the upper half $0 < y \leqq \pi$ of the fundamental region is mapped onto the upper half-plane $0 < \arg w \leqq \pi$, the left half being mapped inside the unit disk $|w| \leqq 1$ and the right half outside (why?). ■

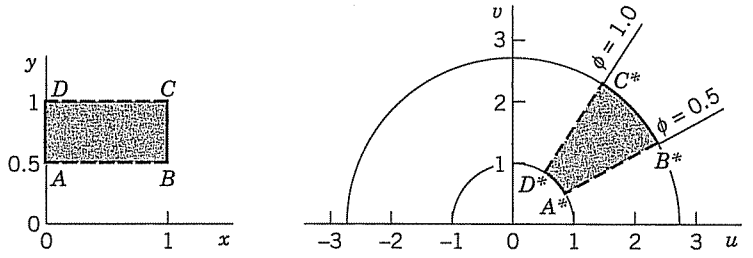


Fig. 382. Mapping by $w = e^z$

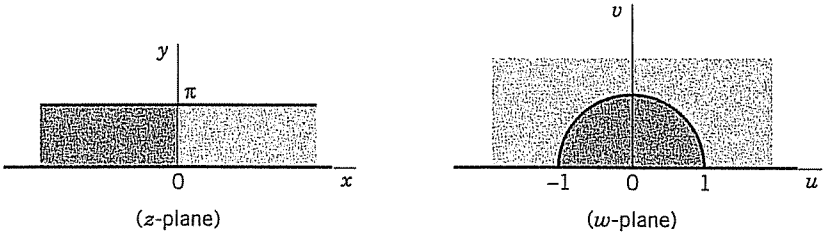


Fig. 383. Mapping by $w = e^z$

EXAMPLE 5 Principle of Inverse Mapping. Mapping $w = \operatorname{Ln} z$

Principle. *The mapping by the inverse $z = f^{-1}(w)$ of $w = f(z)$ is obtained by interchanging the roles of the z-plane and the w-plane in the mapping by $w = f(z)$.*

Now the principal value $w = f(z) = \operatorname{Ln} z$ of the natural logarithm has the inverse $z = f^{-1}(w) = e^w$. From Example 4 (with the notations z and w interchanged!) we know that $f^{-1}(w) = e^w$ maps the fundamental region of the exponential function onto the z-plane without $z = 0$ (because $e^w \neq 0$ for every w). Hence $w = f(z) = \operatorname{Ln} z$ maps the z-plane without the origin and cut along the negative real axis (where $\theta = \operatorname{Im} \operatorname{Ln} z$ jumps by 2π) conformally onto the horizontal strip $-\pi < v \leqq \pi$ of the w-plane, where $w = u + iv$.

Since the mapping $w = \operatorname{Ln} z + 2\pi i$ differs from $w = \operatorname{Ln} z$ by the translation $2\pi i$ (vertically upward), this function maps the z-plane (cut as before and 0 omitted) onto the strip $\pi < v \leqq 3\pi$. Similarly for each of the infinitely many mappings $w = \ln z = \operatorname{Ln} z \pm 2n\pi i$ ($n = 0, 1, 2, \cdots$). The corresponding horizontal strips of width 2π (images of the z-plane under these mappings) together cover the whole w-plane without overlapping. ▩

Magnification Ratio. By the definition of the derivative we have

$$\lim_{z \to z_0} \left| \frac{f(z) - f(z_0)}{z - z_0} \right| = |f'(z_0)|. \tag{4}$$

Therefore, the mapping $w = f(z)$ magnifies (or shortens) the lengths of short lines by approximately the factor $|f'(z_0)|$. The image of a small figure ***conforms*** to the original figure in the sense that it has approximately the same shape. However, since $f'(z)$ varies from point to point, a *large* figure may have an image whose shape is quite different from that of the original figure.

More on the Condition $f'(z) \neq 0$. From (4) in Sec. 13.4 and the Cauchy–Riemann equations we obtain

$$|f'(z)|^2 = \left| \frac{\partial u}{\partial x} + i \frac{\partial v}{\partial x} \right|^2 = \left(\frac{\partial u}{\partial x} \right)^2 + \left(\frac{\partial v}{\partial x} \right)^2 = \frac{\partial u}{\partial x} \frac{\partial v}{\partial y} - \frac{\partial u}{\partial y} \frac{\partial v}{\partial x} \tag{5'}$$

that is,

$$|f'(z)|^2 = \begin{vmatrix} \dfrac{\partial u}{\partial x} & \dfrac{\partial u}{\partial y} \\ \dfrac{\partial v}{\partial x} & \dfrac{\partial v}{\partial y} \end{vmatrix} = \frac{\partial(u, v)}{\partial(x, y)}. \tag{5}$$

This determinant is the so-called **Jacobian** (Sec. 10.3) of the transformation $w = f(z)$ written in real form $u = u(x, y)$, $v = v(x, y)$. Hence $f'(z_0) \neq 0$ implies that the Jacobian is not 0 at z_0. This condition is sufficient that the mapping $w = f(z)$ in a sufficiently small neighborhood of z_0 is one-to-one or injective (different points have different images). See Ref. [GR4] in App. 1.

PROBLEM SET 17.1

1. Verify all calculations in Example 1.
2. Why do the images of the curves $|z| = const$ and $\arg z = const$ under a mapping by an analytic function $f(z)$ intersect at right angles, except at points at which $f'(z) = 0$?
3. Does the mapping $w = \bar{z} = x - iy$ preserve angles in size as well as in sense?

4–6 MAPPING OF CURVES

Find and sketch or graph the image of the given curves under the given mapping.

4. $x = 1, 2, 3, 4$, $y = 1, 2, 3, 4$; $w = z^2$
5. Curves as in Prob. 4, $w = iz$ **(Rotation)**
6. $|z| = 1/3, 1/2, 1, 2, 3$; $\operatorname{Arg} z = 0, \pm\pi/4, \pm\pi/2, \pm 3\pi/2, \pm\pi$; $w = 1/z$

7–15 MAPPING OF REGIONS

Find and sketch or graph the image of the given region under the given mapping.

7. $-\pi/4 < \operatorname{Arg} z < \pi/4$, $|z| < 1/2$, $w = z^3$
8. $x \geqq 1$, $w = 1/z$
9. $|z| > 1$, $w = 3z$
10. $\operatorname{Im} z > 0$, $w = 1 - z$
11. $x \geqq 0$, $y \geqq 0$, $|z| \leqq 4$; $w = z^2$
12. $-1 \leqq x \leqq 1$, $-\pi < y < \pi$; $w = e^z$
13. $\ln 3 < x < \ln 5$, $w = e^z$
14. $-\pi < y \leqq 3\pi$, $w = e^z$
15. $2 \leqq |z| \leqq 3$, $\pi/4 \leqq \theta \leqq \pi/2$; $w = \operatorname{Ln} z$

16. **CAS EXPERIMENT. Orthogonal Nets.** Graph the orthogonal net of the two families of level curves $\operatorname{Re} f(z) = const$ and $\operatorname{Im} f(z) = const$, where (a) $f(z) = z^4$, (b) $f(z) = 1/z$, (c) $f(z) = 1/z^2$, (d) $f(z) = (z + i)/(1 + iz)$. Why do these curves generally intersect at right angles? In your work, experiment to get the best possible graphs. Also do the same for other functions of your own choice. Observe and record shortcomings of your CAS and means to overcome such deficiencies.

17–23 FAILURE OF CONFORMALITY

Find all points at which the following mappings are not conformal.

17. $z(z^4 - 5)$
18. $z^2 + 1/z^2$
19. $\cos \pi z$
20. $\cosh 2z$
21. $z^2 + az + b$
22. $\exp(z^5 - 80z)$
23. $(z - a)^3$, $(z^3 - a)^2$

24–28 MAGNIFICATION RATIO, JACOBIAN

Find the magnification ratio M. Describe what it tells you about the mapping. Where is M equal to 1? Find the Jacobian J.

24. $w = \frac{1}{2}z^2$
25. $w = e^z$
26. $w = z^3$
27. $w = \operatorname{Ln} z$
28. $w = 1/z$

29. **Magnification of Angles.** Let $f(z)$ be analytic at z_0. Suppose that $f'(z_0) = 0, \cdots, f^{(k-1)}(z_0) = 0$. Then the mapping $w = f(z)$ magnifies angles with vertex at z_0 by a factor k. Illustrate this with examples for $k = 2, 3, 4$.
30. Prove the statement in Prob. 29 for general $k = 1, 2, \cdots$. *Hint.* Use the Taylor series.

17.2 Linear Fractional Transformations

Conformal mappings can help in modeling and solving boundary value problems by first mapping regions conformally onto another. We shall explain this for standard regions (disks, half-planes, strips) in the next section. For this it is useful to know properties of special basic mappings. Accordingly, let us begin with the following very important class.

Linear fractional transformations (or **Möbius transformations**) are mappings

$$(1) \qquad w = \frac{az + b}{cz + d} \qquad (ad - bc \neq 0)$$

where a, b, c, d are complex or real numbers. Differentiation gives

$$(2) \qquad w' = \frac{a(cz + d) - c(az + b)}{(cz + d)^2} = \frac{ad - bc}{(cz + d)^2}.$$

This motivates our requirement $ad - bc \neq 0$. It implies conformality for all z and excludes the totally uninteresting case $w' \equiv 0$ once and for all. Special cases of (1) are

$$(3) \qquad \begin{aligned} &w = z + b && (\textit{Translations}) \\ &w = az \quad \text{with } |a| = 1 && (\textit{Rotations}) \\ &w = az + b && (\textit{Linear transformations}) \\ &w = 1/z && (\textit{Inversion in the unit circle}). \end{aligned}$$

EXAMPLE 1 **Properties of the Inversion $w = 1/z$ (Fig. 384)**

In polar forms $z = re^{i\theta}$ and $w = Re^{i\phi}$ the inversion $w = 1/z$ is

$$Re^{i\phi} = \frac{1}{re^{i\theta}} = \frac{1}{r}\,e^{-i\theta} \qquad \text{and gives} \qquad R = \frac{1}{r}, \qquad \phi = -\theta.$$

Hence the unit circle $|z| = r = 1$ is mapped onto the unit circle $|w| = R = 1$; $w = e^{i\phi} = e^{-i\theta}$. For a general z the image $w = 1/z$ can be found geometrically by marking $|w| = R = 1/r$ on the segment from 0 to z and then reflecting the mark in the real axis. (Make a sketch.)

Figure 384 shows that $w = 1/z$ maps horizontal and vertical straight lines onto circles or straight lines. Even the following is true.

$w = 1/z$ maps every straight line or circle onto a circle or straight line.

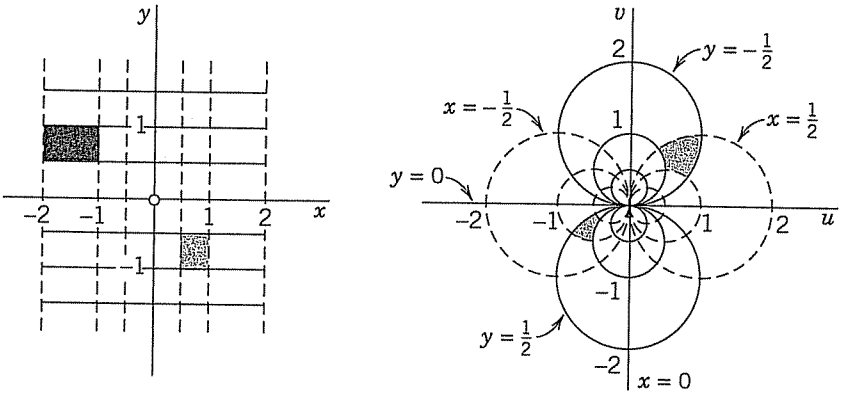


Fig. 384. Mapping (Inversion) $w = 1/z$

Proof. Every straight line or circle in the z-plane can be written

$$A(x^2 + y^2) + Bx + Cy + D = 0 \qquad (A, B\ C, D \text{ real}).$$

$A = 0$ gives a straight line and $A \neq 0$ a circle. In terms of z and $\bar{z}$ this equation becomes

$$Az\bar{z} + B\frac{z + \bar{z}}{2} + C\frac{z - \bar{z}}{2i} + D = 0.$$

Now $w = 1/z$. Substitution of $z = 1/w$ and multiplication by $w\bar{w}$ gives the equation

$$A + B\frac{\bar{w} + w}{2} + C\frac{\bar{w} - w}{2i} + Dw\bar{w} = 0$$

or, in terms of u and v,

$$A + Bu - Cv + D(u^2 + v^2) = 0.$$

This represents a circle (if $D \neq 0$) or a straight line (if $D = 0$) in the w-plane. ∎

The proof in this example suggests the use of z and $\bar{z}$ instead of x and y, a ***general principle*** that is often quite useful in practice.

Surprisingly, *every* linear fractional transformation has the property just proved:

THEOREM 1

Circles and Straight Lines

Every linear fractional transformation (1) *maps the totality of circles and straight lines in the z-plane onto the totality of circles and straight lines in the w-plane.*

PROOF This is trivial for a translation or rotation, fairly obvious for a uniform expansion or contraction, and true for $w = 1/z$, as just proved. Hence it also holds for composites of these special mappings. Now comes the key idea of the proof: represent (1) in terms of these special mappings. When $c = 0$, this is easy. When $c \neq 0$, the representation is

$$w = K\frac{1}{cz + d} + \frac{a}{c} \qquad \text{where} \qquad K = -\frac{ad - bc}{c}.$$

This can be verified by substituting K, taking the common denominator and simplifying; this yields (1). We can now set

$$w_1 = cz, \qquad w_2 = w_1 + d, \qquad w_3 = \frac{1}{w_2}, \qquad w_4 = Kw_3,$$

and see from the previous formula that then $w = w_4 + a/c$. This tells us that (1) is indeed a composite of those special mappings and completes the proof. ∎

Extended Complex Plane

The extended complex plane (the complex plane together with the point ∞ in Sec. 16.2) can now be motivated even more naturally by linear fractional transformations as follows.

To each z for which $cz + d \neq 0$ there corresponds a unique w in (1). Now let $c \neq 0$. Then for $z = -d/c$ we have $cz + d = 0$, so that no w corresponds to this z. This suggests that we let $w = \infty$ be the image of $z = -d/c$.

Also, the **inverse mapping** of (1) is obtained by solving (1) for z; this gives again a linear fractional transformation

$$z = \frac{dw - b}{-cw + a}. \tag{4}$$

When $c \neq 0$, then $cw - a = 0$ for $w = a/c$, and we let a/c be the image of $z = \infty$. With these settings, the linear fractional transformation (1) is now a one-to-one mapping of the extended z-plane onto the extended w-plane. We also say that every linear fractional transformation maps "the extended complex plane in a one-to-one manner onto itself."

Our discussion suggests the following.

General Remark. If $z = \infty$, then the right side of (1) becomes the meaningless expression $(a \cdot \infty + b)/(c \cdot \infty + d)$. We assign to it the value $w = a/c$ if $c \neq 0$ and $w = \infty$ if $c = 0$.

Fixed Points

Fixed points of a mapping $w = f(z)$ are points that are mapped onto themselves, are "kept fixed" under the mapping. Thus they are obtained from

$$w = f(z) = z.$$

The **identity mapping** $w = z$ has every point as a fixed point. The mapping $w = \bar{z}$ has infinitely many fixed points, $w = 1/z$ has two, a rotation has one, and a translation none in the finite plane. (Find them in each case.) For (1), the fixed-point condition $w = z$ is

$$z = \frac{az + b}{cz + d}, \qquad \text{thus} \qquad cz^2 - (a - d)z - b = 0. \tag{5}$$

This is a quadratic equation in z whose coefficients all vanish if and only if the mapping is the identity mapping $w = z$ (in this case, $a = d \neq 0$, $b = c = 0$). Hence we have

THEOREM 2

Fixed Points

A linear fractional transformation, not the identity, has at most two fixed points. If a linear fractional transformation is known to have three or more fixed points, it must be the identity mapping $w = z$.

To make our present general discussion of linear fractional transformations even more useful from a practical point of view, we extend it by further facts and typical examples, in the problem set as well as in the next section.

PROBLEM SET 17.2

1. Verify the calculations in the proof of Theorem 1.
2. **(Composition of LFTs)** Show that substituting a linear fractional transformation (LFT) into a LFT gives a LFT.
3. **(Matrices)** If you are familiar with 2×2 matrices, prove that the coefficient matrices of (1) and (4) are inverses of each other, provided $ad - bc = 1$, and that the composition of LFTs corresponds to the multiplication of the coefficient matrices.

4–7 INVERSE

Find the inverse $z = z(w)$. Check the result by solving $z(w)$ for w.

4. $w = \dfrac{4z + i}{-3iz + 1}$
5. $w = \dfrac{3z}{2z - i}$
6. $w = \dfrac{z + i}{z - i}$
7. $w = \dfrac{2z + 5i}{4z}$

8–14 FIXED POINTS

Find the fixed points.

8. $w = 81z^5$
9. $w = (4 + i)z$
10. $w = z + 4i$
11. $w = (z - i)^2$
12. $w = \dfrac{z - 1}{z + 1}$
13. $w = \dfrac{2iz - 1}{z + 2i}$
14. $w = \dfrac{3z + 2}{z - 1}$
15. Find a LFT whose (only) fixed points are -2 and 2.
16. Find a LFT (not $w = z$) with fixed points 0 and 1.
17. Find all LFTs with fixed points -1 and 1.
18. Find all LFTs whose only fixed point is 0.
19. Find all LFTs with fixed points 0 and ∞.
20. Find all LFTs without fixed points in the finite plane.

17.3 Special Linear Fractional Transformations

In this section we shall see how to determine linear fractional transformations

$$(1) \qquad w = \frac{az + b}{cz + d} \qquad (ad - bc \neq 0)$$

for mapping certain standard domains onto others and how to discuss properties of (1).

A mapping (1) is determined by a, b, c, d, actually by the ratios of three of these constants to the fourth because we can drop or introduce a common factor. This makes it plausible that three conditions determine a unique mapping (1):

THEOREM 1

Three Points and Their Images Given

Three given distinct points z_1, z_2, z_3 can always be mapped onto three prescribed distinct points w_1, w_2, w_3 by one, and only one, linear fractional transformation $w = f(z)$. This mapping is given implicitly by the equation

$$\frac{w - w_1}{w - w_3} \cdot \frac{w_2 - w_3}{w_2 - w_1} = \frac{z - z_1}{z - z_3} \cdot \frac{z_2 - z_3}{z_2 - z_1}. \tag{2}$$

(If one of these points is the point ∞, the quotient of the two differences containing this point must be replaced by 1.)

PROOF Equation (2) is of the form $F(w) = G(z)$ with linear fractional F and G. Hence $w = F^{-1}(G(z)) = f(z)$, where F^{-1} is the inverse of F and is linear fractional (see (4) in Sec. 17.2) and so is the composite $F^{-1}(G(z))$ (by Prob. 21), that is, $w = f(z)$ is linear fractional. Now if in (2) we set $w = w_1, w_2, w_3$ on the left and $z = z_1, z_2, z_3$ on the right, we see that

$$F(w_1) = 0, \quad F(w_2) = 1, \quad F(w_3) = \infty$$
$$G(z_1) = 0, \quad G(z_2) = 1, \quad G(z_3) = \infty.$$

From the first column, $F(w_1) = G(z_1)$, thus $w_1 = F^{-1}(G(z_1)) = f(z_1)$. Similarly, $w_2 = f(z_2)$, $w_3 = f(z_3)$. This proves the existence of the desired linear fractional transformation.

To prove uniqueness, let $w = g(z)$ be a linear fractional transformation, which also maps z_j onto w_j, $j = 1, 2, 3$. Thus $w_j = g(z_j)$. Hence $g^{-1}(w_j) = z_j$, where $w_j = f(z_j)$. Together, $g^{-1}(f(z_j)) = z_j$, a mapping with the three fixed points z_1, z_2, z_3. By Theorem 2 in Sec. 17.2, this is the identity mapping, $g^{-1}(f(z)) = z$ for all z. Thus $f(z) = g(z)$ for all z, the uniqueness.

The last statement of Theorem 1 follows from the General Remark in Sec. 17.2. ∎

Mapping of Standard Domains by Theorem 1

Using Theorem 1, we can now find linear fractional transformations according to the following

Principle. Prescribe three boundary points z_1, z_2, z_3 of the domain D in the z-plane. Choose their images w_1, w_2, w_3 on the boundary of the image D^* of D in the w-plane. Obtain the mapping from (2). Make sure that D is mapped onto D^*, not onto its complement. In the latter case, interchange two w-points. (Why does this help?)

EXAMPLE 1 **Mapping of a Half-Plane onto a Disk (Fig. 385)**

Find the linear fractional transformation (1) that maps $z_1 = -1$, $z_2 = 0$, $z_3 = 1$ onto $w_1 = -1$, $w_2 = -i$, $w_3 = 1$, respectively.

Solution. From (2) we obtain

$$\frac{w - (-1)}{w - 1} \cdot \frac{-i - 1}{-i - (-1)} = \frac{z - (-1)}{z - 1} \cdot \frac{0 - 1}{0 - (-1)},$$

thus

$$w = \frac{z - i}{-iz + 1}.$$

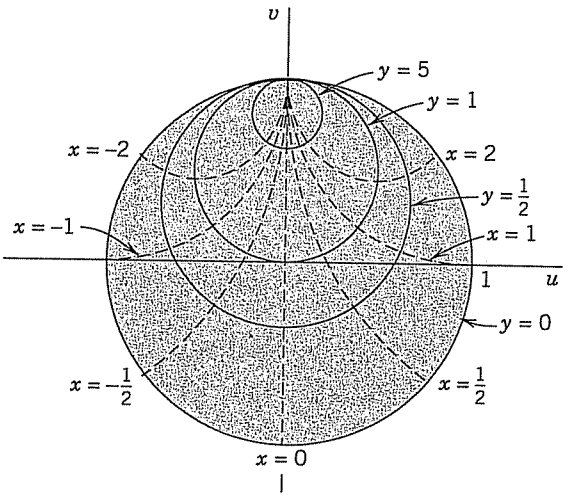


Fig. 385. Linear fractional transformation in Example 1

Let us show that we can determine the specific properties of such a mapping without much calculation. For $z = x$ we have $w = (x - i)/(-ix + 1)$, thus $|w| = 1$, so that the x-axis maps onto the unit circle. Since $z = i$ gives $w = 0$, the upper half-plane maps onto the interior of that circle and the lower half-plane onto the exterior. $z = 0, i, \infty$ go onto $w = -i, 0, i$, so that the positive imaginary axis maps onto the segment S: $u = 0$, $-1 \leqq v \leqq 1$. The vertical lines $x = const$ map onto circles (by Theorem 1, Sec. 17.2) through $w = i$ (the image of $z = \infty$) and perpendicular to $|w| = 1$ (by conformality; see Fig. 385). Similarly, the horizontal lines $y = const$ map onto circles through $w = i$ and perpendicular to S (by conformality). Figure 385 gives these circles for $y \geqq 0$, and for $y < 0$ they lie outside the unit disk shown. ■

EXAMPLE 2 **Occurrence of ∞**

Determine the linear fractional transformation that maps $z_1 = 0$, $z_2 = 1$, $z_3 = \infty$ onto $w_1 = -1$, $w_2 = -i$, $w_3 = 1$, respectively.

Solution. From (2) we obtain the desired mapping

$$w = \frac{z - i}{z + i}.$$

This is sometimes called the *Cayley transformation.*[2] In this case, (2) gave at first the quotient $(1 - \infty)/(z - \infty)$, which we had to replace by 1. ■

EXAMPLE 3 **Mapping of a Disk onto a Half-Plane**

Find the linear fractional transformation that maps $z_1 = -1$, $z_2 = i$, $z_3 = 1$ onto $w_1 = 0$, $w_2 = i$, $w_3 = \infty$, respectively, such that the unit disk is mapped onto the right half-plane. (Sketch disk and half-plane.)

Solution. From (2) we obtain, after replacing $(i - \infty)/(w - \infty)$ by 1,

$$w = -\frac{z + 1}{z - 1}.$$ ■

Mapping half-planes onto half-planes is another task of practical interest. For instance, we may wish to map the upper half-plane $y \geqq 0$ onto the upper half-plane $v \geqq 0$. Then the x-axis must be mapped onto the u-axis.

[2]ARTHUR CAYLEY (1821–1895), English mathematician and professor at Cambridge, is known for his important work in algebra, matrix theory, and differential equations.

EXAMPLE 4 **Mapping of a Half-Plane onto a Half-Plane**

Find the linear fractional transformation that maps $z_1 = -2$, $z_2 = 0$, $z_3 = 2$ onto $w_1 = \infty$, $w_2 = 1/4$, $w_3 = 3/8$, respectively.

Solution. You may verify that (2) gives the mapping function

$$w = \frac{z + 1}{2z + 4}$$

What is the image of the x-axis? Of the y-axis?

Mappings of disks onto disks is a third class of practical problems. We may readily verify that the unit disk in the z-plane is mapped onto the unit disk in the w-plane by the following function, which maps z_0 onto the center $w = 0$.

(3) $$w = \frac{z - z_0}{cz - 1}, \qquad c = \bar{z}_0, \qquad |z_0| < 1.$$

To see this, take $|z| = 1$, obtaining, with $c = \bar{z}_0$ as in (3),

$$\begin{aligned} |z - z_0| &= |\bar{z} - c| \\ &= |z|\,|\bar{z} - c| \\ &= |z\bar{z} - cz| = |1 - cz| = |cz - 1|. \end{aligned}$$

Hence

$$|w| = |z - z_0|/|cz - 1| = 1$$

from (3), so that $|z| = 1$ maps onto $|w| = 1$, as claimed, with z_0 going onto 0, as the numerator in (3) shows.

Formula (3) is illustrated by the following example. Another interesting case will be given in Prob. 10 of Sec. 18.2.

EXAMPLE 5 **Mapping of the Unit Disk onto the Unit Disk**

Taking $z_0 = \frac{1}{2}$ in (3), we obtain (verify!)

$$w = \frac{2z - 1}{z - 2}$$

(Fig. 386).

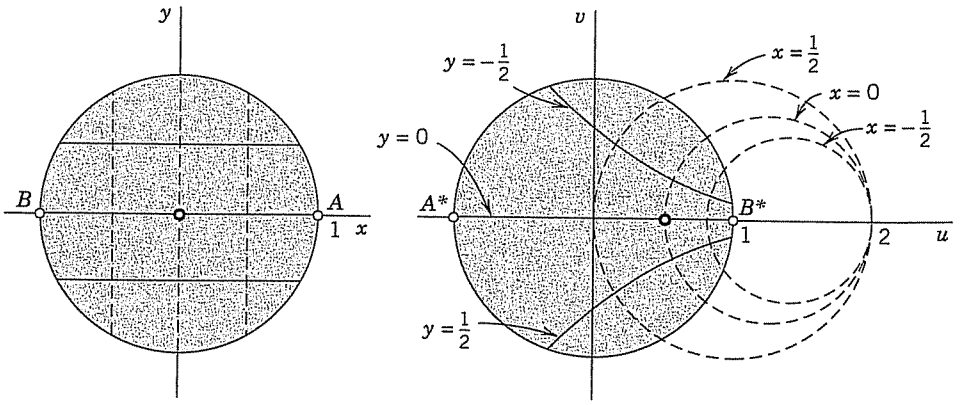


Fig. 386. Mapping in Example 5

EXAMPLE 6 Mapping of an Angular Region onto the Unit Disk

Certain mapping problems can be solved by combining linear fractional transformations with others. For instance, to map the angular region D: $-\pi/6 \leqq \arg z \leqq \pi/6$ (Fig. 387) onto the unit disk $|w| \leqq 1$, we may map D by $Z = z^3$ onto the right Z-half-plane and then the latter onto the disk $|w| \leqq 1$ by

$$w = i\,\frac{Z-1}{Z+1}, \qquad \text{combined} \qquad w = i\,\frac{z^3-1}{z^3+1}.$$

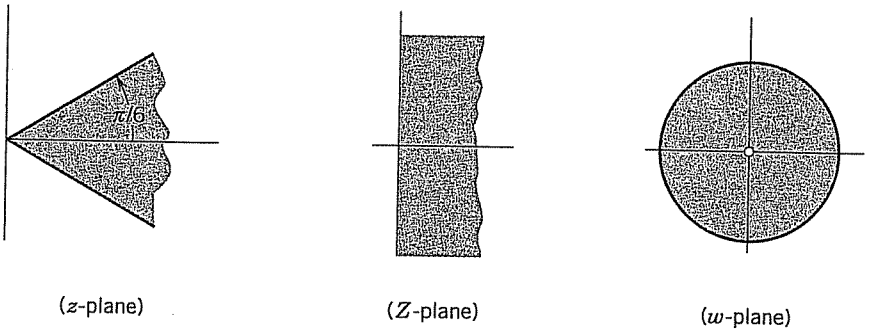


Fig. 387. Mapping in Example 6

This is the end of our discussion of linear fractional transformations. In the next section we turn to conformal mappings by other analytic functions (sine, cosine, etc.).

PROBLEM SET 17.3

1. Derive the mapping in Example 2 from (2).
2. **(Inverse)** Find the inverse of the mapping in Example 1. Show that under that inverse the lines $x = const$ are the images of circles in the w-plane with centers on the line $v = 1$.
3. Verify the formula (3) for disks.
4. Derive the mapping in Example 4 from (2). Find its inverse and prove by calculation that it has the same fixed points as the mapping itself. Is this surprising?
5. **(Inverse)** If $w = f(z)$ is any transformation that has an inverse, prove the (trivial!) fact that f and its inverse have the same fixed points.
6. **CAS EXPERIMENT. Linear Fractional Transformations (LFTs).** **(a)** Graph typical regions (squares, disks, etc.) and their images under the LFTs in Examples 1–5.

   (b) Make an experimental study of the continuous dependence of LFTs on their coefficients. For instance, change the LFT in Example 4 continuously and graph the changing image of a fixed region (applying animation if available).

7–15 LFTs FROM THREE POINTS AND THEIR IMAGES

Find the LFT that maps the given three points onto the three given points in the respective order.

7. $-1, 0, 1$ onto $-0.6 - 0.8i, -1, -0.6, +0.8i$
8. $0, 1, 2$ onto $1, \frac{1}{2}, \frac{1}{3}$
9. $2i, -2i, 4$ onto $-4 + 2i, -4 - 2i, 0$
10. $i, -1, 1$ onto $-1, -i, i$
11. $0, 1, \infty$ onto $\infty, 1, 0$
12. $0, -i, i$ onto $-1, 0, \infty$
13. $2i, i, 0$ onto $\frac{5}{2}i, 2i, \infty$
14. $0, 2i, -2i$ onto $-1, 0, \infty$
15. $-1, 0, 1$ onto $0, 1, -1$
16. Find all LFTs $w(z)$ that map the x-axis onto the u-axis.
17. Find a LFT that maps $|z| \leqq 1$ onto $|w| \leqq 1$ so that $z = i/2$ is mapped onto $w = 0$. Sketch the images of the lines $x = const$ and $y = const$.
18. Find an analytic function that maps the second quadrant of the z-plane onto the interior of the unit circle in the w-plane.
19. Find an analytic function $w = f(z)$ that maps the region $0 \leqq \arg z \leqq \pi/4$ onto the unit disk $|w| \leqq 1$.
20. **(Composite)** Show that the composite of two LFTs is a LFT.

17.4 Conformal Mapping by Other Functions

So far we have discussed the mapping by z^n, e^z (Sec. 17.1) and linear fractional transformations (Secs. 17.2, 17.3), and we shall now turn to the mapping by trigonometric and hyperbolic analytic functions.

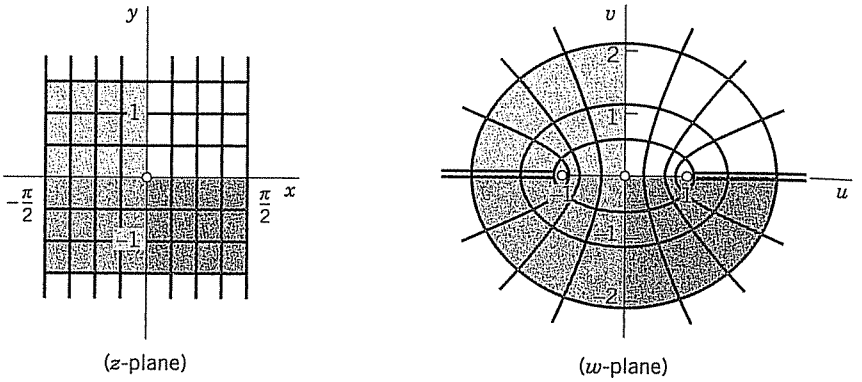


Fig. 388. Mapping $w = u + iv = \sin z$

Sine Function. Figure 388 shows the mapping by

(1) $$w = u + iv = \sin z = \sin x \cosh y + i \cos x \sinh y \qquad \text{(Sec. 13.6).}$$

Hence

(2) $$u = \sin x \cosh y, \qquad v = \cos x \sinh y.$$

Since $\sin z$ is periodic with period 2π, the mapping is certainly not one-to-one if we consider it in the full z-plane. We restrict z to the vertical strip S: $-\frac{1}{2}\pi \leqq x \leqq \frac{1}{2}\pi$ in Fig. 388. Since $f'(z) = \cos z = 0$ at $z = \pm\frac{1}{2}\pi$, the mapping is not conformal at these two critical points. We claim that the rectangular net of straight lines $x = const$ and $y = const$ in Fig. 388 is mapped onto a net in the w-plane consisting of hyperbolas (the images of the vertical lines $x = const$) and ellipses (the images of the horizontal lines $y = const$) intersecting the hyperbolas at right angles (conformality!). Corresponding calculations are simple. From (2) and the relations $\sin^2 x + \cos^2 x = 1$ and $\cosh^2 y - \sinh^2 y = 1$ we obtain

$$\frac{u^2}{\sin^2 x} - \frac{v^2}{\cos^2 x} = \cosh^2 y - \sinh^2 y = 1 \qquad \text{(Hyperbolas)}$$

$$\frac{u^2}{\cosh^2 y} + \frac{v^2}{\sinh^2 y} = \sin^2 x + \cos^2 x = 1 \qquad \text{(Ellipses).}$$

Exceptions are the vertical lines $x = \pm\frac{1}{2}\pi$, which are "folded" onto $u \leqq -1$ and $u \geqq 1$ $(v = 0)$, respectively.

Figure 389 illustrates this further. The upper and lower sides of the rectangle are mapped onto semi-ellipses and the vertical sides onto $-\cosh 1 \leqq u \leqq -1$ and $1 \leqq u \leqq \cosh 1$ $(v = 0)$, respectively. An application to a potential problem will be given in Prob. 5 of Sec. 18.2.

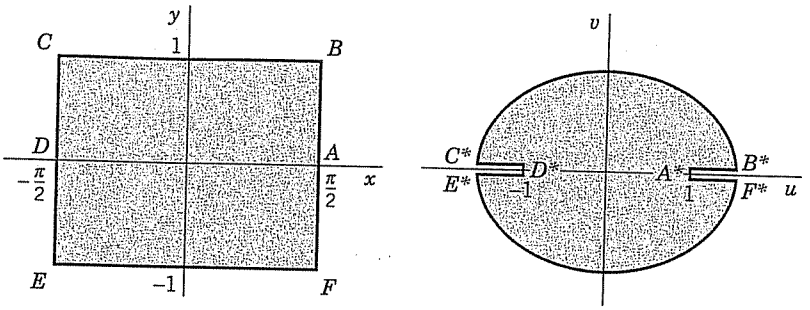


Fig. 389. Mapping by $w = \sin z$

Cosine Function. The mapping $w = \cos z$ could be discussed independently, but since

$$w = \cos z = \sin (z + \tfrac{1}{2}\pi), \tag{3}$$

we see at once that this is the same mapping as $\sin z$ preceded by a translation to the right through $\frac{1}{2}\pi$ units.

Hyperbolic Sine. Since

$$w = \sinh z = -i \sin (iz), \tag{4}$$

the mapping is a counterclockwise rotation $Z = iz$ through $\frac{1}{2}\pi$ (i.e., 90°), followed by the sine mapping $Z^* = \sin Z$, followed by a clockwise 90°-rotation $w = -iZ^*$.

Hyperbolic Cosine. This function

$$w = \cosh z = \cos (iz) \tag{5}$$

defines a mapping that is a rotation $Z = iz$ followed by the mapping $w = \cos Z$.

Figure 390 shows the mapping of a semi-infinite strip onto a half-plane by $w = \cosh z$. Since $\cosh 0 = 1$, the point $z = 0$ is mapped onto $w = 1$. For real $z = x \geqq 0$, $\cosh z$ is real and increases with increasing x in a monotone fashion, starting from 1. Hence the positive x-axis is mapped onto the portion $u \geqq 1$ of the u-axis.

For pure imaginary $z = iy$ we have $\cosh iy = \cos y$. Hence the left boundary of the strip is mapped onto the segment $1 \geqq u \geqq -1$ of the u-axis, the point $z = \pi i$ corresponding to

$$w = \cosh i\pi = \cos \pi = -1.$$

On the upper boundary of the strip, $y = \pi$, and since $\sin \pi = 0$ and $\cos \pi = -1$, it follows that this part of the boundary is mapped onto the portion $u \leqq -1$ of the u-axis. Hence the boundary of the strip is mapped onto the u-axis. It is not difficult to see that the interior of the strip is mapped onto the upper half of the w-plane, and the mapping is one-to-one.

This mapping in Fig. 390 has applications in potential theory, as we shall see in Prob. 12 of Sec. 18.3.

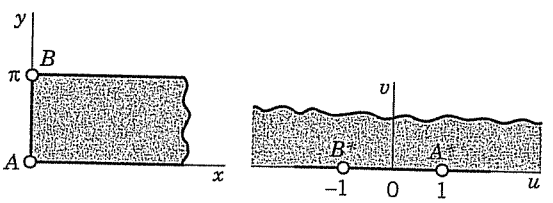


Fig. 390. Mapping by $w = \cosh z$

Tangent Function. Figure 391 shows the mapping of a vertical infinite strip onto the unit circle by $w = \tan z$, accomplished in three steps as suggested by the representation (Sec. 13.6)

$$w = \tan z = \frac{\sin z}{\cos z} = \frac{(e^{iz} - e^{-iz})/i}{e^{iz} + e^{-iz}} = \frac{(e^{2iz} - 1)/i}{e^{2iz} + 1}.$$

Hence if we set $Z = e^{2iz}$ and use $1/i = -i$, we have

$$(6) \qquad w = \tan z = -iW, \qquad W = \frac{Z - 1}{Z + 1}, \qquad Z = e^{2iz}.$$

We now see that $w = \tan z$ is a linear fractional transformation preceded by an exponential mapping (see Sec. 17.1) and followed by a clockwise rotation through an angle $\frac{1}{2}\pi$ (90°).

The strip is S: $-\frac{1}{4}\pi < x < \frac{1}{4}\pi$, and we show that it is mapped onto the unit disk in the w-plane. Since $Z = e^{2iz} = e^{-2y+2ix}$, we see from (10) in Sec. 13.5 that $|Z| = e^{-2y}$, $\operatorname{Arg} Z = 2x$. Hence the vertical lines $x = -\pi/4$, 0, $\pi/4$ are mapped onto the rays $\operatorname{Arg} Z = -\pi/2$, 0, $\pi/2$, respectively. Hence S is mapped onto the right Z-half-plane. Also $|Z| = e^{-2y} < 1$ if $y > 0$ and $|Z| > 1$ if $y < 0$. Hence the upper half of S is mapped inside the unit circle $|Z| = 1$ and the lower half of S outside $|Z| = 1$, as shown in Fig. 391.

Now comes the linear fractional transformation in (6), which we denote by $g(Z)$:

$$(7) \qquad W = g(Z) = \frac{Z - 1}{Z + 1}.$$

For real Z this is real. Hence the real Z-axis is mapped onto the real W-axis. Furthermore, the imaginary Z-axis is mapped onto the unit circle $|W| = 1$ because for pure imaginary $Z = iY$ we get from (7)

$$|W| = |g(iY)| = \left|\frac{iY - 1}{iY + 1}\right| = 1.$$

The right Z-half-plane is mapped inside this unit circle $|W| = 1$, not outside, because $Z = 1$ has its image $g(1) = 0$ inside that circle. Finally, the unit circle $|Z| = 1$ is mapped

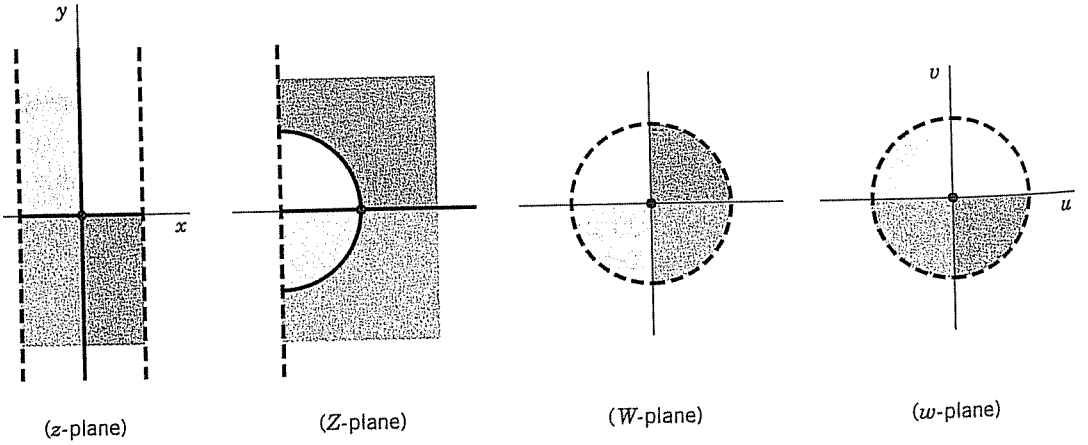


Fig. 391. Mapping by $w = \tan z$

onto the imaginary W-axis, because this circle is $Z = e^{i\phi}$, so that (7) gives a pure imaginary expression, namely,

$$g(e^{i\phi}) = \frac{e^{i\phi} - 1}{e^{i\phi} + 1} = \frac{e^{i\phi/2} - e^{-i\phi/2}}{e^{i\phi/2} + e^{-i\phi/2}} = \frac{i \sin(\phi/2)}{\cos(\phi/2)}.$$

From the W-plane we get to the w-plane simply by a clockwise rotation through $\pi/2$; see (6).

Together we have shown that $w = \tan z$ maps S: $-\pi/4 < \operatorname{Re} z < \pi/4$ onto the unit disk $|w| = 1$, with the four quarters of S mapped as indicated in Fig. 391. This mapping is conformal and one-to-one.

PROBLEM SET 17.4

1–7 CONFORMAL MAPPING $w = e^z$

Find and sketch the image of the given region under $w = e^z$.

1. $0 \leqq x \leqq 2$, $-\pi \leqq y \leqq \pi$
2. $-1 \leqq x \leqq 0$, $0 \leqq y \leqq \pi/2$
3. $-0.5 < x < 0.5$, $3\pi/4 < y < 5\pi/4$
4. $-3 < x < 3$, $\pi/4 < y < 3\pi/4$
5. $0 < x < 1$, $0 < y < \pi$
6. $x < 0$, $-\pi/2 < y < \pi/2$
7. x arbitrary, $0 \leqq y \leqq 2\pi$
8. **CAS EXPERIMENT. Conformal Mapping.** If your CAS can do conformal mapping, use it to solve Prob. 5. Then increase y beyond π, say, to 50π or 100π. State what you expected. See what you get as the image. Explain.

9–12 CONFORMAL MAPPING $w = \sin z$

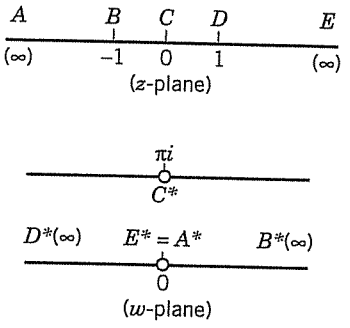
Find and sketch or graph the image of the given region under $w = \sin z$.

9. $0 \leqq x \leqq \pi$, $0 \leqq y \leqq 1$
10. $0 < x < \pi/6$, y arbitrary
11. $0 < x < 2\pi$, $1 < y < 5$
12. $-\pi/4 < x < \pi/4$, $0 < y < 3$
13. Determine all points at which $w = \sin z$ is not conformal.
14. Find and sketch or graph the images of the lines $x = 0$, $\pm\pi/6$, $\pm\pi/3$, $\pm\pi/2$ under the mapping $w = \sin z$.
15. Find an analytic function that maps the region R bounded by the positive x- and y-axes and the hyperbola $xy = \pi/2$ in the first quadrant onto the upper half-plane. *Hint.* First map the region onto a horizontal strip.
16. Describe the mapping $w = \cosh z$ in terms of the mapping $w = \sin z$ and rotations and translations.
17. Find all points at which the mapping $w = \cosh \pi z$ is not conformal.

18–22 CONFORMAL MAPPING $w = \cos z$

Find and sketch or graph the image of the given region under $w = \cos z$.

18. $0 < x < \pi/2$, $0 < y < 2$
19. $0 < x < \pi$, $0 < y < 1$
20. $-1 \leqq x \leqq 1$, $0 \leqq y \leqq 1$
21. $\pi < x < 2\pi$, $y < 0$
22. $0 < x < 2\pi$, $1/2 < y < 1$
23. Find the images of the lines $y = c = const$ under the mapping $w = \cos z$.
24. Show that $w = \operatorname{Ln} \dfrac{z - 1}{z + 1}$ maps the upper half-plane onto the horizontal strip $0 \leqq \operatorname{Im} w \leqq \pi$ as shown in the figure.



Problem 24

25. Find and sketch the image of R: $2 \leqq |z| \leqq 3$, $\pi/4 \leqq \theta \leqq \pi/2$ under the mapping $w = \operatorname{Ln} z$.

17.5 Riemann Surfaces. *Optional*

Riemann surfaces are surfaces on which multivalued relations, such as $w = \sqrt{z}$ or $w = \ln z$, become single-valued, that is, functions in the usual sense. We explain the idea, which is simple—but ingenious, one of the greatest in complex analysis.

The mapping given by

$$w = u + iv = z^2 \tag{1}$$

(Sec. 17.1)

is conformal, except at $z = 0$, where $w' = 2z = 0$. At $z = 0$, angles are doubled under the mapping. Thus the right z-half-plane (including the positive y-axis) is mapped onto the full w-plane, cut along the negative half of the u-axis; this mapping is one-to-one. Similarly for the left z-half-plane (including the negative y-axis). Hence the image of the full z-plane under $w = z^2$ "covers the w-plane twice" in the sense that every $w \neq 0$ is the image of two z-points; if z_1 is one, the other is $-z_1$. For example, $z = i$ and $-i$ are both mapped onto $w = -1$.

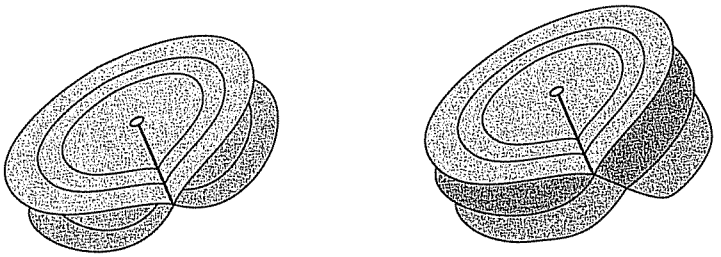
Now comes the crucial idea. We place those two copies of the cut w-plane upon each other so that the upper sheet is the image of the right half z-plane R and the lower sheet is the image of the left half z-plane L. We join the two sheets crosswise along the cuts (along the negative u-axis) so that if z moves from R to L, its image can move from the upper to the lower sheet. The two origins are fastened together because $w = 0$ is the image of just one z-point, $z = 0$. The surface obtained is called a **Riemann surface** (Fig. 392a). $w = 0$ is called a "winding point" or **branch point.** $w = z^2$ maps the full z-plane onto this surface in a one-to-one manner.

By interchanging the roles of the variables z and w it follows that the double-valued relation

$$w = \sqrt{z} \tag{2}$$

(Sec. 13.2)

becomes single-valued on the Riemann surface in Fig. 392a, that is, a function in the usual sense. We can let the upper sheet correspond to the principal value of $\sqrt{z}$. Its image is the right w-half-plane. The other sheet is then mapped onto the left w-half-plane.



(a) Riemann surface of $\sqrt{z}$ (b) Riemann surface of $\sqrt[3]{z}$

Fig. 392. Riemann surfaces

Similarly, the triple-valued relation $w = \sqrt[3]{z}$ becomes single-valued on the three-sheeted Riemann surface in Fig. 392b, which also has a branch point at $z = 0$.

The infinitely many-valued natural logarithm (Sec. 13.7)

$$w = \ln z = \operatorname{Ln} z + 2n\pi i \qquad (n = 0, \pm 1, \pm 2, \cdots)$$

becomes single-valued on a Riemann surface consisting of infinitely many sheets. $w = \operatorname{Ln} z$ corresponds to one of them. This sheet is cut along the negative x-axis and the upper edge of the slit is joined to the lower edge of the next sheet, which corresponds to the argument $\pi < \theta \leqq 3\pi$, that is, to

$$w = \operatorname{Ln} z + 2\pi i.$$

The principal value $\operatorname{Ln} z$ maps its sheet onto the horizontal strip $-\pi < v \leqq \pi$. The function $w = \operatorname{Ln} z + 2\pi i$ maps its sheet onto the neighboring strip $\pi < v \leqq 3\pi$, and so on. The mapping of the points $z \neq 0$ of the Riemann surface onto the points of the w-plane is one-to-one. See also Example 5 in Sec. 17.1.

PROBLEM SET 17.5

1. Consider $w = \sqrt{z}$. Find the path of the image point w of a point z that moves twice around the unit circle, starting from the initial position $z = 1$.
2. Show that the Riemann surface of $w = \sqrt[n]{z}$ consists of n sheets and has a branch point at $z = 0$.
3. Make a sketch, similar to Fig. 392, of the Riemann surface of $\sqrt[4]{z}$.
4. Show that the Riemann surface of $w = \sqrt{(z-1)(z-2)}$ has branch points at $z = 1$ and $z = 2$ and consists of two sheets that may be cut along the line segment from 1 to 2 and joined crosswise. *Hint.* Introduce polar coordinates $z - 1 = r_1 e^{i\theta_1}$, $z - 2 = r_2 e^{i\theta_2}$.

5–10 RIEMANN SURFACES

Find the branch points and the number of sheets of the Riemann surface.

5. $\sqrt{3z + 5}$
6. $\sqrt{(1 - z^2)(4 - z^2)}$
7. $5 + \sqrt[3]{2z + i}$
8. $\ln (3z - 4i)$
9. $e^{\sqrt{z}}$
10. $\sqrt{e^z}$

CHAPTER 17 REVIEW QUESTIONS AND PROBLEMS

1. How did we define the angle of intersection of two oriented curves, and what does it mean to say that a mapping is conformal?
2. At what points is a mapping $w = f(z)$ by an analytic function not conformal? Give examples.
3. What happens to angles at z_0 under a mapping $w = f(z)$ if $f'(z_0) = 0$, $f''(z_0) = 0$, $f'''(z_0) \neq 0$?
4. What do "surjective," "injective," and "bijective" mean?
5. What mapping gave the Joukowski airfoil?
6. What are linear fractional transformations (LFTs)? Why are they important in connection with the extended complex plane?
7. Why did we require that $ad - bc \neq 0$ for a LFT?
8. What are fixed points of a mapping? Give examples.
9. Can you remember mapping properties of $w = \sin z$? $\cos z$? e^z?
10. What is a Riemann surface? Why was it introduced? Explain the simplest example.

11–16 MAPPING $w = z^2$

Find and sketch the image of the given curve or region under $w = z^2$.

11. $y = -1$, $y = 1$
12. $xy = -4$
13. $|z| = 4.5$, $|\arg z| < \pi/4$
14. $0 < y < 2$
15. $\frac{1}{2} < x < 1$
16. $\operatorname{Im} z > 0$

17–22 MAPPING $w = 1/z$

Find and sketch the image of the given curve or region under $w = 1/z$.

17. $x = -1$
18. $y = 1$

19. $|z - \frac{1}{2}| = \frac{1}{2}$

20. $|z| < \frac{1}{2}$, $y < 0$

21. $|\arg z| < \pi/4$

22. $|z| < 1$, $x < 0$, $y > 0$

23–28 FAILURE OF CONFORMALITY

Where is the mapping by the given function not conformal? (Give reason.)

23. $5z^7 + 7z^5$

24. $\cosh 2z$

25. $\sin 2z + \cos 2z$

26. $\cos \pi z^2$

27. $\exp(z^4 + z^2)$

28. $z + 1/z$ $(z \neq 0)$

29–34 LINEAR FRACTIONAL TRANSFORMATIONS (LFTs)

Find the LFT that maps

29. 0, 1, 2 onto 0, i, $2i$, respectively

30. -1, 1, 2 onto 0, 2, 3/2, respectively

31. 1, -1, $-i$ onto 1, -1, i, respectively

32. -1, $-i$, i onto $1 - i$, 2, 0, respectively

33. 0, ∞, -2 onto 0, 1, ∞, respectively

34. 0, i, $2i$ onto 0, ∞, $2i$

35–40 Fixed Points. Find all fixed points of

35. $w = \dfrac{z + 2}{z + 1}$

36. $w = \dfrac{2iz - 1}{z + 2i}$

37. $w = \dfrac{3z + 2}{z - 1}$

38. $w = \dfrac{iz + 5}{5z + i}$

39. $w = \dfrac{(2 + i)z + 1}{z - i}$

40. $w = z^4 + z - 81$

41–45 GIVEN REGIONS

Find an analytic function $w = f(z)$ that maps:

41. The infinite strip $0 < y < \pi/3$ onto the upper half-plane $v > 0$.

42. The interior of the unit circle $|z| = 1$ onto the exterior of the circle $|w + 1| = 5$.

43. The region $x > 0$, $y > 0$, $xy < k$ onto the strip $0 < v < 1$.

44. The semi-disk $|z| < 1$, $x > 0$ onto the exterior of the unit circle $|w| = 1$.

45. The sector $0 < \arg z < \pi/3$ onto the region $u < 1$.

SUMMARY OF CHAPTER 17

Conformal Mapping

A complex function $w = f(z)$ gives a **mapping** of its domain of definition in the complex z-plane onto its range of values in the complex w-plane. If $f(z)$ is *analytic*, this mapping is **conformal,** that is, angle-preserving: the images of any two intersecting curves make the same angle of intersection, in both magnitude and sense, as the curves themselves (Sec. 17.1). Exceptions are the points at which $f'(z) = 0$ (**"critical points,"** e.g. $z = 0$ for $w = z^2$).

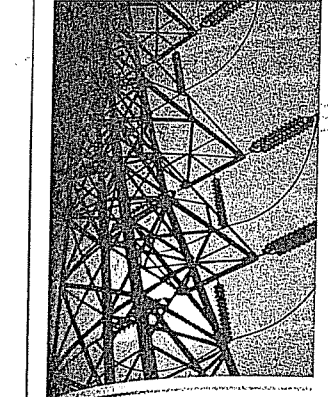
For mapping properties of e^z, $\cos z$, $\sin z$, etc. see Secs. 17.1 and 17.4.

Linear fractional transformations, also called *Möbius transformations*

$$(1) \qquad w = \frac{az + b}{cz + d} \qquad \text{(Secs. 17.2, 17.3)}$$

$(ad - bc \neq 0)$ map the extended complex plane (Sec. 17.2) onto itself. They solve the problems of mapping half-planes onto half-planes or disks, and disks onto disks or half-planes. Prescribing the images of three points determines (1) uniquely.

Riemann surfaces (Sec. 17.5) consist of several sheets connected at certain points called *branch points*. On them, multivalued relations become single-valued, that is, functions in the usual sense. *Examples.* For $w = \sqrt{z}$ we need two sheets (with branch point 0) since this relation is doubly-valued. For $w = \ln z$ we need infinitely many sheets since this relation is infinitely many-valued (see Sec. 13.7).



CHAPTER 18

Complex Analysis and Potential Theory

Laplaces's equation $\nabla^2\Phi = 0$ is one of the most important PDEs in engineering mathematics, because it occurs in gravitation (Secs. 9.7, 12.10), electrostatics (Sec. 9.7), steady-state heat conduction (Sec. 12.5), incompressible fluid flow, etc. The theory of solutions of this equation is called **potential theory** (although "potential" is also used in a more general sense in connection with gradients; see Sec. 9.7).

In the "two-dimensional case" when Φ depends only on two Cartesian coordinates x and y, Laplace's equation becomes

$$\nabla^2\Phi = \Phi_{xx} + \Phi_{yy} = 0.$$

From Sec. 13.4 we know that then its solutions Φ are closely related to complex analytic functions $\Phi + i\,\Psi$. This relation is the main reason for the importance of complex analysis in physics and engineering. (We use the notation $\Phi + i\,\Psi$ since $u + iv$ will be needed in conformal mapping.)

In this chapter we shall consider this connection and its consequences in detail and illustrate it by modeling typical examples from electrostatics (Secs. 18.1, 18.2), heat conduction (Sec. 18.3), and hydrodynamics (Sec. 18.4). This will lead to **boundary value problems,** some of which involving functions whose mapping properties we have studied in Chap. 17. Further relating to that chapter, in Sec. 18.2 we explain conformal mapping as a method in potential theory.

In Sec. 18.5 we derive the important Poisson formula for potentials in a circular disk.

Finally, in Sec. 18.6 we show that results on analytic functions can be used to characterize *general* properties of **harmonic functions** (solutions of Laplace's equation whose second partial derivatives are continuous).

Prerequisite: Chaps. 13, 14, 17.
References and Answers to Problems: App. 1 Part D, App. 2.

18.1 Electrostatic Fields

The electrical force of attraction or repulsion between charged particles is governed by Coulomb's law. This force is the gradient of a function Φ, called the **electrostatic potential.** At any points free of charges, Φ is a solution of Laplace's equation

$$\nabla^2\Phi = 0.$$

The surfaces $\Phi = const$ are called **equipotential surfaces.** At each point P at which the gradient of Φ is not the zero vector, it is perpendicular to the surface $\Phi = const$ through P; that is, the electrical force has the direction perpendicular to the equipotential surface. (See also Secs. 9.7 and 12.10.)

The problems we shall discuss in this entire chapter are **two-dimensional** (for the reason just given in the chapter opening), that is, they model physical systems that lie in three-dimensional space (of course!), but are such that the potential Φ is independent of one of the space coordinates, so that Φ depends only on two coordinates, which we call x and y. Then **Laplace's equation** becomes

$$\nabla^2\Phi = \frac{\partial^2\Phi}{\partial x^2} + \frac{\partial^2\Phi}{\partial y^2} = 0. \tag{1}$$

Equipotential surfaces now appear as **equipotential lines** (curves) in the xy-plane.

Let us illustrate these ideas by a few simple basic examples.

EXAMPLE 1 **Potential Between Parallel Plates**

Find the potential Φ of the field between two parallel conducting plates extending to infinity (Fig. 393), which are kept at potentials Φ_1 and Φ_2, respectively.

Solution. From the shape of the plates it follows that Φ depends only on x, and Laplace's equation becomes $\Phi'' = 0$. By integrating twice we obtain $\Phi = ax + b$, where the constants a and b are determined by the given boundary values of Φ on the plates. For example, if the plates correspond to $x = -1$ and $x = 1$, the solution is

$$\Phi(x) = \tfrac{1}{2}(\Phi_2 - \Phi_1)x + \tfrac{1}{2}(\Phi_2 + \Phi_1).$$

The equipotential surfaces are parallel planes. ∎

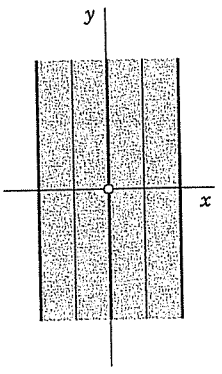


Fig. 393. Potential in Example 1

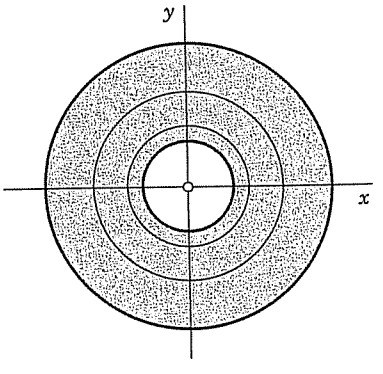


Fig. 394. Potential in Example 2

EXAMPLE 2 **Potential Between Coaxial Cylinders**

Find the potential Φ between two coaxial conducting cylinders extending to infinity on both ends (Fig. 394) and kept at potentials Φ_1 and Φ_2, respectively.

Solution. Here Φ depends only on $r = \sqrt{x^2 + y^2}$, for reasons of symmetry, and Laplace's equation $r^2 u_{rr} + r u_r + u_{\theta\theta} = 0$ [(5), Sec. 12.9] with $u_{\theta\theta} = 0$ and $u = \Phi$ becomes $r\Phi'' + \Phi' = 0$. By separating variables and integrating we obtain

$$\frac{\Phi''}{\Phi'} = -\frac{1}{r}, \qquad \ln \Phi' = -\ln r + \tilde{a}, \qquad \Phi' = \frac{a}{r}, \qquad \Phi = a \ln r + b$$

and a and b are determined by the given values of Φ on the cylinders. Although no infinitely extended conductors exist, the field in our idealized conductor will approximate the field in a long finite conductor in that part which is far away from the two ends of the cylinders. ■

EXAMPLE 3 **Potential in an Angular Region**

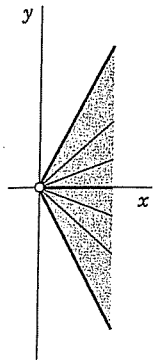


Fig. 395. Potential in Example 3

Find the potential Φ between the conducting plates in Fig. 395, which are kept at potentials Φ_1 (the lower plate) and Φ_2, and make an angle α, where $0 < \alpha \leqq \pi$. (In the figure we have $\alpha = 120° = 2\pi/3$.)

Solution. $\theta = \operatorname{Arg} z$ ($z = x + iy \neq 0$) is constant on rays $\theta = const$. It is harmonic since it is the imaginary part of an analytic function, $\operatorname{Ln} z$ (Sec. 13.7). Hence the solution is

$$\Phi(x, y) = a + b \operatorname{Arg} z$$

with a and b determined from the two boundary conditions (given values on the plates)

$$a + b(-\tfrac{1}{2}\alpha) = \Phi_1, \qquad a + b(\tfrac{1}{2}\alpha) = \Phi_2.$$

Thus $a = (\Phi_2 + \Phi_1)/2$, $b = (\Phi_2 - \Phi_1)/\alpha$. The answer is

$$\Phi(x, y) = \frac{1}{2}(\Phi_2 + \Phi_1) + \frac{1}{\alpha}(\Phi_2 - \Phi_1)\theta, \qquad \theta = \arctan \frac{y}{x}. \quad ■$$

Complex Potential

Let $\Phi(x, y)$ be harmonic in some domain D and $\Psi(x, y)$ a harmonic conjugate of Φ in D. (See Sec. 13.4, where we wrote u and v, now needed in conformal mapping from the next section on; hence the change to Φ and Ψ.) Then

$$F(z) = \Phi(x, y) + i\Psi(x, y) \tag{2}$$

is an analytic function of $z = x + iy$. This function F is called the **complex potential** corresponding to the real potential Φ. Recall from Sec. 13.4 that for given Φ, a conjugate Ψ is uniquely determined except for an additive real constant. Hence we may say ***the*** complex potential, without causing misunderstandings.

The use of F has two advantages, a technical one and a physical one. Technically, F is easier to handle than real or imaginary parts, in connection with methods of complex analysis. Physically, Ψ has a meaning. By conformality, the curves $\Psi = const$ intersect the equipotential lines $\Phi = const$ in the xy-plane at right angles [except where $F'(z) = 0$]. Hence they have the direction of the electrical force and, therefore, are called **lines of force.** They are the paths of moving charged particles (electrons in an electron microscope, etc.).

EXAMPLE 4 **Complex Potential**

In Example 1, a conjugate is $\Psi = ay$. It follows that the complex potential is

$$F(z) = az + b = ax + b + iay,$$

and the lines of force are horizontal straight lines $y = const$ parallel to the x-axis.

EXAMPLE 5 **Complex Potential**

In Example 2 we have $\Phi = a \ln r + b = a \ln |z| + b$. A conjugate is $\Psi = a \operatorname{Arg} z$. Hence the complex potential is

$$F(z) = a \operatorname{Ln} z + b$$

and the lines of force are straight lines through the origin. $F(z)$ may also be interpreted as the complex potential of a source line (a wire perpendicular to the xy-plane) whose trace in the xy-plane is the origin.

EXAMPLE 6 **Complex Potential**

In Example 3 we get $F(z)$ by noting that $i \operatorname{Ln} z = i \ln |z| - \operatorname{Arg} z$, multiplying this by $-b$, and adding a:

$$F(z) = a - ib \operatorname{Ln} z = a + b \operatorname{Arg} z - ib \ln |z|.$$

We see from this that the lines of force are concentric circles $|z| = const$. Can you sketch them?

Superposition

More complicated potentials can often be obtained by superposition.

EXAMPLE 7 **Potential of a Pair of Source Lines (a Pair of Charged Wires)**

Determine the potential of a pair of oppositely charged source lines of the same strength at the points $z = c$ and $z = -c$ on the real axis.

Solution. From Examples 2 and 5 it follows that the potential of each of the source lines is

$$\Phi_1 = K \ln |z - c| \qquad \text{and} \qquad \Phi_2 = -K \ln |z + c|,$$

respectively. Here the real constant K measures the strength (amount of charge). These are the real parts of the complex potentials

$$F_1(z) = K \operatorname{Ln} (z - c) \qquad \text{and} \qquad F_2(z) = -K \operatorname{Ln} (z + c).$$

Hence the complex potential of the combination of the two source lines is

$$F(z) = F_1(z) + F_2(z) = K [\operatorname{Ln} (z - c) - \operatorname{Ln} (z + c)]. \tag{3}$$

The **equipotential lines** are the curves

$$\Phi = \operatorname{Re} F(z) = K \ln \left| \frac{z - c}{z + c} \right| = const, \qquad \text{thus} \qquad \left| \frac{z - c}{z + c} \right| = const.$$

These are circles, as you may show by direct calculation. The **lines of force** are

$$\Psi = \operatorname{Im} F(z) = K[\operatorname{Arg} (z - c) - \operatorname{Arg} (z + c)] = const.$$

We write this briefly (Fig. 396)

$$\Psi = K(\theta_1 - \theta_2) = const.$$

Now $\theta_1 - \theta_2$ is the angle between the line segments from z to c and $-c$ (Fig. 396). Hence the lines of force are the curves along each of which the line segment S: $-c \leqq x \leqq c$ appears under a constant angle. These curves are the totality of circular arcs over S, as is (or should be) known from elementary geometry. Hence the lines of force are circles. Figure 397 shows some of them together with some equipotential lines.

In addition to the interpretation as the potential of two source lines, this potential could also be thought of as the potential between two circular cylinders whose axes are parallel but do not coincide, or as the potential between two equal cylinders that lie outside each other, or as the potential between a cylinder and a plane wall. Explain this, using Fig. 397.

The idea of the complex potential as just explained is the key to a close relation of potential theory to complex analysis and will recur in heat flow and fluid flow.

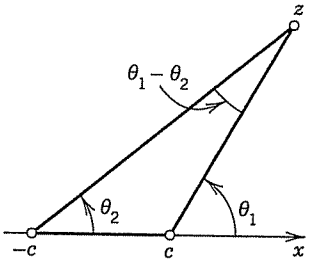


Fig. 396. Arguments in Example 7

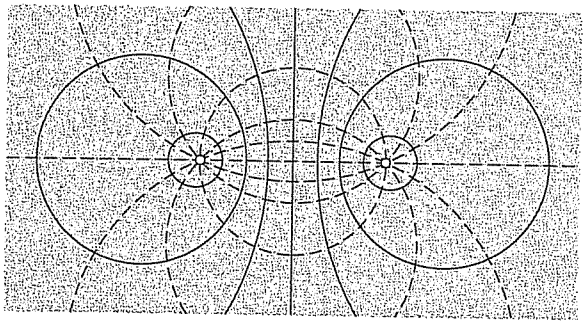


Fig. 397. Equipotential lines and lines of force (dashed) in Example 7

PROBLEM SET 18

1–4 POTENTIAL

Find and sketch the potential. Find the complex potential:

1. Between parallel plates at $x = -3$ and 3, potentials 140 V and 260 V, respectively
2. Between parallel plates at $x = -4$ and 10, potentials 4.4 kV and 10 kV, respectively
3. Between the axes (potential 110 V) and the hyperbola $xy = 1$ (potential 60 V)
4. Between parallel plates at $y = x$ and $x + k$, potentials 0 and 100 V, respectively

5–8 COAXIAL CYLINDERS

Find the potential between two infinite coaxial cylinders of radii r_1 and r_2 having potentials U_1 and U_2, respectively. Find the complex potential.

5. $r_1 = 0.5$, $r_2 = 2.0$, $U_1 = -110$ V, $U_2 = 110$ V
6. $r_1 = 1$, $r_2 = 10$, $U_1 = 100$ V, $U_2 = 1$ kV
7. $r_1 = 1$, $r_2 = 4$, $U_1 = 200$ V, $U_2 = 0$
8. $r_1 = 0.1$, $r_2 = 10$, $U_1 = 150$ V, $U_2 = 50$ V

9. Show that $\Phi = \theta/\pi = (1/\pi) \arctan (y/x)$ is harmonic in the upper half-plane and satisfies the boundary condition $\Phi(x, 0) = 1$ if $x < 0$ and 0 if $x > 0$, and the corresponding complex potential is $F(z) = -(i/\pi)\ \mathrm{Ln}\ z$.
10. Map the upper half z-plane onto the unit disk $|w| \leqq 1$ so that $0, \infty, -1$ are mapped onto $1, i, -i$, respectively. What are the boundary conditions on $|w| = 1$ resulting from the potential in Prob. 9? What is the potential at $w = 0$?
11. Verify by calculation that the equipotential lines in Example 7 are circles.
12. **CAS EXPERIMENT. Complex Potentials.** Graph the equipotential lines and lines of force in (a)–(d) (four graphs, Re $F(z)$ and Im $F(z)$ on the same axes). Then explore further complex potentials of your choice with the purpose of discovering configurations that might be of practical interest.
    (a) $F(z) = z^2$ (b) $F(z) = iz^2$
    (c) $F(z) = 1/z$ (d) $F(z) = i/z$

13–15 POTENTIALS FOR OTHER CONFIGURATIONS

13. Show that $F(z) = \arccos z$ (defined in Problem Set 13.7) gives the potential in Figs. 398 and 399.

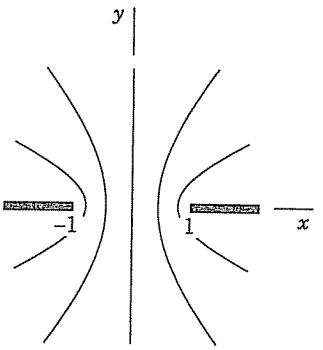


Fig. 398. Slit

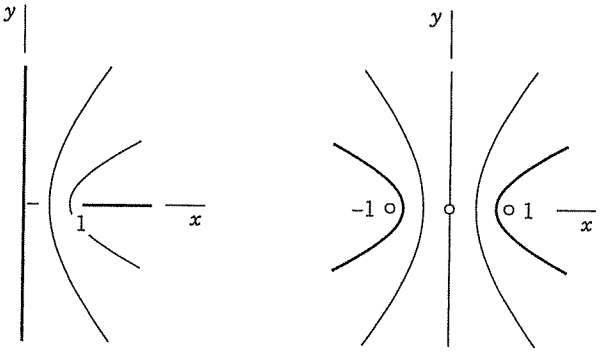


Fig. 399. Other apertures

14. Find the real and complex potentials in the sector $-\pi/6 \leqq \theta \leqq \pi/6$ between the boundary $\theta = \pm\pi/6$ (kept at 0) and the curve $x^3 - 3xy^2 = 1$, kept at 110 V.
15. Find the potential in the first quadrant of the xy-plane between the axes (having potential 220 V) and the hyperbola $xy = 1$ (having potential 110 V).

18.2 Use of Conformal Mapping. Modeling

Complex potentials relate potential theory closely to complex analysis, as we have just seen. Another close relation results from the use of conformal mapping in modeling and solving **boundary value problems** for the Laplace equation, that is, in finding a solution of the equation in some domain assuming given values on the boundary (**"Dirichlet problem"**; see also Sec. 12.5). Then conformal mapping is used to map a given domain onto one for which the solution is known or can be found more easily. This solution is then mapped back to the given domain. This is the idea. That it works is due to the fact that harmonic functions remain harmonic under conformal mapping:

THEOREM 1

Harmonic Functions Under Conformal Mapping

Let Φ^* be harmonic in a domain D^* in the w-plane. Suppose that $w = u + iv = f(z)$ is analytic in a domain D in the z-plane and maps D conformally onto D^*. Then the function

$$\Phi(x, y) = \Phi^*(u(x, y), v(x, y)) \tag{1}$$

is harmonic in D.

PROOF The composite of analytic functions is analytic, as follows from the chain rule. Hence, taking a harmonic conjugate $\Psi^*(u, v)$ of Φ^*, as defined in Sec. 13.4, and forming the analytic function $F^*(w) = \Phi^*(u, v) + i\Psi^*(u, v)$, we conclude that $F(z) = F^*(f(z))$ is analytic in D. Hence its real part $\Phi(x, y) = \operatorname{Re} F(z)$ is harmonic in D. This completes the proof.

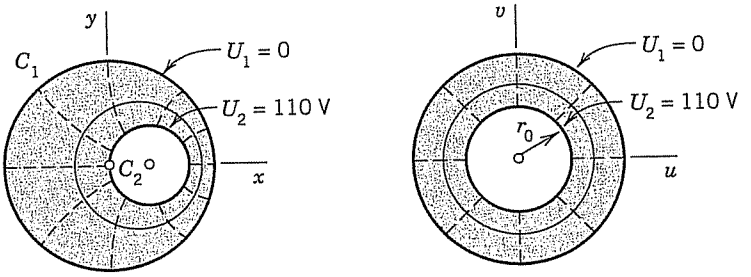
We mention without proof that if D^* is simply connected (Sec. 14.2), then a harmonic conjugate of Φ^* exists. Another proof of Theorem 1 without the use of a harmonic conjugate is given in App. 4. ∎

EXAMPLE 1 **Potential Between Noncoaxial Cylinders**

Model the electrostatic potential between the cylinders C_1: $|z| = 1$ and C_2: $|z - 2/5| = 2/5$ in Fig. 400. Then give the solution for the case that C_1 is grounded, $U_1 = 0$ V, and C_2 has the potential $U_2 = 110$ V.

Solution. We map the unit disk $|z| = 1$ onto the unit disk $|w| = 1$ in such a way that C_2 is mapped onto some cylinder C_2^*: $|w| = r_0$. By (3), Sec. 17.3, a linear fractional transformation mapping the unit disk onto the unit disk is

$$w = \frac{z - b}{bz - 1} \tag{2}$$



(a) z-plane (b) w-plane

Fig. 400. Example 1

where we have chosen $b = z_0$ real without restriction. z_0 is of no immediate help here because centers of circles do not map onto centers of the images, in general. However, we now have two free constants b and r_0 and shall succeed by imposing two reasonable conditions, namely, that 0 and 4/5 (Fig. 400) should be mapped onto r_0 and $-r_0$, respectively. This gives by (2)

$$r_0 = \frac{0 - b}{0 - 1} = b, \qquad \text{and with this,} \qquad -r_0 = \frac{4/5 - b}{4b/5 - 1} = \frac{4/5 - r_0}{4r_0/5 - 1},$$

a quadratic equation in r_0 with solutions $r_0 = 2$ (no good because $r_0 < 1$) and $r_0 = 1/2$. Hence our mapping function (2) with $b = 1/2$ becomes that in Example 5 of Sec. 17.3,

$$w = f(z) = \frac{2z - 1}{z - 2}. \tag{3}$$

From Example 5 in Sec. 18.1, writing w for z we have as the complex potential in the w-plane the function $F^*(w) = a \operatorname{Ln} w + k$ and from this the real potential

$$\Phi^*(u, v) = \operatorname{Re} F^*(w) = a \ln |w| + k.$$

This is our model. We now determine a and k from the boundary conditions. If $|w| = 1$, then $\Phi^* = a \ln 1 + k = 0$, hence $k = 0$. If $|w| = r_0 = 1/2$, then $\Phi^* = a \ln (1/2) = 110$, hence $a = 110/\ln (1/2) = -158.7$. Substitution of (3) now gives the desired solution in the given domain in the z-plane

$$F(z) = F^*(f(z)) = a \operatorname{Ln} \frac{2z - 1}{z - 2}.$$

The real potential is

$$\Phi(x, y) = \operatorname{Re} F(z) = a \ln \left| \frac{2z - 1}{z - 2} \right|, \qquad a = -158.7.$$

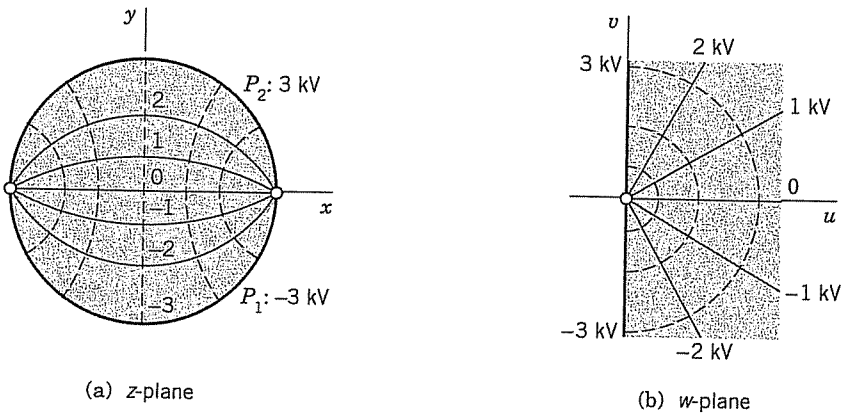
Can we "see" this result? Well, $\Phi(x, y) = const$ if and only if $|(2z - 1)/(z - 2)| = const$, that is, $|w| = const$ by (2) with $b = 1/2$. These circles are images of circles in the z-plane because the inverse of a linear fractional transformation is linear fractional (see (4), Sec. 17.2), and any such mapping maps circles onto circles (or straight lines), by Theorem 1 in Sec. 17.2. Similarly for the rays $\arg w = const$. Hence the equipotential lines $\Phi(x, y) = const$ are circles, and the lines of force are circular arcs (dashed in Fig. 400). These two families of curves intersect orthogonally, that is, at right angles, as shown in Fig. 400. ▩

EXAMPLE 2 Potential Between Two Semicircular Plates

Model the potential between two semicircular plates P_1 and P_2 in Fig. 401a having potentials -3000 V and 3000 V, respectively. Use Example 3 in Sec. 18.1 and conformal mapping.

Solution. ***Step 1.*** We map the unit disk in Fig. 401a onto the right half of the w-plane (Fig. 401b) by using the linear fractional transformation in Example 3, Sec. 17.3:

$$w = f(z) = \frac{1 + z}{1 - z}.$$



(a) z-plane (b) w-plane

Fig. 401. Example 2

The boundary $|z| = 1$ is mapped onto the boundary $u = 0$ (the v-axis), with $z = -1, i, 1$ going onto $w = 0, i, \infty$, respectively, and $z = -i$ onto $w = -i$. Hence the upper semicircle of $|z| = 1$ is mapped onto the upper half, and the lower semicircle onto the lower half of the v-axis, so that the boundary conditions in the w-plane are as indicated in Fig. 401b.

Step 2. We determine the potential $\Phi^*(u, v)$ in the right half-plane of the w-plane. Example 3 in Sec. 18.1 with $\alpha = \pi$, $U_1 = -3000$, and $U_2 = 3000$ [with $\Phi^*(u, v)$ instead of $\Phi(x, y)$] yields

$$\Phi^*(u, v) = \frac{6000}{\pi}\varphi, \qquad \varphi = \arctan\frac{v}{u}.$$

On the positive half of the imaginary axis ($\varphi = \pi/2$), this equals 3000 and on the negative half -3000, as it should be. Φ^* is the real part of the complex potential

$$F^*(w) = -\frac{6000\,i}{\pi}\operatorname{Ln} w.$$

Step 3. We substitute the mapping function into F^* to get the complex potential $F(z)$ in Fig. 401a in the form

$$F(z) = F^*(f(z)) = -\frac{6000\,i}{\pi}\operatorname{Ln}\frac{1+z}{1-z}.$$

The real part of this is the potential we wanted to determine:

$$\Phi(x, y) = \operatorname{Re} F(z) = \frac{6000}{\pi}\operatorname{Im}\operatorname{Ln}\frac{1+z}{1-z} = \frac{6000}{\pi}\operatorname{Arg}\frac{1+z}{1-z}.$$

As in Example 1 we conclude that the equipotential lines $\Phi(x, y) = const$ are circular arcs because they correspond to Arg $[(1 + z)/(1 - z)] = const$, hence to Arg $w = const$. Also, Arg $w = const$ are rays from 0 to ∞, the images of $z = -1$ and $z = 1$, respectively. Hence the equipotential lines all have -1 and 1 (the points where the boundary potential jumps) as their endpoints (Fig. 401a). The lines of force are circular arcs, too, and since they must be orthogonal to the equipotential lines, their centers can be obtained as intersections of tangents to the unit circle with the x-axis, (Explain!)

Further examples can easily be constructed. Just take any mapping $w = f(z)$ in Chap. 17, a domain D in the z-plane, its image D^* in the w-plane, and a potential Φ^* in D^*. Then (1) gives a potential in D. Make up some examples of your own, involving, for instance, linear fractional transformations.

Basic Comment on Modeling

We formulated the examples in this section as models on the electrostatic potential. It is quite important to realize that this is accidental. We could equally well have phrased everything in terms of (time-independent) heat flow; then instead of voltages we would have had temperatures, the equipotential lines would have become isotherms (= lines of constant temperature), and the lines of the electrical force would have become lines along which heat flows from higher to lower temperatures (more on this in the next section). Or we could have talked about fluid flow; then the electrostatic lines of force would have become streamlines (more on this in Sec. 18.4). What we again see here is the ***unifying power of mathematics:*** different phenomena and systems from different areas in physics having the same types of model can be treated by the same mathematical methods. What differs from area to area is just the kinds of problems that are of practical interest.

PROBLEM SET 18.2

1. Verify Theorem 1 for $\Phi^*(u, v) = u^2 - v^2$, $w = f(z) = e^z$ and any domain D.
2. Verify Theorem 1 for $\Phi^*(u, v) = uv$, $w = f(z) = e^z$, and D: $x \leqq 0$, $0 \leqq y \leqq \pi$. Sketch D and D^*.
3. Carry out all steps of the second proof of Theorem 1 (given in App. 4) in detail.
4. Derive (3) from (2).
5. Let D^* be the image of the rectangle D: $0 \leqq x \leqq \frac{1}{2}\pi$, $0 \leqq y \leqq 1$ under $w = \sin z$, and $\Phi^*(u, v) = u^2 - v^2$. Find the corresponding potential Φ in D and its boundary values.
6. What happens in Prob. 5 if you replace the potential by the conjugate $\Phi^* = 2uv$? Sketch or graph some of the equipotential lines $\Phi = const.$
7. **CAS PROJECT. Graphing Potential Fields.**
   (a) Graph equipotential lines in Probs. 1 and 2.
   (b) Graph equipotential lines if the complex potential is $F(z) = iz^2$, $F(z) = e^z$, $F(z) = ie^z$, $F(z) = e^{iz}$.
   (c) Graph equipotential surfaces corresponding to $F(z) = \ln z$ as cylinders in space.
8. **TEAM PROJECT. Noncoaxial Cylinders.** Find the potential between the cylinders C_1: $|z| = 1$ (potential $U_1 = 0$) and C_2: $|z - c| = c$ ($U_2 = 110$ V), where $0 < c < \frac{1}{2}$. Sketch or graph the equipotential curves and their orthogonal trajectories for $c = 0.1, 0.2, 0.3, 0.4$. Try to think of the further extension C_1: $|z| = 1$, C_2: $|z - c| = \rho \neq c$.
9. Find the potential Φ in the region R in the first quadrant of the z-plane bounded by the axes (having potential U_1) and the hyperbola $y = 1/x$ (having potential 0) in two ways, (i) directly, (ii) by mapping R onto a suitable infinite strip.
10. **(Extension of Example 2)** Find the linear fractional transformation $z = g(Z)$ that maps $|Z| \leqq 1$ onto $|z| \leqq 1$ with $Z = i/2$ being mapped onto $z = 0$. Show that $Z_1 = 0.6 + 0.8i$ is mapped onto $z = -1$ and $Z_2 = -0.6 + 0.8i$ onto $z = 1$, so that the equipotential lines of Example 2 look in $|Z| \leqq 1$ as shown in Fig. 402.

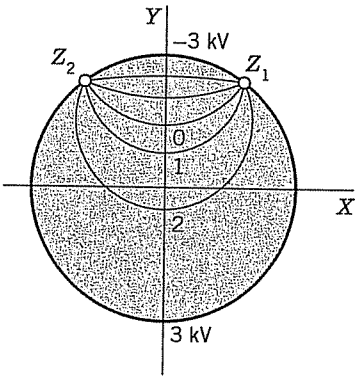


Fig. 402. Problem 10

11. The equipotential lines in Prob. 10 are circles. Why?
12. Show that in Example 2 the y-axis is mapped onto the unit circle in the w-plane.
13. Find the complex and real potentials in the upper half-plane with boundary values 0 if $x < 4$ and 10 kV if $x > 4$ on the x-axis.
14. **(Angular region)** Applying a suitable conformal mapping, obtain from Fig. 401b the potential Φ in the angular region $-\frac{1}{4}\pi < \text{Arg } z < \frac{1}{4}\pi$ such that $\Phi = -3$ kV if $\text{Arg } z = -\frac{1}{4}\pi$ and $\Phi = 3$ kV if $\text{Arg } z = \frac{1}{4}\pi$.
15. At $z = \pm 1$ in Fig. 401a the tangents to the equipotential lines shown make equal angles ($\pi/6$). Why?

18.3 Heat Problems

Laplace's equation also governs heat flow problems that are **steady,** that is, time-independent. Indeed, heat conduction in a body of homogeneous material is modeled by the **heat equation**

$$T_t = c^2 \nabla^2 T$$

where the function T is temperature, $T_t = \partial T/\partial t$, t is time, and c^2 is a positive constant (depending on the material of the body; see Sec. 12.5). Hence if a problem is steady, so that $T_t = 0$, and two-dimensional, then the heat equation reduces to the two-dimensional Laplace equation

$$\nabla^2 T = T_{xx} + T_{yy} = 0, \tag{1}$$

so that the problem can be treated by our present methods.

$T(x, y)$ is called the **heat potential.** It is the real part of the **complex heat potential**

$$F(z) = T(x, y) + i\Psi(x, y).$$

The curves $T(x, y) = const$ are called **isotherms** (= lines of constant temperature) and the curves $\Psi(x, y) = const$ **heat flow lines,** because along them, heat flows from higher to lower temperatures.

It follows that all the examples considered so far (Secs. 18.1, 18.2) can now be reinterpreted as problems on heat flow. The electrostatic equipotential lines $\Phi(x, y) = const$ now become isotherms $T(x, y) = const$, and the lines of electrical force become lines of heat flow, as in the following two problems.

EXAMPLE 1 **Temperature Between Parallel Plates**

Find the temperature between two parallel plates $x = 0$ and $x = d$ in Fig. 403 having temperatures 0 and 100°C, respectively.

Solution. As in Example 1 of Sec. 18.1 we conclude that $T(x, y) = ax + b$. From the boundary conditions, $b = 0$ and $a = 100/d$. The answer is

$$T(x, y) = \frac{100}{d}\, x \ [°C].$$

The corresponding complex potential is $F(z) = (100/d)z$. Heat flows horizontally in the negative x-direction along the lines $y = const$. ■

EXAMPLE 2 **Temperature Distribution Between a Wire and a Cylinder**

Find the temperature field around a long thin wire of radius $r_1 = 1$ mm that is electrically heated to $T_1 = 500°F$ and is surrounded by a circular cylinder of radius $r_2 = 100$ mm, which is kept at temperature $T_2 = 60°F$ by cooling it with air. See Fig. 404. (The wire is at the origin of the coordinate system.)

Solution. T depends only on r, for reasons of symmetry. Hence, as in Sec. 18.1 (Example 2),

$$T(x, y) = a \ln r + b.$$

The boundary conditions are

$$T_1 = 500 = a \ln 1 + b, \qquad T_2 = 60 = a \ln 100 + b.$$

Hence $b = 500$ (since $\ln 1 = 0$) and $a = (60 - b)/\ln 100 = -95.54$. The answer is

$$T(x, y) = 500 - 95.54 \ln r \ [°F].$$

The isotherms are concentric circles. Heat flows from the wire radially outward to the cylinder. Sketch T as a function of r. Does it look physically reasonable? ■

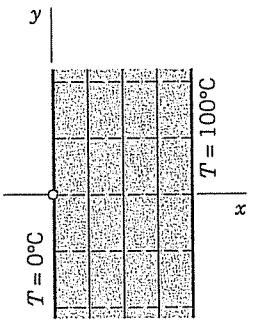


Fig. 403. Example 1

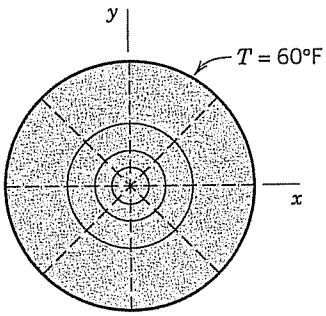


Fig. 404. Example 2

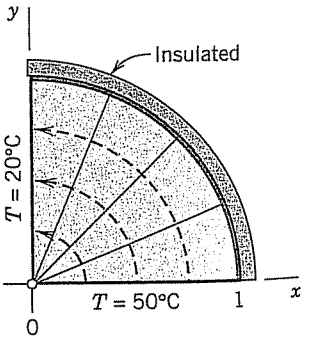


Fig. 405. Example 3

Mathematically the calculations remain the same in the transition to another field of application. Physically, new problems may arise, with boundary conditions that would make no sense physically or would be of no practical interest. This is illustrated by the next two examples.

EXAMPLE 3 **A Mixed Boundary Value Problem**

Find the temperature distribution in the region in Fig. 405 (cross section of a solid quarter-cylinder), whose vertical portion of the boundary is at 20°C, the horizontal portion at 50°C, and the circular portion is insulated.

Solution. The insulated portion of the boundary must be a heat flow line, since by the insulation, heat is prevented from crossing such a curve, hence heat must flow along the curve. Thus the isotherms must meet such a curve at right angles. Since T is constant along an isotherm, this means that

$$\frac{\partial T}{\partial n} = 0 \qquad \text{along an insulated portion of the boundary.} \tag{2}$$

Here $\partial T/\partial n$ is the **normal derivative** of T, that is, the directional derivative (Sec. 9.7) in the direction normal (perpendicular) to the insulated boundary. Such a problem in which T is prescribed on one portion of the boundary and $\partial T/\partial n$ on the other portion is called a **mixed boundary value problem.**

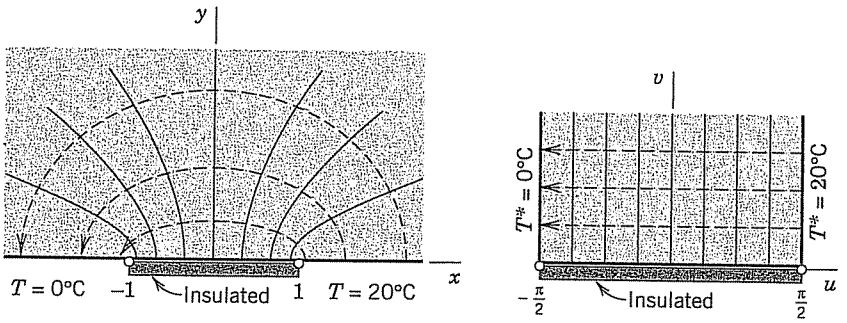
In our case, the normal direction to the insulated circular boundary curve is the radial direction toward the origin. Hence (2) becomes $\partial T/\partial r = 0$, meaning that along this curve the solution must not depend on r. Now $\operatorname{Arg} z = \theta$ satisfies (1), as well as this condition, and is constant (0 and $\pi/2$) on the straight portions of the boundary. Hence the solution is of the form

$$T(x, y) = a\theta + b.$$

The boundary conditions yield $a \cdot \pi/2 + b = 20$ and $a \cdot 0 + b = 50$. This gives

$$T(x, y) = 50 - \frac{60}{\pi}\,\theta, \qquad \theta = \arctan\frac{y}{x}.$$

The isotherms are portions of rays $\theta = const$. Heat flows from the x-axis along circles $r = const$ (dashed in Fig. 405) to the y-axis. ■



(a) z-plane (b) w-plane

Fig. 406. Example 4

EXAMPLE 4 **Another Mixed Boundary Value Problem in Heat Conduction**

Find the temperature field in the upper half-plane when the x-axis is kept at $T = 0°C$ for $x < -1$, is insulated for $-1 < x < 1$, and is kept at $T = 20°C$ for $x > 1$ (Fig. 406a).

Solution. We map the half-plane in Fig. 406a onto the vertical strip in Fig. 406b, find the temperature $T^*(u, v)$ there, and map it back to get the temperature $T(x, y)$ in the half-plane.

The idea of using that strip is suggested by Fig. 388 in Sec. 17.4 with the roles of $z = x + iy$ and $w = u + iv$ interchanged. The figure shows that $z = \sin w$ maps our present strip onto our half-plane in Fig. 406a. Hence the inverse function

$$w = f(z) = \arcsin z$$

maps that half-plane onto the strip in the w-plane. This is the mapping function that we need according to Theorem 1 in Sec. 18.2.

The insulated segment $-1 < x < 1$ on the x-axis maps onto the segment $-\pi/2 < u < \pi/2$ on the u-axis. The rest of the x-axis maps onto the two vertical boundary portions $u = -\pi/2$ and $\pi/2$, $v > 0$, of the strip. This gives the transformed boundary conditions in Fig. 406b for $T^*(u, v)$, where on the insulated horizontal boundary, $\partial T^*/\partial n = \partial T^*/\partial v = 0$ because v is a coordinate normal to that segment.

Similarly to Example 1 we obtain

$$T^*(u, v) = 10 + \frac{20}{\pi} u$$

which satisfies all the boundary conditions. This is the real part of the complex potential $F^*(w) = 10 + (20/\pi)w$. Hence the complex potential in the z-plane is

$$F(z) = F^*(f(z)) = 10 + \frac{20}{\pi} \arcsin z$$

and $T(x, y) = \text{Re } F(z)$ is the solution. The isotherms are $u = const$ in the strip and the hyperbolas in the z-plane, perpendicular to which heat flows along the dashed ellipses from the 20°-portion to the cooler 0°-portion of the boundary, a physically very reasonable result. ■

This section and the last one show the usefulness of conformal mappings and complex potentials. The latter will also play a role in the next section on fluid flow.

PROBLEM SET 18.3

1. **CAS PROJECT. Isotherms.** Graph isotherms and lines of heat flow in Examples 2–4. Can you see from the graphs where the heat flow is very rapid?
2. Find the temperature and the complex potential in an infinite plate with edges $y = x - 2$ and $y = x + 2$ kept at -10°C and 20°C, respectively.
3. Find the temperature between two parallel plates $y = 0$ and $y = d$ kept at temperatures 0°C and 100°C, respectively. (i) Proceed directly. (ii) Use Example 1 and a suitable mapping.
4. Find the temperature T in the sector $0 \leqq \text{Arg } z \leqq \pi/3$, $|z| \leqq 1$ if $T = 20$°C on the x-axis, $T = 50$°C on $y = \sqrt{3}\, x$, and the curved portion is insulated.
5. Find the temperature in Fig. 405 if $T = -20$°C on the y-axis, $T = 100$°C on the x-axis, and the circular portion of the boundary is insulated as before.
6. Interpret Prob. 10 in Sec. 18.2 as a heat flow problem (with boundary temperatures, say, 20°C and 300°C). Along what curves does the heat flow?
7. Find the temperature and the complex potential in the first quadrant of the z-plane if the y-axis is kept at 100°C, the segment $0 < x < 1$ of the x-axis is insulated and the portion $x > 1$ of the x-axis is kept at 200°C. *Hint.* Use Example 4.
8. **TEAM PROJECT. Piecewise Constant Boundary Temperatures. (a) A basic building block** is shown in Fig. 407. Find the corresponding temperature and complex potential in the upper half-plane.

   (b) Conformal mapping. What temperature in the first quadrant of the z-plane is obtained from (a) by the mapping $w = a + z^2$ and what are the transformed boundary conditions?

   (c) Superposition. Find the temperature T^* and the complex potential F^* in the upper half-plane satisfying the boundary condition in Fig. 408.

   (d) Semi-infinite strip. Applying $w = \cosh z$ to (c), obtain the solution of the boundary value problem in Fig. 409.

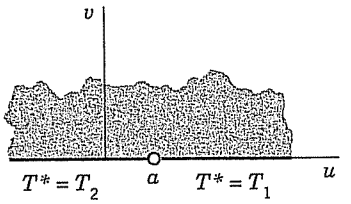


Fig. 407. Team Project 8(a)

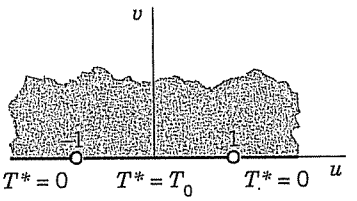


Fig. 408. Team Project 8(c)

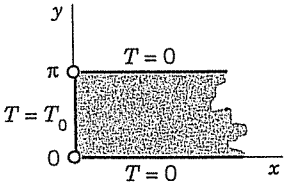
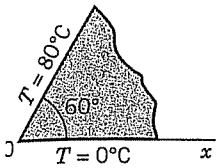


Fig. 409. Team Project 8(d)

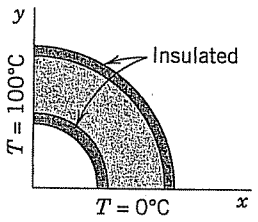
9–14 TEMPERATURE DISTRIBUTIONS IN PLATES

Find the temperature $T(x, y)$ in the given thin metal plate whose faces are insulated and whose edges are kept at the indicated temperatures or are insulated as shown.

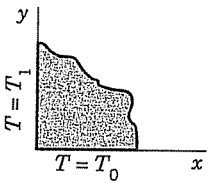
9.



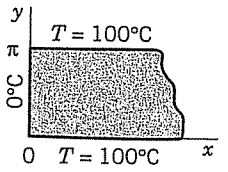
10.



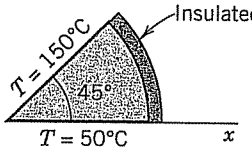
11.



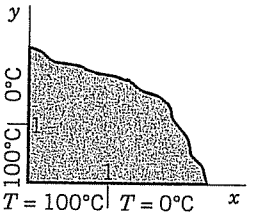
12.



13.



14.



18.4 Fluid Flow

Laplace's equation also plays a basic role in hydrodynamics, in steady nonviscous fluid flow under physical conditions discussed later in this section. In order that methods of complex analysis can be applied, our problems will be ***two-dimensional,*** so that the **velocity vector** V by which the motion of the fluid can be given depends only on two space variables x and y and the motion is the same in all planes parallel to the xy-plane.

Then we can use for the velocity vector V a complex function

$$V = V_1 + iV_2 \tag{1}$$

giving the magnitude $|V|$ and direction Arg V of the velocity at each point $z = x + iy$. Here V_1 and V_2 are the components of the velocity in the x and y directions. V is tangential to the path of the moving particles, called a **streamline** of the motion (Fig. 410).

We show that under suitable assumptions (explained in detail following the examples), for a given flow there exists an analytic function

$$F(z) = \Phi(x, y) + i\Psi(x, y), \tag{2}$$

called the **complex potential** of the flow, such that the streamlines are given by $\Psi(x, y) = const,$ and the velocity vector or, briefly, the **velocity** is given by

$$V = V_1 + iV_2 = \overline{F'(z)} \tag{3}$$

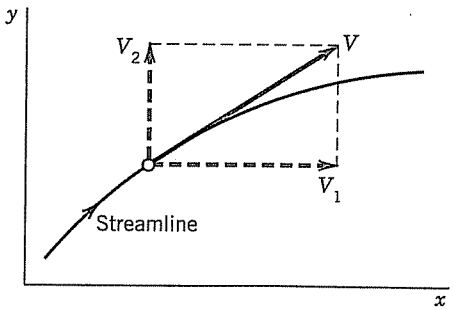


Fig. 410. Velocity

where the bar denotes the complex conjugate. Ψ is called the **stream function.** The function Φ is called the **velocity potential.** The curves $\Phi(x, y) = const$ are called **equipotential lines.** The velocity vector V is the **gradient** of Φ; by definition, this means that

$$(4) \qquad V_1 = \frac{\partial \Phi}{\partial x}, \qquad V_2 = \frac{\partial \Phi}{\partial y}.$$

Indeed, for $F = \Phi + i\Psi$, Eq. (4) in Sec. 13.4 is $F' = \Phi_x + i\Psi_x$ with $\Psi_x = -\Phi_y$ by the second Cauchy–Riemann equation. Together we obtain (3):

$$\overline{F'(z)} = \Phi_x - i\Psi_x = \Phi_x + i\Phi_y = V_1 + iV_2 = V.$$

Furthermore, since $F(z)$ is analytic, Φ and Ψ satisfy Laplace's equation

$$(5) \qquad \nabla^2\Phi = \frac{\partial^2 \Phi}{\partial x^2} + \frac{\partial^2 \Phi}{\partial y^2} = 0, \qquad \nabla^2\Psi = \frac{\partial^2 \Psi}{\partial x^2} + \frac{\partial^2 \Psi}{\partial y^2} = 0.$$

Whereas in electrostatics the boundaries (conducting plates) are equipotential lines, in fluid flow the boundaries across which fluid cannot flow must be streamlines. Hence in fluid flow the stream function is of particular importance.

Before discussing the conditions for the validity of the statements involving (2)–(5), let us consider two flows of practical interest, so that we first see what is going on from a practical point of view. Further flows are included in the problem set.

EXAMPLE 1 **Flow Around a Corner**

The complex potential $F(z) = z^2 = x^2 - y^2 + 2ixy$ models a flow with

Equipotential lines	$\Phi = x^2 - y^2 = const$	(Hyperbolas)
Streamlines	$\Psi = 2xy = const$	(Hyperbolas).

From (3) we obtain the velocity vector

$$V = 2\bar{z} = 2(x - iy), \qquad \text{that is,} \qquad V_1 = 2x, \qquad V_2 = -2y.$$

The speed (magnitude of the velocity) is

$$|V| = \sqrt{V_1^2 + V_2^2} = 2\sqrt{x^2 + y^2}.$$

The flow may be interpreted as the flow in a channel bounded by the positive coordinates axes and a hyperbola, say, $xy = 1$ (Fig. 411). We note that the speed along a streamline S has a minimum at the point P where the cross section of the channel is large. ■

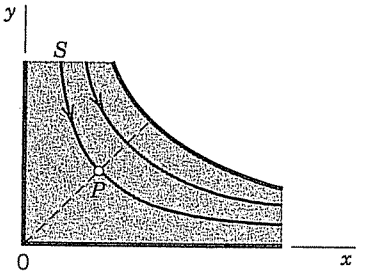


Fig. 411. Flow around a corner (Example 1)

EXAMPLE 2 Flow Around a Cylinder

Consider the complex potential

$$F(z) = \Phi(x, y) + i\Psi(x, y) = z + \frac{1}{z}.$$

Using the polar form $z = re^{i\theta}$, we obtain

$$F(z) = re^{i\theta} + \frac{1}{r}e^{-i\theta} = \left(r + \frac{1}{r}\right)\cos\theta + i\left(r - \frac{1}{r}\right)\sin\theta.$$

Hence the streamlines are

$$\Psi(x, y) = \left(r - \frac{1}{r}\right)\sin\theta = const.$$

In particular, $\Psi(x, y) = 0$ gives $r - 1/r = 0$ or $\sin\theta = 0$. Hence this streamline consists of the unit circle ($r = 1/r$ gives $r = 1$) and the x-axis ($\theta = 0$ and $\theta = \pi$). For large $|z|$ the term $1/z$ in $F(z)$ is small in absolute value, so that for these z the flow is nearly uniform and parallel to the x-axis. Hence we can interpret this as a flow around a long circular cylinder of unit radius that is perpendicular to the z-plane and intersects it in the unit circle $|z| = 1$ and whose axis corresponds to $z = 0$.

The flow has two **stagnation points** (that is, points at which the velocity V is zero), at $z = \pm 1$. This follows from (3) and

$$F'(z) = 1 - \frac{1}{z^2}, \qquad \text{hence} \qquad z^2 - 1 = 0. \qquad \text{(See Fig. 412.)}$$

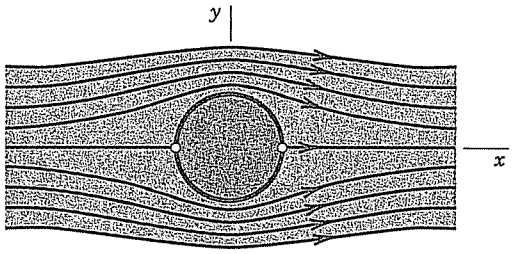


Fig. 412. Flow around a cylinder (Example 2)

Assumptions and Theory Underlying (2)–(5)

THEOREM 1

Complex Potential of a Flow

If the domain of flow is simply connected and the flow is irrotational and incompressible, then the statements involving (2)–(5) *hold. In particular, then the flow has a complex potential* $F(z)$, *which is an analytic function.* (Explanation of terms below.)

PROOF We prove this theorem, along with a discussion of basic concepts related to fluid flow.

(a) ***First Assumption: Irrotational.*** Let C be any smooth curve in the z-plane given by $z(s) = x(s) + iy(s)$, where s is the arc length of C. Let the real variable V_t be the component of the velocity V tangent to C (Fig. 413). Then the value of the real line integral

(6)
$$\int_C V_t\, ds$$

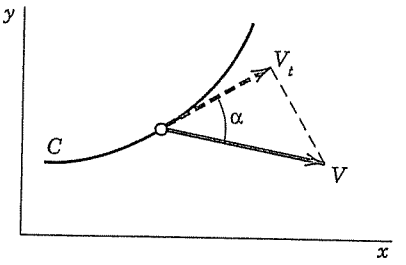


Fig. 413. Tangential component of the velocity with respect to a curve C

taken along C in the sense of increasing s is called the **circulation** of the fluid along C, a name that will be motivated as we proceed in this proof. Dividing the circulation by the length of C, we obtain the *mean velocity*[1] of the flow along the curve C. Now

$$V_t = |V| \cos \alpha \qquad \text{(Fig. 413).}$$

Hence V_t is the dot product (Sec. 9.2) of V and the tangent vector dz/ds of C (Sec. 17.1); thus in (6),

$$V_t\, ds = \left(V_1 \frac{dx}{ds} + V_2 \frac{dy}{ds}\right) ds = V_1\, dx + V_2\, dy.$$

The circulation (6) along C now becomes

$$\int_C V_t\, ds = \int_C (V_1\, dx + V_2\, dy). \tag{7}$$

As the next idea, let C be a ***closed curve*** satisfying the assumption as in Green's theorem (Sec. 10.4), and let C be the boundary of a simply connected domain D. Suppose further that V has continuous partial derivatives in a domain containing D and C. Then we can use Green's theorem to represent the circulation around C by a double integral,

$$\oint_C (V_1\, dx + V_2\, dy) = \iint_D \left(\frac{\partial V_2}{\partial x} - \frac{\partial V_1}{\partial y}\right) dx\, dy. \tag{8}$$

The integrand of this double integral is called the **vorticity** of the flow. The vorticity divided by 2 is called the **rotation**

$$\omega(x, y) = \frac{1}{2}\left(\frac{\partial V_2}{\partial x} - \frac{\partial V_1}{\partial y}\right). \tag{9}$$

---

[1] *Definitions:* $\dfrac{1}{b-a}\displaystyle\int_a^b f(x)\, dx =$ **mean value** of f on the interval $a \leqq x \leqq b$,

$\dfrac{1}{L}\displaystyle\int_C f(s)\, ds =$ **mean value** of f on C (L = length of C),

$\dfrac{1}{A}\displaystyle\iint_D f(x, y)\, dx\, dy =$ **mean value** of f on D (A = area of D).

We assume the flow to be **irrotational,** that is, $\omega(x, y) \equiv 0$ throughout the flow; thus,

$$\frac{\partial V_2}{\partial x} - \frac{\partial V_1}{\partial y} = 0. \tag{10}$$

To understand the physical meaning of vorticity and rotation, take for C in (8) a circle. Let r be the radius of C. Then the circulation divided by the length $2\pi r$ of C is the mean velocity of the fluid along C. Hence by dividing this by r we obtain the mean **angular velocity** ω_0 of the fluid about the center of the circle:

$$\omega_0 = \frac{1}{2\pi r^2} \iint_D \left(\frac{\partial V_2}{\partial x} - \frac{\partial V_1}{\partial y} \right) dx\, dy = \frac{1}{\pi r^2} \iint_D \omega(x, y)\, dx\, dy.$$

If we now let $r \to 0$, the limit of ω_0 is the value of ω at the center of C. Hence $\omega(x, y)$ is the limiting angular velocity of a circular element of the fluid as the circle shrinks to the point (x, y). Roughly speaking, *if a spherical element of the fluid were suddenly solidified and the surrounding fluid simultaneously annihilated, the element would rotate with the angular velocity* ω.

(b) ***Second Assumption: Incompressible.*** Our second assumption is that the fluid is incompressible. (Fluids include liquids, which are incompressible, and gases, such as air, which are compressible.) Then

$$\frac{\partial V_1}{\partial x} + \frac{\partial V_2}{\partial y} = 0 \tag{11}$$

in every region that is free of **sources** or **sinks,** that is, points at which fluid is produced or disappears, respectively. The expression in (11) is called the **divergence** of V and is denoted by div V. (See also (7) in Sec. 9.8.)

(c) ***Complex Velocity Potential.*** If the domain D of the flow is simply connected (Sec. 14.2) and the flow is irrotational, then (10) implies that the line integral (7) is independent of path in D (by Theorem 3 in Sec. 10.2, where $F_1 = V_1$, $F_2 = V_2$, $F_3 = 0$, and z is the third coordinate in space and has nothing to do with our present z). Hence if we integrate from a fixed point (a, b) in D to a variable point (x, y) in D, the integral becomes a function of the point (x, y), say, $\Phi(x, y)$:

$$\Phi(x, y) = \int_{(a, b)}^{(x, y)} (V_1\, dx + V_2\, dy). \tag{12}$$

We claim that the flow has a velocity potential Φ, which is given by (12). To prove this, all we have to do is to show that (4) holds. Now since the integral (7) is independent of path, $V_1\, dx + V_2\, dy$ is exact (Sec. 10.2), namely, the differential of Φ, that is,

$$V_1\, dx + V_2\, dy = \frac{\partial \Phi}{\partial x}\, dx + \frac{\partial \Phi}{\partial y}\, dy.$$

From this we see that $V_1 = \partial\Phi/\partial x$ and $V_2 = \partial\Phi/\partial y$, which gives (4).

That Φ is harmonic follows at once by substituting (4) into (11), which gives the first Laplace equation in (5).

We finally take a harmonic conjugate Ψ of Φ. Then the other equation in (5) holds. Also, since the second partial derivatives of Φ and Ψ are continuous, we see that the complex function

$$F(z) = \Phi(x, y) + i\Psi(x, y)$$

is analytic in D. Since the curves $\Psi(x, y) = const$ are perpendicular to the equipotential curves $\Phi(x, y) = const$ (except where $F'(z) = 0$), we conclude that $\Psi(x, y) = const$ are the streamlines. Hence Ψ is the stream function and $F(z)$ is the complex potential of the flow. This completes the proof of Theorem 1 as well as our discussion of the important role of complex analysis in compressible fluid flow. ■

PROBLEM SET 18.4

1–15 FLOW PATTERNS: STREAMLINES, COMPLEX POTENTIAL

These problems should encourage you to experiment with various functions $F(z)$, many of which model interesting flow patterns.

1. **(Parallel flow)** Show that $F(z) = -iKz$ (K positive real) describes a uniform flow upward, which can be interpreted as a uniform flow between two parallel lines (parallel planes in three-dimensional space). See Fig. 414. Find the velocity vector, the streamlines, and the equipotential lines.

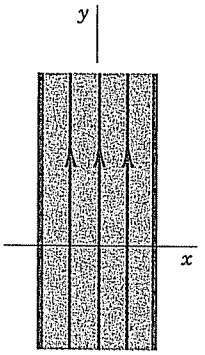


Fig. 414. Parallel flow in Problem 1

2. **(Conformal mapping)** Obtain the flow in Example 1 from that in Prob. 1 by a suitable conformal mapping.
3. Find the complex potential of a uniform flow parallel to the x-axis in the positive x-direction.
4. What happens to the flow in Prob. 1 if you replace z by $ze^{-i\alpha}$ with constant α, e.g., $\alpha = \pi/4$?
5. What is the complex potential of an upward parallel flow in the direction of $y = 2x$?
6. **(Extension of Example 1)** Sketch or graph the flow in Example 1 on the whole upper half-plane. Show that you can interpret it as as flow against a horizontal wall (the x-axis).
7. What $F(z)$ would be suitable in Example 1 if the angle of the corner were $\pi/3$?
8. Sketch or graph the streamlines and equipotential lines of $F(z) = iz^3$. Find V. Find all points at which V is parallel to the x-axis.
9. Find and graph the streamlines of $F(z) = z^2 + 2z$. Interpret the flow.
10. Show that $F(z) = iz^2$ models a flow around a corner. Sketch the streamlines and equipotential lines. Find V.
11. **(Potential $F(z) = 1/z$)** Show that the streamlines of $F(z) = 1/z$ are circles through the origin.
12. **(Cylinder)** What happens in Example 2 if you replace z by z^2? Sketch and interpret the resulting flow in the first quadrant.
13. Change $F(z)$ in Example 2 slightly to obtain a flow around a cylinder of radius r_0 that gives the flow in Example 2 if $r_0 \to 1$.
14. **(Aperture)** Show that $F(z) = \text{arccosh } z$ gives confocal hyperbolas as streamlines, with foci at $z = \pm 1$, and the flow may be interpreted as a flow through an aperture (Fig. 415).
15. **(Elliptical cylinder)** Show that $F(z) = \arccos z$ gives confocal ellipses as streamlines, with foci at $z = \pm 1$, and that the flow circulates around an elliptic cylinder or a plate (the segment from -1 to 1 in Fig. 416).

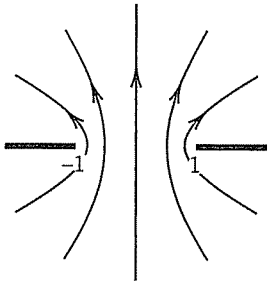


Fig. 415. Flow through an aperture in Problem 14

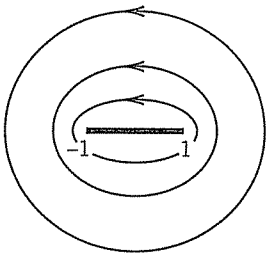


Fig. 416. Flow around a plate in Problem 15

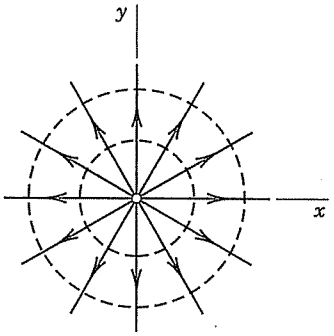


Fig. 417. Point source

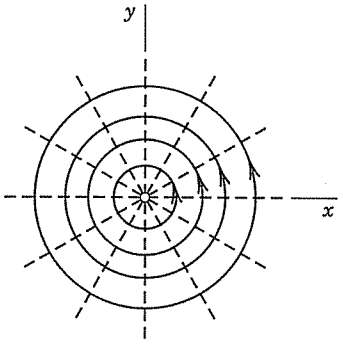


Fig. 418. Vortex flow

16. **TEAM PROJECT. Role of the Natural Logarithm in Modeling Flows. (a) Basic flows: Source and sink.** Show that $F(z) = (c/2\pi) \ln z$ with constant positive real c gives a flow directed radially outward (Fig. 417), so that F models a **point source** at $z = 0$ (that is, a **source line** $x = 0$, $y = 0$ in space) at which fluid is produced. c is called the **strength** or **discharge** of the source. If c is negative real, show that the flow is directed radially inward, so that F models a **sink** at $z = 0$, a point at which fluid disappears. Note that $z = 0$ is the singular point of $F(z)$.

(b) Basic flows: Vortex. Show that $F(z) = -(Ki/2\pi) \ln z$ with positive real K gives a flow circulating counterclockwise around $z = 0$ (Fig. 418). $z = 0$ is called a **vortex.** Note that each time we travel around the vortex, the potential increases by K.

(c) Addition of flows. Show that addition of the velocity vectors of two flows gives a flow whose complex potential is obtained by adding the complex potentials of those flows.

(d) Source and sink combined. Find the complex potentials of a flow with a source of strength 1 at $z = -a$ and of a flow with a sink of strength 1 at $z = a$. Add both and sketch or graph the streamlines. Show that for small $|a|$ these lines look similar to those in Prob. 11.

(e) Flow with circulation around a cylinder. Add the potential in (b) to that in Example 2. Show that this gives a flow for which the cylinder wall $|z| = 1$ is a streamline. Find the speed and show that the stagnation points are

$$z = \frac{iK}{4\pi} \pm \sqrt{\frac{-K^2}{16\pi^2} + 1};$$

if $K = 0$ they are at ± 1; as K increases they move up on the unit circle until they unite at $z = i$ ($K = 4\pi$, see Fig. 419), and if $K > 4\pi$ they lie on the imaginary axis (one lies in the field of flow and the other one lies inside the cylinder and has no physical meaning).

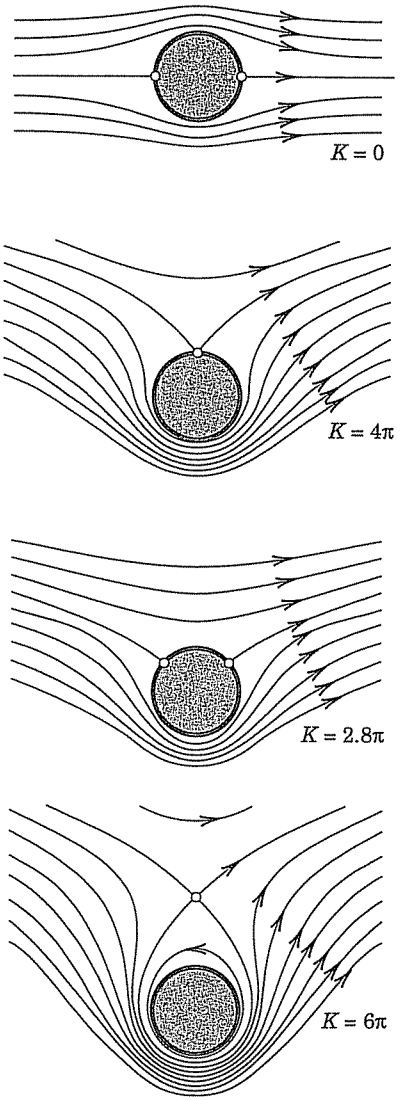


Fig. 419. Flow around a cylinder without circulation ($K = 0$) and with circulation

18.5 Poisson's Integral Formula for Potentials

So far in this chapter we have seen that complex analysis offers powerful methods for modeling and solving two-dimensional potential problems based on conformal mappings and complex potentials. A further method results from complex integration. As a most important result it yields Poisson's integral formula (5) for potentials in a standard domain (a circular disk) and from (5) a useful series (7) for these potentials. Hence we can solve problems for disks and then map solutions conformally onto other domains.

Poisson's formula will follow from Cauchy's integral formula (Sec. 14.3)

$$(1) \qquad F(z) = \frac{1}{2\pi i} \oint_C \frac{F(z^*)}{z^* - z}\, dz^*.$$

Here C is the circle $z^* = Re^{i\alpha}$ (counterclockwise, $0 \leqq \alpha \leqq 2\pi$), and we assume that $F(z^*)$ is analytic in a domain containing C and its full interior. Since $dz^* = iRe^{i\alpha}\, d\alpha = iz^*\, d\alpha$, we obtain from (1)

$$(2) \qquad F(z) = \frac{1}{2\pi} \int_0^{2\pi} F(z^*)\, \frac{z^*}{z^* - z}\, d\alpha \qquad (z^* = Re^{i\alpha},\ z = re^{i\theta}).$$

Now comes a little trick. If instead of z inside C we take a Z outside C, the integrals (1) and (2) are zero by Cauchy's integral theorem (Sec. 14.2). We choose $Z = z^*\bar{z}^*/\bar{z} = R^2/\bar{z}$, which is outside C because $|Z| = R^2/|z| = R^2/r > R$. From (2) we thus have

$$0 = \frac{1}{2\pi} \int_0^{2\pi} F(z^*)\, \frac{z^*}{z^* - Z}\, d\alpha = \frac{1}{2\pi} \int_0^{2\pi} F(z^*)\, \frac{z^*}{z^* - \dfrac{z^*\bar{z}^*}{\bar{z}}}\, d\alpha$$

and by straightforward simplification of the last expression on the right,

$$0 = \frac{1}{2\pi} \int_0^{2\pi} F(z^*)\, \frac{\bar{z}}{\bar{z} - \bar{z}^*}\, d\alpha.$$

We subtract this from (2) and use the following formula that you can verify by direct calculation ($\bar{z}z^*$ cancels):

$$(3) \qquad \frac{z^*}{z^* - z} - \frac{\bar{z}}{\bar{z} - \bar{z}^*} = \frac{z^*\bar{z}^* - z\bar{z}}{(z^* - z)(\bar{z}^* - \bar{z})}.$$

We then have

$$(4) \qquad F(z) = \frac{1}{2\pi} \int_0^{2\pi} F(z^*)\, \frac{z^*\bar{z}^* - z\bar{z}}{(z^* - z)(\bar{z}^* - \bar{z})}\, d\alpha.$$

From the polar representations of z and z^* we see that the quotient in the integrand is real and equal to

$$\frac{R^2 - r^2}{(Re^{i\alpha} - re^{i\theta})(Re^{-i\alpha} - re^{-i\theta})} = \frac{R^2 - r^2}{R^2 - 2Rr\cos(\theta - \alpha) + r^2}.$$

We now write $F(z) = \Phi(r, \theta) + i\Psi(r, \theta)$ and take the real part on both sides of (4). Then we obtain **Poisson's integral formula**[2]

$$\Phi(r, \theta) = \frac{1}{2\pi} \int_0^{2\pi} \Phi(R, \alpha) \frac{R^2 - r^2}{R^2 - 2Rr \cos(\theta - \alpha) + r^2} \, d\alpha. \tag{5}$$

This formula represents the harmonic function Φ in the disk $|z| \leqq R$ in terms of its values $\Phi(R, \alpha)$ on the boundary (the circle) $|z| = R$.

Formula (5) is still valid if the boundary function $\Phi(R, \alpha)$ is merely piecewise continuous (as is practically often the case; see Fig. 401 in Sec. 18.2 for an example). Then (5) gives a function harmonic in the open disk, and on the circle $|z| = R$ equal to the given boundary function, except at points where the latter is discontinuous. A proof can be found in Ref. [D1] in App. 1.

Series for Potentials in Disks

From (5) we may obtain an important series development of Φ in terms of simple harmonic functions. We remember that the quotient in the integrand of (5) was derived from (3). We claim that the right side of (3) is the real part of

$$\frac{z^* + z}{z^* - z} = \frac{(z^* + z)(\bar{z}^* - \bar{z})}{(z^* - z)(\bar{z}^* - \bar{z})} = \frac{z^*\bar{z}^* - z\bar{z} - z^*\bar{z} + z\bar{z}^*}{|z^* - z|^2}.$$

Indeed, the last denominator is real and so is $z^*\bar{z}^* - z\bar{z}$ in the numerator, whereas $-z^*\bar{z} + z\bar{z}^* = 2i \operatorname{Im}(z\bar{z}^*)$ in the numerator is pure imaginary. This verifies our claim. Now by the use of the geometric series we obtain (develop the denominator)

$$\frac{z^* + z}{z^* - z} = \frac{1 + (z/z^*)}{1 - (z/z^*)} = \left(1 + \frac{z}{z^*}\right) \sum_{n=0}^{\infty} \left(\frac{z}{z^*}\right)^n = 1 + 2 \sum_{n=1}^{\infty} \left(\frac{z}{z^*}\right)^n. \tag{6}$$

Since $z = re^{i\theta}$ and $z^* = Re^{i\alpha}$, we have

$$\operatorname{Re}\left[\left(\frac{z}{z^*}\right)^n\right] = \operatorname{Re}\left[\frac{r^n}{R^n} e^{in\theta} e^{-in\alpha}\right] = \left(\frac{r}{R}\right)^n \cos(n\theta - n\alpha).$$

On the right, $\cos(n\theta - n\alpha) = \cos n\theta \cos n\alpha + \sin n\theta \sin n\alpha$. Hence from (6) we obtain

$$\begin{aligned} \operatorname{Re} \frac{z^* + z}{z^* - z} &= 1 + 2 \sum_{n=1}^{\infty} \operatorname{Re}\left(\frac{z}{z^*}\right)^n \\ &= 1 + 2 \sum_{n=1}^{\infty} \left(\frac{r}{R}\right)^n (\cos n\theta \cos n\alpha + \sin n\theta \sin n\alpha). \end{aligned} \tag{6*}$$

[2]SIMÉON DENIS POISSON (1781–1840), French mathematician and physicist, professor in Paris from 1809. His work includes potential theory, partial differential equations (Poisson equation, Sec. 12.1), and probability (Sec. 24.7).

This expression is equal to the quotient in (5), as we have mentioned before, and by inserting it into (5) and integrating term by term with respect to α from 0 to 2π we obtain

$$\Phi(r, \theta) = a_0 + \sum_{n=1}^{\infty} \left(\frac{r}{R}\right)^n (a_n \cos n\theta + b_n \sin n\theta) \tag{7}$$

where the coefficients are [the 2 in (6*) cancels the 2 in $1/(2\pi)$ in (5)]

$$a_0 = \frac{1}{2\pi} \int_0^{2\pi} \Phi(R, \alpha)\, d\alpha, \qquad a_n = \frac{1}{\pi} \int_0^{2\pi} \Phi(R, \alpha) \cos n\alpha\, d\alpha,$$
$$b_n = \frac{1}{\pi} \int_0^{2\pi} \Phi(R, \alpha) \sin n\alpha\, d\alpha, \qquad n = 1, 2, \cdots, \tag{8}$$

the Fourier coefficients of $\Phi(R, \alpha)$; see Sec. 11.1. Now for $r = R$ the series (7) becomes the Fourier series of $\Phi(R, \alpha)$. Hence the representation (7) will be valid whenever the given $\Phi(R, \alpha)$ on the boundary can be represented by a Fourier series.

EXAMPLE 1 **Dirichlet Problem for the Unit Disk**

Find the electrostatic potential $\Phi(r, \theta)$ in the unit disk $r < 1$ having the boundary values

$$\Phi(1, \alpha) = \begin{cases} -\alpha/\pi & \text{if } -\pi < \alpha < 0 \\ \alpha/\pi & \text{if } 0 < \alpha < \pi \end{cases} \qquad \text{(Fig. 420).}$$

Solution. Since $\Phi(1, \alpha)$ is even, $b_n = 0$, and from (8) we obtain $a_0 = \frac{1}{2}$ and

$$a_n = \frac{1}{\pi} \left[-\int_{-\pi}^{0} \frac{\alpha}{\pi} \cos n\alpha\, d\alpha + \int_0^{\pi} \frac{\alpha}{\pi} \cos n\alpha\, d\alpha \right] = \frac{2}{n^2\pi^2} (\cos n\pi - 1).$$

Hence, $a_n = -4/(n^2\pi^2)$ if n is odd, $a_n = 0$ if $n = 2, 4, \cdots$, and the potential is

$$\Phi(r, \theta) = \frac{1}{2} - \frac{4}{\pi^2} \left[r \cos\theta + \frac{r^3}{3^2} \cos 3\theta + \frac{r^5}{5^2} \cos 5\theta + \cdots \right].$$

Figure 421 shows the unit disk and some of the equipotential lines (curves $\Phi = const$). ∎

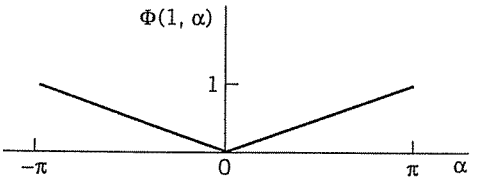


Fig. 420. Boundary values in Example 1

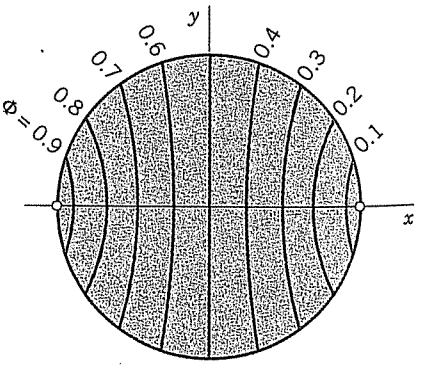


Fig. 421. Potential in Example 1

PROBLEM SET 18.5

1. Verify (3).
2. Show that every term in (7) is a harmonic function in the disk $r < R$.
3. Give the details of the derivation of the series (7) from the Poisson formula (5).

4–13 HARMONIC FUNCTIONS IN A DISK

Using (7), find the potential $\Phi(r, \theta)$ in the unit disk $r < 1$ having the given boundary values $\Phi(1, \theta)$. Using the sum of the first few terms of the series, compute some values of Φ and sketch a figure of the equipotential lines.

4. $\Phi(1, \theta) = \sin 2\theta$
5. $\Phi(1, \theta) = 2 \sin^2 \theta$
6. $\Phi(1, \theta) = \cos^2 5\theta$
7. $\Phi(1, \theta) = \theta$ if $-\pi < \theta < \pi$
8. $\Phi(1, \theta) = \theta$ if $0 < \theta < 2\pi$
9. $\Phi(1, \theta) = \sin^3 2\theta$
10. $\Phi(1, \theta) = \cos^4 \theta$
11. $\Phi(1, \theta) = \theta^2$ if $-\pi < \theta < \pi$
12. $\Phi(1, \theta) = 1$ if $-\frac{1}{2}\pi < \theta < \frac{1}{2}\pi$, $\Phi(1, \theta) = 0$ if $\frac{1}{2}\pi < \theta < \frac{3}{2}\pi$
13. $\Phi(1, \theta) = \theta$ if $-\frac{1}{2}\pi < \theta < \frac{1}{2}\pi$, $\Phi(1, \theta) = \pi - \theta$ if $\frac{1}{2}\pi < \theta < \frac{3}{2}\pi$
14. **TEAM PROJECT. Potential in a Disk. (a) Mean value property.** Show that the value of a harmonic function Φ at the center of a circle C equals the mean of the value of Φ on C (see Sec. 18.4, footnote 1, for definitions of mean values).

    (b) Separation of variables. Show that the terms of (7) appear as solutions in separating the Laplace equation in polar coordinates.

    (c) Harmonic conjugate. Find a series for a harmonic conjugate Ψ of Φ from (7).

    (d) Power series. Find a series for $F(z) = \Phi + i\Psi$.
15. **CAS EXPERIMENT. Series (7).** Write a program for series developments (7). Experiment on accuracy by computing values from partial sums and comparing them with values that you obtain from your CAS graph. Do this **(a)** for Example 1 and Fig. 421, **(b)** for Φ in Prob. 8 (which is discontinuous on the boundary!), **(c)** for a Φ of your choice with continuous boundary values, **(d)** for Φ with discontinuous boundary values.

18.6 General Properties of Harmonic Functions

General properties of harmonic functions can often be obtained from properties of analytic functions in a simple fashion. Specifically, important mean value properties of harmonic functions follow readily from those of analytic functions. The details are as follows.

THEOREM 1

Mean Value Property of Analytic Functions

Let $f(z)$ be analytic in a simply connected domain D. Then the value of $F(z)$ at a point z_0 in D is equal to the mean value of $F(z)$ on any circle in D with center at z_0.

PROOF In Cauchy's integral formula (Sec. 14.3)

$$F(z_0) = \frac{1}{2\pi i} \oint_C \frac{F(z)}{z - z_0}\, dz \tag{1}$$

we choose for C the circle $z = z_0 + re^{i\alpha}$ in D. Then $z - z_0 = re^{i\alpha}$, $dz = ire^{i\alpha}\, d\alpha$, and (1) becomes

$$F(z_0) = \frac{1}{2\pi} \int_0^{2\pi} F(z_0 + re^{i\alpha})\, d\alpha. \tag{2}$$

The right side is the mean value of F on the circle (= value of the integral divided by the length 2π of the interval of integration). This proves the theorem. ∎

For harmonic functions, Theorem 1 implies

THEOREM 2

Two Mean Value Properties of Harmonic Functions

Let $\Phi(x, y)$ be harmonic in a simply connected domain D. Then the value of $\Phi(x, y)$ at a point (x_0, y_0) in D is equal to the mean value of $\Phi(x, y)$ on any circle in D with center at (x_0, y_0). This value is also equal to the mean value of $\Phi(x, y)$ on any circular disk in D with center (x_0, y_0). [See footnote 1 in Sec. 18.4.]

PROOF The first part of the theorem follows from (2) by taking the real parts on both sides,

$$\Phi(x_0, y_0) = \operatorname{Re} F(x_0 + iy_0) = \frac{1}{2\pi} \int_0^{2\pi} \Phi(x_0 + r\cos\alpha, y_0 + r\sin\alpha)\, d\alpha.$$

The second part of the theorem follows by integrating this formula over r from 0 to r_0 (the radius of the disk) and dividing by $r_0^2/2$,

$$\Phi(x_0, y_0) = \frac{1}{\pi r_0^2} \int_0^{r_0} \int_0^{2\pi} \Phi(x_0 + r\cos\alpha, y_0 + r\sin\alpha) r\, d\alpha\, dr. \tag{3}$$

The right side is the indicated mean value (integral divided by the area of the region of integration). ∎

Returning to analytic functions, we state and prove another famous consequence of Cauchy's integral formula. The proof is indirect and shows quite a nice idea of applying the *ML*-inequality. (A *bounded region* is a region that lies entirely in some circle about the origin.)

THEOREM 3

Maximum Modulus Theorem for Analytic Functions

Let $F(z)$ be analytic and nonconstant in a domain containing a bounded region R and its boundary. Then the absolute value $|F(z)|$ cannot have a maximum at an interior point of R. Consequently, the maximum of $|F(z)|$ is taken on the boundary of R. If $F(z) \neq 0$ in R, the same is true with respect to the minimum of $|F(z)|$.

PROOF We assume that $|F(z)|$ has a maximum at an interior point z_0 of R and show that this leads to a contradiction. Let $|F(z_0)| = M$ be this maximum. Since $F(z)$ is not constant, $|F(z)|$ is not constant, as follows from Example 3 in Sec. 13.4. Consequently, we can find a circle C of radius r with center at z_0 such that the interior of C is in R and $|F(z)|$ is smaller than M at some point P of C. Since $|F(z)|$ is continuous, it will be smaller than M on an arc C_1 of C that contains P (see Fig. 422), say,

$$|F(z)| \leqq M - k \quad (k > 0) \qquad \text{for all } z \text{ on } C_1.$$

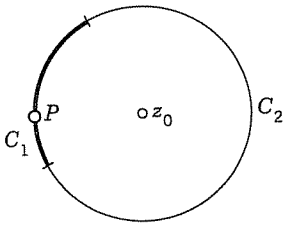


Fig. 422. Proof of Theorem 3

Let C_1 have the length L_1. Then the complementary arc C_2 of C has the length $2\pi r - L_1$. We now apply the *ML*-inequality (Sec. 14.1) to (1) and note that $|z - z_0| = r$. We then obtain (using straightforward calculation in the second line of the formula)

$$M = |F(z_0)| \leqq \frac{1}{2\pi}\left|\int_{C_1} \frac{F(z)}{z - z_0}\,dz\right| + \frac{1}{2\pi}\left|\int_{C_2} \frac{F(z)}{z - z_0}\,dz\right|$$

$$\leqq \frac{1}{2\pi}\left(\frac{M - k}{r}\right) L_1 + \frac{1}{2\pi}\left(\frac{M}{r}\right)(2\pi r - L_1) = M - \frac{kL_1}{2\pi r} < M$$

that is, $M < M$, which is impossible. Hence our assumption is false and the first statement is proved.

Next we prove the second statement. If $F(z) \neq 0$ in R, then $1/F(z)$ is analytic in R. From the statement already proved it follows that the maximum of $1/|F(z)|$ lies on the boundary of R. But this maximum corresponds to the minimum of $|F(z)|$. This completes the proof. ∎

This theorem has several fundamental consequences for harmonic functions, as follows.

THEOREM 4

Harmonic Functions

Let $\Phi(x, y)$ be harmonic in a domain containing a simply connected bounded region R and its boundary curve C. Then:

(I) (Maximum principle) *If $\Phi(x, y)$ is not constant, it has neither a maximum nor a minimum in R. Consequently, the maximum and the minimum are taken on the boundary of R.*

(II) *If $\Phi(x, y)$ is constant on C, then $\Phi(x, y)$ is a constant.*

(III) *If $h(x, y)$ is harmonic in R and on C and if $h(x, y) = \Phi(x, y)$ on C, then $h(x, y) = \Phi(x, y)$ everywhere in R.*

PROOF (I) Let $\Psi(x, y)$ be a conjugate harmonic function of $\Phi(x, y)$ in R. Then the complex function $F(z) = \Phi(x, y) + i\Psi(x, y)$ is analytic in R, and so is $G(z) = e^{F(z)}$. Its absolute value is

$$|G(z)| = e^{\operatorname{Re} F(z)} = e^{\Phi(x, y)}.$$

From Theorem 3 it follows that $|G(z)|$ cannot have a maximum at an interior point of R. Since e^{Φ} is a monotone increasing function of the real variable Φ, the statement about the

maximum of Φ follows. From this, the statement about the minimum follows by replacing Φ by $-\Phi$.

(II) By (I) the function $\Phi(x, y)$ takes its maximum and its minimum on C. Thus, if $\Phi(x, y)$ is constant on C, its minimum must equal its maximum, so that $\Phi(x, y)$ must be a constant.

(III) If h and Φ are harmonic in R and on C, then $h - \Phi$ is also harmonic in R and on C, and by assumption, $h - \Phi = 0$ everywhere on C. By (II) we thus have $h - \Phi = 0$ everywhere in R, and (III) is proved. ∎

The last statement of Theorem 4 is very important. It means that a *harmonic function is uniquely determined in R by its values on the boundary of R.* Usually, $\Phi(x, y)$ is required to be harmonic in R and continuous on the boundary of R, that is,

$$\lim_{\substack{x \to x_0 \\ y \to y_0}} \Phi(x, y) = \Phi(x_0, y_0), \text{ where } (x_0, y_0) \text{ is on the boundary and } (x, y) \text{ is in } R.$$

Under these assumptions the maximum principle (I) is still applicable. The problem of determining $\Phi(x, y)$ when the boundary values are given is called the **Dirichlet problem** for the Laplace equation in two variables, as we know. From (III) we thus have, as a highlight of our discussion,

THEOREM 5

Uniqueness Theorem for the Dirichlet Problem

If for a given region and given boundary values the Dirichlet problem for the Laplace equation in two variables has a solution, the solution is unique.

PROBLEM SET 18.6

1. Integrate $|z|^2$ around the unit circle. Does your result contradict Theorem 1?

2–4 **VERIFY THEOREM 1** for the given $F(z)$, z_0, and circle of radius 1.

2. $(z + 1)^3$, $z_0 = 2$
3. $(z - 2)^2$, $z_0 = \frac{1}{3}$
4. $10z^4$, $z_0 = 0$

5–7 **VERIFY THEOREM 2** for the given $\Phi(x, y)$, (x_0, y_0) and circle of radius 1.

5. $(x - 2)(y - 2)$, $(4, -4)$
6. $x^2 - y^2$, $(3, 8)$
7. $x^3 - 3xy^2$, $(1, 1)$

8. Derive Theorem 2 from Poisson's integral formula.
9. **CAS EXPERIMENT. Graphing Potentials.** Graph the potentials in Probs. 5 and 7 and for three other functions of your choice as surfaces over a rectangle or a disk in the xy-plane. Find the locations of maxima and minima by inspecting these graphs.
10. **TEAM PROJECT. Maximum Modulus of Analytic Functions. (a)** Verify Theorem 3 for (i) $F(z) = z^2$ and the square $4 \leqq x \leqq 6$, $2 \leqq y \leqq 4$, (ii) $F(z) = e^{3z}$ and any bounded domain, (iii) $F(z) = \sin z$ and the unit disk.

(b) $F(x) = \cos x$ (x real) has a maximum 1 at 0. How does it follow that this cannot be a maximum of $|F(z)| = |\cos z|$ in a domain containing $z = 0$?

(c) $F(z) = 1 + |z|^2$ is not zero in the disk $|z| \leqq 4$ and has a minimum at an interior point. Does this contradict Theorem 3?

(d) If $F(z)$ is analytic and not constant in the closed unit disk D: $|z| \leqq 1$ and $|F(z)| = c = const$ on the unit circle, show that $F(z)$ must have a zero in D. Can you extend this to an arbitrary simple closed curve?

11–13 MAXIMUM MODULUS

Find the location and size of the maximum of $|F(z)|$ in the unit disk $|z| \leqq 1$.

11. $F(z) = z^2 - 1$
12. $F(z) = az + b$ (a, b complex)
13. $F(z) = \cos 2z$
14. Verify the maximum principle for $\Phi(x, y) = e^x \cos y$ and the rectangle $a \leqq x \leqq b$, $0 \leqq y \leqq 2\pi$.
15. **(Conjugate)** Do Φ and a harmonic conjugate Ψ of Φ in a region R have their maximum at the same point of R?
16. **(Conformal mapping)** Find the location (u_1, v_1) of the maximum of $\Phi^* = e^u \cos v$ in R^*: $|w| \leqq 1$, $v \geqq 0$, where $w = u + iv$. Find the region R that is mapped onto R^* by $w = f(z) = z^2$. Find the potential in R resulting from Φ^* and the location (x_1, y_1) of the maximum. Is (u_1, v_1) the image of (x_1, y_1)? If so, is this just by chance?

CHAPTER 18 REVIEW QUESTIONS AND PROBLEMS

1. Why can potential problems be modeled and solved by complex analysis? For what dimensions?
2. What is a harmonic function? A harmonic conjugate?
3. Give a few examples of potential problems considered in this chapter.
4. What is a complex potential? What does it give physically?
5. How can conformal mapping be used in connection with the Dirichlet problem?
6. What heat problems reduce to potential problems? Give a few examples.
7. Write a short essay on potential theory in fluid flow from memory.
8. What is a mixed boundary value problem? Where did it occur?
9. State Poisson's formula and its derivation from Cauchy's formula.
10. State the maximum modulus theorem and mean value theorems for harmonic functions.
11. Find the potential and complex potential between the plates $y = x$ and $y = x + 10$ kept at 10 V and 110 V, respectively.
12. Find the potential between the cylinders $|z| = 1$ cm having potential 0 and $|z| = 10$ cm having potential 20 kV.
13. Find the complex potential in Prob. 12.
14. Find the equipotential line $U = 0$ V between the cylinders $|z| = 0.25$ cm and $|z| = 4$ cm kept at -220 V and 220 V, respectively. (Guess first.)
15. Find the potential between the cylinders $|z| = 10$ cm and $|z| = 100$ cm kept at the potentials 10 kV and 0, respectively.
16. Find the potential in the angular region between the plates $\operatorname{Arg} z = \pi/6$, kept at 8 kV, and $\operatorname{Arg} z = \pi/3$, kept at 6 kV.
17. Find the equipotential lines of $F(z) = i \operatorname{Ln} z$.
18. Find and sketch the equipotential lines of $F(z) = (1 + i)/z$.
19. What is the complex potential in the upper half-plane if the negative half of the x-axis has potential 1 kV and the positive half is grounded?
20. Find the potential on the ray $y = x$, $x > 0$, and on the positive half of the x-axis if the positive half of the y-axis is at 1200 V and the negative half is grounded.
21. Interpret Prob. 20 as a problem in heat conduction.
22. Find the temperature in the upper half-plane if the portion $x > 2$ of the x-axis is kept at 50°C and the other portion at 0°C.
23. Show that the isotherms of $F(z) = -iz^2 + z$ are hyperbolas.
24. If the region between two concentric cylinders of radii 2 cm and 10 cm contains water and the outer cylinder is kept at 20°C, to what temperature must we heat the inner cylinder in order to have 30°C at distance 5 cm from the axis?
25. What are the streamlines of $F(z) = i/z$?
26. What is the complex potential of a flow around a cylinder of radius 4 without circulation?
27. Find the complex potential of a source at $z = 5$. What are the streamlines?
28. Find the temperature in the unit disk $|z| \leqq 1$ in the form of an infinite series if the left semicircle of $|z| = 1$ has the temperature of 50°C and the right semicircle has the temperature 0°C.
29. Same task as in Prob. 28 if the upper semicircle is at 40°C and the lower at 0°C.
30. Find a series for the potential in the unit disk with boundary values $\Phi(1, \theta) = \theta^2$ $(-\pi < \theta < \pi)$.

SUMMARY OF CHAPTER 18

Complex Analysis and Potential Theory

Potential theory is the theory of solutions of **Laplace's equation**

$$\nabla^2\Phi = 0. \tag{1}$$

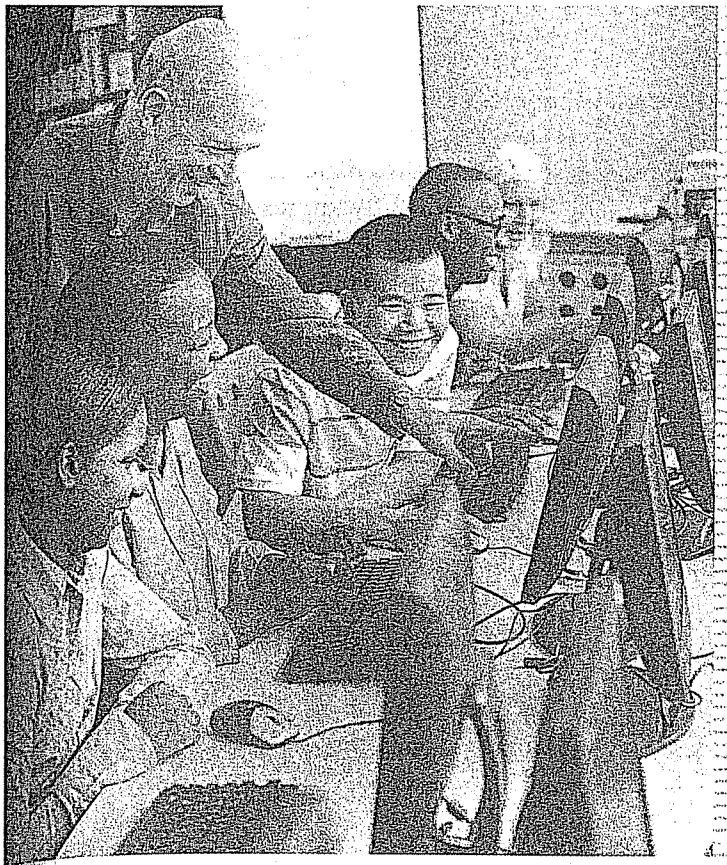
Solutions whose second partial derivatives are ***continuous*** are called **harmonic functions.** Equation (1) is the most important PDE in physics, where it is of interest in two and three dimensions. It appears in electrostatics (Sec. 18.1), steady-state heat problems (Sec. 18.3), fluid flow (Sec. 18.4), gravity, etc. Whereas the three-dimensional case requires other methods (see Chap. 12), two-dimensional potential theory can be handled by complex analysis, since the real and imaginary parts of an analytic function are harmonic (Sec. 13.4). They remain harmonic under conformal mapping (Sec. 18.2), so that **conformal mapping** becomes a powerful tool in solving boundary value problems for (1), as is illustrated in this chapter. With a real potential Φ in (1) we can associate a **complex potential**

$$F(z) = \Phi + i\Psi \qquad \text{(Sec. 18.1).} \tag{2}$$

Then both families of curves $\Phi = const$ and $\Psi = const$ have a physical meaning. In electrostatics, they are equipotential lines and lines of electrical force (Sec. 18.1). In heat problems, they are isotherms (curves of constant temperature) and lines of heat flow (Sec. 18.3). In fluid flow, they are equipotential lines of the velocity potential and streamlines (Sec. 18.4).

For the disk, the solution of the Dirichlet problem is given by the **Poisson formula** (Sec. 18.5) or by a series that on the boundary circle becomes the Fourier series of the given boundary values (Sec. 18.5).

Harmonic functions, like analytic functions, have a number of general properties; particularly important are the **mean value property** and the **maximum modulus property** (Sec. 18.6), which implies the uniqueness of the solution of the Dirichlet problem (Theorem 5 in Sec. 18.6).



PART E

Numeric Analysis

Software (p. 778–779)

CHAPTER 19 **Numerics in General**

CHAPTER 20 **Numeric Linear Algebra**

CHAPTER 21 **Numerics for ODEs and PDEs**

Numeric analysis, more briefly also called **numerics,** concerns **numeric methods,** that is, methods for solving problems in terms of numbers or corresponding graphical representations. It also includes the investigation of the range of applicability and of the accuracy and stability of these methods.

Typical tasks for numerics are the evaluation of definite integrals, the solution of equations and linear systems, the solution of differential or integral equations for which there are no solution formulas, and the evaluation of experimental data for which we want to obtain, for example, an approximating polynomial.

Numeric methods then provide the transition from the mathematical model to an **algorithm,** which is a detailed stepwise recipe for solving a problem of the indicated kind to be programmed on your computer, using your CAS (computer algebra system) or other software, or on your programmable calculator.

In this and the next two chapters we explain and illustrate the most frequently used basic numeric methods in algorithmic form. Chapter 19 concerns numerics in general; Chap. 20 numeric linear algebra, in particular, methods for linear systems and matrix eigenvalue problems; and Chap. 21 numerics for ODEs and PDEs.

The algorithms are given in a form that seems best for showing how a method works. We suggest that you also make use of programs from public-domain or commercial software listed on pp. 778–779 or obtainable on the Internet.

Numerics has increased in importance to the engineer more than any other field of mathematics owing to the ongoing development of powerful software resulting from great research activity in numerics: new methods are invented, existing methods are improved and adapted, and old methods—impractical in precomputer times—are rediscovered. A main goal in these activities is the development of well-structured software. And in large-scale work—millions of equations or steps of iteration—even small algorithmic improvements may have a large effect on computing time, storage demand, accuracy, and stability.

On average this makes the algorithms used in practice more and more complicated. ***However, the more sophisticated modern software will become, the more important it will be to understand concepts and algorithms in a basic form that shows original motivating ideas of recent developments.***

To avoid misunderstandings: Various simple classical methods are still very useful in many routine situations and produce satisfactory results. In other words, not *everything* has become more sophisticated.

Software

See also http://www.wiley.com/college/kreyszig/

The following list will help you if you wish to find software. You may also obtain information on known and new software from magazines, such as *Byte Magazine* or *PC Magazine,* from articles published by the *American Mathematical Society* (see also their website at www.ams.org), the *Society for Industrial and Applied Mathematics* (SIAM, at www.siam.org), the *Association for Computing Machinery* (ACM, at www.acm.org), or the *Institute of Electrical and Electronics Engineers* (IEEE, at www.ieee.org). Consult also your library, Computer Science Department, or Mathematics Department.

Derive. Texas Instruments, Inc., Dallas, TX. Phone 1-800-842-2737 or (972) 917-8324, website at www.derive.com or www.education.ti.com.

EISPACK. See LAPACK.

GAMS (Guide to Available Mathematical Software). Website at http://gams.nist.gov. On-line cross-index of software development by NIST, with links to IMSL, NAG, and NETLIB.

IMSL (International Mathematical and Statistical Library). Visual Numerics, Inc., Houston, TX. Phone 1-800-222-4675 or (713) 784-3131, website at www.vni.com. Mathematical and statistical Fortran routines with graphics.

LAPACK. Fortran 77 routines for linear algebra. This software package supersedes LINPACK and EISPACK. You can download the routines
(see http://cm.bell-labs.com/netlib/bib/mirrors.html) or order them directly from NAG. The LAPACK User's Guide is available at www.netlib.org.

LINPACK see LAPACK

Maple. Waterloo Maple, Inc., Waterloo, ON, Canada. Phone 1-800-267-6583 or (519) 747-2373, website at www.maplesoft.com.

Maple Computer Guide. For Advanced Engineering Mathematics, 9th edition. By E. Kreyszig and E. J. Norminton. J. Wiley and Sons, Inc., Hoboken, NJ. Phone 1-800-225-5945 or (201) 748-6000.

Mathcad. MathSoft, Inc., Cambridge, MA. Phone 1-800-628-4223 or (617) 444-8000, website at www.mathcad.com or www.mathsoft.com.

Mathematica. Wolfram Research, Inc., Champaign, IL. Phone 1-800-965-3726 or (217) 398-0700, website at www.wolframresearch.com.

Mathematica Computer Guide. For Advanced Engineering Mathematics, 9th edition. By E. Kreyszig and E. J. Norminton. J. Wiley and Sons, Inc., Hoboken, NJ. Phone 1-800-225-5945 or (201) 748-6000.

Matlab. The MathWorks, Inc., Natick, MA. Phone (508) 647-7000, website at www.mathworks.com.

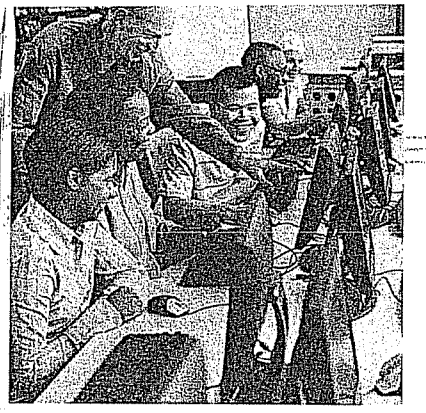
NAG. Numerical Algorithms Group, Inc., Downders Grove, IL. Phone (630) 971-2337, website at www.nag.com. Numeric routines in Fortran 77, Fortran 90, and C.

NETLIB. Extensive library of public-domain software. See at www.netlib.org and http://cm.bell-labs.com/netlib/.

NIST. National Institute of Standards and Technology, Gaithersburg, MD. Phone (301) 975-2000, website at www.nist.gov. For Mathematical and Computational Science Division phone (301) 975-3800. See also http://math.nist.gov.

Numerical Recipes. Cambridge University Press, New York, NY. Phone (212) 924-3900, website at www.us.cambridge.org. Books (also source codes on CD ROM and discettes) containing numeric routines in *C*, C^{++}, Fortran 77, and Fortran 90. To order, call office at West Nyack, NY, at 1-800-872-7423 or (845) 353-7500 or online at www.numerical-recipes.com.

FURTHER SOFTWARE IN STATISTICS. See Part G.



CHAPTER 19

Numerics in General

This first chapter on numerics begins with an explanation of some general concepts, such as floating point, roundoff errors, and general numeric errors and their propagation. In Sec. 19.2 we discuss methods for solving equations. Interpolation methods, including splines, follow in Secs. 19.3 and 19.4. The last section (19.5) concerns numeric integration and differentiation.

The purpose of this chapter is twofold. First, for all these tasks the student should become familiar with the most basic (but not too complicated) numeric solution methods. These are indispensable for the engineer, because for many problems there is no solution formula (think of a complicated integral or a polynomial of high degree or the interpolation of values obtained by measurements). In other cases a complicated solution formula may exist but may be practically useless.

Second, the student should learn to understand some basic ideas and concepts that are important throughout numerics, such as the practical form of algorithms, the estimation of errors, and the order of convergence.

Prerequisite: Elementary calculus
References and Answers to Problems: App. 1 Part E, App. 2

19.1 Introduction

Numeric methods are used to solve problems on computers or calculators by numeric calculations, resulting in a table of numbers and/or graphical representations (figures). The steps from a given situation (in engineering, economics, etc.) to the final answer are usually as follows.

1. **Modeling.** We set up a mathematical model of our problem, such as an integral, a system of equations, or a differential equation.
2. **Choosing a numeric method** and parameters (e.g., step size), perhaps with a preliminary error estimation.
3. **Programming.** We use the algorithm to write a corresponding program in a CAS, such as Maple, Mathematica, Matlab, or Mathcad, or, say, in Fortran, *C*, or *C*++, selecting suitable routines from a software system as needed.
4. **Doing the computation.**
5. **Interpreting the results** in physical or other terms, also deciding to rerun if further results are needed.

Steps 1 and 2 are related. A slight change of the model may often admit of a more efficient method. To choose methods, we must first get to know them. Chapters 19–21 contain efficient algorithms for the most important classes of problems occurring frequently in practice.

In Step 3 the program consists of the given data and a sequence of instructions to be executed by the computer in a certain order for producing the answer in numeric or graphic form.

To create a good understanding of the nature of numeric work, we continue in this section with some simple general remarks.

Floating-Point Form of Numbers

We know that in decimal notation, every real number is represented by a finite or an infinite sequence of decimal digits. Now most computers have two ways of representing numbers, called *fixed point* and *floating point.* In a **fixed-point** system all numbers are given with a fixed number of decimals after the decimal point; for example, numbers given with 3 decimals are 62.358, 0.014, 1.000. In a text we would write, say, 3 decimals as 3D. Fixed-point representations are impractical in most scientific computations because of their limited range (explain!) and will not concern us.

In a **floating-point** system we write, for instance,

$$0.6247 \cdot 10^{3}, \qquad 0.1735 \cdot 10^{-13}, \qquad -0.2000 \cdot 10^{-1}$$

or sometimes also

$$6.247 \cdot 10^{2}, \qquad 1.735 \cdot 10^{-14}, \qquad -2.000 \cdot 10^{-2}.$$

We see that in this system the number of significant digits is kept fixed, whereas the decimal point is "floating." Here, a **significant digit** of a number c is any given digit of c, except possibly for zeros to the left of the first nonzero digit; these zeros serve only to fix the position of the decimal point. (Thus any other zero is a significant digit of c.) For instance, each of the numbers

$$1360, \quad 1.360, \quad 0.001360$$

has 4 significant digits. In a text we indicate, say, 4 significant digits, by 4S.

The use of exponents permits us to represent very large and very small numbers. Indeed, theoretically any nonzero number a can be written as

$$(1) \qquad a = \pm m \cdot 10^{n}, \qquad 0.1 \leqq |m| < 1, \qquad n \text{ integer.}$$

On the computer, m is limited to k digits (e.g., $k = 8$) and n is limited, giving representations (for finitely many numbers only!)

$$(2) \qquad \bar{a} = \pm \overline{m} \cdot 10^{n}, \qquad \overline{m} = 0.d_1 d_2 \cdots d_k, \qquad d_1 > 0.$$

These numbers $\bar{a}$ are often called *k-digit decimal machine numbers.* Their fractional part m (or $\overline{m}$) is called the *mantissa.* This has nothing to do with "mantissa" as used for logarithms. n is called the *exponent* of $\bar{a}$.

Underflow and Overflow. The range of exponents that a typical computer can handle is very large. The IEEE (Institute of Electrical and Electronics Engineers) floating-point standard for **single precision** (the usual number of digits in calculations) is about $-38 < n < 38$ (about $-125 < n^* < 125$ for the exponent in **binary representations,** i.e., representations in base 2). [For so-called **double precision** it is about $-308 < n < 308$ (about $-1020 < n^* < 1020$ for binary).] If in a computation a number outside that range occurs, this is called **underflow** when the number is smaller and **overflow** when it is larger. In the case of underflow the result is usually set to zero and computation continues. Overflow causes the computer to halt. Standard codes (by IMSL, NAG, etc.) are written to avoid overflow. Error messages on overflow may then indicate programming errors (incorrect input data, etc.).

Roundoff

An error is caused by **chopping** (= discarding all decimals from some decimal on) or **rounding.** This error is called **roundoff error,** regardless of whether we chop or round. The rule for rounding off a number to k decimals is as follows. (The rule for rounding off to k significant digits is the same, with "decimal" replaced by "significant digit.")

Roundoff Rule. Discard the $(k + 1)$th and all subsequent decimals. (a) If the number thus discarded is less than half a unit in the kth place, leave the kth decimal unchanged (*"rounding down"*). (b) If it is greater than half a unit in the kth place, add one to the kth decimal (*"rounding up"*). (c) If it is exactly half a unit, round off to the nearest *even* decimal. (Example: Rounding off 3.45 and 3.55 to 1 decimal gives 3.4 and 3.6, respectively.)

The last part of the rule is supposed to ensure that in discarding exactly half a decimal, rounding up and rounding down happens about equally often, on the average.

If we round off 1.2535 to 3, 2, 1 decimals, we get 1.254, 1.25, 1.3, but if 1.25 is rounded off to one decimal, without further information, we get 1.2.

Chopping is not recommended because the corresponding error can be larger than that in rounding, and is systematic. (Nevertheless, some computers use it because it is simpler and faster. On the other hand, some computers and calculators improve accuracy of results by doing intermediate calculations using one or more extra digits, called *guarding digits.*)

Error in Rounding. Let $\bar{a} = fl(a)$ in (2) be the floating-point computer approximation of a in (1) obtained by rounding, where fl suggests **floating.** Then the roundoff rule gives (by dropping exponents) $|m - \bar{m}| \leqq \frac{1}{2} \cdot 10^{-k}$. Since $|m| \geqq 0.1$, this implies (when $a \neq 0$)

$$\left|\frac{a - \bar{a}}{a}\right| \approx \left|\frac{m - \bar{m}}{m}\right| \leqq \frac{1}{2} \cdot 10^{1-k}. \tag{3}$$

The right side $u = \frac{1}{2} \cdot 10^{1-k}$ is called the **rounding unit.** If we write $\bar{a} = a(1 + \delta)$, we have by algebra $(\bar{a} - a)/a = \delta$, hence $|\delta| \leqq u$ by (3). *This shows that the rounding unit u is an error bound in rounding.*

Rounding errors may ruin a computation completely, even a small computation. In general, these errors become the more dangerous the more arithmetic operations (perhaps several millions!) we have to perform. It is therefore important to analyze computational programs for expected rounding errors and to find an arrangement of the computations such that the effect of rounding errors is as small as possible.

The arithmetic in a computer is not exact either and causes further errors; however, these will not be relevant to our discussion.

Accuracy in Tables. Although available software has rendered various tables of function values superfluous, some tables (of higher functions, of coefficients of integration formulas, etc.) will still remain in occasional use. If a table shows k significant digits, it is conventionally assumed that any value $\tilde{a}$ in the table deviates from the exact value a by at most $\pm\frac{1}{2}$ unit of the kth digit.

Algorithm. Stability

Numeric methods can be formulated as algorithms. An **algorithm** is a step-by-step procedure that states a numeric method in a form (a **"pseudocode"**) understandable to humans. (Turn pages to see what algorithms look like.) The algorithm is then used to write a program in a programming language that the computer can understand so that it can execute the numeric method. Important algorithms follow in the next sections. For routine tasks your CAS or some other software system may contain programs that you can use or include as parts of larger programs of your own.

Stability. To be useful, an algorithm should be **stable;** that is, small changes in the initial data should cause only small changes in the final results. However, if small changes in the initial data can produce large changes in the final results, we call the algorithm **unstable.**

This *"numeric instability,"* which in most cases can be avoided by choosing a better algorithm, must be distinguished from *"mathematical instability"* of a problem, which is called *"ill-conditioning,"* a concept we discuss in the next section.

Some algorithms are stable only for certain initial data, so that one must be careful in such a case.

Errors of Numeric Results

Final results of computations of unknown quantities generally are **approximations;** that is, they are not exact but involve errors. Such an error may result from a combination of the following effects. **Roundoff errors** result from rounding, as discussed on p. 782. **Experimental errors** are errors of given data (probably arising from measurements). **Truncating errors** result from truncating (prematurely breaking off), for instance, if we replace a Taylor series with the sum of its first few terms. These errors depend on the computational method used and must be dealt with individually for each method. ["Truncating" is sometimes used as a term for chopping off (see before), a terminology that is not recommended.]

Formulas for Errors. If $\tilde{a}$ is an approximate value of a quantity whose exact value is a, we call the difference

(4) $$\epsilon = a - \tilde{a}$$

the **error** of $\tilde{a}$. Hence

(4*) $$a = \tilde{a} + \epsilon, \qquad \text{True value} = \text{Approximation} + \text{Error}.$$

For instance, if $\tilde{a} = 10.5$ is an approximation of $a = 10.2$, its error is $\epsilon = -0.3$. The error of an approximation $\tilde{a} = 1.60$ of $a = 1.82$ is $\epsilon = 0.22$.

CAUTION! In the literature $|a - \tilde{a}|$ ("absolute error") or $\tilde{a} - a$ are sometimes also used as definitions of error.

The **relative error** ϵ_r of $\tilde{a}$ is defined by

$$\epsilon_r = \frac{\epsilon}{a} = \frac{a - \tilde{a}}{a} = \frac{\text{Error}}{\text{True value}} \qquad (a \neq 0). \tag{5}$$

This looks useless because a is unknown. But if $|\epsilon|$ is much less than $|\tilde{a}|$, then we can use $\tilde{a}$ instead of a and get

$$\epsilon_r \approx \frac{\epsilon}{\tilde{a}}. \tag{5'}$$

This still looks problematic because ϵ is unknown—if it were known, we could get $a = \tilde{a} + \epsilon$ from (4) and we would be done. But what one often can obtain in practice is an **error bound** for $\tilde{a}$, that is, a number β such that

$$|\epsilon| \leqq \beta, \qquad \text{hence} \qquad |a - \tilde{a}| \leqq \beta.$$

This tells us how far away from our computed $\tilde{a}$ the unknown a can at most lie. Similarly, for the relative error, an error bound is a number β_r such that

$$|\epsilon_r| \leqq \beta_r, \qquad \text{hence} \qquad \left|\frac{a - \tilde{a}}{a}\right| \leqq \beta_r.$$

Error Propagation

This is an important matter. It refers to how errors at the beginning and in later steps (roundoff, for example) propagate into the computation and affect accuracy, sometimes very drastically. We state here what happens to error bounds. Namely, bounds for the *error* add under addition and subtraction, whereas bounds for the *relative error* add under multiplication and division. You do well to keep this in mind.

THEOREM 1

Error Propagation

(a) *In addition and subtraction, an error bound for the results is given by the sum of the error bounds for the terms.*

(b) *In multiplication and division, an error bound for the relative error of the results is given (approximately) by the sum of the bounds for the relative errors of the given numbers.*

PROOF **(a)** We use the notations $x = \tilde{x} + \epsilon_1$, $y = \tilde{y} + \epsilon_2$, $|\epsilon_1| \leqq \beta_1$, $|\epsilon_2| \leqq \beta_2$. Then for the error ϵ of the *difference* we obtain

$$\begin{aligned} |\epsilon| &= |x - y - (\tilde{x} - \tilde{y})| \\ &= |x - \tilde{x} - (y - \tilde{y})| \\ &= |\epsilon_1 - \epsilon_2| \leqq |\epsilon_1| + |\epsilon_2| \leqq \beta_1 + \beta_2. \end{aligned}$$

The proof for the *sum* is similar and is left to the student.

(b) For the relative error ϵ_r of $\tilde{x}\tilde{y}$ we get from the relative errors ϵ_{r1} and ϵ_{r2} of $\tilde{x}$, $\tilde{y}$ and bounds β_{r1}, β_{r2}

$$|\epsilon_r| = \left|\frac{xy - \tilde{x}\tilde{y}}{xy}\right| = \left|\frac{xy - (x - \epsilon_1)(y - \epsilon_2)}{xy}\right| = \left|\frac{\epsilon_1 y + \epsilon_2 x - \epsilon_1\epsilon_2}{xy}\right|$$

$$\approx \left|\frac{\epsilon_1 y + \epsilon_2 x}{xy}\right| \leqq \left|\frac{\epsilon_1}{x}\right| + \left|\frac{\epsilon_2}{y}\right| = |\epsilon_{r1}| + |\epsilon_{r2}| \leqq \beta_{r1} + \beta_{r2}.$$

This proof shows what "approximately" means: we neglected $\epsilon_1\epsilon_2$ as small in absolute value compared to $|\epsilon_1|$ and $|\epsilon_2|$. The proof for the quotient is similar but slightly more tricky (see Prob. 15). ∎

Basic Error Principle

Every numeric method should be accompanied by an error estimate. If such a formula is lacking, is extremely complicated, or is impractical because it involves information (for instance, on derivatives) that is not available, the following may help.

Error Estimation by Comparison. *Do a calculation twice with different accuracy. Regard the difference $\tilde{a}_2 - \tilde{a}_1$ of the results $\tilde{a}_1$, $\tilde{a}_2$ as a (perhaps crude) estimate of the error ϵ_1 of the inferior result $\tilde{a}_1$.* Indeed, $\tilde{a}_1 + \epsilon_1 = \tilde{a}_2 + \epsilon_2$ by formula (4*). This implies $\tilde{a}_2 - \tilde{a}_1 = \epsilon_1 - \epsilon_2 \approx \epsilon_1$ because $\tilde{a}_2$ is generally more accurate than $\tilde{a}_1$, so that $|\epsilon_2|$ is small compared to $|\epsilon_1|$.

Loss of Significant Digits

This means that a result of a calculation has fewer correct digits than the numbers from which it was obtained. This happens if we subtract two numbers of about the same size, for example, $0.1439 - 0.1426$ ("subtractive cancellation"). It may occur in simple problems, but it can be avoided in most cases by simple changes of the algorithm—if one is aware of it! Let us illustrate this with the following basic problem.

EXAMPLE 1 **Quadratic Equation. Loss of Significant Digits**

Find the roots of the equation

$$x^2 - 40x + 2 = 0,$$

using 4 significant digits (abbreviated 4S) in the computation.

Solution. A formula for the roots x_1, x_2 of a quadratic equation $ax^2 + bx + c = 0$ is

$$(6) \qquad x_1 = \frac{1}{2a}(-b + \sqrt{b^2 - 4ac}), \qquad x_2 = \frac{1}{2a}(-b - \sqrt{b^2 - 4ac}).$$

Furthermore, since $x_1x_2 = c/a$, another formula for those roots is

$$(7) \qquad x_1 \text{ as before}, \qquad x_2 = \frac{c}{ax_1}.$$

From (6) we obtain $x = 20 \pm \sqrt{398} = 20.00 \pm 19.95$. This gives $x_1 = 20.00 + 19.95 = 39.95$, involving no difficulty, whereas $x_2 = 20.00 - 19.95 = 0.05$ is poor because it involves loss of significant digits.

In contrast, (7) gives $x_1 = 39.95$, $x_2 = 2.000/39.95 = 0.05006$, in error by less than one unit of the last digit, as a computation with more digits shows. (The 10S-value is 0.0500 626 5674.)

Comment. To avoid misunderstandings: 4S was used for convenience; (7) is better than (6) regardless of the number of digits used. For instance, the 8S-computation by (6) is $x_1 = 39.949\,937$, $x_2 = 0.050\,063$, which is poor, and by (7) it is x_1 as before, $x_2 = 2/x_1 = 0.050\,062\,657$.

In a quadratic equation with real roots, if x_2 is absolutely largest (because $b > 0$), use (6) for x_2 and then $x_1 = c/(ax_2)$. ■

PROBLEM SET 19.2

1. **(Floating point)** Write 98.17, −100.987, 0.0057869, −13600 in floating-point form, rounded to 4S (4 significant digits).
2. Write −0.0286403, 11.25845, −31681.55 in floating-point form rounded to 6S.
3. **Small differences of large numbers** may be particularly strongly affected by rounding errors. Illustrate this by computing 0.36443/(17.862 − 17.798) as given with 5S, then rounding stepwise to 4S, 3S, and 2S, where "stepwise" means: round the *rounded* numbers, not the given ones.
4. Do the work in Prob. 3 with numbers of your choice that give even more drastically different results. How can you avoid such difficulties?
5. The quotient in Prob. 3 is of the form $a/(b - c)$. Write it as $a(b + c)/(b^2 - c^2)$. Compute it first with 5S, then rounding numerator 12.996 and denominator 2.28 stepwise as in Prob. 3. Compare and comment.
6. **(Quadratic equation)** Solve $x^2 - 20x + 1 = 0$ by (6) and by (7), using 6S in the computation. Compare and comment.
7. Do the computations in Prob. 6 with 4S and 2S.
8. Solve $x^2 + 100x + 2 = 0$ by (6) and (7) with 5S and compare.
9. Calculate $1/e = 0.367879$ (6S) from the partial sums of 5 to 10 terms of the Maclaurin series (a) of e^{-x} with $x = 1$, (b) of e^x with $x = 1$ and then taking the reciprocal. Which is more accurate?
10. Addition with a fixed number of significant digits depends on the order in which you add the numbers. Illustrate this with an example. Find an empirical rule for the best order.
11. **Approximations of π** $= 3.141\,592\,653\,589\,79\cdots$ are 22/7 and 355/113. Determine the corresponding errors and relative errors to 3 significant digits.
12. Compute π by **Machin's approximation**
    $16 \arctan(1/5) - 4 \arctan(1/239)$ to 10S (which are correct). (In 1986, D. H. Bailey computed almost 30 million decimals of π on a CRAY-2 in less than 30 hours. The race for more and more decimals is continuing.)
13. **(Rounding and adding)** Let $a_1, \cdots, a_n$ be numbers with a_j correctly rounded to D_j decimals. In calculating the sum $a_1 + \cdots + a_n$, retaining $D = \min D_j$ decimals, is it essential that we first add and then round the result or that we first round each number to D decimals and then add?
14. **(Theorems on errors)** Prove Theorem 1(a) for addition.
15. Prove Theorem 1(b) for division.
16. Show that in Example 1 the absolute value of the error of $x_2 = 2.000/39.95 = 0.05006$ is less than 0.00001.
17. **Overflow and underflow** can sometimes be avoided by simple changes in a formula. Explain this in terms of $\sqrt{x^2 + y^2} = x\sqrt{1 + (y/x)^2}$ with $x^2 \geqq y^2$ and x so large that x^2 would cause overflow. Invent examples of your own.
18. **(Nested form)** Evaluate
    $$f(x) = x^3 - 7.5x^2 + 11.2x + 2.8$$
    $$= ((x - 7.5)x + 11.2)x + 2.8$$
    at $x = 3.94$ using 3S arithmetic and rounding, in both of the given forms. The latter, called the *nested form,* is usually preferable since it minimizes the number of operations and thus the effect of rounding.
19. **CAS EXPERIMENT. Chopping and Rounding.**
    (a) Let $x = 4/7$ and $y = 1/3$. Find the errors ϵ_{chop}, ϵ_{round} and the relative errors $\epsilon_{\text{r,ch}}$, $\epsilon_{\text{r,rd}}$ of $x + y$, $x - y$, xy, x/y in chopping and rounding to 5S. Experiment with other fractions of your choice.

    (b) Graph ϵ_{chop} and ϵ_{round} (for 5S) of $k/21$ as a function of $k = 1, 2, \cdots, 21$ on common axes. What average value can you read from the graph for ϵ_{chop}? For ϵ_{round}? Experiment with other integers that give similar graphs. Different types of graphs. Can you characterize the different types in terms of prime factors?

    (c) How does the situation in (b) change if you take 4S instead of 5S?

    (d) Write programs for the work in (a)–(c).
20. **WRITING PROJECT. Numerics.** In your own words write about the overall role of numeric methods in applied mathematics, why they are important, where and when they must be used or can be used, and how they are influenced by the use of the computer in engineering and other work.

19.2 Solution of Equations by Iteration

From here on, each section will be devoted to some basic kind of problem and corresponding solution methods. We begin with methods of finding solutions of a single equation

$$f(x) = 0 \tag{1}$$

where f is a given function. For this task there are practically no formulas (except in a few simple cases), so that one depends almost entirely on numeric algorithms. A **solution** of (1) is a number $x = s$ such that $f(s) = 0$. Here, s suggests "solution," but we shall also use other letters.

Examples are $x^3 + x = 1$, $\sin x = 0.5x$, $\tan x = x$, $\cosh x = \sec x$, $\cosh x \cos x = -1$, which can all be written in the form (1). The first concerns an **algebraic equation** because the corresponding f is a polynomial, and in this case the solutions are also called **roots** of the equation. The other equations are **transcendental equations** because they involve transcendental functions. Solving equations (1) is a task of prime importance because engineering applications abound: some occur in Chaps. 2, 4, 8 (characteristic equations), 6 (partial fractions), 12 (eigenvalues, zeros of Bessel functions), and 16 (integration), but there are many, many others.

To solve (1) when there is no formula for the exact solution, we can use an approximation method, in particular an **iteration method,** that is, a method in which we start from an initial guess x_0 (which may be poor) and compute step by step (in general better and better) approximations $x_1, x_2, \cdots$ of an unknown solution of (1). We discuss three such methods that are of particular practical importance and mention two others in the problem set. These methods and the underlying principles are basic for understanding the diverse methods in software packages.

In general, iteration methods are easy to program because the computational operations are the same in each step—just the data change from step to step—and, more important, if in a concrete case a method converges, it is stable (see Sec. 19.1) in general.

Fixed-Point Iteration for Solving Equations $f(x) = 0$

Our present use of the word "fixed point" has absolutely nothing to do with that in the last section.

In one way or another we transform (1) *algebraically* into the form

$$x = g(x). \tag{2}$$

Then we choose an x_0 and compute $x_1 = g(x_0)$, $x_2 = g(x_1)$, and in general

$$x_{n+1} = g(x_n) \qquad (n = 0, 1, \cdots). \tag{3}$$

A solution of (2) is called a **fixed point** of g, motivating the name of the method. This is a solution of (1), since from $x = g(x)$ we can return to the original form $f(x) = 0$. From (1) we may get several different forms of (2). The behavior of corresponding iterative sequences $x_0, x_1, \cdots$ may differ, in particular, with respect to their speed of convergence. Indeed, some of them may not converge at all. Let us illustrate these facts with a simple example.

EXAMPLE 1 **An Iteration Process (Fixed-Point Iteration)**

Set up an iteration process for the equation $f(x) = x^2 - 3x + 1 = 0$. Since we know the solutions

$$x = 1.5 \pm \sqrt{1.25}, \quad \text{thus} \quad 2.618\,034 \quad \text{and} \quad 0.381\,966,$$

we can watch the behavior of the error as the iteration proceeds.

Solution. The equation may be written

$$x = g_1(x) = \tfrac{1}{3}(x^2 + 1), \quad \text{thus} \quad x_{n+1} = \tfrac{1}{3}(x_n{}^2 + 1). \tag{4a}$$

If we choose $x_0 = 1$, we obtain the sequence (Fig. 423a; computed with 6S and then rounded)

$$x_0 = 1.000, \quad x_1 = 0.667, \quad x_2 = 0.481, \quad x_3 = 0.411, \quad x_4 = 0.390, \cdots$$

which seems to approach the smaller solution. If we choose $x_0 = 2$, the situation is similar. If we choose $x_0 = 3$, we obtain the sequence (Fig. 423a, upper part)

$$x_0 = 3.000, \quad x_1 = 3.333, \quad x_2 = 4.037, \quad x_3 = 5.766, \quad x_4 = 11.415, \cdots$$

which diverges.

Our equation may also be written (divide by x)

$$x = g_2(x) = 3 - \frac{1}{x}, \quad \text{thus} \quad x_{n+1} = 3 - \frac{1}{x_n}, \tag{4b}$$

and if we choose $x_0 = 1$, we obtain the sequence (Fig. 423b)

$$x_0 = 1.000, \quad x_1 = 2.000, \quad x_2 = 2.500, \quad x_3 = 2.600, \quad x_4 = 2.615, \cdots$$

which seems to approach the larger solution. Similarly, if we choose $x_0 = 3$, we obtain the sequence (Fig. 423*b*)

$$x_0 = 3.000, \quad x_1 = 2.667, \quad x_2 = 2.625, \quad x_3 = 2.619, \quad x_4 = 2.618, \cdots.$$

Our figures show the following. In the lower part of Fig. 423a the slope of $g_1(x)$ is less than the slope of $y = x$, which is 1, thus $|g_1'(x)| < 1$, and we seem to have convergence. In the upper part, $g_1(x)$ is steeper ($g_1'(x) > 1$) and we have divergence. In Fig. 423b the slope of $g_2(x)$ is less near the intersection point ($x = 2.618$, fixed point of g_2, solution of $f(x) = 0$), and both sequences seem to converge. From all this we conclude that convergence seems to depend on the fact that in a neighborhood of a solution the curve of $g(x)$ is less steep than the straight line $y = x$, and we shall now see that this condition $|g'(x)| < 1$ (= slope of $y = x$) is sufficient for convergence. ■

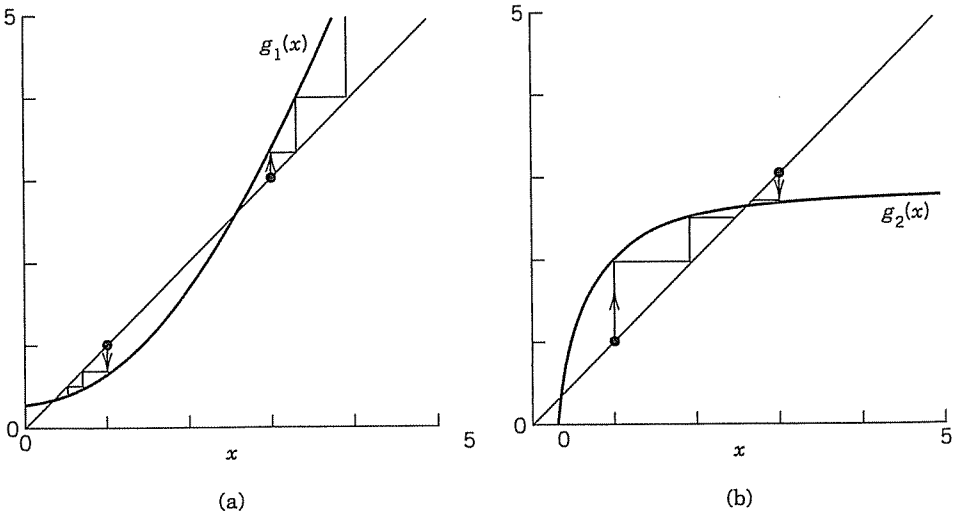


Fig. 423. Example 1, iterations (4a) and (4b)

An iteration process defined by (3) is called **convergent** for an x_0 if the corresponding sequence $x_0, x_1, \cdots$ is convergent.

A sufficient condition for convergence is given in the following theorem, which has various practical applications.

THEOREM 1

Convergence of Fixed-Point Iteration

Let $x = s$ be a solution of $x = g(x)$ and suppose that g has a continuous derivative in some interval J containing s. Then if $|g'(x)| \leqq K < 1$ in J, the iteration process defined by (3) converges for any x_0 in J, and the limit of the sequence $\{x_n\}$ is s.

PROOF By the mean value theorem of differential calculus there is a t between x and s such that

$$g(x) - g(s) = g'(t)\,(x - s) \qquad (x \text{ in } J).$$

Since $g(s) = s$ and $x_1 = g(x_0)$, $x_2 = g(x_1), \cdots$, we obtain from this and the condition on $|g'(x)|$ in the theorem

$$|x_n - s| = |g(x_{n-1}) - g(s)| = |g'(t)||x_{n-1} - s| \leqq K|x_{n-1} - s|.$$

Applying this inequality n times, for $n, n-1, \cdots, 1$ gives

$$|x_n - s| \leqq K|x_{n-1} - s| \leqq K^2|x_{n-2} - s| \leqq \cdots \leqq K^n|x_0 - s|.$$

Since $K < 1$, we have $K^n \to 0$; hence $|x_n - s| \to 0$ as $n \to \infty$. ∎

We mention that a function g satisfying the condition in Theorem 1 is called a **contraction** because $|g(x) - g(v)| \leqq K|x - v|$, where $K < 1$. Furthermore, K gives information on the speed of convergence. For instance, if $K = 0.5$, then the accuracy increases by at least 2 digits in only 7 steps because $0.5^7 < 0.01$.

EXAMPLE 2 **An Iteration Process. Illustration of Theorem 1**

Find a solution of $f(x) = x^3 + x - 1 = 0$ by iteration.

Solution. A sketch shows that a solution lies near $x = 1$. We may write the equation as $(x^2 + 1)x = 1$ or

$$x = g_1(x) = \frac{1}{1 + x^2}, \qquad \text{so that} \qquad x_{n+1} = \frac{1}{1 + x_n^2}. \qquad \text{Also} \qquad |g_1'(x)| = \frac{2|x|}{(1 + x^2)^2} < 1$$

for any x because $4x^2/(1 + x^2)^4 = 4x^2/(1 + 4x^2 + \cdots) < 1$, so that by Theorem 1 we have convergence for any x_0. Choosing $x_0 = 1$, we obtain (Fig. 424 on p. 790)

$$x_1 = 0.500, \quad x_2 = 0.800, \quad x_3 = 0.610, \quad x_4 = 0.729, \quad x_5 = 0.653, \quad x_6 = 0.701, \cdots.$$

The solution exact to 6D is $s = 0.682\,328$.

The given equation may also be written

$$x = g_2(x) = 1 - x^3. \qquad \text{Then} \qquad |g_2'(x)| = 3x^2$$

and this is greater than 1 near the solution, so that we cannot apply Theorem 1 and assert convergence. Try $x_0 = 1$, $x_0 = 0.5$, $x_0 = 2$ and see what happens.

The example shows that the transformation of a given $f(x) = 0$ into the form $x = g(x)$ with g satisfying $|g'(x)| \leqq K < 1$ may need some experimentation. ∎

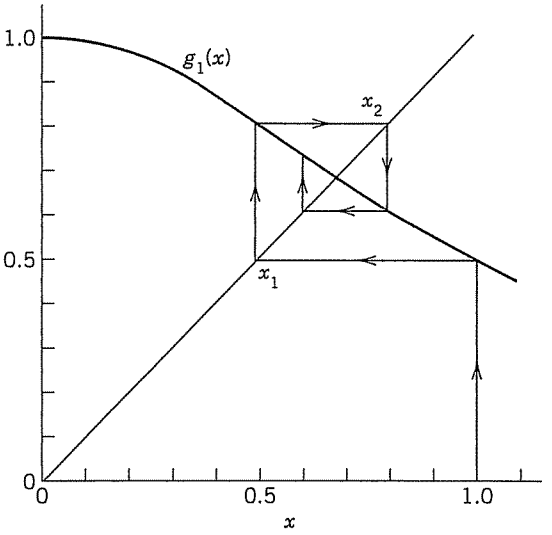


Fig. 424. Iteration in Example 2

Newton's Method for Solving Equations $f(x) = 0$

Newton's method, also known as **Newton–Raphson's method,**[1] is another iteration method for solving equations $f(x) = 0$, where f is assumed to have a continuous derivative f'. The method is commonly used because of its simplicity and great speed. The underlying idea is that we approximate the graph of f by suitable tangents. Using an approximate value x_0 obtained from the graph of f, we let x_1 be the point of intersection of the x-axis and the tangent to the curve of f at x_0 (see Fig. 425). Then

$$\tan \beta = f'(x_0) = \frac{f(x_0)}{x_0 - x_1}, \qquad \text{hence} \qquad x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

In the second step we compute $x_2 = x_1 - f(x_1)/f'(x_1)$, in the third step x_3 from x_2 again by the same formula, and so on. We thus have the algorithm shown in Table 19.1. Formula (5) in this algorithm can also be obtained if we algebraically solve Taylor's formula

$$(5^*) \qquad f(x_{n+1}) \approx f(x_n) + (x_{n+1} - x_n) f'(x_n) = 0.$$

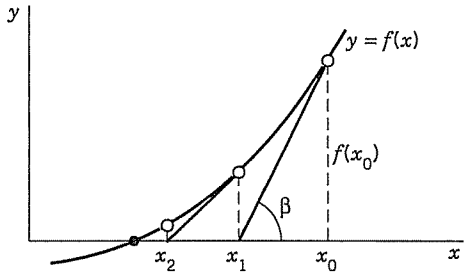


Fig. 425. Newton's method

[1]JOSEPH RAPHSON (1648–1715), English mathematician who published a method similar to Newton's method. For historical details, see Ref. [GR2], p. 203, listed in App. 1.

Table 19.1 Newton's Method for Solving Equations $f(x) = 0$

ALGORITHM NEWTON $(f, f', x_0, \epsilon, N)$

This algorithm computes a solution of $f(x) = 0$ given an initial approximation x_0 (starting value of the iteration). Here the function $f(x)$ is continuous and has a continuous derivative $f'(x)$.

INPUT: f, f', initial approximation x_0, tolerance $\epsilon > 0$, maximum number of iterations N.

OUTPUT: Approximate solution x_n $(n \leqq N)$ or message of failure.

For $n = 0, 1, 2, \cdots, N - 1$ do:

1 Compute $f'(x_n)$.

2 If $f'(x_n) = 0$ then OUTPUT "Failure". Stop.

[*Procedure completed unsuccessfully*]

3 Else compute

(5) $$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}.$$

4 If $|x_{n+1} - x_n| \leqq \epsilon |x_n|$ then OUTPUT x_{n+1}. Stop.

[*Procedure completed successfully*]

End

5 OUTPUT "Failure". Stop.

[*Procedure completed unsuccessfully after N iterations*]

End NEWTON

If it happens that $f'(x_n) = 0$ for some n (see line 2 of the algorithm), then try another starting value x_0. Line 3 is the heart of Newton's method.

The inequality in line 4 is a **termination criterion.** If the sequence of the x_n converges and the criterion holds, we have reached the desired accuracy and stop. In this line the factor $|x_n|$ is needed in the case of zeros of very small (or very large) absolute value because of the high density (or of the scarcity) of machine numbers for those x.

WARNING! The criterion by itself does not imply convergence. *Example.* The harmonic series diverges, although its partial sums $x_n = \sum_{k=1}^{n} 1/k$ satisfy the criterion because $\lim (x_{n+1} - x_n) = \lim (1/(n + 1)) = 0$.

Line 5 gives another termination criterion and is needed because Newton's method may diverge or, due to a poor choice of x_0, may not reach the desired accuracy by a reasonable number of iterations. Then we may try another x_0. If $f(x) = 0$ has more than one solution, different choices of x_0 may produce different solutions. Also, an iterative sequence may sometimes converge to a solution different from the expected one.

EXAMPLE 3 Square Root

Set up a Newton iteration for computing the square root x of a given positive number c and apply it to $c = 2$.

Solution. We have $x = \sqrt{c}$, hence $f(x) = x^2 - c = 0$, $f'(x) = 2x$, and (5) takes the form

$$x_{n+1} = x_n - \frac{x_n^2 - c}{2x_n} = \frac{1}{2}\left(x_n + \frac{c}{x_n}\right).$$

For $c = 2$, choosing $x_0 = 1$, we obtain

$$x_1 = 1.500\,000, \qquad x_2 = 1.416\,667, \qquad x_3 = 1.414\,216, \qquad x_4 = 1.414\,214, \cdots.$$

x_4 is exact to 6D. ▩

EXAMPLE 4 Iteration for a Transcendental Equation

Find the positive solution of $2 \sin x = x$.

Solution. Setting $f(x) = x - 2 \sin x$, we have $f'(x) = 1 - 2 \cos x$, and (5) gives

$$x_{n+1} = x_n - \frac{x_n - 2 \sin x_n}{1 - 2 \cos x_n} = \frac{2(\sin x_n - x_n \cos x_n)}{1 - 2 \cos x_n} = \frac{N_n}{D_n}.$$

From the graph of f we conclude that the solution is near $x_0 = 2$. We compute:

n	x_n	N_n	D_n	x_{n+1}
0	2.00000	3.48318	1.83229	1.90100
1	1.90100	3.12470	1.64847	1.89552
2	1.89552	3.10500	1.63809	1.89550
3	1.89550	3.10493	1.63806	1.89549

$x_4 = 1.89549$ is exact to 5D since the solution to 6D is 1.895 494. ▩

EXAMPLE 5 Newton's Method Applied to an Algebraic Equation

Apply Newton's method to the equation $f(x) = x^3 + x - 1 = 0$.

Solution. From (5) we have

$$x_{n+1} = x_n - \frac{x_n^3 + x_n - 1}{3x_n^2 + 1} = \frac{2x_n^3 + 1}{3x_n^2 + 1}.$$

Starting from $x_0 = 1$, we obtain

$$x_1 = 0.750\,000, \qquad x_2 = 0.686\,047, \qquad x_3 = 0.682\,340, \qquad x_4 = 0.682\,328, \cdots$$

where x_4 has the error $-1 \cdot 10^{-6}$. A comparison with Example 2 shows that the present convergence is much more rapid. This may motivate the concept of the *order of an iteration process,* to be discussed next. ▩

Order of an Iteration Method. Speed of Convergence

The quality of an iteration method may be characterized by the speed of convergence, as follows.

Let $x_{n+1} = g(x_n)$ define an iteration method, and let x_n approximate a solution s of $x = g(x)$. Then $x_n = s - \epsilon_n$, where ϵ_n is the error of x_n. Suppose that g is differentiable a number of times, so that the Taylor formula gives

$$\begin{aligned} x_{n+1} = g(x_n) &= g(s) + g'(s)(x_n - s) + \tfrac{1}{2}g''(s)(x_n - s)^2 + \cdots \\ &= g(s) - g'(s)\epsilon_n + \tfrac{1}{2}g''(s)\epsilon_n^2 + \cdots. \end{aligned} \tag{6}$$

The exponent of ϵ_n in the first nonvanishing term after $g(s)$ is called the **order** of the iteration process defined by g. The order measures the speed of convergence.

To see this, subtract $g(s) = s$ on both sides of (6). Then on the left you get $x_{n+1} - s = -\epsilon_{n+1}$, where ϵ_{n+1} is the error of x_{n+1}. And on the right the remaining expression equals approximately its first nonzero term because $|\epsilon_n|$ is small in the case of convergence. Thus

$$\text{(7)} \qquad \begin{aligned} &\text{(a)} \quad \epsilon_{n+1} \approx +g'(s)\epsilon_n && \text{in the case of first order,} \\ &\text{(b)} \quad \epsilon_{n+1} \approx -\tfrac{1}{2}g''(s)\epsilon_n^{\,2} && \text{in the case of second order,} \qquad \text{etc.} \end{aligned}$$

Thus if $\epsilon_n = 10^{-k}$ in some step, then for second order, $\epsilon_{n+1} = c \cdot (10^{-k})^2 = c \cdot 10^{-2k}$, so that the number of significant digits is about doubled in each step.

Convergence of Newton's Method

In Newton's method, $g(x) = x - f(x)/f'(x)$. By differentiation,

$$\text{(8)} \qquad g'(x) = 1 - \frac{f'(x)^2 - f(x)f''(x)}{f'(x)^2} = \frac{f(x)f''(x)}{f'(x)^2}.$$

Since $f(s) = 0$, this shows that also $g'(s) = 0$. Hence Newton's method is at least of second order. If we differentiate again and set $x = s$, we find that

$$\text{(8*)} \qquad g''(s) = \frac{f''(s)}{f'(s)}$$

which will not be zero in general. This proves

THEOREM 2

Second-Order Convergence of Newton's Method

If $f(x)$ is three times differentiable and f' and f'' are not zero at a solution s of $f(x) = 0$, then for x_0 sufficiently close to s, Newton's method is of second order.

Comments. For Newton's method, (7b) becomes, by (8*),

$$\text{(9)} \qquad \epsilon_{n+1} \approx -\frac{f''(s)}{2f'(s)}\,\epsilon_n^{\,2}.$$

For the rapid convergence of the method indicated in Theorem 2 it is important that s be a *simple* zero of $f(x)$ (thus $f'(s) \neq 0$) and that x_0 be close to s, because in Taylor's formula we took only the linear term [see (5*)], assuming the quadratic term to be negligibly small. (With a bad x_0 the method may even diverge!)

EXAMPLE 6 Prior Error Estimate of the Number of Newton Iteration Steps

Use $x_0 = 2$ and $x_1 = 1.901$ in Example 4 for estimating how many iteration steps we need to produce the solution to 5D accuracy. This is an **a priori estimate** or **prior estimate** because we can compute it after only one iteration, prior to further iterations.

Solution. We have $f(x) = x - 2 \sin x = 0$. Differentiation gives

$$\frac{f''(s)}{2f'(s)} \approx \frac{f''(x_1)}{2f'(x_1)} = \frac{2 \sin x_1}{2(1 - 2 \cos x_1)} \approx 0.57.$$

Hence (9) gives

$$|\epsilon_{n+1}| \approx 0.57\epsilon_n{}^2 \approx 0.57(0.57\epsilon_{n-1}^2)^2 = 0.57^3\epsilon_{n-1}^4 \approx \cdots \approx 0.57^M\epsilon_0^{M+1} \leqq 5 \cdot 10^{-6}$$

where $M = 2^n + 2^{n-1} + \cdots + 2 + 1 = 2^{n+1} - 1$. We show below that $\epsilon_0 \approx -0.11$. Consequently, our condition becomes

$$0.57^M 0.11^{M+1} \leqq 5 \cdot 10^{-6}.$$

Hence $n = 2$ is the smallest possible n, according to this crude estimate, in good agreement with Example 4.

$\epsilon_0 \approx -0.11$ is obtained from $\epsilon_1 - \epsilon_0 = (\epsilon_1 - s) - (\epsilon_0 - s) = -x_1 + x_0 \approx 0.10$, hence $\epsilon_1 = \epsilon_0 + 0.10 \approx -0.57\epsilon_0{}^2$ or $0.57\epsilon_0{}^2 + \epsilon_0 + 0.10 \approx 0$, which gives $\epsilon_0 \approx -0.11$. ▩

Difficulties in Newton's Method. Difficulties may arise if $|f'(x)|$ is very small near a solution s of $f(x) = 0$, for instance, if s is a zero of $f(x)$ of second (or higher) order (so that Newton's method converges only linearly, as an application of l'Hôpital's rule to (8) shows). Geometrically, small $|f'(x)|$ means that the tangent of $f(x)$ near s almost coincides with the x-axis (so that double precision may be needed to get $f(x)$ and $f'(x)$ accurately enough). Then for values $x = \tilde{s}$ far away from s we can still have small function values

$$R(\tilde{s}) = f(\tilde{s}).$$

In this case we call the equation $f(x) = 0$ **ill-conditioned.** $R(\tilde{s})$ is called the **residual** of $f(x) = 0$ at $\tilde{s}$. Thus a small residual guarantees a small error of $\tilde{s}$ only if the equation is ***not*** ill-conditioned.

EXAMPLE 7 An Ill-Conditioned Equation

$f(x) = x^5 + 10^{-4}x = 0$ is ill-conditioned. $x = 0$ is a solution. $f'(0) = 10^{-4}$ is small. At $\tilde{s} = 0.1$ the residual $f(0.1) = 2 \cdot 10^{-5}$ is small, but the error -0.1 is larger in absolute value by a factor 5000. Invent a more drastic example of your own. ▩

Secant Method for Solving $f(x) = 0$

Newton's method is very powerful but has the disadvantage that the derivative f' may sometimes be a far more difficult expression than f itself and its evaluation therefore computationally expensive. This situation suggests the idea of replacing the derivative with the difference quotient

$$f'(x_n) \approx \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}.$$

Then instead of (5) we have the formula of the popular secant method

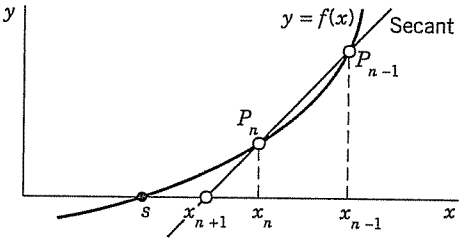


Fig. 426. Secant method

(10)
$$x_{n+1} = x_n - f(x_n)\frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}.$$

Geometrically, we intersect the x-axis at x_{n+1} with the secant of $f(x)$ passing through P_{n-1} and P_n in Fig. 426. We need two starting values x_0 and x_1. Evaluation of derivatives is now avoided. It can be shown that convergence is **superlinear** (that is, more rapid than linear, $|\epsilon_{n+1}| \approx const \cdot |\epsilon_n|^{1.62}$; see [E5] in App. 1), almost quadratic like Newton's method. The algorithm is similar to that of Newton's method, as the student may show.

CAUTION! It is ***not*** good to write (10) as

$$x_{n+1} = \frac{x_{n-1}f(x_n) - x_n f(x_{n-1})}{f(x_n) - f(x_{n-1})},$$

because this may lead to loss of significant digits if x_n and x_{n-1} are about equal. (Can you see this from the formula?)

EXAMPLE 8 Secant Method

Find the positive solution of $f(x) = x - 2\sin x = 0$ by the secant method, starting from $x_0 = 2$, $x_1 = 1.9$.

Solution. Here, (10) is

$$x_{n+1} = x_n - \frac{(x_n - 2\sin x_n)(x_n - x_{n-1})}{x_n - x_{n-1} + 2(\sin x_{n-1} - \sin x_n)} = x_n - \frac{N_n}{D_n}.$$

Numerical values are:

n	x_{n-1}	x_n	N_n	D_n	$x_{n+1} - x_n$
1	2.000 000	1.900 000	−0.000 740	−0.174 005	−0.004 253
2	1.900 000	1.895 747	−0.000 002	−0.006 986	−0.000 252
3	1.895 747	1.895 494	0		0

$x_3 = 1.895\,494$ is exact to 6D. See Example 4. ■

Summary of Methods. The methods for computing solutions s of $f(x) = 0$ with given continuous (or differentiable) $f(x)$ start with an initial approximation x_0 of s and generate a sequence $x_1, x_2, \cdots$ by **iteration. Fixed point methods** solve $f(x) = 0$ written as $x = g(x)$, so that s is a *fixed point* of g, that is, $s = g(s)$. For $g(x) = x - f(x)/f'(x)$ this is **Newton's method,** which for good x_0 and simple zeros converges quadratically (and for multiple zeros linearly). From Newton's method the **secant method** follows by replacing $f'(x)$ by a difference quotient. The **bisection method** and the **method of false position** in Problem Set 19.2 always converge, but often slowly.

PROBLEM SET 19.2

1–7 FIXED-POINT ITERATION

Apply fixed-point iteration and answer related questions where indicated. Show details of your work.

1. $x = 1.4 \sin x$, $x_0 = 1.4$
2. Do the iterations indicated at the end of Example 2. Sketch a figure similar to Fig. 424.
3. Why do we obtain a monotone sequence in Example 1, but not in Example 2?
4. $f = x^4 - x + 0.2 = 0$, the root near 1, $x_0 = 1$
5. f as in Prob. 4, the root near 0, $x_0 = 0$
6. Find the smallest positive solution of $\sin x = e^{-0.5x}$, $x_0 = 1$.
7. **(Bessel functions, drumhead)** A partial sum of the Maclaurin series of $J_0(x)$ (Sec. 5.5) is $f(x) = 1 - \frac{1}{4}x^2 + \frac{1}{64}x^4 - \frac{1}{2304}x^6$. Conclude from a sketch that $f(x) = 0$ near $x = 2$. Write $f(x) = 0$ as $x = g(x)$ (by dividing $f(x)$ by $\frac{1}{4}x$ and taking the resulting x-term to the other side). Find the zero. (See Sec. 12.9 for the importance of these zeros.)
8. **CAS PROJECT. Fixed-Point Iteration. (a) Existence.** Prove that if g is continuous in a closed interval I and its range lies in I, then the equation $x = g(x)$ has at least one solution in I. Illustrate that it may have more than one solution in I.

   (b) Convergence. Let $f(x) = x^3 + 2x^2 - 3x - 4 = 0$. Write this as $x = g(x)$, for g choosing (1) $(x^3 - f)^{1/3}$, (2) $(x^2 - \frac{1}{2}f)^{1/2}$, (3) $x + \frac{1}{3}f$, (4) $x(1 + \frac{1}{4}f)$, (5) $(x^3 - f)/x^2$, (6) $(2x^2 - f)/2x$, (7) $x - f/f'$ and in each case $x_0 = 1.5$. Find out about convergence and divergence and the number of steps to reach exact 6S-values of a root.

9–18 NEWTON'S METHOD

Apply Newton's method (6D accuracy). First sketch the function(s) to see what is going on.

9. $\sin x = \cot x$, $x_0 = 1$
10. $x = \cos x$, $x_0 = 1$
11. $x^3 - 5x + 3 = 0$, $x_0 = 2$
12. $x + \ln x = 2$, $x_0 = 2$
13. **(Vibrating beam)** Find the solution of $\cos x \cosh x = 1$ near $x = \frac{3}{2}\pi$. (This determines a frequency of a vibrating beam; see Problem Set 12.3.)
14. **(Heating, cooling)** At what time x (4S-accuracy only) will the processes governed by $f_1(x) = 100(1 - e^{-0.2x})$ and $f_2(x) = 40e^{-0.01x}$ reach the same temperature? Also find the latter.
15. **(Associated Legendre functions)** Find the smallest positive zero of $P_4^2 = (1 - x^2)P_4'' = \frac{15}{2}(-7x^4 + 8x^2 - 1)$ (Sec. 5.3) (a) by Newton's method, (b) exactly, by solving a quadratic equation.
16. **(Legendre polynomials)** Find the largest root of the Legendre polynomial $P_5(x)$ given by $P_5(x) = \frac{1}{8}(63x^5 - 70x^3 + 15x)$ (Sec. 5.3) (to be needed in Gauss integration in Sec. 19.5) (a) by Newton's method, (b) from a quadratic equation.
17. Design a Newton iteration for cube roots and compute $\sqrt[3]{7}$ (6D, $x_0 = 2$).
18. Design a Newton iteration for $\sqrt[k]{c}$ $(c > 0)$. Use it to compute $\sqrt{2}$, $\sqrt[3]{2}$, $\sqrt[4]{2}$, $\sqrt[5]{2}$ (6D, $x_0 = 1$).
19. **TEAM PROJECT. Bisection Method.** This simple but slowly convergent method for finding a solution of $f(x) = 0$ with continuous f is based on the **intermediate value theorem,** which states that if a continuous function f has opposite signs at some $x = a$ and $x = b$ $(> a)$, that is, either $f(a) < 0$, $f(b) > 0$ or $f(a) > 0$, $f(b) < 0$, then f must be 0 somewhere on $[a, b]$. The solution is found by repeated bisection of the interval and in each iteration picking that half which also satisfies that sign condition.

   (a) Algorithm. Write an algorithm for the method.

   (b) Comparison. Solve $x = \cos x$ by Newton's method and by bisection. Compare.

   (c) Solve $e^{-x} = \ln x$ and $e^x + x^4 + x = 2$ by bisection.
20. **TEAM PROJECT. Method of False Position (Regula falsi).** Figure 427 shows the idea. We assume that f is continuous. We compute the x-intercept c_0 of the line through $(a_0, f(a_0))$, $(b_0, f(b_0))$. If $f(c_0) = 0$, we are done. If $f(a_0)f(c_0) < 0$ (as in Fig. 427), we set $a_1 = a_0$, $b_1 = c_0$ and repeat to get c_1, etc. If $f(a_0)f(c_0) > 0$, then $f(c_0)f(b_0) < 0$ and we set $a_1 = c_0$, $b_1 = b_0$, etc.

   (a) Algorithm. Show that

$$c_0 = \frac{a_0 f(b_0) - b_0 f(a_0)}{f(b_0) - f(a_0)}$$

   and write an algorithm for the method.

   (b) Comparison. Solve $x^3 = 5x + 6$ by Newton's method, the secant method, and the method of false position. Compare.

   (c) Solve $x^4 = 2$, $\cos x = \sqrt{x}$, and $x + \ln x = 2$ by the method of false position.

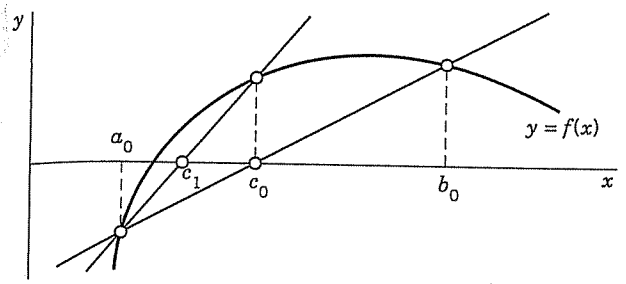


Fig. 427. Method of false position

21–24 SECANT METHOD

Solve, using x_0 and x_1 as indicated.

21. Prob. 11, $x_0 = 0.5$, $x_1 = 2.0$

22. $e^{-x} - \tan x = 0$, $x_0 = 1$, $x_1 = 0.7$

23. Prob. 9, $x_0 = 1$, $x_1 = 0.5$

24. Prob. 10, $x_0 = 0.5$, $x_1 = 1$

25. WRITING PROJECT. Solution of Equations. Compare the methods in this section and problem set, discussing advantages and disadvantages using examples of your own.

19.3 Interpolation

Interpolation means finding (approximate) values of a function $f(x)$ for an x between ***different*** x-values $x_0, x_1, \cdots, x_n$ at which the values of $f(x)$ are given. These values may come from a "mathematical" function, such as a logarithm or a Bessel function, or, perhaps more frequently, they may be measured or automatically recorded values of an "empirical" function, such as the air resistance of a car or an airplane at different speeds, or the yield of a chemical process at different temperatures, or the size of the U.S. population as it appears from censuses taken at 10-year intervals. We write these given values of a function f in the form

$$f_0 = f(x_0), \qquad f_1 = f(x_1), \qquad \cdots, \qquad f_n = f(x_n)$$

or as ordered pairs

$$(x_0, f_0), \qquad (x_1, f_1), \qquad \cdots, \qquad (x_n, f_n).$$

A standard idea in interpolation now is to find a polynomial $p_n(x)$ of degree n (or less) that assumes the given values; thus

$$\text{(1)} \qquad p_n(x_0) = f_0, \quad p_n(x_1) = f_1, \quad \cdots, \quad p_n(x_n) = f_n.$$

We call this p_n an **interpolation polynomial** and $x_0, \cdots, x_n$ the **nodes.** And if $f(x)$ is a mathematical function, we call p_n an **approximation** of f (or a **polynomial approximation,** because there are other kinds of approximations, as we shall see later). We use p_n to get (approximate) values of f for x's between x_0 and x_n (**"interpolation"**) or sometimes outside this interval $x_0 \leqq x \leqq x_n$ (**"extrapolation"**).

Motivation. Polynomials are convenient to work with because we can readily differentiate and integrate them, again obtaining polynomials. Moreover, they approximate *continuous* functions with any desired accuracy. That is, for any continuous $f(x)$ on an interval J: $a \leqq x \leqq b$ and error bound $\beta > 0$, there is a polynomial $p_n(x)$ (of sufficiently high degree n) such that

$$|f(x) - p_n(x)| < \beta \quad \text{for all } x \text{ on } J.$$

This is the famous **Weierstrass approximation theorem** (for a proof see Ref. [GR7], p. 280; see App. 1).

Existence and Uniqueness. p_n satisfying (1) for given data exists—we give formulas for it below. p_n is unique. Indeed, if another polynomial q_n also satisfies $q_n(x_0) = f_0, \cdots, q_n(x_n) = f_n$, then $p_n(x) - q_n(x) = 0$ at $x_0, \cdots, x_n$, but a polynomial $p_n - q_n$ of degree n (or less) with $n + 1$ roots must be identically zero, as we know from algebra; thus $p_n(x) = q_n(x)$ for all x, which means uniqueness. ∎

How to Find p_n? This is the important practical question. We answer it by explaining several standard methods. For given data, these methods give the same polynomial, by the uniqueness just proved (which is thus of practical interest!), but expressed in several forms suitable for different purposes.

Lagrange Interpolation

Given $(x_0, f_0), (x_1, f_1), \cdots, (x_n, f_n)$ with arbitrarily spaced x_j, Lagrange had the idea of multiplying each f_j by a polynomial that is 1 at x_j and 0 at the other n nodes and then taking the sum of these $n + 1$ polynomials. Clearly, this gives the unique interpolation polynomial of degree n or less. Beginning with the simplest case, let us see how this works.

Linear interpolation is interpolation by the straight line through (x_0, f_0), (x_1, f_1); see Fig. 428. Thus the linear Lagrange polynomial p_1 is a sum $p_1 = L_0 f_0 + L_1 f_1$ with L_0 the linear polynomial that is 1 at x_0 and 0 at x_1; similarly, L_1 is 0 at x_0 and 1 at x_1. Obviously,

$$L_0(x) = \frac{x - x_1}{x_0 - x_1}, \qquad L_1(x) = \frac{x - x_0}{x_1 - x_0}.$$

This gives the linear Lagrange polynomial

$$(2) \qquad p_1(x) = L_0(x)f_0 + L_1(x)f_1 = \frac{x - x_1}{x_0 - x_1} \cdot f_0 + \frac{x - x_0}{x_1 - x_0} \cdot f_1.$$

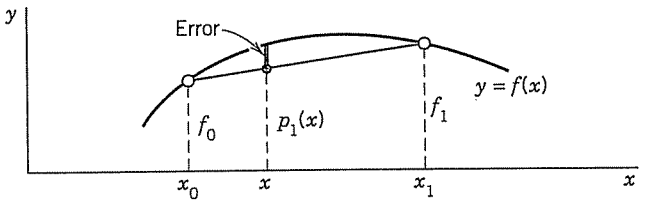


Fig. 428. Linear interpolation

EXAMPLE 1 **Linear Lagrange Interpolation**

Compute a 4D-value of ln 9.2 from ln 9.0 = 2.1972, ln 9.5 = 2.2513 by linear Lagrange interpolation and determine the error, using ln 9.2 = 2.2192 (4D).

Solution. $x_0 = 9.0$, $x_1 = 9.5$, $f_0 = \ln 9.0$, $f_1 = \ln 9.5$. In (2) we need

$$L_0(x) = \frac{x - 9.5}{-0.5} = -2.0(x - 9.5), \qquad L_0(9.2) = -2.0(-0.3) = 0.6$$

$$L_1(x) = \frac{x - 9.0}{0.5} = 2.0(x - 9.0), \qquad L_1(9.2) = 2 \cdot 0.2 = 0.4$$

(see Fig. 429) and obtain the answer

$$\ln 9.2 \approx p_1(9.2) = L_0(9.2)f_0 + L_1(9.2)f_1 = 0.6 \cdot 2.1972 + 0.4 \cdot 2.2513 = 2.2188.$$

The error is $\epsilon = a - \tilde{a} = 2.2192 - 2.2188 = 0.0004$. Hence linear interpolation is not sufficient here to get 4D-accuracy; it would suffice for 3D-accuracy. ■

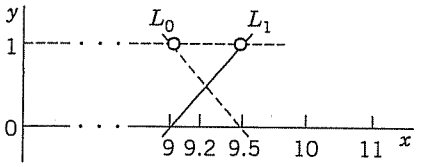


Fig. 429. L_0 and L_1 in Example 1

Quadratic interpolation is interpolation of given (x_0, f_0), (x_1, f_1), (x_2, f_2) by a second-degree polynomial $p_2(x)$, which by Lagrange's idea is

(3a)
$$p_2(x) = L_0(x)f_0 + L_1(x)f_1 + L_2(x)f_2$$

with $L_0(x_0) = 1$, $L_1(x_1) = 1$, $L_2(x_2) = 1$, and $L_0(x_1) = L_0(x_2) = 0$, etc. We claim that

(3b)
$$L_0(x) = \frac{l_0(x)}{l_0(x_0)} = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)}$$

$$L_1(x) = \frac{l_1(x)}{l_1(x_1)} = \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)}$$

$$L_2(x) = \frac{l_2(x)}{l_2(x_2)} = \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)}.$$

How did we get this? Well, the numerator makes $L_k(x_j) = 0$ if $j \neq k$. And the denominator makes $L_k(x_k) = 1$ because it equals the numerator at $x = x_k$.

EXAMPLE 2 Quadratic Lagrange Interpolation

Compute ln 9.2 by (3) from the data in Example 1 and the additional third value ln 11.0 = 2.3979.

Solution. In (3),

$$L_0(x) = \frac{(x - 9.5)(x - 11.0)}{(9.0 - 9.5)(9.0 - 11.0)} = x^2 - 20.5x + 104.5, \qquad L_0(9.2) = 0.5400,$$

$$L_1(x) = \frac{(x - 9.0)(x - 11.0)}{(9.5 - 9.0)(9.5 - 11.0)} = -\frac{1}{0.75}(x^2 - 20x + 99), \qquad L_1(9.2) = 0.4800,$$

$$L_2(x) = \frac{(x - 9.0)(x - 9.5)}{(11.0 - 9.0)(11.0 - 9.5)} = \frac{1}{3}(x^2 - 18.5x + 85.5), \qquad L_2(9.2) = -0.0200,$$

(see Fig. 430), so that (3a) gives, exact to 4D,

$$\ln 9.2 \approx p_2(9.2) = 0.5400 \cdot 2.1972 + 0.4800 \cdot 2.2513 - 0.0200 \cdot 2.3979 = 2.2192.$$ ■

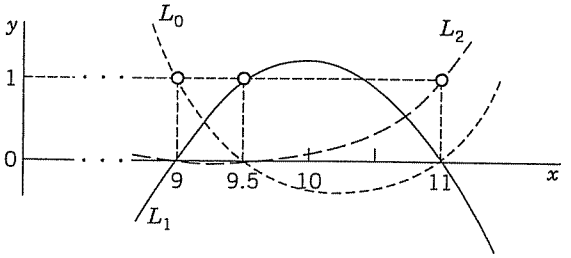


Fig. 430. L_0, L_1, L_2 in Example 2

General Lagrange Interpolation Polynomial. For general n we obtain

$$\text{(4a)} \qquad f(x) \approx p_n(x) = \sum_{k=0}^{n} L_k(x) f_k = \sum_{k=0}^{n} \frac{l_k(x)}{l_k(x_k)} f_k$$

where $L_k(x_k) = 1$ and L_k is 0 at the other nodes, and the L_k are independent of the function f to be interpolated. We get (4a) if we take

$$\text{(4b)} \qquad \begin{aligned} l_0(x) &= (x - x_1)(x - x_2) \cdots (x - x_n), \\ l_k(x) &= (x - x_0) \cdots (x - x_{k-1})(x - x_{k+1}) \cdots (x - x_n), \qquad 0 < k < n, \\ l_n(x) &= (x - x_0)(x - x_1) \cdots (x - x_{n-1}). \end{aligned}$$

We can easily see that $p_n(x_k) = f_k$. Indeed, inspection of (4b) shows that $l_k(x_j) = 0$ if $j \neq k$, so that for $x = x_k$, the sum in (4a) reduces to the single term $(l_k(x_k)/l_k(x_k))f_k = f_k$.

Error Estimate. If f is itself a polynomial of degree n (or less), it must coincide with p_n because the $n + 1$ data $(x_0, f_0), \cdots, (x_n, f_n)$ determine a polynomial uniquely, so the error is zero. Now the special f has its $(n + 1)$st derivative identically zero. This makes it plausible that for a *general* f its $(n + 1)$st derivative $f^{(n+1)}$ should measure the error

$$\epsilon_n(x) = f(x) - p_n(x).$$

It can be shown that this is true if $f^{(n+1)}$ exists and is continuous. Then, with a suitable t between x_0 and x_n (or between x_0, x_n, and x if we extrapolate),

$$\text{(5)} \qquad \epsilon_n(x) = f(x) - p_n(x) = (x - x_0)(x - x_1) \cdots (x - x_n) \frac{f^{(n+1)}(t)}{(n + 1)!}.$$

Thus $|\epsilon_n(x)|$ is 0 at the nodes and small near them, because of continuity. The product $(x - x_0) \cdots (x - x_n)$ is large for x away from the nodes. This makes extrapolation risky. And interpolation at an x will be best if we choose nodes on both sides of that x. Also, we get error bounds by taking the smallest and the largest value of $f^{(n+1)}(t)$ in (5) on the interval $x_0 \leqq t \leqq x_n$ (or on the interval also containing x if we *extra*polate).

Most importantly, since p_n is unique, as we have shown, we have

THEOREM 1

Error of Interpolation

Formula (5) *gives the error for* ***any*** *polynomial interpolation method if* $f(x)$ *has a continuous* $(n + 1)$*st derivative.*

Practical error estimate. If the derivative in (5) is difficult or impossible to obtain, apply the Error Principle (Sec. 19.1), that is, take another node and the Lagrange polynomial $p_{n+1}(x)$ and regard $p_{n+1}(x) - p_n(x)$ as a (crude) error estimate for $p_n(x)$.

EXAMPLE 3 **Error Estimate (5) of Linear Interpolation. Damage by Roundoff. Error Principle**

Estimate the error in Example 1 first by (5) directly and then by the Error Principle (Sec. 19.1).

Solution. (A) ***Estimation by* (5).** We have $n = 1$, $f(t) = \ln t$, $f'(t) = 1/t$, $f''(t) = -1/t^2$. Hence

$$\epsilon_1(x) = (x - 9.0)(x - 9.5)\frac{(-1)}{2t^2}, \qquad \text{thus} \qquad \epsilon_1(9.2) = \frac{0.03}{t^2}.$$

$t = 9.0$ gives the maximum $0.03/9^2 = 0.00037$ and $t = 9.5$ gives the minimum $0.03/9.5^2 = 0.00033$, so that we get $0.00033 \leqq \epsilon_1(9.2) \leqq 0.00037$, or better, 0.00038 because $0.3/81 = 0.003\,703 \cdots$.

But the error 0.0004 in Example 1 disagrees, and we can learn something! Repetition of the computation there with 5D instead of 4D gives

$$\ln 9.2 \approx p_1(9.2) = 0.6 \cdot 2.19722 + 0.4 \cdot 2.25129 = 2.21885$$

with an actual error $\epsilon = 2.21920 - 2.21885 = 0.00035$, which lies nicely near the middle between our two error bounds.

This shows that the discrepancy (0.0004 vs. 0.00035) was caused by rounding, which is not taken into account in (5).

(B) *Estimation by the Error Principle.* We calculate $p_1(9.2) = 2.21885$ as before and then $p_2(9.2)$ as in Example 2 but with 5D, obtaining

$$p_2(9.2) = 0.54 \cdot 2.19722 + 0.48 \cdot 2.25129 - 0.02 \cdot 2.39790 = 2.21916.$$

The difference $p_2(9.2) - p_1(9.2) = 0.00031$ is the approximate error of $p_1(9.2)$ that we wanted to obtain; this is an approximation of the actual error 0.00035 given above. ▩

Newton's Divided Difference Interpolation

For given data $(x_0, f_0), \cdots, (x_n, f_n)$ the interpolation polynomial $p_n(x)$ satisfying (1) is unique, as we have shown. But for different purposes we may use $p_n(x)$ in different forms. **Lagrange's form** just discussed is useful for deriving formulas in numeric differentiation (approximation formulas for derivatives) and integration (Sec. 19.5).

Practically more important are Newton's forms of $p_n(x)$, which we shall also use for solving ODEs (in Sec. 21.2). They involve fewer arithmetic operations than Lagrange's form. Moreover, it often happens that we have to increase the degree n to reach a required accuracy. Then in Newton's forms we can use all the previous work and just add another term, a possibility without counterpart for Lagrange's form. This also simplifies the application of the Error Principle (used in Example 3 for Lagrange). The details of these ideas are as follows.

Let $p_{n-1}(x)$ be the $(n-1)$st Newton polynomial (whose form we shall determine); thus $p_{n-1}(x_0) = f_0, p_{n-1}(x_1) = f_1, \cdots, p_{n-1}(x_{n-1}) = f_{n-1}$. Furthermore, let us write the nth Newton polynomial as

$$p_n(x) = p_{n-1}(x) + g_n(x); \tag{6}$$

hence

$$g_n(x) = p_n(x) - p_{n-1}(x). \tag{6'}$$

Here $g_n(x)$ is to be determined so that $p_n(x_0) = f_0, p_n(x_1) = f_1, \cdots, p_n(x_n) = f_n$.

Since p_n and p_{n-1} agree at $x_0, \cdots, x_{n-1}$, we see that g_n is zero there. Also, g_n will generally be a polynomial of nth degree because so is p_n, whereas p_{n-1} can be of degree $n - 1$ at most. Hence g_n must be of the form

$$g_n(x) = a_n(x - x_0)(x - x_1) \cdots (x - x_{n-1}). \tag{6''}$$

We determine the constant a_n. For this we set $x = x_n$ and solve (6″) algebraically for a_n. Replacing $g_n(x_n)$ according to (6′) and using $p_n(x_n) = f_n$, we see that this gives

$$a_n = \frac{f_n - p_{n-1}(x_n)}{(x_n - x_0)(x_n - x_1) \cdots (x_n - x_{n-1})}. \tag{7}$$

We write a_k instead of a_n and show that a_k equals the **kth divided difference,** recursively denoted and defined as follows:

$$a_1 = f[x_0, x_1] = \frac{f_1 - f_0}{x_1 - x_0}$$

$$a_2 = f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0}$$

and in general

$$a_k = f[x_0, \cdots, x_k] = \frac{f[x_1, \cdots, x_k] - f[x_0, \cdots, x_{k-1}]}{x_k - x_0}. \tag{8}$$

PROOF If $n = 1$, then $p_{n-1}(x_n) = p_0(x_1) = f_0$ because $p_0(x)$ is constant and equal to f_0, the value of $f(x)$ at x_0. Hence (7) gives

$$a_1 = \frac{f_1 - p_0(x_1)}{x_1 - x_0} = \frac{f_1 - f_0}{x_1 - x_0} = f[x_0, x_1],$$

and (6) and (6″) give the Newton interpolation polynomial of the first degree

$$p_1(x) = f_0 + (x - x_0)f[x_0, x_1].$$

If $n = 2$, then this p_1 and (7) give

$$a_2 = \frac{f_2 - p_1(x_2)}{(x_2 - x_0)(x_2 - x_1)} = \frac{f_2 - f_0 - (x_2 - x_0)f[x_0, x_1]}{(x_2 - x_0)(x_2 - x_1)} = f[x_0, x_1, x_2]$$

where the last equality follows by straightforward calculation and comparison with the definition of the right side. (Verify it; be patient.) From (6) and (6″) we thus obtain the second Newton polynomial

$$p_2(x) = f_0 + (x - x_0)f[x_0, x_1] + (x - x_0)(x - x_1)f[x_0, x_1, x_2].$$

For $n = k$, formula (6) gives

$$p_k(x) = p_{k-1}(x) + (x - x_0)(x - x_1) \cdots (x - x_{k-1})f[x_0, \cdots, x_k]. \tag{9}$$

With $p_0(x) = f_0$ by repeated application with $k = 1, \cdots, n$ this finally gives **Newton's divided difference interpolation formula**

$$\begin{aligned} f(x) \approx\ & f_0 + (x - x_0)f[x_0, x_1] + (x - x_0)(x - x_1)f[x_0, x_1, x_2] \\ & + \cdots + (x - x_0)(x - x_1) \cdots (x - x_{n-1})f[x_0, \cdots, x_n]. \end{aligned} \tag{10}$$

An algorithm is shown in Table 19.2. The first do-loop computes the divided differences and the second the desired value $p_n(\hat{x})$.

Example 4 shows how to arrange differences near the values from which they are obtained; the latter always stand a half-line above and a half-line below in the preceding column. Such an arrangement is called a (divided) **difference table.** ■

Table 19.2 Newton's Divided Difference Interpolation

ALGORITHM INTERPOL $(x_0, \cdots, x_n; f_0, \cdots, f_n; \hat{x})$

This algorithm computes an approximation $p_n(\hat{x})$ of $f(\hat{x})$ at $\hat{x}$.

INPUT: Data $(x_0, f_0), (x_1, f_1), \cdots, (x_n, f_n)$; $\hat{x}$

OUTPUT: Approximation $p_n(\hat{x})$ of $f(\hat{x})$

Set $f[x_j] = f_j \quad (j = 0, \cdots, n)$.

For $m = 1, \cdots, n - 1$ do:

  For $j = 0, \cdots, n - m$ do:

$$f[x_j, \cdots, x_{j+m}] = \frac{f[x_{j+1}, \cdots, x_{j+m}] - f[x_j, \cdots, x_{j+m-1}]}{x_{j+m} - x_j}$$

  End

End

Set $p_0(x) = f_0$.

For $k = 1, \cdots, n$ do:

$$p_k(\hat{x}) = p_{k-1}(\hat{x}) + (\hat{x} - x_0) \cdots (\hat{x} - x_{k-1}) f[x_0, \cdots, x_k]$$

End

OUTPUT $p_n(\hat{x})$

End INTERPOL

EXAMPLE 4 Newton's Divided Difference Interpolation Formula

Compute $f(9.2)$ from the values shown in the first two columns of the following table.

x_j	$f_j = f(x_j)$	$f[x_j, x_{j+1}]$	$f[x_j, x_{j+1}, x_{j+2}]$	$f[x_j, \cdots, x_{j+3}]$
8.0	(2.079 442)			
		(0.117 783)		
9.0	2.197 225		(−0.006 433)	
		0.108 134		(0.000 411)
9.5	2.251 292		−0.005 200	
		0.097 735		
11.0	2.397 895			

Solution. We compute the divided differences as shown. Sample computation:

$$(0.097\,735 - 0.108\,134)/(11 - 9) = -0.005\,200.$$

The values we need in (10) are circled. We have

$$f(x) \approx p_3(x) = 2.079\,442 + 0.117\,783(x - 8.0) - 0.006\,433(x - 8.0)(x - 9.0) + 0.000\,411(x - 8.0)(x - 9.0)(x - 9.5).$$

At $x = 9.2$,

$$f(9.2) \approx 2.079\,442 + 0.141\,340 - 0.001\,544 - 0.000\,030 = 2.219\,208.$$

The value exact to 6D is $f(9.2) = \ln 9.2 = 2.219\,203$. Note that we can nicely see how the accuracy increases from term to term:

$$p_1(9.2) = 2.220\,782, \qquad p_2(9.2) = 2.219\,238, \qquad p_3(9.2) = 2.219\,208.$$

Equal Spacing: Newton's Forward Difference Formula

Newton's formula (10) is valid for *arbitrarily spaced* nodes as they may occur in practice in experiments or observations. However, in many applications the x_j's are *regularly spaced*—for instance, in measurements taken at regular intervals of time. Then, denoting the distance by h, we can write

$$\textbf{(11)} \qquad x_0, \quad x_1 = x_0 + h, \quad x_2 = x_0 + 2h, \quad \cdots, \quad x_n = x_0 + nh.$$

We show how (8) and (10) now simplify considerably!

To get started, let us define the *first forward difference* of f at x_j by

$$\Delta f_j = f_{j+1} - f_j,$$

the *second forward difference* of f at x_j by

$$\Delta^2 f_j = \Delta f_{j+1} - \Delta f_j,$$

and, continuing in this way, the ***k*th forward difference** of f at x_j by

$$\textbf{(12)} \qquad \Delta^k f_j = \Delta^{k-1} f_{j+1} - \Delta^{k-1} f_j \qquad (k = 1, 2, \cdots).$$

Examples and an explanation of the name "forward" follow on p. 806. What is the point of this? We show that if we have regular spacing (11), then

$$\textbf{(13)} \qquad f[x_0, \cdots, x_k] = \frac{1}{k!\,h^k}\,\Delta^k f_0.$$

We prove (13) by induction. It is true for $k = 1$ because $x_1 = x_0 + h$, so that

$$f[x_0, x_1] = \frac{f_1 - f_0}{x_1 - x_0} = \frac{1}{h}(f_1 - f_0) = \frac{1}{1!\,h}\,\Delta f_0.$$

Assuming (13) to be true for all forward differences of order k, we show that (13) holds for $k+1$. We use (8) with $k+1$ instead of k; then we use $(k+1)h = x_{k+1} - x_0$, resulting from (11), and finally (12) with $j = 0$, that is, $\Delta^{k+1} f_0 = \Delta^k f_1 - \Delta^k f_0$. This gives

$$\begin{aligned} f[x_0, \cdots, x_{k+1}] &= \frac{f[x_1, \cdots, x_{k+1}] - f[x_0, \cdots, x_k]}{(k+1)h} \\ &= \frac{1}{(k+1)h} \left[\frac{1}{k!\, h^k} \Delta^k f_1 - \frac{1}{k!\, h^k} \Delta^k f_0 \right] \\ &= \frac{1}{(k+1)!\, h^{k+1}} \Delta^{k+1} f_0 \end{aligned}$$

which is (13) with $k+1$ instead of k. Formula (13) is proved. ∎

In (10) we finally set $x = x_0 + rh$. Then $x - x_0 = rh$, $x - x_1 = (r-1)h$ since $x_1 - x_0 = h$, and so on. With this and (13), formula (10) becomes **Newton's** (or *Gregory*[2]*–Newton's*) **forward difference interpolation formula**

$$\begin{aligned} f(x) \approx p_n(x) &= \sum_{s=0}^{n} \binom{r}{s} \Delta^s f_0 \qquad (x = x_0 + rh, \quad r = (x - x_0)/h) \\ &= f_0 + r\Delta f_0 + \frac{r(r-1)}{2!} \Delta^2 f_0 + \cdots + \frac{r(r-1)\cdots(r-n+1)}{n!} \Delta^n f_0 \end{aligned} \tag{14}$$

where the **binomial coefficients** in the first line are defined by

$$\binom{r}{0} = 1, \quad \binom{r}{s} = \frac{r(r-1)(r-2)\cdots(r-s+1)}{s!} \qquad (s > 0, \text{ integer}) \tag{15}$$

and $s! = 1 \cdot 2 \cdots s$.

Error. From (5) we get, with $x - x_0 = rh$, $x - x_1 = (r-1)h$, etc.,

$$\epsilon_n(x) = f(x) - p_n(x) = \frac{h^{n+1}}{(n+1)!} r(r-1)\cdots(r-n) f^{(n+1)}(t) \tag{16}$$

with t as characterized in (5).

Formula (16) is an exact formula for the error, but it involves the unknown t. In Example 5 (below) we show how to use (16) for obtaining an error estimate and an interval in which the true value of $f(x)$ must lie.

Comments on Accuracy. **(A)** The order of magnitude of the error $\epsilon_n(x)$ is about equal to that of the next difference not used in $p_n(x)$.

(B) One should choose $x_0, \cdots, x_n$ such that the x at which one interpolates is as well centered between $x_0, \cdots, x_n$ as possible.

---

[2]JAMES GREGORY (1638–1675), Scots mathematician, professor at St. Andrews and Edinburgh. Δ in (14) and ∇^2 (on p. 807) have nothing to do with the Laplacian.

The reason for (A) is that in (16),

$$f^{n+1}(t) \approx \frac{\Delta^{n+1} f(t)}{h^{n+1}}, \qquad \frac{|r(r-1)\cdots(r-n)|}{1\cdot 2\cdots(n+1)} \leqq 1 \quad \text{if} \quad |r| \leqq 1$$

(and actually for any r as long as we do not *extrapolate*). The reason for (B) is that $|r(r-1)\cdots(r-n)|$ becomes smallest for that choice.

EXAMPLE 5 Newton's Forward Difference Formula. Error Estimation

Compute cosh 0.56 from (14) and the four values in the following table and estimate the error.

j	x_j	$f_j = \cosh x_j$	Δf_j	$\Delta^2 f_j$	$\Delta^3 f_j$
0	0.5	(1.127 626)			
			(0.057 839)		
1	0.6	1.185 465		(0.011 865)	
			0.069 704		(0.000 697)
2	0.7	1.255 169		0.012 562	
			0.082 266		
3	0.8	1.337 435			

Solution. We compute the forward differences as shown in the table. The values we need are circled. In (14) we have $r = (0.56 - 0.50)/0.1 = 0.6$, so that (14) gives

$$\begin{aligned}\cosh 0.56 &\approx 1.127\,626 + 0.6 \cdot 0.057\,839 + \frac{0.6(-0.4)}{2} \cdot 0.011\,865 + \frac{0.6(-0.4)(-1.4)}{6} \cdot 0.000\,697 \\ &= 1.127\,626 + 0.034\,703 - 0.001\,424 + 0.000\,039 \\ &= 1.160\,944.\end{aligned}$$

Error estimate. From (16), since the fourth derivative is $\cosh^{(4)} t = \cosh t$,

$$\begin{aligned}\epsilon_3(0.56) &= \frac{0.1^4}{4!} \cdot 0.6(-0.4)(-1.4)(-2.4) \cosh t \\ &= A \cosh t,\end{aligned}$$

where $A = -0.000\,003\,36$ and $0.5 \leqq t \leqq 0.8$. We do not know t, but we get an inequality by taking the largest and smallest $\cosh t$ in that interval:

$$A \cosh 0.8 \leqq \epsilon_3(0.62) \leqq A \cosh 0.5.$$

Since

$$f(x) = p_3(x) + \epsilon_3(x),$$

this gives

$$p_3(0.56) + A \cosh 0.8 \leqq \cosh 0.56 \leqq p_3(0.56) + A \cosh 0.5.$$

Numeric values are

$$1.160\,939 \leqq \cosh 0.56 \leqq 1.160\,941.$$

The exact 6D-value is $\cosh 0.56 = 1.160\,941$. It lies within these bounds. Such bounds are not always so tight. Also, we did not consider roundoff errors, which will depend on the number of operations. ■

This example also explains the name "*forward* difference formula": we see that the differences in the formula slope forward in the difference table.

Equal Spacing: Newton's Backward Difference Formula

Instead of forward-sloping differences we may also employ backward-sloping differences. The difference table remains the same as before (same numbers, in the same positions), except for a very harmless change of the running subscript j (which we explain in Example 6, below). Nevertheless, purely for reasons of convenience it is standard to introduce a second name and notation for differences as follows. We define the *first backward difference* of f at x_j by

$$\nabla f_j = f_j - f_{j-1},$$

the *second backward difference* of f at x_j by

$$\nabla^2 f_j = \nabla f_j - \nabla f_{j-1},$$

and, continuing in this way, the ***k*th backward difference** of f at x_j by

$$\nabla^k f_j = \nabla^{k-1} f_j - \nabla^{k-1} f_{j-1} \qquad (k = 1, 2, \cdots). \tag{17}$$

A formula similar to (14) but involving backward differences is **Newton's** (or *Gregory–Newton's*) **backward difference interpolation formula**

$$\begin{aligned} f(x) \approx p_n(x) &= \sum_{s=0}^{n} \binom{r+s-1}{s} \nabla^s f_0 \qquad (x = x_0 + rh,\ r = (x - x_0)/h) \\ &= f_0 + r\nabla f_0 + \frac{r(r+1)}{2!}\nabla^2 f_0 + \cdots + \frac{r(r+1)\cdots(r+n-1)}{n!}\nabla^n f_0. \end{aligned} \tag{18}$$

EXAMPLE 6 **Newton's Forward and Backward Interpolations**

Compute a 7D-value of the Bessel function $J_0(x)$ for $x = 1.72$ from the four values in the following table, using (a) Newton's forward formula (14), (b) Newton's backward formula (18).

j_{for}	j_{back}	x_j	$J_0(x_j)$	1st Diff.	2nd Diff.	3rd Diff.
0	−3	1.7	0.397 9849			
				−0.057 9985		
1	−2	1.8	0.339 9864		−0.000 1693	
				−0.058 1678		0.000 4093
2	−1	1.9	0.281 8186		0.000 2400	
				−0.057 9278		
3	0	2.0	0.223 8908			

Solution. The computation of the differences is the same in both cases. Only their notation differs.

(a) Forward. In (14) we have $r = (1.72 - 1.70)/0.1 = 0.2$, and j goes from 0 to 3 (see first column). In each column we need the first given number, and (14) thus gives

$$J_0(1.72) \approx 0.397\,9849 + 0.2(-0.057\,9985) + \frac{0.2(-0.8)}{2}(-0.000\,1693) + \frac{0.2(-0.8)(-1.8)}{6} \cdot 0.000\,4093$$

$$= 0.397\,9849 - 0.011\,5997 + 0.000\,0135 + 0.000\,0196 = 0.386\,4183,$$

which is exact to 6D, the exact 7D-value being 0.386 4185.

(b) Backward. For (18) we use j shown in the second column, and in each column the last number. Since $r = (1.72 - 2.00)/0.1 = -2.8$, we thus get from (18)

$$J_0(1.72) \approx 0.223\,8908 - 2.8(-0.057\,9278) + \frac{-2.8(-1.8)}{2} \cdot 0.000\,2400 + \frac{-2.8(-1.8)(-0.8)}{6} \cdot 0.000\,4093$$

$$= 0.223\,8908 + 0.162\,1978 + 0.000\,6048 - 0.000\,2750$$

$$= 0.386\,4184.$$

Central Difference Notation

This is a third notation for differences. The first central difference of $f(x)$ at x_j is defined by

$$\delta f_j = f_{j+1/2} - f_{j-1/2}$$

and the ***k*th central difference** of $f(x)$ at x_j by

$$\delta^k f_j = \delta^{k-1} f_{j+1/2} - \delta^{k-1} f_{j-1/2} \qquad (j = 2, 3, \cdots). \tag{19}$$

Thus in this notation a difference table, for example, for f_{-1}, f_0, f_1, f_2, looks as follows:

x_{-1}	f_{-1}			
		$\delta f_{-1/2}$		
x_0	f_0		$\delta^2 f_0$	
		$\delta f_{1/2}$		$\delta^3 f_{1/2}$
x_1	f_1		$\delta^2 f_1$	
		$\delta f_{3/2}$		
x_2	f_2			

Central differences are used in numeric differentiation (Sec. 19.5), differential equations (Chap. 21), and centered interpolation formulas (e.g., Everett's formula in Team Project 22). These are formulas that use function values "symmetrically" located on both sides of the interpolation point x. Such values are available near the middle of a given table, where centered interpolation formulas tend to give better results than those of Newton's formulas, which do not have that "symmetry" property.

PROBLEM SET 19.2

1. **(Linear interpolation)** Calculate $p_1(x)$ in Example 1. Compute from it $\ln 9.4 \approx p_1(9.4)$.
2. Estimate the error in Prob. 1 by (5).
3. **(Quadratic interpolation)** Calculate the Lagrange polynomial $p_2(x)$ for the 4D-values of the Gamma function [(24), App. 3.1] $\Gamma(1.00) = 1.0000$, $\Gamma(1.02) = 0.9888$, $\Gamma(1.04) = 0.9784$, and from it approximations of $\Gamma(x)$ for $x = 1.005, 1.010, 1.015, 1.025, 1.030, 1.035$.
4. **(Error bounds)** Derive error bounds for $p_2(9.2)$ in Example 2 from (5).
5. **(Error function)** Calculate the Lagrange polynomial $p_2(x)$ for the 5D-values of the error function $f(x) = \operatorname{erf} x = (2/\sqrt{\pi}) \int_0^x e^{-w^2}\, dw$, namely, $f(0.25) = 0.27633$, $f(0.5) = 0.52050$, $f(1) = 0.84270$, and from p_2 an approximation of $f(0.75)$ (= 0.71116, 5D).
6. Derive an error bound in Prob. 5 from (5).
7. **(Sine integral)** Calculate the Lagrange polynomial $p_2(x)$ for the 4D-values of the sine integral $\mathrm{Si}(x)$ [(40) in App. 3.1], namely, $\mathrm{Si}(0) = 0$, $\mathrm{Si}(1) = 0.9461$, $\mathrm{Si}(2) = 1.6054$, and from p_2 approximations of $\mathrm{Si}(0.5)$ (= 0.4931, 4D) and $\mathrm{Si}(1.5)$ (= 1.3247, 4D).
8. **(Linear and quadratic interpolation)** Find $e^{-0.25}$ and $e^{-0.75}$ by linear interpolation with $x_0 = 0$, $x_1 = 0.5$ and $x_0 = 0.5$, $x_1 = 1$, respectively. Then find $p_2(x)$

interpolating e^{-x} with $x_0 = 0$, $x_1 = 0.5$, $x_2 = 1$ and from it $e^{-0.25}$ and $e^{-0.75}$. Compare the errors of these linear and quadratic interpolations. Use 4D-values of e^{-x}.

9. **(Cubic Lagrange interpolation)** Calculate and sketch or graph L_0, L_1, L_2, L_3 for $x = 0, 1, 2, 3$ on common axes. Find $p_3(x)$ for the data

(0, 1)

(1, 0.765198)

(2, 0.223891)

(3, −0.260052)

[values of the Bessel function $J_0(x)$]. Find p_3 for $x = 0.5, 1.5, 2.5$ by interpolation.

10. **(Interpolation and extrapolation)** Calculate $p_2(x)$ in Example 2. Compute from it approximations of ln 9.4, ln 10, ln 10.5, ln 11.5, ln 12, compute the errors by using exact 4D-values, and comment.

11. **(Extrapolation)** Does a sketch or graph of the product of the $(x - x_j)$ in (5) for the data in Prob. 10 indicate that extrapolation is likely to involve larger errors than interpolation does?

12. **(Lower degree)** Find the degree of the interpolation polynomial for the data

(−2, 33)

(0, 5)

(2, 9)

(4, 45)

(6, 113).

13. **(Newton's forward difference formula)** Set up (14) for the data in Prob. 7 and derive $p_2(x)$ from (14).

14. Set up Newton's forward difference formula for the data in Prob. 3 and compute $\Gamma(1.01)$, $\Gamma(1.03)$, $\Gamma(1.05)$.

15. **(Newton's divided difference formula)** Compute $f(0.8)$ and $f(0.9)$ from

$$f(0.5) = 0.479$$

$$f(1.0) = 0.841$$

$$f(2.0) = 0.909$$

by quadratic interpolation.

16. Compute $f(6.5)$ from

$$f(6.0) = 0.1506$$

$$f(7.0) = 0.3001$$

$$f(7.5) = 0.2663$$

$$f(7.7) = 0.2346$$

by cubic interpolation, using (10).

17. **(Central differences)** Write the difference in the table in Example 5 in central difference notation.

18. **(Subtabulation)** Compute the Bessel function $J_1(x)$ for $x = 0.1, 0.3, \cdots, 0.9$ from $J_1(0) = 0$, $J_1(0.2) = 0.09950$, $J_1(0.4) = 0.19603$, $J_1(0.6) = 0.28670$, $J_1(0.8) = 0.36884$, $J_1(1.0) = 0.44005$. Use (14) with $n = 5$.

19. **(Notations)** Compute a difference table of $f(x) = x^3$ for $x = 0, 1, 2, 3, 4, 5$. Choose $x_0 = 2$ and write all occurring numbers in terms of the notations (a) for central differences, (b) for forward differences, (c) for backward differences.

20. **CAS EXPERIMENT. Adding Terms in Newton Formulas.** Write a program for the forward formula (14). Experiment on the increase of accuracy by successively adding terms. As data use values of some function of your choice for which your CAS gives the values needed in determining errors.

21. **WRITING PROJECT. Interpolation: Comparison of Methods.** Make a list of 5–6 ideas that you feel are most basic in this section. Arrange them in the best logical order. Discuss them in a 2–3 page report.

22. **TEAM PROJECT. Interpolation and Extrapolation. (a) Lagrange practical error estimate** (after Theorem 1). Apply this to $p_1(9.2)$ and $p_2(9.2)$ for the data $x_0 = 9.0$, $x_1 = 9.5$, $x_2 = 11.0$, $f_0 = \ln x_0$, $f_1 = \ln x_1$, $f_2 = \ln x_2$ (6S-values).

(b) Extrapolation. Given $(x_j, f(x_j)) = (0.2, 0.9980)$, $(0.4, 0.9686)$, $(0.6, 0.8443)$, $(0.8, 0.5358)$, $(1.0, 0)$. Find $f(0.7)$ from the quadratic interpolation polynomials based on (α) 0.6, 0.8, 1.0, (β) 0.4, 0.6, 0.8, (γ) 0.2, 0.4, 0.6. Compare the errors and comment. [Exact $f(x) = \cos(\frac{1}{2}\pi x^2)$, $f(0.7) = 0.7181$ (4S).]

(c) Graph the product of factors $(x - x_j)$ in the error formula (5) for $n = 2, \cdots, 10$ separately. What do these graphs show regarding accuracy of interpolation and extrapolation?

(d) Central differences. Show that $\delta^2 f_m = f_{m+1} - 2f_m + f_{m-1}$, and, furthermore $\delta^3 f_{m+1/2} = f_{m+2} - 3f_{m+1} + 3f_m - f_{m-1}$, $\delta^n f_m = \Delta^n f_{m-n/2} = \nabla^n f_{m+n/2}$.

(e) Everett's interpolation formula

$$f(x) \approx (1 - r)f_0 + rf_1 + \frac{(2 - r)(1 - r)(-r)}{3!}\delta^2 f_0 + \frac{(r + 1)r(r - 1)}{3!}\delta^2 f_1 \tag{20}$$

is an example of a formula involving only even-order differences. Use it to compute the Bessel function $J_0(x)$ for $x = 1.72$ from $J_0(1.60) = 0.455\,4022$ and $J_0(1.7)$, $J_0(1.8)$, $J_0(1.9)$ in Example 6.

19.4 Spline Interpolation

Given data (function values, points in the xy-plane) $(x_0, f_0), (x_1, f_1), \cdots, (x_n, f_n)$ can be interpolated by a polynomial $P_n(x)$ of degree n or less so that the curve of $P_n(x)$ passes through these $n + 1$ points (x_j, f_j); here $f_0 = f(x_0), \cdots, f_n = f(x_n)$. See Sec. 19.3.

Now if n is large, there may be trouble: $P_n(x)$ may tend to oscillate for x between the **nodes** $x_0, \cdots, x_n$. Hence we must be prepared for **numeric instability** (Sec. 19.1). Figure 431 shows a famous example by C. Runge[3] for which the maximum error even approaches ∞ as $n \to \infty$ (with the nodes kept equidistant and their number increased). Figure 432 illustrates the increase of the oscillation with n for some other function that is piecewise linear.

Those undesirable oscillations are avoided by the method of splines initiated by I. J. Schoenberg in 1946 (*Quarterly of Applied Mathematics* **4,** pp. 45–99, 112–141). This method is widely used in practice. It also laid the foundation for much of modern **CAD (computer-aided design).** Its name is borrowed from a *draftman's spline,* which is an elastic rod bent to pass through given points and held in place by weights. The mathematical idea of the method is as follows:

Instead of using a single high-degree polynomial P_n over the entire interval $a \leqq x \leqq b$ in which the nodes lie, that is,

$$a = x_0 < x_1 < \cdots < x_n = b, \tag{1}$$

we use n low-degree, e.g., cubic, polynomials

$$q_0(x), \qquad q_1(x), \qquad \cdots, \qquad q_{n-1}(x),$$

one over each subinterval between adjacent nodes, hence q_0 from x_0 to x_1, then q_1 from

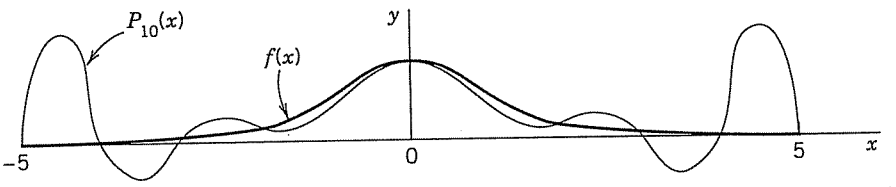


Fig. 431. Runge's example $f(x) = 1/(1 + x^2)$ and interpolating polynomial $P_{10}(x)$

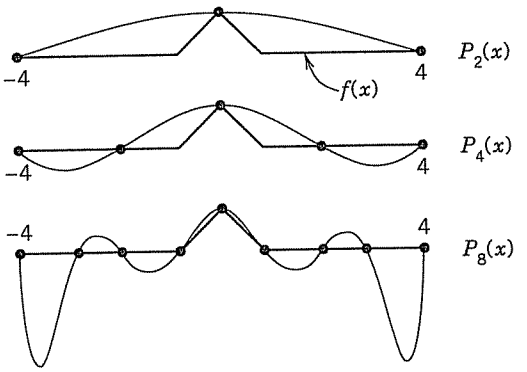


Fig. 432. Piecewise linear function $f(x)$ and interpolation polynomials of increasing degrees

[3]CARL RUNGE (1856–1927), German mathematician, also known for his work on ODEs (Sec. 21.1).

x_1 to x_2, and so on. From this we compose an interpolation function $g(x)$, called a **spline,** by fitting these polynomials together into a single continuous curve passing through the data points, that is,

$$(2)\quad g(x_0) = f(x_0) = f_0, \qquad g(x_1) = f(x_1) = f_1, \qquad \cdots, \qquad g(x_n) = f(x_n) = f_n.$$

Note that $g(x) = q_0(x)$ when $x_0 \leqq x \leqq x_1$, then $g(x) = q_1(x)$ when $x_1 \leqq x \leqq x_2$, and so on, according to our construction of g.

Thus **spline interpolation** *is piecewise polynomial interpolation.*

The simplest q_j's would be linear polynomials. However, the curve of a piecewise linear continuous function has corners and would be of little interest in general—think of designing the body of a car or a ship.

We shall consider cubic splines because these are the most important ones in applications. By definition, a **cubic spline** $g(x)$ interpolating given data $(x_0, f_0), \cdots, (x_n, f_n)$ is a continuous function on the interval $a = x_0 \leqq x \leqq x_n = b$ that has continuous first and second derivatives and satisfies the interpolation condition (2); furthermore, between adjacent nodes, $g(x)$ is given by a polynomial $q_j(x)$ of degree 3 or less.

We claim that there is such a cubic spline. And if in addition to (2) we also require that

$$(3)\qquad g'(x_0) = k_0, \qquad g'(x_n) = k_n$$

(given tangent directions of $g(x)$ at the two endpoints of the interval $a \leqq x \leqq b$), then we have a uniquely determined cubic spline. This is the content of the following existence and uniqueness theorem, whose proof will also suggest the actual determination of splines. (Condition (3) will be discussed after the proof.)

THEOREM 1

Existence and Uniqueness of Cubic Splines

Let $(x_0, f_0), (x_1, f_1), \cdots, (x_n, f_n)$ with arbitrarily spaced given x_j [see (1)] *and given $f_j = f(x_j), j = 0, 1, \cdots, n$. Let k_0 and k_n be any given numbers. Then there is one and only one cubic spline $g(x)$ corresponding to* (1) *and satisfying* (2) *and* (3).

PROOF By definition, on every subinterval I_j given by $x_j \leqq x \leqq x_{j+1}$ the spline $g(x)$ must agree with a polynomial $q_j(x)$ of degree not exceeding 3 such that

$$(4)\qquad q_j(x_j) = f(x_j), \qquad q_j(x_{j+1}) = f(x_{j+1}) \qquad (j = 0, 1, \cdots, n-1).$$

For the derivatives we write

$$(5)\qquad q_j'(x_j) = k_j, \qquad q_j'(x_{j+1}) = k_{j+1} \qquad (j = 0, 1, \cdots, n-1)$$

with k_0 and k_n given and $k_1, \cdots, k_{n-1}$ to be determined later. Equations (4) and (5) are four conditions for each $q_j(x)$. By direct calculation, using the notation

$$(6^*)\qquad c_j = \frac{1}{h_j} = \frac{1}{x_{j+1} - x_j} \qquad (j = 0, 1, \cdots, n-1)$$

we can verify that the unique cubic polynomial $q_j(x)$ $(j = 0, 1, \cdots, n-1)$ satisfying

(4) and (5) is

$$
\begin{aligned}
q_j(x) = {} & f(x_j)c_j^2(x - x_{j+1})^2[1 + 2c_j(x - x_j)] \\
& + f(x_{j+1})c_j^2(x - x_j)^2[1 - 2c_j(x - x_{j+1})] \\
& + k_j c_j^2(x - x_j)(x - x_{j+1})^2 \\
& + k_{j+1}c_j^2(x - x_j)^2(x - x_{j+1}).
\end{aligned} \tag{6}
$$

Differentiating twice, we obtain

$$q_j''(x_j) = -6c_j^2 f(x_j) + 6c_j^2 f(x_{j+1}) - 4c_j k_j - 2c_j k_{j+1} \tag{7}$$

$$q_j''(x_{j+1}) = 6c_j^2 f(x_j) - 6c_j^2 f(x_{j+1}) + 2c_j k_j + 4c_j k_{j+1}. \tag{8}$$

By definition, $g(x)$ has continuous second derivatives. This gives the conditions

$$q_{j-1}''(x_j) = q_j''(x_j) \qquad (j = 1, \cdots, n-1).$$

If we use (8) with j replaced by $j - 1$, and (7), these $n - 1$ equations become

$$c_{j-1}k_{j-1} + 2(c_{j-1} + c_j)k_j + c_j k_{j+1} = 3\left[c_{j-1}^2 \nabla f_j + c_j^2 \nabla f_{j+1}\right] \tag{9}$$

where $\nabla f_j = f(x_j) - f(x_{j-1})$ and $\nabla f_{j+1} = f(x_{j+1}) - f(x_j)$ and $j = 1, \cdots, n-1$, as before. This linear system of $n - 1$ equations has a unique solution $k_1, \cdots, k_{n-1}$ since the coefficient matrix is strictly diagonally dominant (that is, in each row the (positive) diagonal entry is greater than the sum of the other (positive) entries). Hence the determinant of the matrix cannot be zero (as follows from Theorem 3 in Sec. 20.7), so that we may determine unique values $k_1, \cdots, k_{n-1}$ of the first derivative of $g(x)$ at the nodes. This proves the theorem. ∎

Storage and Time Demands in solving (9) are modest, since the matrix of (9) is **sparse** (has few nonzero entries) and **tridiagonal** (may have nonzero entries only on the diagonal and on the two adjacent "parallels" above and below it). Pivoting (Sec. 7.3) is not necessary because of that dominance. This makes splines efficient in solving large problems with thousands of nodes or more. For some literature and some critical comments, see *American Mathematical Monthly* **105** (1998), 929–941.

Condition (3) includes the **clamped conditions**

$$g'(x_0) = f'(x_0), \qquad g'(x_n) = f'(x_n), \tag{10}$$

in which the tangent directions $f'(x_0)$ and $f'(x_n)$ at the ends are given. Other conditions of practical interest are the **free** or **natural conditions**

$$g''(x_0) = 0, \qquad g''(x_n) = 0 \tag{11}$$

(geometrically: zero curvature at the ends, as for the draftman's spline), giving a **natural spline.** These names are motivated by Fig. 290 in Problem Set 12.3.

Determination of Splines. Let k_0 and k_n be given. Obtain $k_1, \cdots, k_{n-1}$ by solving the linear system (9). Recall that the spline $g(x)$ to be found consists of n cubic polynomials $q_0, \cdots, q_{n-1}$. We write these polynomials in the form

$$q_j(x) = a_{j0} + a_{j1}(x - x_j) + a_{j2}(x - x_j)^2 + a_{j3}(x - x_j)^3 \tag{12}$$

where $j = 0, \cdots, n - 1$. Using Taylor's formula, we obtain

$$\begin{aligned}
a_{j0} &= q_j(x_j) = f_j && \text{by (2),} \\
a_{j1} &= q_j'(x_j) = k_j && \text{by (5),} \\
a_{j2} &= \frac{1}{2}\, q_j''(x_j) = \frac{3}{h^2}\,(f_{j+1} - f_j) - \frac{1}{h}\,(k_{j+1} + 2k_j) && \text{by (7),} \\
a_{j3} &= \frac{1}{6}\, q_j'''(x_j) = \frac{2}{h^3}\,(f_j - f_{j+1}) + \frac{1}{h^2}\,(k_{j+1} + k_j)
\end{aligned} \tag{13}$$

with a_{j3} obtained by calculating $q_j''(x_{j+1})$ from (12) and equating the result to (8), that is,

$$q_j''(x_{j+1}) = 2a_{j2} + 6a_{j3}h = \frac{6}{h^2}\,(f_j - f_{j+1}) + \frac{2}{h}\,(k_j + 2k_{j+1}),$$

and now subtracting from this $2a_{j2}$ as given in (13) and simplifying.

Note that for ***equidistant nodes*** of distance $h_j = h$ we can write $c_j = c = 1/h$ in (6*) and have from (9) simply

$$k_{j-1} + 4k_j + k_{j+1} = \frac{3}{h}\,(f_{j+1} - f_{j-1}) \qquad (j = 1, \cdots, n - 1). \tag{14}$$

EXAMPLE 1 **Spline Interpolation. Equidistant Nodes**

Interpolate $f(x) = x^4$ on the interval $-1 \leqq x \leqq 1$ by the cubic spline $g(x)$ corresponding to the nodes $x_0 = -1$, $x_1 = 0$, $x_2 = 1$ and satisfying the clamped conditions $g'(-1) = f'(-1)$, $g'(1) = f'(1)$.

Solution. In our standard notation the given data are $f_0 = f(-1) = 1$, $f_1 = f(0) = 0$, $f_2 = f(1) = 1$. We have $h = 1$ and $n = 2$, so that our spline consists of $n = 2$ polynomials

$$q_0(x) = a_{00} + a_{01}(x + 1) + a_{02}(x + 1)^2 + a_{03}(x + 1)^3 \qquad (-1 \leqq x \leqq 0),$$

$$q_1(x) = a_{10} + a_{11}x + a_{12}x^2 + a_{13}x^3 \qquad (0 \leqq x \leqq 1).$$

We determine the k_j from (14) (equidistance!) and then the coefficients of the spline from (13). Since $n = 2$, the system (14) is a single equation (with $j = 1$ and $h = 1$)

$$k_0 + 4k_1 + k_2 = 3(f_2 - f_0).$$

Here $f_0 = f_2 = 1$ (the value of x^4 at the ends) and $k_0 = -4$, $k_2 = 4$, the values of the derivative $4x^3$ at the ends -1 and 1. Hence

$$-4 + 4k_1 + 4 = 3(1 - 1) = 0, \qquad k_1 = 0.$$

From (13) we can now obtain the coefficients of q_0, namely, $a_{00} = f_0 = 1$, $a_{01} = k_0 = -4$, and

$$a_{02} = \frac{3}{1^2}(f_1 - f_0) - \frac{1}{1}(k_1 + 2k_0) = 3(0 - 1) - (0 - 8) = \quad 5$$

$$a_{03} = \frac{2}{1^3}(f_0 - f_1) + \frac{1}{1^2}(k_1 + k_0) = 2(1 - 0) + (0 - 4) = -2.$$

Similarly, for the coefficients of q_1 we obtain from (13) the values $a_{10} = f_1 = 0$, $a_{11} = k_1 = 0$, and

$$a_{12} = 3(f_2 - f_1) - (k_2 + 2k_1) = 3(1 - 0) - (4 + 0) = -1$$

$$a_{13} = 2(f_1 - f_2) + (k_2 + \ k_1) \ = 2(0 - 1) + (4 + 0) = 2.$$

This gives the polynomials of which the spline $g(x)$ consists, namely,

$$g(x) = \begin{cases} q_0(x) = 1 - 4(x + 1) + 5(x + 1)^2 - 2(x + 1)^3 = -x^2 - 2x^3 & \text{if} \quad -1 \leqq x \leqq 0 \\ q_1(x) = -x^2 + 2x^3 & \text{if} \quad 0 \leqq x \leqq 1. \end{cases}$$

Figure 433 shows $f(x)$ and this spline. Do you see that we could have saved over half of our work by using symmetry?

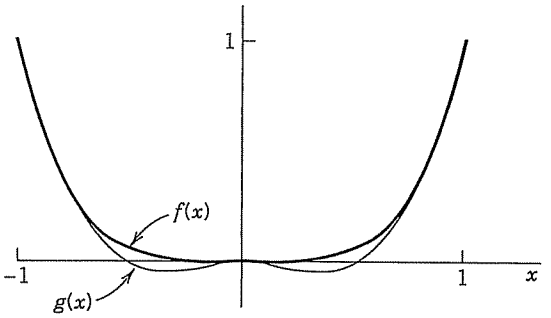


Fig. 433. Function $f(x) = x^4$ and cubic spline $g(x)$ in Example 1

EXAMPLE 2 **Natural Spline. Arbitrarily Spaced Nodes**

Find a spline approximation and a polynomial approximation for the curve of the cross section of the circular-shaped Shrine of the Book in Jerusalem shown in Fig. 434.

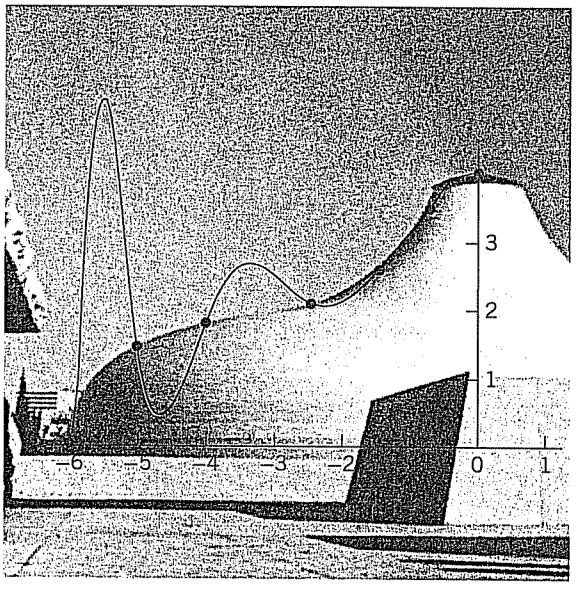


Fig. 434. Shrine of the Book in Jerusalem (Architects F. Kissler and A. M. Bartus)

Solution. Thirteen points, about equally distributed along the contour (not along the x-axis!), give these data:

x_j	−5.8	−5.0	−4.0	−2.5	−1.5	−0.8	0	0.8	1.5	2.5	4.0	5.0	5.8
f_j	0	1.5	1.8	2.2	2.7	3.5	3.9	3.5	2.7	2.2	1.8	1.5	0

The figure shows the corresponding interpolation polynomial of 12th degree, which is useless because of its oscillation. (Because of roundoff your software will also give you small error terms involving odd powers of x.) The polynomial is

$$P_{12}(x) = 3.9000 - 0.65083x^2 + 0.033858x^4 + 0.011041x^6 - 0.0014010x^8 + 0.000055595x^{10} - 0.00000071867x^{12}.$$

The spline follows practically the contour of the roof, with a small error near the nodes −0.8 and 0.8. The spline is symmetric. Its six polynomials corresponding to positive x have the following coefficients of their representations (12). (Note well that (12) is in terms of powers of $x - x_j$, not x!)

j	x-interval	a_{j0}	a_{j1}	a_{j2}	a_{j3}
0	0.0...0.8	3.9	0.00	−0.61	−0.015
1	0.8...1.5	3.5	−1.01	−0.65	0.66
2	1.5...2.5	2.7	−0.95	0.73	−0.27
3	2.5...4.0	2.2	−0.32	−0.091	0.084
4	4.0...5.0	1.8	−0.027	0.29	−0.56
5	5.0...5.8	1.5	−1.13	−1.39	0.58

PROBLEM SET 19.4

1. **WRITING PROJECT. Splines.** In your own words, and using as few formulas as possible, write a short report on spline interpolation, its motivation, a comparison with polynomial interpolation, and its applications.
2. **(Individual polynomial q_j)** Show that $q_j(x)$ in (6) satisfies the interpolation condition (4) as well as the derivative condition (5).
3. Verify the differentiations that give (7) and (8) from (6).
4. **(System for derivatives)** Derive the basic linear system (9) for $k_1, \cdots, k_{n-1}$ as indicated in the text.
5. **(Equidistant nodes)** Derive (14) from (9).
6. **(Coefficients)** Give the details of the derivation of a_{j2} and a_{j3} in (13).
7. Verify the computations in Example 1.
8. **(Comparison)** Compare the spline g in Example 1 with the quadratic interpolation polynomial over the whole interval. Find the maximum deviations of g and p_2 from f. Comment.
9. **(Natural spline condition)** Using the given coefficients, verify that the spline in Example 2 satisfies $g''(x) = 0$ at the ends.

10–16 DETERMINATION OF SPLINES

Find the cubic spline $g(x)$ for the given data with k_0 and k_n as given.

10. $f(-2) = f(-1) = f(1) = f(2) = 0$, $f(0) = 1$, $k_0 = k_4 = 0$
11. If we started from the piecewise linear function in Fig. 435, we would obtain $g(x)$ in Prob. 10 as the spline satisfying $g'(-2) = f'(-2) = 0$, $g'(2) = f'(2) = 0$. Find and sketch or graph the corresponding interpolation polynomial of 4th degree and compare it with the spline. Comment.

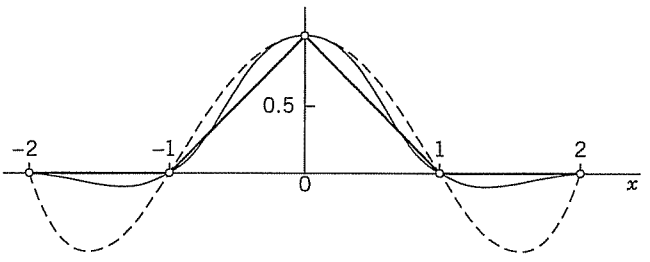


Fig. 435. Spline and interpolation polynomial in Problems 10 and 11

12. $f_0 = f(0) = 1$, $f_1 = f(2) = 9$, $f_2 = f(4) = 41$, $f_3 = f(6) = 41$, $k_0 = 0$, $k_3 = -12$

13. $f_0 = f(-1) = 0$, $f_1 = f(0) = 4$, $f_2 = f(1) = 0$, $k_0 = 0$, $k_2 = 0$. Is $g(x)$ even? (Give reason.)
14. $f_0 = f(0) = 0$, $f_1 = f(1) = 1$, $f_2 = f(2) = 6$, $f_3 = f(3) = 10$, $k_0 = 0$, $k_3 = 0$
15. $f_0 = f(0) = 1$, $f_1 = f(1) = 0$, $f_2 = f(2) = -1$, $f_3 = f(3) = 0$, $k_0 = 0$, $k_3 = -6$
16. It can happen that a spline is given by the same polynomial in two adjacent subintervals. To illustrate this, find the cubic spline $g(x)$ for $f(x) = \sin x$ corresponding to the partition $x_0 = -\pi/2$, $x_1 = 0$, $x_2 = \pi/2$ of the interval $-\pi/2 \leqq x \leqq \pi/2$ and satisfying $g'(-\pi/2) = f'(-\pi/2)$ and $g'(\pi/2) = f'(\pi/2)$.
17. **(Natural conditions)** Explain the remark after (11).
18. **CAS EXPERIMENT. Spline versus Polynomial.** If your CAS gives natural splines, find the natural splines when x is integer from $-m$ to m, and $y(0) = 1$ and all other y equal to 0. Graph each such spline along with the interpolation polynomial p_{2m}. Do this for $m = 2$ to 10 (or more). What happens with increasing m?
19. If a cubic spline is three times continuously differentiable (that is, it has continuous first, second, and third derivatives), show that it must be a single polynomial.
20. **TEAM PROJECT. Hermite Interpolation and Bezier Curves.** In **Hermite interpolation** we are looking for a polynomial $p(x)$ (of degree $2n + 1$ or less) such that $p(x)$ and its derivative $p'(x)$ have given values at $n + 1$ nodes. (More generally, $p(x)$, $p'(x)$, $p''(x)$, $\cdots$ may be required to have given values at the nodes.)

(a) Curves with given endpoints and tangents. Let C be a curve in the xy-plane parametrically represented by $\mathbf{r}(t) = [x(t), y(t)]$, $0 \leqq t \leqq 1$ (see Sec. 9.5). Show that for given initial and terminal points of a curve and given initial and terminal tangents, say,

$$\begin{aligned} A\text{:}\quad \mathbf{r}_0 &= [x(0), y(0)] \\ &= [x_0, y_0], \\ B\text{:}\quad \mathbf{r}_1 &= [x(1), y(1)] \\ &= [x_1, y_1] \\ \mathbf{v}_0 &= [x'(0), y'(0)] \\ &= [x_0', y_0'], \\ \mathbf{v}_1 &= [x'(1), y'(1)] \\ &= [x_1', y_1'] \end{aligned}$$

we can find a curve C, namely,

$$\begin{aligned} \mathbf{r}(t) &= \mathbf{r}_0 + \mathbf{v}_0 t \\ &\quad + (3(\mathbf{r}_1 - \mathbf{r}_0) - (2\mathbf{v}_0 + \mathbf{v}_1))t^2 \\ &\quad + (2(\mathbf{r}_0 - \mathbf{r}_1) + \mathbf{v}_0 + \mathbf{v}_1)t^3; \end{aligned} \tag{15}$$

in components,

$$\begin{aligned} x(t) &= x_0 + x_0' t + (3(x_1 - x_0) - (2x_0' + x_1'))t^2 \\ &\quad + (2(x_0 - x_1) + x_0' + x_1')t^3 \\ y(t) &= y_0 + y_0' t + (3(y_1 - y_0) - (2y_0' + y_1'))t^2 \\ &\quad + (2(y_0 - y_1) + y_0' + y_1')t^3. \end{aligned}$$

Note that this is a cubic Hermite interpolation polynomial, and $n = 1$ because we have two nodes (the endpoints of C). (This has nothing to do with the Hermite polynomials in Sec. 5.8.) The two points

$$\begin{aligned} G_A\text{:}\ \mathbf{g}_0 &= \mathbf{r}_0 + \mathbf{v}_0 \\ &= [x_0 + x_0', y_0 + y_0'] \end{aligned}$$

and

$$\begin{aligned} G_B\text{:}\ \mathbf{g}_1 &= \mathbf{r}_1 - \mathbf{v}_1 \\ &= [x_1 - x_1', y_1 - y_1'] \end{aligned}$$

are called **guidepoints** because the segments AG_A and BG_B specify the tangents graphically. A, B, G_A, G_B determine C, and C can be changed quickly by moving the points. A curve consisting of such Hermite interpolation polynomials is called a **Bezier curve,** after the French engineer P. Bezier of the Renault Automobile Company, who introduced them in the early 1960s in designing car bodies. Bezier curves (and surfaces) are used in computer-aided design (CAD) and computer-aided manufacturing (CAM). (For more details, see Ref. [E21] in App. 1.)

(b) Find and graph the Bezier curve and its guidepoints if A: [0, 0], B: [1, 0], $\mathbf{v}_0 = [\frac{1}{2}, \frac{1}{2}]$, $\mathbf{v}_1 = [-\frac{1}{2}, -\frac{1}{4}\sqrt{3}]$.

(c) Changing guidepoints changes C. Moving guidepoints farther away makes C "staying near the tangents for a longer time." Confirm this by changing $\mathbf{v}_0$ and $\mathbf{v}_1$ in (b) to $2\mathbf{v}_0$ and $2\mathbf{v}_1$ (see Fig. 436).

(d) Make experiments of your own. What happens if you change $\mathbf{v}_1$ in (b) to $-\mathbf{v}_1$. If you rotate the tangents? If you multiply $\mathbf{v}_0$ and $\mathbf{v}_1$ by positive factors less than 1?

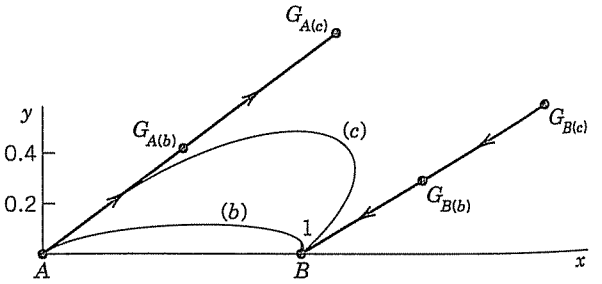


Fig. 436. Team Project 20(b) and (c): Bezier curves

19.5 Numeric Integration and Differentiation

Numeric integration means the numeric evaluation of integrals

$$J = \int_a^b f(x)\,dx$$

where a and b are given and f is a function given analytically by a formula or empirically by a table of values. Geometrically, J is the area under the curve of f between a and b (Fig. 437).

We know that if f is such that we can find a differentiable function F whose derivative is f, then we can evaluate J by applying the familiar formula

$$J = \int_a^b f(x)\,dx = F(b) - F(a) \qquad [F'(x) = f(x)].$$

Tables of integrals or a CAS (Mathematica, Maple, etc.) may be helpful for this purpose.

However, applications often lead to integrals whose analytic evaluation would be very difficult or even impossible, or whose integrand is an empirical function given by recorded numeric values. Then we may obtain approximate numeric values of the integral by a numeric integration method.

Rectangular Rule. Trapezoidal Rule

Numeric integration methods are obtained by approximating the integrand f by functions that can easily be integrated.

The simplest formula, the **rectangular rule,** is obtained if we subdivide the interval of integration $a \leqq x \leqq b$ into n subintervals of equal length $h = (b - a)/n$ and in each subinterval approximate f by the constant $f(x_j^*)$, the value of f at the midpoint x_j^* of the jth subinterval (Fig. 438). Then f is approximated by a **step function** (piecewise constant function), the n rectangles in Fig. 438 have the areas $f(x_1^*)h, \cdots, f(x_n^*)h$, and the **rectangular rule** is

$$\text{(1)} \qquad J = \int_a^b f(x)\,dx \approx h[f(x_1^*) + f(x_2^*) + \cdots + f(x_n^*)] \qquad \left(h = \frac{b-a}{n}\right).$$

The **trapezoidal rule** is generally more accurate. We obtain it if we take the same subdivision as before and approximate f by a broken line of segments (chords) with endpoints $[a, f(a)], [x_1, f(x_1)], \cdots, [b, f(b)]$ on the curve of f (Fig. 439). Then the area under the curve of f between a and b is approximated by n trapezoids of areas

$$\tfrac{1}{2}[f(a) + f(x_1)]h, \qquad \tfrac{1}{2}[f(x_1) + f(x_2)]h, \qquad \cdots, \qquad \tfrac{1}{2}[f(x_{n-1}) + f(b)]h.$$

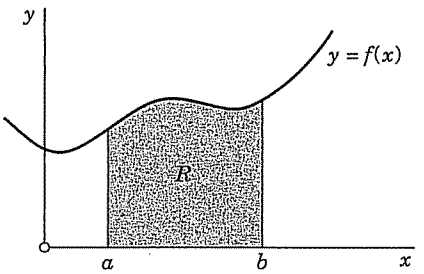


Fig. 437. Geometric interpretation of a definite integral

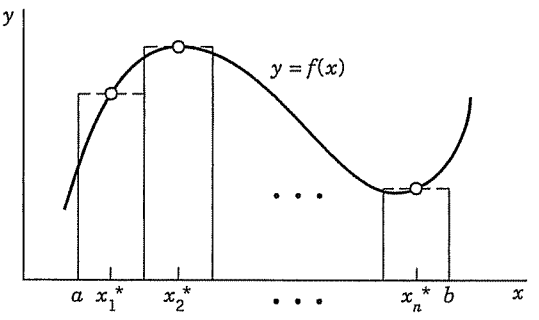


Fig. 438. Rectangular rule

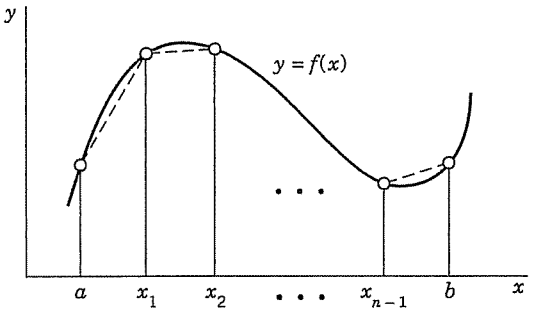


Fig. 439. Trapezoidal rule

By taking their sum we obtain the **trapezoidal rule**

$$(2) \qquad J = \int_a^b f(x)\,dx \approx h\left[\tfrac{1}{2}f(a) + f(x_1) + f(x_2) + \cdots + f(x_{n-1}) + \tfrac{1}{2}f(b)\right]$$

where $h = (b - a)/n$, as in (1). The x_j's and a and b are called **nodes.**

EXAMPLE 1 **Trapezoidal Rule**

Evaluate $J = \int_0^1 e^{-x^2}\,dx$ by means of (2) with $n = 10$.

Solution. $J \approx 0.1(0.5 \cdot 1.367\,879 + 6.778\,167) = 0.746\,211$ from Table 19.3. ∎

Table 19.3 Computations in Example 1

j	x_j	x_j^2	$e^{-x_j^2}$	
0	0	0	1.000 000	
1	0.1	0.01		0.990 050
2	0.2	0.04		0.960 789
3	0.3	0.09		0.913 931
4	0.4	0.16		0.852 144
5	0.5	0.25		0.778 801
6	0.6	0.36		0.697 676
7	0.7	0.49		0.612 626
8	0.8	0.64		0.527 292
9	0.9	0.81		0.444 858
10	1.0	1.00	0.367 879	
Sums			1.367 879	6.778 167

Error Bounds and Estimate for the Trapezoidal Rule

An error estimate for the trapezoidal rule can be derived from (5) in Sec. 19.3 with $n = 1$ by integration as follows. For a single subinterval we have

$$f(x) - p_1(x) = (x - x_0)(x - x_1)\,\frac{f''(t)}{2}$$

with a suitable t depending on x, between x_0 and x_1. Integration over x from $a = x_0$ to $x_1 = x_0 + h$ gives

$$\int_{x_0}^{x_0+h} f(x)\,dx - \frac{h}{2}\,[f(x_0) + f(x_1)] = \int_{x_0}^{x_0+h} (x - x_0)(x - x_0 - h)\,\frac{f''(t(x))}{2}\,dx.$$

Setting $x - x_0 = v$ and applying the mean value theorem of integral calculus, which we can use because $(x - x_0)(x - x_0 - h)$ does not change sign, we find that the right side equals

$$\text{(3*)} \qquad \int_0^h v(v - h)\,dv\,\frac{f''(\tilde{t})}{2} = \left(\frac{h^3}{3} - \frac{h^3}{2}\right)\frac{f''(\tilde{t})}{2} = -\frac{h^3}{12}\,f''(\tilde{t})$$

where $\tilde{t}$ is a (suitable, unknown) value between x_0 and x_1. This is the error for the trapezoidal rule with $n = 1$, often called the **local error.**

Hence the **error** ϵ of (2) with any n is the sum of such contributions from the n subintervals; since $h = (b - a)/n$, $nh^3 = n(b - a)^3/n^3$, and $(b - a)^2 = n^2h^2$, we obtain

$$\text{(3)} \qquad \epsilon = -\frac{(b - a)^3}{12n^2}\,f''(\hat{t}) = -\frac{(b - a)}{12}\,h^2 f''(\hat{t})$$

with (suitable, unknown) $\hat{t}$ between a and b.

Because of (3) the trapezoidal rule (2) is also written

$$\text{(2*)} \quad J = \int_a^b f(x)\,dx = h\left[\tfrac{1}{2}f(a) + f(x_1) + \cdots + f(x_{n-1}) + \tfrac{1}{2}f(b)\right] - \frac{b - a}{12}\,h^2 f''(\hat{t}).$$

Error Bounds are now obtained by taking the largest value for f'', say, M_2, and the smallest value, $M_2{}^*$, in the interval of integration. Then (3) gives (note that K is negative)

$$\text{(4)} \qquad KM_2 \leqq \epsilon \leqq KM_2{}^* \qquad \text{where} \qquad K = -\frac{(b - a)^3}{12n^2} = -\frac{b - a}{12}\,h^2.$$

Error Estimation by Halving h is advisable if h'' is very complicated or unknown, for instance, in the case of experimental data. Then we may apply the Error Principle of Sec. 19.1. That is, we calculate by (2), first with h, obtaining, say, $J = J_h + \epsilon_h$, and then with $\frac{1}{2}h$, obtaining $J = J_{h/2} + \epsilon_{h/2}$. Now if we replace h^2 in (3) with $(\frac{1}{2}h)^2$, the error is multiplied by 1/4. Hence $\epsilon_{h/2} \approx \frac{1}{4}\epsilon_h$ (not exactly because $\hat{t}$ may differ). Together, $J_{h/2} + \epsilon_{h/2} = J_h + \epsilon_h \approx J_h + 4\epsilon_{h/2}$. Thus $J_{h/2} - J_h = (4 - 1)\epsilon_{h/2}$. Division by 3 gives the error formula for $J_{h/2}$

$$\text{(5)} \qquad \epsilon_{h/2} \approx \frac{1}{3}\,(J_{h/2} - J_h).$$

EXAMPLE 2 Error Estimation for the Trapezoidal Rule by (4) and (5)

Estimate the error of the approximate value in Example 1 by (4) and (5).

Solution. (A) ***Error bounds by*** **(4).** By differentiation, $f''(x) = 2(2x^2 - 1)e^{-x^2}$. Also, $f'''(x) > 0$ if $0 < x < 1$, so that the minimum and maximum occur at the ends of the interval. We compute

$M_2 = f''(1) = 0.735\,759$ and $M_2^* = f''(0) = -2$. Furthermore, $K = -1/1200$, and (4) gives

$$-0.000\,614 \leqq \epsilon \leqq 0.001\,667.$$

Hence the exact value of J must lie between

$$0.746\,211 - 0.000\,614 = 0.745\,597 \qquad \text{and} \qquad 0.746\,211 + 0.001\,667 = 0.747\,878.$$

Actually, $J = 0.746\,824$, exact to 6D.

(B) *Error estimate by* (5). $J_h = 0.746211$ in Example 1. Also,

$$J_{h/2} = 0.05\left[\sum_{j=1}^{19} e^{-(j/20)^2} + \frac{1}{2}(1 + 0.367879)\right] = 0.746671.$$

Hence $\epsilon_{h/2} = \frac{1}{3}(J_{h/2} - J_h) = 0.000153$ and $J_{h/2} + \epsilon_{h/2} = 0.746824$, exact to 6D. ■

Simpson's Rule of Integration

Piecewise constant approximation of f led to the rectangular rule (1), piecewise linear approximation to the trapezoidal rule (2), and piecewise quadratic approximation will lead to Simpson's rule, which is of great practical importance because it is sufficiently accurate for most problems, but still sufficiently simple.

To derive Simpson's rule, we divide the interval of integration $a \leqq x \leqq b$ into an ***even number*** of equal subintervals, say, into $n = 2m$ subintervals of length $h = (b - a)/(2m)$, with endpoints $x_0\ (= a), x_1, \cdots, x_{2m-1}, x_{2m}\ (= b)$; see Fig. 440. We now take the first two subintervals and approximate $f(x)$ in the interval $x_0 \leqq x \leqq x_2 = x_0 + 2h$ by the Lagrange polynomial $p_2(x)$ through $(x_0, f_0), (x_1, f_1), (x_2, f_2)$, where $f_j = f(x_j)$. From (3) in Sec. 19.3 we obtain

$$(6)\quad p_2(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} f_0 + \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} f_1 + \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} f_2.$$

The denominators in (6) are $2h^2$, $-h^2$, and $2h^2$, respectively. Setting $s = (x - x_1)/h$, we have

$$x - x_1 = sh, \qquad x - x_0 = x - (x_1 - h) = (s + 1)h$$
$$x - x_2 = x - (x_1 + h) = (s - 1)h$$

and we obtain

$$p_2(x) = \tfrac{1}{2}s(s - 1)f_0 - (s + 1)(s - 1)f_1 + \tfrac{1}{2}(s + 1)sf_2.$$

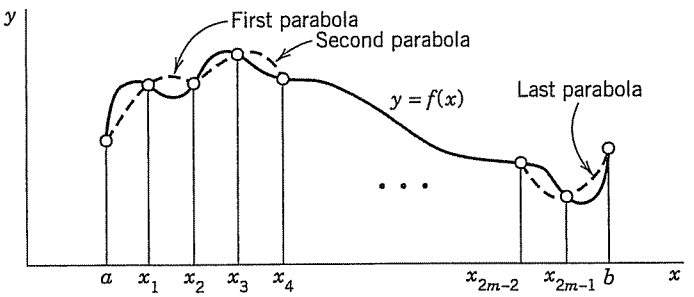


Fig. 440. Simpson's rule

We now integrate with respect to x from x_0 to x_2. This corresponds to integrating with respect to s from -1 to 1. Since $dx = h\,ds$, the result is

$$(7^*) \qquad \int_{x_0}^{x_2} f(x)\,dx \approx \int_{x_0}^{x_2} p_2(x)\,dx = h\left(\frac{1}{3}f_0 + \frac{4}{3}f_1 + \frac{1}{3}f_2\right).$$

A similar formula holds for the next two subintervals from x_2 to x_4, and so on. By summing all these m formulas we obtain **Simpson's rule**[4]

$$(7) \qquad \int_a^b f(x)\,dx \approx \frac{h}{3}(f_0 + 4f_1 + 2f_2 + 4f_3 + \cdots + 2f_{2m-2} + 4f_{2m-1} + f_{2m}),$$

where $h = (b-a)/(2m)$ and $f_j = f(x_j)$. Table 19.4 shows an algorithm for Simpson's rule.

Table 19.4 Simpson's Rule of Integration

ALGORITHM SIMPSON $(a, b, m, f_0, f_1, \cdots, f_{2m})$

This algorithm computes the integral $J = \int_a^b f(x)\,dx$ from given values $f_j = f(x_j)$ at equidistant $x_0 = a$, $x_1 = x_0 + h, \cdots, x_{2m} = x_0 + 2mh = b$ by Simpson's rule (7), where $h = (b-a)/(2m)$.

INPUT: $a, b, m, f_0, \cdots, f_{2m}$

OUTPUT: Approximate value $\tilde{J}$ of J

Compute $s_0 = f_0 + f_{2m}$

$s_1 = f_1 + f_3 + \cdots + f_{2m-1}$

$s_2 = f_2 + f_4 + \cdots + f_{2m-2}$

$h = (b-a)/2m$

$\tilde{J} = \frac{h}{3}(s_0 + 4s_1 + 2s_2)$

OUTPUT $\tilde{J}$. Stop.

End SIMPSON

Error of Simpson's Rule (7). If the fourth derivative $f^{(4)}$ exists and is continuous on $a \leqq x \leqq b$, the **error** of (7), call it ϵ_S, is

$$(8) \qquad \epsilon_S = -\frac{(b-a)^5}{180(2m)^4} f^{(4)}(\hat{t}) = -\frac{(b-a)}{180} h^4 f^{(4)}(\hat{t});$$

[4]THOMAS SIMPSON (1710–1761), self-taught English mathematician, author of several popular textbooks. Simpson's rule was used much earlier by Torricelli, Gregory (in 1668), and Newton (in 1676).

here $\hat{t}$ is a suitable unknown value between a and b. This is obtained similarly to (3). With this we may also write Simpson's rule (7) as

$$(7^{**}) \qquad \int_a^b f(x)\,dx = \frac{h}{3}\,(f_0 + 4f_1 + \cdots + f_{2m}) - \frac{(b-a)}{180}\,h^4 f^{(4)}(\hat{t}).$$

Error Bounds. By taking for $f^{(4)}$ in (8) the maximum M_4 and minimum M_4^* on the interval of integration we obtain from (8) the error bounds (note that C is negative)

$$(9) \qquad CM_4 \leqq \epsilon_S \leqq CM_4^* \qquad \text{where} \qquad C = -\frac{(b-a)^5}{180(2m)^4} = -\frac{(b-a)}{180}\,h^4.$$

Degree of Precision (DP) *of an integration formula.* This is the maximum degree of arbitrary polynomials for which the formula gives exact values of integrals over any intervals.

Hence for the trapezoidal rule,

$$\text{DP} = 1$$

because we approximate the curve of f by portions of straight lines (linear polynomials).

For Simpson's rule we might expect DP = 2 (why?). Actually,

$$\text{DP} = 3$$

by (9) because $f^{(4)}$ is identically zero for a cubic polynomial. This makes Simpson's rule sufficiently accurate for most practical problems and accounts for its popularity.

Numeric Stability *with respect to rounding* is another important property of Simpson's rule. Indeed, for the sum of the roundoff errors ϵ_j of the $2m + 1$ values f_j in (7) we obtain, since $h = (b - a)/2m$,

$$\frac{h}{3}\,|\epsilon_0 + 4\epsilon_1 + \cdots + \epsilon_{2m}| \leqq \frac{(b-a)}{3 \cdot 2m}\,6mu = (b-a)u$$

where u is the rounding unit ($u = \frac{1}{2} \cdot 10^{-6}$ if we round off to 6D; see Sec. 19.1). Also $6 = 1 + 4 + 1$ is the sum of the coefficients for a pair of intervals in (7); take $m = 1$ in (7) to see this. The bound $(b - a)u$ is independent of m, so that it cannot increase with increasing m, that is, with decreasing h. This proves stability. ∎

Newton–Cotes Formulas. We mention that the trapezoidal and Simpson rules are special *closed Newton–Cotes formulas,* that is, integration formulas in which $f(x)$ is interpolated at equally spaced nodes by a polynomial of degree n ($n = 1$ for trapezoidal, $n = 2$ for Simpson), and **closed** means that a and b are nodes ($a = x_0$, $b = x_n$). $n = 3$ (the three-eighths rule; Review Prob. 33) and a higher n are used occasionally. From $n = 8$ on, some of the coefficients become negative, so that a positive f_j could make a negative contribution to an integral, which is absurd. For more on this topic see Ref. [E25] in App. 1.

EXAMPLE 3 Simpson's Rule. Error Estimate

Evaluate $J = \int_0^1 e^{-x^2}\,dx$ by Simpson's rule with $2m = 10$ and estimate the error.

Solution. Since $h = 0.1$, Table 19.5 gives

$$J \approx \frac{0.1}{3}\,(1.367\,879 + 4 \cdot 3.740\,266 + 2 \cdot 3.037\,901) = 0.746\,825.$$

Estimate of error. Differentiation gives $f^{(4)}(x) = 4(4x^4 - 12x^2 + 3)e^{-x^2}$. By considering the derivative $f^{(5)}$ of $f^{(4)}$ we find that the largest value of $f^{(4)}$ in the interval of integration occurs at 0 and the smallest value at $x^* = (2.5 - 0.5\sqrt{10})^{1/2}$. Computation gives the values $M_4 = f^{(4)}(0) = 12$ and $M_4{}^* = f^{(4)}(x^*) = -7.419$. Since $2m = 10$ and $b - a = 1$, we obtain $C = -1/1\,800\,000 = -0.000\,000\,56$. Therefore, from (9),

$$-0.000\,007 \leqq \epsilon_S \leqq 0.000\,005.$$

Hence J must lie between $0.746\,825 - 0.000\,007 = 0.746\,818$ and $0.746\,825 + 0.000\,005 = 0.746\,830$, so that at least four digits of our approximate value are exact. Actually, the value 0.746 825 is exact to 5D because $J = 0.746\,824$ (exact to 6D).

Thus our result is much better than that in Example 1 obtained by the trapezoidal rule, whereas the number of operations is nearly the same in both cases. ▩

Table 19.5 Computations in Example 3

j	x_j	$x_j{}^2$		$e^{-x_j{}^2}$	
0	0	0	1.000 000		
1	0.1	0.01		0.990 050	
2	0.2	0.04			0.960 789
3	0.3	0.09		0.913 931	
4	0.4	0.16			0.852 144
5	0.5	0.25		0.778 801	
6	0.6	0.36			0.697 676
7	0.7	0.49		0.612 626	
8	0.8	0.64			0.527 292
9	0.9	0.81		0.444 858	
10	1.0	1.00	0.367 879		
Sums			1.367 879	3.740 266	3.037 901

Instead of picking an $n = 2m$ and then estimating the error by (9), as in Example 3, it is better to require an accuracy (e.g., 6D) and then determine $n = 2m$ from (9).

EXAMPLE 4 Determination of $n = 2m$ in Simpson's Rule from the Required Accuracy

What n should we choose in Example 3 to get 6D-accuracy?

Solution. Using $M_4 = 12$ (which is bigger in absolute value than $M_4{}^*$), we get from (9), with $b - a = 1$ and the required accuracy,

$$|CM_4| = \frac{12}{180(2m)^4} = \frac{1}{2} \cdot 10^{-6}, \qquad \text{thus} \qquad m = \left[\frac{2 \cdot 10^6 \cdot 12}{180 \cdot 2^4}\right]^{1/4} = 9.55.$$

Hence we should choose $n = 2m = 20$. Do the computation, which parallels that in Example 3.

Note that the error bounds in (4) or (9) may sometimes be loose, so that in such a case a smaller $n = 2m$ may already suffice. ▩

Error Estimation for Simpson's Rule by Halving h. The idea is the same as in (5) and gives

$$\epsilon_{h/2} \approx \frac{1}{15}(J_{h/2} - J_h). \tag{10}$$

J_h is obtained by using h and $J_{h/2}$ by using $\frac{1}{2}h$, and $\epsilon_{h/2}$ is the error of $J_{h/2}$.

Derivation. In (5) we had $\frac{1}{3}$ as the reciprocal of $3 = 4 - 1$ and $\frac{1}{4} = (\frac{1}{2})^2$ resulted from h^2 in (3) by replacing h with $\frac{1}{2}h$. In (10) we have $\frac{1}{15}$ as the reciprocal of $15 = 16 - 1$ and $\frac{1}{16} = (\frac{1}{2})^4$ results from h^4 in (8) by replacing h with $\frac{1}{2}h$.

EXAMPLE 5 Error Estimation for Simpson's Rule by Halving

Integrate $f(x) = \frac{1}{4}\pi x^4 \cos\frac{1}{4}\pi x$ from 0 to 2 with $h = 1$ and apply (10).

Solution. The exact 5D-value of the integral is $J = 1.25953$. Simpson's rule gives

$$J_h = \tfrac{1}{3}[f(0) + 4f(1) + f(2)] = \tfrac{1}{3}(0 + 4 \cdot 0.555360 + 0) = 0.740480,$$

$$J_{h/2} = \frac{1}{6}\left[f(0) + 4f\left(\frac{1}{2}\right) + 2f(1) + 4f\left(\frac{3}{2}\right) + f(2)\right]$$

$$= \frac{1}{6}[0 + 4 \cdot 0.045351 + 2 \cdot 0.555361 + 4 \cdot 1.521579 + 0] = 1.22974.$$

Hence (10) gives $\epsilon_{h/2} = \frac{1}{15}(1.22974 - 0.74048) = 0.032617$ and thus $J \approx J_{h/2} + \epsilon_{h/2} = 1.26236$, with an error -0.00283, which is less in absolute value than $\frac{1}{10}$ of the error 0.02979 of $J_{h/2}$. Hence the use of (10) was well worthwhile. ■

Adaptive Integration

The idea is to adapt step h to the variability of $f(x)$. That is, where f varies but little, we can proceed in large steps without causing a substantial error in the integral, but where f varies rapidly, we have to take small steps in order to stay everywhere close enough to the curve of f.

Changing h is done systematically, usually by halving h, and automatically (not "by hand") depending on the size of the (estimated) error over a subinterval. The subinterval is halved if the corresponding error is still too large, that is, larger than a given **tolerance** TOL (maximum admissible absolute error), or is not halved if the error is less than or equal to TOL.

Adapting is one of the techniques typical of modern software. In connection with integration it can be applied to various methods. We explain it here for Simpson's rule. In Table 19.6 a star means that for that subinterval, TOL has been reached.

EXAMPLE 6 Adaptive Integration with Simpson's Rule

Integrate $f(x) = \frac{1}{4}\pi x^4 \cos\frac{1}{4}\pi x$ from $x = 0$ to 2 by adaptive integration and with Simpson's rule and TOL[0, 2] = 0.0002.

Solution. Table 19.6 shows the calculations. Figure 441 shows the integrand $f(x)$ and the adapted intervals used. The first two intervals ([0, 0.5], [0.5, 1.0]) have length 0.5, hence $h = 0.25$ [because we use $2m = 2$ subintervals in Simpson's rule (7**)]. The next two intervals ([1.00, 1.25], [1.25, 1.50]) have length 0.25 (hence $h = 0.125$) and the last four intervals have length 0.125. *Sample computations.* For 0.740480 see Example 5. Formula (10) gives $(0.123716 - 0.122794)/15 = 0.000061$. Note that 0.123716 refers to [0, 0.5] and [0.5, 1], so that we must subtract the value corresponding to [0, 1] in the line before. Etc. TOL[0, 2] = 0.0002 gives

0.0001 for subintervals of length 1, 0.00005 for length 0.5, etc. The value of the integral obtained is the sum of the values marked by an asterisk (for which the error estimate has become less than TOL). This gives

$$J \approx 0.123716 + 0.528895 + 0.388263 + 0.218483 = 1.25936.$$

The exact 5D-value is $J = 1.25953$. Hence the error is 0.00017. This is about 1/200 of the absolute value of that in Example 5. Our more extensive computation has produced a much better result. ▩

Table 19.6 Computations in Example 6

Interval	Integral	Error (10)	TOL	Comment
[0, 2]	0.740480		0.0002	
[0, 1]	0.122794			
[1, 2]	1.10695			
	Sum = 1.22974	0.032617	0.0002	Divide further
[0.0, 0.5]	0.004782			
[0.5, 1.0]	0.118934			
	Sum = 0.123716*	0.000061	0.0001	TOL reached
[1.0, 1.5]	0.528176			
[1.5, 2.0]	0.605821			
	Sum = 1.13300	0.001803	0.0001	Divide further
[1.00, 1.25]	0.200544			
[1.25, 1.50]	0.328351			
	Sum = 0.528895*	0.000048	0.00005	TOL reached
[1.50, 1.75]	0.388235			
[1.75, 2.00]	0.218457			
	Sum = 0.606692	0.000058	0.00005	Divide further
[1.500, 1.625]	0.196244			
[1.625, 1.750]	0.192019			
	Sum = 0.388263*	0.000002	0.000025	TOL reached
[1.750, 1.875]	0.153405			
[1.875, 2.000]	0.065078			
	Sum = 0.218483*	0.000002	0.000025	TOL reached

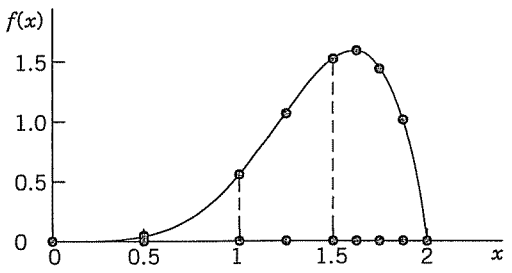


Fig. 441. Adaptive integration in Example 6

Gauss Integration Formulas Maximum Degree of Precision

Our integration formulas discussed so far use function values at *predetermined* (equidistant) x-values (nodes) and give exact results for polynomials not exceeding a certain degree [called the *degree of precision;* see after (9)]. But we can get much more accurate integration formulas as follows. We set

$$\int_{-1}^{1} f(t)\,dt \approx \sum_{j=1}^{n} A_j f_j \qquad [f_j = f(t_j)] \tag{11}$$

with fixed n, and $t = \pm 1$ obtained from $x = a, b$ by setting $x = \frac{1}{2}[a(t-1) + b(t+1)]$. Then we determine the n coefficients $A_1, \cdots, A_n$ and n nodes $t_1, \cdots, t_n$ so that (11) gives exact results for polynomials of degree k as high as possible. Since $n + n = 2n$ is the number of coefficients of a polynomial of degree $2n - 1$, it follows that $k \leqq 2n - 1$.

Gauss has shown that exactness for polynomials of degree not exceeding $2n - 1$ (instead of $n - 1$ for predetermined nodes) can be attained, and he has given the location of the t_j (= the jth zero of the Legendre polynomial P_n in Sec. 5.3) and the coefficients A_j which depend on n but not on $f(t)$, and are obtained by using Lagrange's interpolation polynomial, as shown in Ref. [E5] listed in App. 1. With these t_j and A_j, formula (11) is called a **Gauss integration formula** or *Gauss quadrature formula.* Its degree of precision is $2n - 1$, as just explained. Table 19.7 gives the values needed for $n = 2, \cdots, 5$. (For larger n, see pp. 916–919 of Ref. [GR1] in App. 1.)

Table 19.7 Gauss Integration: Nodes t_j and Coefficients A_j

n	Nodes t_j	Coefficients A_j	Degree of Precision
2	−0.57735 02692	1	3
	0.57735 02692	1	
3	−0.77459 66692	0.55555 55556	5
	0	0.88888 88889	
	0.77459 66692	0.55555 55556	
4	−0.86113 63116	0.34785 48451	7
	−0.33998 10436	0.65214 51549	
	0.33998 10436	0.65214 51549	
	0.86113 63116	0.34785 48451	
5	−0.90617 98459	0.23692 68851	9
	−0.53846 93101	0.47862 86705	
	0	0.56888 88889	
	0.53846 93101	0.47862 86705	
	0.90617 98459	0.23692 68851	

EXAMPLE 7 Gauss Integration Formula with $n = 3$

Evaluate the integral in Example 3 by the Gauss integration formula (11) with $n = 3$.

Solution. We have to convert our integral from 0 to 1 into an integral from −1 to 1. We set $x = \frac{1}{2}(t + 1)$. Then $dx = \frac{1}{2}\,dt$, and (11) with $n = 3$ and the above values of the nodes and the coefficients yields

$$\int_0^1 \exp(-x^2)\,dx = \frac{1}{2}\int_{-1}^1 \exp\left(-\frac{1}{4}(t+1)^2\right)dt$$

$$\approx \frac{1}{2}\left[\frac{5}{9}\exp\left(-\frac{1}{4}\left(1-\sqrt{\frac{3}{5}}\right)^2\right) + \frac{8}{9}\exp\left(-\frac{1}{4}\right) + \frac{5}{9}\exp\left(-\frac{1}{4}\left(1+\sqrt{\frac{3}{5}}\right)^2\right)\right] = 0.746\,815$$

(exact to 6D: 0.746 825), which is almost as accurate as the Simpson result obtained in Example 3 with a much larger number of arithmetic operations. With 3 function values (as in this example) and Simpson's rule we would get $\frac{1}{6}(1 + 4e^{-0.25} + e^{-1}) = 0.747\,180$, with an error over 30 times that of the Gauss integration. ■

EXAMPLE 8 Gauss Integration Formula with $n = 4$ and 5

Integrate $f(x) = \frac{1}{4}\pi x^4 \cos\frac{1}{4}\pi x$ from $x = 0$ to 2 by Gauss. Compare with the adaptive integration in Example 6 and comment.

Solution. $x = t + 1$ gives $f(t) = \frac{1}{4}\pi(t+1)^4 \cos(\frac{1}{4}\pi(t+1))$, as needed in (11). For $n = 4$ we calculate (6S)

$$J \approx A_1 f_1 + \cdots + A_4 f_4 = A_1(f_1 + f_4) + A_2(f_2 + f_3)$$

$$= 0.347855(0.000290309 + 1.02570) + 0.652145(0.129464 + 1.25459) = 1.25950.$$

The error is 0.00003 because $J = 1.25953$ (6S). Calculating with 10S and $n = 4$ gives the same result; so the error is due to the formula, not rounding. For $n = 5$ and 10S we get $J \approx 1.25952\,6185$, too large by the amount 0.00000 0250 because $J = 1.25952\,5935$ (10S). The accuracy is impressive, particularly if we compare the amount of work with that in Example 6. ■

Gauss integration is of considerable practical importance. Whenever the integrand f is given by a formula (not just by a table of numbers) or when experimental measurements can be set at times t_j (or whatever t represents) shown in Table 19.7 or in Ref. [GR1], then the great accuracy of Gauss integration outweighs the disadvantage of the complicated t_j and A_j (which may have to be stored). Also, Gauss coefficients A_j are positive for all n, in contrast with some of the Newton–Cotes coefficients for larger n.

Of course, there are frequent applications with equally spaced nodes, so that Gauss integration does not apply (or has no great advantage if one first has to get the t_j in (11) by interpolation).

Since the endpoints -1 and 1 of the interval of integration in (11) are not zeros of P_n, they do not occur among $t_0, \cdots, t_n$, and the Gauss formula (11) is called, therefore, an **open formula,** in contrast with a **closed formula,** in which the endpoints of the interval of integration are t_0 and t_n. [For example, (2) and (7) are closed formulas.]

Numeric Differentiation

Numeric differentiation is the computation of values of the derivative of a function f from given values of f. Numeric differentiation should be avoided whenever possible, because, whereas *integration* is a *smoothing* process and is not affected much by small inaccuracies in function values, *differentiation* tends to make matters *rough* and generally gives values of f' much less accurate than those of f—remember that the derivative is the limit of the difference quotient, and in the latter you usually have a small difference of large quantities that you then divide by a small quantity. However, the formulas to be obtained will be basic in the numeric solution of differential equations.

We use the notations $f_j' = f'(x_j)$, $f_j'' = f''(x_j)$, etc., and may obtain rough approximation formulas for derivatives by remembering that

$$f'(x) = \lim_{h\to 0}\frac{f(x+h) - f(x)}{h}.$$

This suggests

(12) $$f'_{1/2} \approx \frac{\delta f_{1/2}}{h} = \frac{f_1 - f_0}{h}.$$

Similarly, for the second derivative we obtain

(13) $$f''_1 \approx \frac{\delta^2 f_1}{h^2} = \frac{f_2 - 2f_1 + f_0}{h^2}, \qquad \text{etc.}$$

More accurate approximations are obtained by differentiating suitable Lagrange polynomials. Differentiating (6) and remembering that the denominators in (6) are $2h^2$, $-h^2$, $2h^2$, we have

$$f'(x) \approx p'_2(x) = \frac{2x - x_1 - x_2}{2h^2} f_0 - \frac{2x - x_0 - x_2}{h^2} f_1 + \frac{2x - x_0 - x_1}{2h^2} f_2.$$

Evaluating this at x_0, x_1, x_2, we obtain the "three-point formulas"

(14)
$$\text{(a)} \quad f'_0 \approx \frac{1}{2h}(-3f_0 + 4f_1 - f_2),$$
$$\text{(b)} \quad f'_1 \approx \frac{1}{2h}(-f_0 + f_2),$$
$$\text{(c)} \quad f'_2 \approx \frac{1}{2h}(f_0 - 4f_1 + 3f_2).$$

Applying the same idea to the Lagrange polynomial $p_4(x)$, we obtain similar formulas, in particular,

(15) $$f'_2 \approx \frac{1}{12h}(f_0 - 8f_1 + 8f_3 - f_4).$$

Some examples and further formulas are included in the problem set as well as in Ref. [E5] listed in App. 1.

PROBLEM SET 19

1. **(Rectangular rule)** Evaluate the integral in Example 1 by the rectangular rule (1) with a subinterval of length 0.1.
2. Derive a formula for lower and upper bounds for the rectangular rule and apply it to Prob. 1.

3–8 TRAPEZOIDAL AND SIMPSON'S RULES

Evaluate the integrals numerically as indicated and determine the error by using an integration formula known from calculus.

$$F(x) = \int_1^x \frac{dx^*}{x^*}, \qquad G(x) \int_0^x \frac{dx^*}{\cos^2 x^*},$$

$$H(x) = \int_0^x x^* e^{-x^*}\, dx^*$$

3. $F(2)$ by (2), $n = 10$
4. $F(2)$ by (7), $n = 10$
5. $G(1)$ by (2), $n = 10$
6. $G(1)$ by (7), $n = 10$
7. $H(4)$ by (2), $n = 10$
8. $H(4)$ by (7), $n = 10$

9–12 HALVING

Estimate the error by halving.

9. In Prob. 5
10. In Prob. 6
11. In Prob. 7
12. In Prob. 8

13–19 NONELEMENTARY INTEGRALS

The following integrals cannot be evaluated by the usual methods of calculus. Evaluate them as indicated.

$$\mathrm{Si}(x) = \int_0^x \frac{\sin x^*}{x^*}\, dx^*,$$

$$\mathrm{S}(x) = \int_0^x \sin\left(x^{*2}\right) dx^*, \quad \mathrm{C}(x) = \int_0^x \cos\left(x^{*2}\right) dx^*$$

$\mathrm{Si}(x)$ is the **sine integral.** $\mathrm{S}(x)$ and $\mathrm{C}(x)$ are the **Fresnel integrals.** (See App. 3.1.)

13. Si(1) by (2), $n = 5$, $n = 10$

14. Using the values in Prob. 13, obtain a better value for Si(1). *Hint.* Use (5).

15. Si(1) by (7), $2m = 2$, $2m = 4$

16. Obtain a better value in Prob. 15. *Hint.* Use (10).

17. Si(1) by (7), $2m = 10$

18. S(1.25) by (7), $2m = 10$

19. C(1.25) by (7), $2m = 10$

20. (Stability) Prove that the trapezoidal rule is stable with respect to rounding.

21–24 GAUSS INTEGRATION

Integrate by (11) with $n = 5$:

21. $1/x$ from 1 to 3

22. $\cos x$ from 0 to $\frac{1}{2}\pi$

23. e^{-x^2} from 0 to 1

24. $\sin(x^2)$ from 0 to 1.25

25. (Given TOL) Find the smallest n in computing the integral of $1/x$ from 1 to 2 for which 5D-accuracy is guaranteed (a) by (4) in the use of (2), (b) by (9) in the use of (7). Compare and comment.

26. TEAM PROJECT. Romberg Integration (W. Romberg, *Norske Videnskab. Trondheim, Førh.* 28, Nr. 7, 1955). This method uses the trapezoidal rule and gains precision stepwise by halving h and adding an error estimate. Do this for the integral of $f(x) = e^{-x}$ from $x = 0$ to $x = 2$ with TOL $= 10^{-3}$, as follows.

Step 1. Apply the trapezoidal rule (2) with $h = 2$ (hence $n = 1$) to get an approximation J_{11}. Halve h and use (2) to get J_{21} and an error estimate

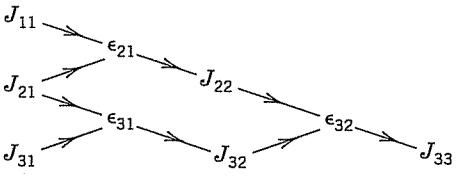
$$\epsilon_{21} = \frac{1}{2^2 - 1}\,(J_{21} - J_{11}).$$

If $|\epsilon_{21}| \leqq$ TOL, stop. The result is $J_{22} = J_{21} + \epsilon_{21}$.

Step 2. Show that $\epsilon_{21} = -0.066596$, hence $|\epsilon_{21}| >$ TOL and go on. Use (2) with $h/4$ to get J_{31} and add to it the error estimate $\epsilon_{31} = \frac{1}{3}(J_{31} - J_{21})$ to get the better $J_{32} = J_{31} + \epsilon_{31}$. Calculate

$$\epsilon_{32} = \frac{1}{2^4 - 1}\,(J_{32} - J_{22}) = \frac{1}{15}\,(J_{32} - J_{22}).$$

If $|\epsilon_{32}| \leqq$ TOL, stop. The result is $J_{33} = J_{32} + \epsilon_{32}$. (Why does $2^4 = 16$ come in?) Show that we obtain $\epsilon_{32} = -0.000266$, so that we can stop. Arrange your J- and ϵ-values in a kind of "difference table."



If $|\epsilon_{32}|$ were greater than TOL, you would have to go on and calculate in the next step J_{41} from (2) with $h = \frac{1}{4}$; then

$$J_{42} = J_{41} + \epsilon_{41} \quad \text{with} \quad \epsilon_{41} = \frac{1}{3}\,(J_{41} - J_{31})$$

$$J_{43} = J_{42} + \epsilon_{42} \quad \text{with} \quad \epsilon_{42} = \frac{1}{15}\,(J_{42} - J_{32})$$

$$J_{44} = J_{43} + \epsilon_{43} \quad \text{with} \quad \epsilon_{43} = \frac{1}{63}\,(J_{43} - J_{33})$$

where $63 = 2^6 - 1$. (How does this come in?)

Apply the Romberg method to the integral of $f(x) = \frac{1}{4}\pi x^4 \cos \frac{1}{4}\pi x$ from $x = 0$ to 2 with TOL $= 10^{-4}$.

DIFFERENTIATION

27. Consider $f(x) = x^4$ for $x_0 = 0$, $x_1 = 0.2$, $x_2 = 0.4$, $x_3 = 0.6$, $x_4 = 0.8$. Calculate f_2' from (14a), (14b), (14c), (15). Determine the errors. Compare and comment.

28. A **"four-point formula"** for the derivative is

$$f_2' \approx \frac{1}{6h}\,(-2f_1 - 3f_2 + 6f_3 - f_4).$$

Apply it to $f(x) = x^4$ with $x_1, \cdots, x_4$ as in Prob. 27, determine the error, and compare it with that in the case of (15).

29. The derivative $f'(x)$ can also be approximated in terms of first-order and higher order differences (see Sec. 19.3):

$$f'(x_0) \approx \frac{1}{h}\left(\Delta f_0 - \frac{1}{2}\,\Delta^2 f_0 + \frac{1}{3}\,\Delta^3 f_0 - \frac{1}{4}\,\Delta^4 f_0 + - \cdots\right).$$

Compute $f'(0.4)$ in Prob. 27 from this formula, using differences up to and including first order, second order, third order, fourth order.

30. Derive the formula in Prob. 29 from (14) in Sec. 19.3.

CHAPTER 19 REVIEW QUESTIONS AND PROBLEMS

1. What is a numeric method? How has the computer influenced numeric methods?
2. What is floating-point representation of numbers? Overflow and underflow?
3. How do error and relative error behave under addition? Under multiplication?
4. Why are roundoff errors important? State the rounding rules.
5. What is an algorithm? Which of its properties are important in software implementation?
6. Why is the selection of a *good* method at least as important on a large computer as it is on a small one?
7. Explain methods for solving equations, in particular fixed-point iteration and its convergence.
8. Can the Newton (–Raphson) method diverge? Is it fast? Same questions for the bisection method.
9. What is the advantage of Newton's interpolation formulas over Lagrange's?
10. What do you remember about errors in polynomial interpolation?
11. What is spline interpolation? Its advantage over polynomial interpolation?
12. List and compare numeric integration methods. When would you apply them?
13. In what sense is Gauss integration optimal? Explain details.
14. What does adaptive integration mean? Why is it useful?
15. Why is numeric differentiation generally more delicate than numeric integration?
16. Write -0.35287, 1274.799, -0.00614, 24.9482, $1/3$, $85/7$ in floating-point form with 5S (5 significant digits, properly rounded).
17. Compute $(5.346 - 3.644)/(3.454 - 3.055)$ as given and then rounded stepwise to 3S, 2S, 1S. ("Stepwise" means rounding the four rounded numbers, not the given ones.) Comment on your results.
18. Compute $0.29731/(4.1232 - 4.0872)$ with the numbers as given and then rounded stepwise (that is, rounding the rounded numbers) to 4S, 3S, 2S. Comment.
19. Solve $x^2 - 50x + 1 = 0$ by (6) and by (7) in Sec. 19.1, using 5S in the computation. Compare and comment.
20. Solve $x^2 - 200x + 4 = 0$ by (6) and by (7) in Sec. 19.1, using 5S in the computation. Compare and comment.
21. Let 4.81 and 12.752 be correctly rounded to the number of digits shown. Determine the smallest interval in which the sum (using the true instead of the rounded values) must lie.
22. Answer the question in Prob. 21 for the difference $4.81 - 12.752$.
23. What is the relative error of $n\bar{a}$ in terms of that of $\bar{a}$?
24. Show that the relative error of $\bar{a}^2$ is about twice that of $\bar{a}$.
25. Compute the solution of $x^5 = x + 0.2$ near $x = 0$ by transforming the equation into the form $x = g(x)$ and starting from $x_0 = 0$. (Use 6S.)
26. Solve $\cos x = x$ by iteration (6S, $x_0 = 1$), writing it as $x = (0.74x + \cos x)/1.74$, obtaining $x_4 = 0.739\,085$ (exact to 6S!). Why does this converge so rapidly?
27. Solve $x^4 - x^3 - 2x - 34 = 0$ by Newton's method with $x_0 = 3$ and 6S accuracy.
28. Solve $\cos x - x = 0$ by the method of false position.
29. Compute $f(1.28)$ from

$$\begin{aligned} f(1.0) &= 3.00000 \\ f(1.2) &= 2.98007 \\ f(1.4) &= 2.92106 \\ f(1.6) &= 2.82534 \\ f(1.8) &= 2.69671 \\ f(2.0) &= 2.54030 \end{aligned}$$

by linear interpolation. By quadratic interpolation, using $f(1.2)$, $f(1.4)$, $f(1.6)$.

30. Find the cubic spline for the data

$$\begin{aligned} f(-1) &= 3 \\ f(1) &= 1 \\ f(3) &= 23 \\ f(5) &= 45 \\ k_0 &= k_3 = 3. \end{aligned}$$

31. Compute the integral of x^3 from 0 to 1 by the trapezoidal rule with $n = 5$. What error bounds are obtained from (4) in Sec. 19.5? What is the actual error of the result? Why is this result larger than the exact value?
32. Compute the integral of $\cos(x^2)$ from 0 to 1 by Simpson's rule with $2m = 2$ and $2m = 4$ and estimate the error by (10) in Sec. 19.5. (This is the Fresnel integral (38) in App. 3.1 with $x = 1$.)
33. Compute the integral of $\cos x$ from 0 to $\frac{1}{2}\pi$ by the **three-eights rule**

$$\int_a^b f(x)\,dx = \frac{3}{8}\,h(f_0 + 3f_1 + 3f_2 + f_3) - \frac{1}{80}\,(b-a)h^4 f^{(\mathrm{iv})}(\hat{t})$$

and give error bounds; here $a \leqq \hat{t} \leqq b$ and $x_j = a + (b-a)j/3$, $j = 0, \cdots, 3$.

SUMMARY OF CHAPTER 19
Numerics in General

In this chapter we discussed concepts that are relevant throughout numeric work as a whole and methods of a general nature, as opposed to methods for linear algebra (Chap. 20) or differential equations (Chap. 21).

In scientific computations we use the **floating-point** representation of numbers (Sec. 19.1); fixed-point representation is less suitable in most cases.

Numeric methods give approximate values $\tilde{a}$ of quantities. The **error** ϵ of $\tilde{a}$ is

$$\epsilon = a - \tilde{a} \tag{1}$$

(Sec. 19.1)

where a is the exact value. The *relative error* of $\tilde{a}$ is ϵ/a. Errors arise from rounding, inaccuracy of measured values, truncation (that is, replacement of integrals by sums, series by partial sums), and so on.

An algorithm is called **numerically stable** if small changes in the initial data give only correspondingly small changes in the final results. Unstable algorithms are generally useless because errors may become so large that results will be very inaccurate. Numeric instability of algorithms must not be confused with mathematical instability of problems (*"ill-conditioned problems,"* Sec. 19.2).

Fixed-point iteration is a method for solving equations $f(x) = 0$ in which the equation is first transformed algebraically to $x = g(x)$, an initial guess x_0 for the solution is made, and then approximations $x_1, x_2, \cdots$, are successively computed by iteration from (see Sec. 19.2)

$$x_{n+1} = g(x_n) \qquad (n = 0, 1, \cdots). \tag{2}$$

Newton's method for solving equations $f(x) = 0$ is an iteration

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \tag{3}$$

(Sec. 19.2).

Here x_{n+1} is the x-intercept of the tangent of the curve $y = f(x)$ at the point x_n. This method is of second order (Theorem 2, Sec. 19.2). If we replace f' in (3) by a difference quotient (geometrically: we replace the tangent by a secant), we obtain the **secant method;** see (10) in Sec. 19.2. For the *bisection method* (which converges slowly) and the *method of false position,* see Problem Set 19.2.

Polynomial interpolation means the determination of a polynomial $p_n(x)$ such that $p_n(x_j) = f_j$, where $j = 0, \cdots, n$ and $(x_0, f_0), \cdots, (x_n, f_n)$ are measured or observed values, values of a function, etc. $p_n(x)$ is called an *interpolation polynomial.* For given data, $p_n(x)$ of degree n (or less) is unique. However, it can be written in different forms, notably in **Lagrange's form** (4), Sec. 19.3, or in **Newton's divided difference form** (10), Sec. 19.3, which requires fewer operations. For regularly spaced $x_0, x_1 = x_0 + h, \cdots, x_n = x_0 + nh$ the latter becomes **Newton's forward difference formula** (formula (14) in Sec. 19.3)

(4) $$f(x) \approx p_n(x) = f_0 + r\Delta f_0 + \cdots + \frac{r(r-1)\cdots(r-n+1)}{n!}\,\Delta^n f_0$$

where $r = (x - x_0)/h$ and the forward differences are $\Delta f_j = f_{j+1} - f_j$ and

$$\Delta^k f_j = \Delta^{k-1} f_{j+1} - \Delta^{k-1} f_j \qquad (k = 2, 3, \cdots).$$

A similar formula is *Newton's backward difference interpolation formula* (formula (18) in Sec. 19.3).

Interpolation polynomials may become numerically unstable as n increases, and instead of interpolating and approximating by a single high-degree polynomial it is preferable to use a cubic **spline** $g(x)$, that is, a twice continuously differentiable interpolation function [thus, $g(x_j) = f_j$], which in each subinterval $x_j \leqq x \leqq x_{j+1}$ consists of a cubic polynomial $q_j(x)$; see Sec. 19.4.

Simpson's rule of numeric integration is [see (7), Sec. 19.5]

(5) $$\int_a^b f(x)\,dx \approx \frac{h}{3}\,(f_0 + 4f_1 + 2f_2 + 4f_3 + \cdots + 2f_{2m-2} + 4f_{2m-1} + f_{2m})$$

with equally spaced nodes $x_j = x_0 + jh$, $j = 1, \cdots, 2m$, $h = (b - a)/(2m)$, and $f_j = f(x_j)$. It is simple but accurate enough for many applications. Its degree of precision is DP = 3 because the error (8), Sec. 19.5, involves h^4. A more practical error estimate is (10), Sec. 19.5,

$$\epsilon_{h/2} = \frac{1}{15}\,(J_{h/2} - J_h),$$

obtained by first computing with step h, then with step $h/2$, and then taking 1/15 of the difference of the results.

Simpson's rule is the most important of the **Newton–Cotes formulas,** which are obtained by integrating Lagrange interpolation polynomials, linear ones for the **trapezoidal rule** (2), Sec. 19.5, quadratic for Simpson's rule, cubic for the *three-eights rule* (see the Chap. 19 Review Problems), etc.

Adaptive integration (Sec. 19.5, Example 6) is integration that adjusts (*"adapts"*) the step (automatically) to the variability of $f(x)$.

Romberg integration (Team Project 26, Problem Set 19.5) starts from the trapezoidal rule (2), Sec. 19.5, with h, $h/2$, $h/4$, etc. and improves results by systematically adding error estimates.

Gauss integration (11), Sec. 19.5, is important because of its great accuracy (DP = $2n - 1$, compared to Newton–Cotes's DP = $n - 1$ or n). This is achieved by an optimal choice of the nodes, which are not equally spaced; see Table 19.7, Sec. 19.5.

Numeric differentiation is discussed at the end of Sec. 19.5. (Its main application (to differential equations) follows in Chap. 21.)

CHAPTER 20

Numeric Linear Algebra

In this chapter we explain some of the most important numeric methods for solving linear systems of equations (Secs. 20.1–20.4), for fitting straight lines or parabolas (Sec. 20.5), and for matrix eigenvalue problems (Secs. 20.6–20.9). These methods are of considerable practical importance because many problems in engineering, statistics, and elsewhere lead to mathematical models whose solution requires methods of numeric linear algebra.

COMMENT. ***This chapter is independent of Chap. 19 and can be studied immediately after Chap. 7 or 8.***

Prerequisite: Secs. 7.1, 7.2, 8.1.
Sections that may be omitted in a shorter course: 20.4, 20.5, 20.9
References and Answers to Problems: App. 1 Part E, App. 2

20.1 Linear Systems: Gauss Elimination

A **linear system of *n* equations** in n *unknowns* $x_1 \cdots, x_n$ is a set of equations $E_1, \cdots, E_n$ of the form

$$\begin{aligned} E_1: &\quad a_{11}x_1 + \cdots + a_{1n}x_n = b_1 \\ E_2: &\quad a_{21}x_1 + \cdots + a_{2n}x_n = b_2 \\ &\quad \cdots\cdots\cdots\cdots\cdots\cdots \\ E_n: &\quad a_{n1}x_1 + \cdots + a_{nn}x_n = b_n \end{aligned} \tag{1}$$

where the **coefficients** a_{jk} and the b_j are given numbers. The system is called **homogeneous** if all the b_j are zero; otherwise it is called **nonhomogeneous.** Using matrix multiplication (Sec. 7.2), we can write (1) as a single vector equation

$$\mathbf{Ax} = \mathbf{b} \tag{2}$$

where the **coefficient matrix** $\mathbf{A} = [a_{jk}]$ is the $n \times n$ matrix

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}, \quad \text{and} \quad \mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \quad \text{and} \quad \text{and } \mathbf{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}$$

are column vectors. The following matrix $\tilde{\mathbf{A}}$ is called the **augmented matrix** of the system (1):

$$\tilde{\mathbf{A}} = [\mathbf{A} \quad \mathbf{b}] = \begin{bmatrix} a_{11} & \cdots & a_{1n} & b_1 \\ a_{21} & \cdots & a_{2n} & b_2 \\ \cdot & \cdots & \cdot & \cdot \\ a_{n1} & \cdots & a_{nn} & b_n \end{bmatrix}.$$

A **solution** of (1) is a set of numbers $x_1, \cdots, x_n$ that satisfy all the n equations, and a **solution vector** of (1) is a vector $\mathbf{x}$ whose components constitute a solution of (1).

The method of solving such a system by determinants (Cramer's rule in Sec. 7.7) is not practical, even with efficient methods for evaluating the determinants.

A practical method for the solution of a linear system is the so-called *Gauss elimination*, which we shall now discuss ***(proceeding independently of Sec. 7.3).***

Gauss Elimination

This standard method for solving linear systems (1) is a systematic process of elimination that reduces (1) to **"triangular form"** because the system can then be easily solved by **"back substitution."** For instance, a triangular system is

$$\begin{aligned} 3x_1 + 5x_2 + 2x_3 &= 8 \\ 8x_2 + 2x_3 &= -7 \\ 6x_3 &= 3 \end{aligned}$$

and back substitution gives $x_3 = 3/6 = 1/2$ from the third equation, then

$$x_2 = \tfrac{1}{8}(-7 - 2x_3) = -1$$

from the second equation, and finally from the first equation

$$x_1 = \tfrac{1}{3}(8 - 5x_2 - 2x_3) = 4.$$

How do we reduce a given system (1) to triangular form? In the first step we ***eliminate*** $\boldsymbol{x_1}$ from equation E_2 to E_n in (1). We do this by adding (or subtracting) suitable multiples of E_1 from equations $\mathrm{E}_2, \cdots, \mathrm{E}_n$ and taking the resulting equations, call them $\mathrm{E}_2^*, \cdots, \mathrm{E}_n^*$ as the new equations. The first equation, E_1, is called the **pivot equation** in this step, and a_{11} is called the **pivot.** This equation is left unaltered. In the second step we take the new second equation E_2^* (which no longer contains x_1) as the pivot equation and use it to ***eliminate*** $\boldsymbol{x_2}$ from E_3^* to E_n^*. And so on. After $n - 1$ steps this gives a triangular system that can be solved by back substitution as just shown. In this way we obtain precisely all solutions of the *given* system (as proved in Sec. 7.3).

The pivot a_{kk} (in step k) ***must be*** different from zero and ***should be*** large in absolute value, to avoid roundoff magnification by the multiplication in the elimination. For this we choose as our pivot equation one that has the absolutely largest a_{jk} in column k on or below the main diagonal (actually, the uppermost if there are several such equations). This popular method is called **partial pivoting.** It is used in CASs (e.g., in Maple).

Partial pivoting distinguishes it from **total pivoting,** which involves both row and column interchanges but is hardly used in practice.

Let us illustrate this method with a simple example.

EXAMPLE 1 **Gauss Elimination. Partial Pivoting**

Solve the system

$$\begin{aligned} \mathrm{E}_1: \quad & 8x_2 + 2x_3 = -7 \\ \mathrm{E}_2: \quad 3x_1 + 5x_2 + 2x_3 &= \ \ 8 \\ \mathrm{E}_3: \quad 6x_1 + 2x_2 + 8x_3 &= 26. \end{aligned}$$

Solution. We must pivot since E_1 has no x_1-term. In Column 1, equation E_3 has the largest coefficient. Hence we interchange E_1 and E_3,

$$\begin{aligned} 6x_1 + 2x_2 + 8x_3 &= \ \ 26 \\ 3x_1 + 5x_2 + 2x_3 &= \ \ 8 \\ 8x_2 + 2x_3 &= -7. \end{aligned}$$

Step 1. Elimination of x_1

It would suffice to show the augmented matrix and operate on it. We show both the equations and the augmented matrix. In the first step, the first equation is the pivot equation. Thus

$$\begin{aligned} &\text{Pivot } 6 \longrightarrow \boxed{6x_1} + 2x_2 + 8x_3 = \ \ 26 \\ &\text{Eliminate} \longrightarrow \boxed{3x_1} + 5x_2 + 2x_3 = \ \ 8 \\ &\qquad\qquad 8x_2 + 2x_3 = -7 \end{aligned} \qquad \left[\begin{array}{cccc} 6 & 2 & 8 & 26 \\ 3 & 5 & 2 & 8 \\ 0 & 8 & 2 & -7 \end{array}\right]$$

To eliminate x_1 from the other equations (here, from the second equation), do:

Subtract 3/6 = 1/2 times the pivot equation from the second equation.

The result is

$$\begin{aligned} 6x_1 + 2x_2 + 8x_3 &= \ \ 26 \\ 4x_2 - 2x_3 &= -5 \\ 8x_2 + 2x_3 &= -7 \end{aligned} \qquad \left[\begin{array}{cccc} 6 & 2 & 8 & 26 \\ 0 & 4 & -2 & -5 \\ 0 & 8 & 2 & -7 \end{array}\right]$$

Step 2. Elimination of x_2

The largest coefficient in Column 2 is 8. Hence we take the *new* third equation as the pivot equation, interchanging equations 2 and 3,

$$\begin{aligned} 6x_1 + 2x_2 + 8x_3 &= \ \ 26 \\ &\text{Pivot } 8 \longrightarrow \boxed{8x_2} + 2x_3 = -7 \\ &\text{Eliminate} \longrightarrow \boxed{4x_2} - 2x_3 = -5 \end{aligned} \qquad \left[\begin{array}{cccc} 6 & 2 & 8 & 26 \\ 0 & 8 & 2 & -7 \\ 0 & 4 & -2 & -5 \end{array}\right]$$

To eliminate x_2 from the third equation, do:

Subtract 1/2 times the pivot equation from the third equation.

The resulting triangular system is shown below. This is the end of the forward elimination. Now comes the back substitution.

Back substitution. ***Determination of x_3, x_2, x_1***

The triangular system obtained in Step 2 is

$$\begin{aligned} 6x_1 + 2x_2 + 8x_3 &= \ \ 26 \\ 8x_2 + 2x_3 &= -7 \\ -3x_3 &= -\tfrac{3}{2} \end{aligned} \qquad \left[\begin{array}{cccc} 6 & 2 & 8 & 26 \\ 0 & 8 & 2 & -7 \\ 0 & 0 & -3 & -\frac{3}{2} \end{array}\right]$$

From this system, taking the last equation, then the second equation, and finally the first equation, we compute the solution

$$x_3 = \tfrac{1}{2}$$

$$x_2 = \tfrac{1}{8}(-7 - 2x_3) = -1$$

$$x_1 = \tfrac{1}{6}(26 - 2x_2 - 8x_3) = 4.$$

This agrees with the values given above, before the beginning of the example. ■

The general algorithm for the Gauss elimination is shown in Table 20.1. To help explain the algorithm, we have numbered some of its lines. b_j is denoted by $a_{j,n+1}$, for uniformity. In lines 1 and 2 we look for a possible pivot. [For $k = 1$ we can always find one; otherwise x_1 would not occur in (1).] In line 2 we do pivoting if necessary, picking an a_{jk} of greatest absolute value (the one with the smallest j if there are several) and interchange the corresponding rows. If $|a_{kk}|$ is greatest, we do no pivoting. m_{jk} in line 3 suggests *multiplier*, since these are the factors by which we have to multiply the pivot equation E_k^* in Step k before subtracting it from an equation E_j^* below E_k^* from which we want to eliminate x_k. Here we have written E_k^* and E_j^* to indicate that after Step 1 these are no longer the given equations in (1), but these underwent a change in each step, as indicated in line 4. Accordingly, a_{jk} etc. in lines 1–4 refer to the most recent equations, and $j \geqq k$ in line 1 indicates that we leave untouched all the equations that have served as pivot equations in previous steps. For $p = k$ in line 4 we get 0 on the right, as it should be in the elimination,

$$a_{jk} - m_{jk}a_{kk} = a_{jk} - \frac{a_{jk}}{a_{kk}}\, a_{kk} = 0.$$

In line 5, if the last equation in the *triangular* system is $0 = b_n^* \neq 0$, we have no solution. If it is $0 = b_n^* = 0$, we have no unique solution because we then have fewer equations than unknowns.

EXAMPLE 2 Gauss Elimination in Table 20.1, Sample Computation

In Example 1 we had $a_{11} = 0$, so that pivoting was necessary. The greatest coefficient in Column 1 was a_{31}. Thus $\tilde{j} = 3$ in line 2, and we interchanged E_1 and E_3. Then in lines 3 and 4 we computed $m_{21} = 3/6 = \frac{1}{2}$ and

$$a_{22} = 5 - \tfrac{1}{2} \cdot 2 = 4, \qquad a_{23} = 2 - \tfrac{1}{2} \cdot 8 = -2, \qquad a_{24} = 8 - \tfrac{1}{2} \cdot 26 = -5,$$

and then $m_{31} = 0/6 = 0$, so that the third equation $8x_2 + 2x_3 = -7$ did not change in Step 1. In Step 2 ($k = 2$) we had 8 as the greatest coefficient in Column 2, hence $\tilde{j} = 3$. We interchanged equations 2 and 3, computed $m_{32} = -4/8 = -\frac{1}{2}$ in line 4, and the $a_{33} = -2 - \frac{1}{2} \cdot 2 = -3$, $a_{34} = -5 - \frac{1}{2}(-7) = -\frac{3}{2}$. This produced the triangular form used in the back substitution. ■

If $a_{kk} = 0$ in Step k, ***we must pivot.*** If $|a_{kk}|$ is small, ***we should pivot*** because of roundoff error magnification that may seriously affect accuracy or even produce nonsensical results.

EXAMPLE 3 Difficulty with Small Pivots

The solution of the system

$$0.0004x_1 + 1.402x_2 = 1.406$$

$$0.4003x_1 - 1.502x_2 = 2.501$$

is $x_1 = 10$, $x_2 = 1$. We solve this system by the Gauss elimination, using four-digit floating-point arithmetic. (4D is for simplicity. Make an 8D-arithmetic example that shows the same.)

(a) Picking the first of the given equations as the pivot equation, we have to multiply this equation by $m = 0.4003/0.0004 = 1001$ and subtract the result from the second equation, obtaining

Table 20.1 Gauss Elimination

ALGORITHM GAUSS ($\tilde{\mathbf{A}} = [a_{jk}] = [\mathbf{A} \quad \mathbf{b}]$)

This algorithm computes a unique solution $\mathbf{x} = [x_j]$ of the system (1) or indicates that (1) has no unique solution.

INPUT: Augmented $n \times (n + 1)$ matrix $\tilde{\mathbf{A}} = [a_{jk}]$, where $a_{j,n+1} = b_j$

OUTPUT: Solution $\mathbf{x} = [x_j]$ of (1) or message that the system (1) has no unique solution

For $k = 1, \cdots, n - 1$, do:

1 If $a_{jk} = 0$ for all $j \geqq k$ then OUTPUT "No unique solution exists." Stop

[*Procedure completed unsuccessfully;* $\mathbf{A}$ *is singular*]

2 Else exchange the contents of rows $\tilde{j}$ and k of $\tilde{\mathbf{A}}$ with $\tilde{j}$ the smallest $j \geqq k$ such that $|a_{jk}|$ is maximum in column k.

3 For $j = k + 1, \cdots, n$, do:

$$m_{jk}: = \frac{a_{jk}}{a_{kk}}$$

4 For $p = k + 1, \cdots, n + 1$, do:

$$a_{jp}: = a_{jp} - m_{jk}a_{kp}$$

End

End

End

5 If $a_{nn} = 0$ then OUTPUT "No unique solution exists." Stop

Else

6
$$x_n = \frac{a_{n,n+1}}{a_{nn}} \qquad [\textit{Start back substitution}]$$

For $i = n - 1, \cdots, 1$, do:

7
$$x_i = \frac{1}{a_{ii}} \left(a_{i,n+1} - \sum_{j=i+1}^{n} a_{ij}x_j \right)$$

End

OUTPUT $\mathbf{x} = [x_j]$. Stop

End GAUSS

$$-1405x_2 = -1404.$$

Hence $x_2 = -1404/(-1405) = 0.9993$, and from the first equation, instead of $x_1 = 10$, we get

$$x_1 = \frac{1}{0.0004} (1.406 - 1.402 \cdot 0.9993) = \frac{0.005}{0.0004} = 12.5.$$

This failure occurs because $|a_{11}|$ is small compared with $|a_{12}|$, so that a small roundoff error in x_2 leads to a large error in x_1.

(b) Picking the second of the given equations as the pivot equation, we have to multiply this equation by 0.0004/0.4003 = 0.000 9993 and subtract the result from the first equation, obtaining

$$1.404x_2 = 1.404.$$

Hence $x_2 = 1$, and from the pivot equation $x_1 = 10$. This success occurs because $|a_{21}|$ is not very small compared to $|a_{22}|$, so that a small roundoff error in x_2 would not lead to a large error in x_1. Indeed, for instance, if we had the value $x_2 = 1.002$, we would still have from the pivot equation the good value $x_1 = (2.501 + 1.505)/0.4003 = 10.01$. ▩

Error estimates for the Gauss elimination are discussed in Ref. [E5] listed in App. 1.

Row scaling means the multiplication of each Row j by a suitable scaling factor s_j. It is done in connection with partial pivoting to get more accurate solutions. Despite much research (see Refs. [E9], [E24] in App. 1) and the proposition of several principles, scaling is still not well understood. As a possibility, one can scale for pivot choice only (not in the calculation, to avoid additional roundoff) and take as first pivot the entry a_{j1} for which $|a_{j1}|/|A_j|$ is largest; here A_j is an entry of largest absolute value in Row j. Similarly in the further steps of the Gauss elimination.

For instance, for the system

$$4.0000x_1 + 14020x_2 = 14060$$

$$0.4003x_1 - 1.502x_2 = 2.501$$

we might pick 4 as pivot, but dividing the first equation by 10^4 gives the system in Example 3, for which the second equation is a better pivot equation.

Operation Count

Quite generally, important factors in judging the quality of a numeric method are

Amount of storage

Amount of time (≡ number of operations)

Effect of roundoff error.

For the Gauss elimination, the operation count for a full matrix (a matrix with relatively many nonzero entries) is as follows. In Step k we eliminate x_k from $n - k$ equations. This needs $n - k$ divisions in computing the m_{jk} (line 3) and $(n - k)(n - k + 1)$ multiplications and as many subtractions (both in line 4). Since we do $n - 1$ steps, k goes from 1 to $n - 1$ and thus the total number of operations in this forward elimination is

$$f(n) = \sum_{k=1}^{n-1} (n - k) + 2\sum_{k=1}^{n-1} (n - k)(n - k + 1) \qquad \text{(write } n - k = s\text{)}$$

$$= \sum_{s=1}^{n-1} s + 2\sum_{s=1}^{n-1} s(s + 1) = \tfrac{1}{2}(n - 1)n + \tfrac{2}{3}(n^2 - 1)n \approx \tfrac{2}{3}n^3$$

where $2n^3/3$ is obtained by dropping lower powers of n. We see that $f(n)$ grows about proportional to n^3. We say that $f(n)$ is of *order* n^3 and write

$$f(n) = O(n^3)$$

where O suggests **order.** The general definition of O is as follows. We write

$$f(n) = O(h(n))$$

if the quotient $|f(n)/h(n)|$ remains bounded (does not trail off to infinity) as $n \to \infty$. In our present case, $h(n) = n^3$ and, indeed, $f(n)/n^3 \to 2/3$ because the omitted terms divided by n^3 go to zero as $n \to \infty$.

In the back substitution of x_i we make $n - i$ multiplications and as many subtractions, as well as 1 division. Hence the number of operations in the back substitution is

$$b(n) = 2\sum_{i=1}^{n}(n - i) + n = 2\sum_{s=1}^{n} s + n = n(n + 1) + n = n^2 + 2n = O(n^2).$$

We see that it grows more slowly than the number of operations in the forward elimination of the Gauss algorithm, so that it is negligible for large systems because it is smaller by a factor n, approximately. For instance, if an operation takes 10^{-9} sec, then the times needed are:

Algorithm	$n = 1000$	$n = 10\,000$
Elimination	0.7 sec	11 min
Back substitution	0.001 sec	0.1 sec

PROBLEM SET 20.1

For ***applications*** of linear systems see Secs. 7.1 and 8.2.

1–3 GEOMETRIC INTERPRETATION

Solve graphically and explain geometrically.

1. $4x_1 + x_2 = -4.3$

$3x_1 - 5x_2 = -33.7$

2. $1.820x_1 - 1.183x_2 = 0$

$-12.74x_1 + 8.281x_2 = 0$

3. $7.2x_1 - 3.5x_2 = 16.0$

$-21.6x_1 + 10.5x_2 = -48.5$

4–14 GAUSS ELIMINATION

Solve the following linear systems by Gauss elimination, with partial pivoting if necessary (but without scaling). Show the intermediate steps. Check the result by substitution. If no solution or more than one solution exists, give a reason.

4. $6x_1 + x_2 = -3$

$4x_1 - 2x_2 = 6$

5. $2x_1 - 8x_2 = -4$

$6x_1 + 2x_2 = 14$

6. $25.38x_1 - 15.48x_2 = 30.60$

$-7.05x_1 + 4.30x_2 = -8.50$

7. $6x_2 + 13x_3 = 137.86$

$6x_1 - 8x_3 = -85.88$

$13x_1 - 8x_2 = 178.54$

8. $5x_1 + 3x_2 + x_3 = 2$

$-4x_2 + 8x_3 = -3$

$10x_1 - 6x_2 + 26x_3 = 0$

9. $4x_1 + 10x_2 - 2x_3 = -20$

$-x_1 - 15x_2 + 3x_3 = 30$

$25x_2 - 5x_3 = -50$

10.
$$\begin{aligned} x_1 + 2x_2 - 3x_3 &= -11 \\ 10x_1 + x_2 + x_3 &= 8 \\ 10x_2 + 2x_3 &= 2 \end{aligned}$$

11.
$$\begin{aligned} 3.4x_1 - 6.12x_2 - 2.72x_3 &= 0 \\ -x_1 + 1.80x_2 + 0.80x_3 &= 0 \\ 2.7x_1 - 4.86x_2 - 2.16x_3 &= 0 \end{aligned}$$

12.
$$\begin{aligned} 3x_2 + 5x_3 &= 1.20736 \\ 3x_1 - 4x_2 &= -2.34066 \\ 5x_1 + 6x_3 &= -0.329193 \end{aligned}$$

13.
$$\begin{aligned} 6.4x_1 + 3.2x_2 &= -1.6 \\ 3.2x_1 - 1.6x_2 + 4.8x_3 &= 32.0 \\ 4.8x_2 - 9.6x_3 + 7.2x_4 &= -78.0 \\ 7.2x_3 + 4.8x_4 &= 20.4 \end{aligned}$$

14.
$$\begin{aligned} 4.4x_2 + 3.0x_3 - 6.6x_4 &= -4.65 \\ 0.4x_1 + 3.6x_2 + 8.4x_4 &= 4.62 \\ -2.0x_1 - 6.2x_2 + 5.0x_3 &= -4.35 \\ x_1 - 7.6x_3 + 3.0x_4 &= 5.97 \end{aligned}$$

15. CAS EXPERIMENT. Gauss Elimination. Write a program for the Gauss elimination with pivoting. Apply it to Probs. 11–14. Experiment with systems whose coefficient determinant is small in absolute value. Also investigate the performance of your program for larger systems of your choice, including sparse systems.

16. TEAM PROJECT. Linear Systems and Gauss Elimination. (a) Existence and uniqueness. Find a and b such that $ax_1 + x_2 = b$, $x_1 + x_2 = 3$ has (i) a unique solution, (ii) infinitely many solutions, (iii) no solutions.

(b) Gauss elimination and nonexistence. Apply the Gauss elimination to the following two systems and compare the calculations step by step. Explain why the elimination fails if no solution exists.

$$\begin{aligned} x_1 + x_2 + x_3 &= 3 \\ 4x_1 + 2x_2 - x_3 &= 5 \\ 9x_1 + 5x_2 - x_3 &= 13 \end{aligned}$$

$$\begin{aligned} x_1 + x_2 + x_3 &= 3 \\ 4x_1 + 2x_2 - x_3 &= 5 \\ 9x_1 + 5x_2 - x_3 &= 12. \end{aligned}$$

(c) Zero determinant. Why may a computer program give you the result that a homogeneous linear system has only the trivial solution although you know its coefficient determinant to be zero?

(d) Pivoting. Solve System (A) (below) by the Gauss elimination first without pivoting. Show that for any fixed machine word length and sufficiently small $\epsilon > 0$ the computer gives $x_2 = 1$ and then $x_1 = 0$. What is the exact solution? Its limit as $\epsilon \to 0$? Then solve the system by the Gauss elimination with pivoting. Compare and comment.

(e) Pivoting. Solve System (B) by the Gauss elimination and three-digit rounding arithmetic, choosing (i) the first equation, (ii) the second equation as pivot equation. (Remember to round to 3S after each operation before doing the next, just as would be done on a computer!) Then use four-digit rounding arithmetic in those two calculations. Compare and comment.

(A)
$$\begin{aligned} \epsilon x_1 + x_2 &= 1 \\ x_1 + x_2 &= 2 \end{aligned}$$

(B)
$$\begin{aligned} 4.03x_1 + 2.16x_2 &= -4.61 \\ 6.21x_1 + 3.35x_2 &= -7.19 \end{aligned}$$

20.2 Linear Systems: LU-Factorization, Matrix Inversion

We continue our discussion of numeric methods for solving linear systems of n equations in n unknowns $x_1, \cdots, x_n$,

$$\mathbf{Ax} = \mathbf{b} \tag{1}$$

where $\mathbf{A} = [a_{jk}]$ is the $n \times n$ coefficient matrix and $\mathbf{x}^T = [x_1 \cdots x_n]$ and $\mathbf{b}^T = [b_1 \cdots b_n]$. We present three related methods that are modifications of the Gauss elimination, which

require fewer arithmetic operations. They are named after Doolittle, Crout, and Cholesky and use the idea of the LU-factorization of **A,** which we explain first.

An **LU-factorization** of a given square matrix **A** is of the form

$$\mathbf{A} = \mathbf{LU} \tag{2}$$

where **L** is ***lower triangular*** and **U** is ***upper triangular.*** For example,

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

It can be proved that for any nonsingular matrix (see Sec. 7.8) the rows can be reordered so that the resulting matrix **A** has an LU-factorization (2) in which **L** turns out to be the matrix of the ***multipliers*** m_{jk} of the Gauss elimination, with main diagonal $1, \cdots, 1$, and **U** is the matrix of the triangular system at the end of the Gauss elimination. (See Ref. [E5], pp. 155–156, listed in App. 1.)

The ***crucial idea*** now is that **L** and **U** in (2) can be computed directly, without solving simultaneous equations (thus, without using the Gauss elimination). As a count shows, this needs about $n^3/3$ operations, about half as many as the Gauss elimination, which needs about $2n^3/3$ (see Sec. 20.1). And once we have (2), we can use it for solving $\mathbf{Ax} = \mathbf{b}$ in two steps, involving only about n^2 operations, simply by noting that $\mathbf{Ax} = \mathbf{LUx} = \mathbf{b}$ may be written

$$\text{(3)} \qquad \text{(a)} \quad \mathbf{Ly} = \mathbf{b} \qquad \text{where} \qquad \text{(b)} \quad \mathbf{Ux} = \mathbf{y}$$

and solving first (3a) for **y** and then (3b) for **x.** Here we can require that **L** have main diagonal $1, \cdots, 1$ as stated before; then this is called **Doolittle's method.** Both systems (3a) and (3b) are triangular, so we can solve them as in the back substitution for the Gauss elimination.

A similar method, **Crout's method,** is obtained from (2) if **U** (instead of **L**) is required to have main diagonal $1, \cdots, 1$. In either case the factorization (2) is unique.

EXAMPLE 1 **Doolittle's Method**

Solve the system in Example 1 of Sec. 20.1 by Doolittle's method.

Solution. The decomposition (2) is obtained from

$$\mathbf{A} = [a_{jk}] = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} 3 & 5 & 2 \\ 0 & 8 & 2 \\ 6 & 2 & 8 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ m_{21} & 1 & 0 \\ m_{31} & m_{32} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix}$$

by determining the m_{jk} and u_{jk}, using matrix multiplication. By going through **A** row by row we get successively

$a_{11} = 3 = 1 \cdot u_{11} = u_{11}$	$a_{12} = 5 = 1 \cdot u_{12} = u_{12}$	$a_{13} = 2 = 1 \cdot u_{13} = u_{13}$
$a_{21} = 0 = m_{21}u_{11}$	$a_{22} = 8 = m_{21}u_{12} + u_{22}$	$a_{23} = 2 = m_{21}u_{13} + u_{23}$
$m_{21} = 0$	$u_{22} = 8$	$u_{23} = 2$
$a_{31} = 6 = m_{31}u_{11}$	$a_{32} = 2 = m_{31}u_{12} + m_{32}u_{22}$	$a_{33} = 8 = m_{31}u_{13} + m_{32}u_{23} + u_{33}$
$= m_{31} \cdot 3$	$= 2 \cdot 5 + m_{32} \cdot 8$	$= 2 \cdot 2 - 1 \cdot 2 + u_{33}$
$m_{31} = 2$	$m_{32} = -1$	$u_{33} = 6$

Thus the factorization (2) is

$$\begin{bmatrix} 3 & 5 & 2 \\ 0 & 8 & 2 \\ 6 & 2 & 8 \end{bmatrix} = \mathbf{LU} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & -1 & 1 \end{bmatrix} \begin{bmatrix} 3 & 5 & 2 \\ 0 & 8 & 2 \\ 0 & 0 & 6 \end{bmatrix}.$$

We first solve $\mathbf{Ly} = \mathbf{b}$, determining $y_1 = 8$, then $y_2 = -7$, then y_3 from $2y_1 - y_2 + y_3 = 16 + 7 + y_3 = 26$; thus (note the interchange in $\mathbf{b}$ because of the interchange in $\mathbf{A}$!)

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & -1 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 8 \\ -7 \\ 26 \end{bmatrix}. \qquad \text{Solution} \qquad \mathbf{y} = \begin{bmatrix} 8 \\ -7 \\ 3 \end{bmatrix}.$$

Then we solve $\mathbf{Ux} = \mathbf{y}$, determining $x_3 = 3/6$, then x_2, then x_1, that is,

$$\begin{bmatrix} 3 & 5 & 2 \\ 0 & 8 & 2 \\ 0 & 0 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 8 \\ -7 \\ 3 \end{bmatrix}. \qquad \text{Solution} \qquad \mathbf{x} = \begin{bmatrix} 4 \\ -1 \\ 1/2 \end{bmatrix}.$$

This agrees with the solution in Example 1 of Sec. 20.1. ∎

Our formulas in Example 1 suggest that for general n the entries of the matrices $\mathbf{L} = [m_{jk}]$ (with main diagonal $1, \cdots, 1$ and m_{jk} suggesting "multiplier") and $\mathbf{U} = [u_{jk}]$ in the **Doolittle method** are computed from

$$\begin{aligned} u_{1k} &= a_{1k} && k = 1, \cdots, n \\ m_{j1} &= \frac{a_{j1}}{u_{11}} && j = 2, \cdots, n \\ u_{jk} &= a_{jk} - \sum_{s=1}^{j-1} m_{js} u_{sk} && k = j, \cdots, n; \quad j \geqq 2 \\ m_{jk} &= \frac{1}{u_{kk}} \left(a_{jk} - \sum_{s=1}^{k-1} m_{js} u_{sk} \right) && j = k+1, \cdots, n; \quad k \geqq 2. \end{aligned} \tag{4}$$

Row Interchanges. Matrices, such as

$$\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \qquad \text{or} \qquad \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

have no LU-factorization (try!). This indicates that for obtaining an LU-factorization, row interchanges of $\mathbf{A}$ (and corresponding interchanges in $\mathbf{b}$) may be necessary.

Cholesky's Method

For a ***symmetric, positive definite*** matrix $\mathbf{A}$ (thus $\mathbf{A} = \mathbf{A}^T$, $\mathbf{x}^T\mathbf{Ax} > 0$ for all $\mathbf{x} \neq \mathbf{0}$) we can in (2) even choose $\mathbf{U} = \mathbf{L}^T$, thus $u_{jk} = m_{kj}$ (but cannot impose conditions on the main diagonal entries). For example,

$$\mathbf{A} = \begin{bmatrix} 4 & 2 & 14 \\ 2 & 17 & -5 \\ 14 & -5 & 83 \end{bmatrix} = \mathbf{LL}^T = \begin{bmatrix} 2 & 0 & 0 \\ 1 & 4 & 0 \\ 7 & -3 & 5 \end{bmatrix} \begin{bmatrix} 2 & 1 & 7 \\ 0 & 4 & -3 \\ 0 & 0 & 5 \end{bmatrix}. \tag{5}$$

The popular method of solving $\mathbf{Ax} = \mathbf{b}$ based on this factorization $\mathbf{A} = \mathbf{LL}^T$ is called **Cholesky's method.** In terms of the entries of $\mathbf{L} = [l_{jk}]$ the formulas for the factorization are

$$l_{11} = \sqrt{a_{11}}$$

$$l_{j1} = \frac{a_{j1}}{l_{11}} \qquad j = 2, \cdots, n$$

$$l_{jj} = \sqrt{a_{jj} - \sum_{s=1}^{j-1} l_{js}^2} \qquad j = 2, \cdots, n \tag{6}$$

$$l_{pj} = \frac{1}{l_{jj}} \left(a_{pj} - \sum_{s=1}^{j-1} l_{js} l_{ps} \right) \qquad p = j + 1, \cdots, n; \quad j \geqq 2.$$

If $\mathbf{A}$ is symmetric but not positive definite, this method could still be applied, but then leads to a *complex* matrix $\mathbf{L}$, so that the method becomes impractical.

EXAMPLE 2 Cholesky's Method

Solve by Cholesky's method:

$$\begin{aligned} 4x_1 + 2x_2 + 14x_3 &= 14 \\ 2x_1 + 17x_2 - 5x_3 &= -101 \\ 14x_1 - 5x_2 + 83x_3 &= 155. \end{aligned}$$

Solution. From (6) or from the form of the factorization

$$\begin{bmatrix} 4 & 2 & 14 \\ 2 & 17 & -5 \\ 14 & -5 & 83 \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} l_{11} & l_{21} & l_{31} \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{bmatrix}$$

we compute, in the given order,

$$l_{11} = \sqrt{a_{11}} = 2 \qquad l_{21} = \frac{a_{21}}{l_{11}} = \frac{2}{2} = 1 \qquad l_{31} = \frac{a_{31}}{l_{11}} = \frac{14}{2} = 7$$

$$l_{22} = \sqrt{a_{22} - l_{21}^2} = \sqrt{17 - 1} = 4$$

$$l_{32} = \frac{1}{l_{22}} (a_{32} - l_{31} l_{21}) = \frac{1}{4} (-5 - 7 \cdot 1) = -3$$

$$l_{33} = \sqrt{a_{33} - l_{31}^2 - l_{32}^2} = \sqrt{83 - 7^2 - (-3)^2} = 5.$$

This agrees with (5). We now have to solve $\mathbf{Ly} = \mathbf{b}$, that is,

$$\begin{bmatrix} 2 & 0 & 0 \\ 1 & 4 & 0 \\ 7 & -3 & 5 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 14 \\ -101 \\ 155 \end{bmatrix}. \qquad \text{Solution} \qquad \mathbf{y} = \begin{bmatrix} 7 \\ -27 \\ 5 \end{bmatrix}.$$

As the second step, we have to solve $\mathbf{Ux} = \mathbf{L}^T\mathbf{x} = \mathbf{y}$, that is,

$$\begin{bmatrix} 2 & 1 & 7 \\ 0 & 4 & -3 \\ 0 & 0 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 7 \\ -27 \\ 5 \end{bmatrix}. \qquad \text{Solution} \qquad \mathbf{x} = \begin{bmatrix} 3 \\ -6 \\ 1 \end{bmatrix}. \qquad \blacksquare$$

THEOREM 1

Stability of the Cholesky Factorization

The Cholesky LL^T-factorization is numerically stable (as defined in Sec. 19.1).

PROOF We have $a_{jj} = l_{j1}^2 + l_{j2}^2 + \cdots + l_{jj}^2$ by squaring the third formula in (6) and solving it for a_{jj}. Hence for all l_{jk} (note that $l_{jk} = 0$ for $k > j$) we obtain (the inequality being trivial)

$$l_{jk}^2 \leqq l_{j1}^2 + l_{j2}^2 + \cdots + l_{jj}^2 = a_{jj}.$$

That is, l_{jk}^2 is bounded by an entry of **A,** which means stability against rounding. ∎

Gauss–Jordan Elimination. Matrix Inversion

Another variant of the Gauss elimination is the **Gauss–Jordan elimination,** introduced by W. Jordan in 1920, in which back substitution is avoided by additional computations that reduce the matrix to diagonal form, instead of the triangular form in the Gauss elimination. But this reduction from the Gauss triangular to the diagonal form requires more operations than back substitution does, so that the method is *disadvantageous* for solving systems $\mathbf{Ax} = \mathbf{b}$. But it may be used for matrix inversion, where the situation is as follows.

The **inverse** of a nonsingular square matrix **A** may be determined in principle by solving the n systems

$$\mathbf{Ax} = \mathbf{b}_j \qquad (j = 1, \cdots, n) \tag{7}$$

where $\mathbf{b}_j$ is the jth column of the $n \times n$ unit matrix.

However, it is preferable to produce $\mathbf{A}^{-1}$ by operating on the unit matrix **I** in the same way as the Gauss–Jordan algorithm, reducing **A** to **I.** A typical illustrative example of this method is given in Sec. 7.8.

PROBLEM SET 20.7

1–7 DOOLITTLE'S METHOD

Show the factorization and solve by Doolittle's method.

1.
$$3x_1 + 2x_2 = 15.2$$
$$15x_1 + 11x_2 = 77.3$$

2.
$$2x_1 + 9x_2 = 41$$
$$3x_1 - 5x_2 = -31$$

3.
$$4x_1 - 6x_2 = -34$$
$$8x_1 - 7x_2 = -53$$

4.
$$2x_1 + x_2 + 2x_3 = 0$$
$$-2x_1 + 2x_2 + x_3 = 0$$
$$x_1 + 2x_2 - 2x_3 = 36$$

5.
$$6x_1 + 4x_2 + 3x_3 = 2.0$$
$$4x_1 + 3x_2 + 2x_3 = 0.5$$
$$3x_1 + 4x_2 + 2x_3 = -2.5$$

6.
$$x_1 - x_2 + 2.6x_3 = -9.88$$
$$0.5x_1 - 3.0x_2 + 3.3x_3 = -16.54$$
$$-1.5x_1 - 3.5x_2 - 10.4x_3 = 21.02$$

7.
$$3x_1 + 9x_2 + 6x_3 = 2.3$$
$$18x_1 + 48x_2 + 39x_3 = 13.6$$
$$9x_1 - 27x_2 + 42x_3 = 4.5$$

8. TEAM PROJECT. Crout's method factorizes $\mathbf{A} = \mathbf{LU}$, where **L** is lower triangular and **U** is upper

triangular with diagonal entries $u_{jj} = 1, j = 1, \cdots, n$,

(a) Formulas. Obtain formulas for Crout's method similar to (4).

(b) Examples. Solve Probs. 1 and 7 by Crout's method.

(c) Factor the following matrix by the Doolittle, Crout, and Cholesky methods.

$$\begin{bmatrix} 1 & -4 & 2 \\ -4 & 25 & 4 \\ 2 & 4 & 24 \end{bmatrix}$$

(d) Give the formulas for factoring a tridiagonal matrix by Crout's method.

(e) When can you obtain Crout's factorization from Doolittle's by transposition?

9–13 CHOLESKY'S METHOD

Show the factorization and solve.

9.
$$\begin{aligned} 9x_1 + 6x_2 + 12x_3 &= 87 \\ 6x_1 + 13x_2 + 11x_3 &= 118 \\ 12x_1 + 11x_2 + 26x_3 &= 154 \end{aligned}$$

10.
$$\begin{aligned} 0.04x_1 + 0.12x_3 &= 1.4 \\ 0.64x_2 + 0.32x_3 &= 1.6 \\ 0.12x_1 + 0.32x_2 + 0.56x_3 &= 5.4 \end{aligned}$$

11.
$$\begin{aligned} 4x_1 + 6x_2 + 8x_3 &= 0 \\ 6x_1 + 34x_2 + 52x_3 &= -80 \\ 8x_1 + 52x_2 + 129x_3 &= -226 \end{aligned}$$

12.
$$\begin{aligned} x_1 - x_2 + 3x_3 + 2x_4 &= 30 \\ -x_1 + 5x_2 - 5x_3 - 2x_4 &= -70 \\ 3x_1 - 5x_2 + 19x_3 + 3x_4 &= 188 \\ 2x_1 - 2x_2 + 3x_3 + 21x_4 &= 2 \end{aligned}$$

13.
$$\begin{aligned} 4x_1 + 2x_2 + 4x_3 &= 20 \\ 2x_1 + 2x_2 + 3x_3 + 2x_4 &= 36 \\ 4x_1 + 3x_2 + 6x_3 + 3x_4 &= 60 \\ 2x_2 + 3x_3 + 9x_4 &= 122 \end{aligned}$$

14. CAS PROJECT. Cholesky's Method. (a) Write a program for solving linear systems by Cholesky's method and apply it to Example 2 in the text, to Probs. 9–11, and to systems of your choice.

(b) Splines. Apply the factorization part of the program to the following matrices (as they occur in (9), Sec. 19.4 (with $c_j = 1$), in connection with splines).

$$\begin{bmatrix} 2 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 2 \end{bmatrix}, \qquad \begin{bmatrix} 2 & 1 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 1 & 2 \end{bmatrix}.$$

15. (Definiteness) Let $\mathbf{A}$ and $\mathbf{B}$ be positive definite $n \times n$ matrices. Are $-\mathbf{A}, \mathbf{A}^T, \mathbf{A} + \mathbf{B}, \mathbf{A} - \mathbf{B}$ positive definite?

16–19 INVERSE

Find the inverse by the Gauss–Jordan method, showing the details.

16. In Prob 4.

17. In Prob. 5.

18. In Prob. 6.

19. In Prob. 7.

20. (Rounding) For the following matrix $\mathbf{A}$ find $\det \mathbf{A}$. What happens if you round off the given entries to (a) 5S, (b) 4S, (c) 3S, (d) 2S, (e) 1S? What is the practical implication of your work?

$$\mathbf{A} = \begin{bmatrix} 1/3 & 1/4 & 2 \\ -1/9 & 1 & 1/7 \\ 4/63 & -3/28 & 13/49 \end{bmatrix}$$

20.3 Linear Systems: Solution by Iteration

The Gauss elimination and its variants in the last two sections belong to the **direct methods** for solving linear systems of equations; these are methods that give solutions after an amount of computation that can be specified in advance. In contrast, in an **indirect** or **iterative method** we start from an approximation to the true solution and, if successful, obtain better and better approximations from a computational cycle repeated as often as may be necessary for achieving a required accuracy, so that the amount of arithmetic depends upon the accuracy required and varies from case to case.

We apply iterative methods if the convergence is rapid (if matrices have large main diagonal entries, as we shall see), so that we save operations compared to a direct method. We also use iterative methods if a large system is **sparse,** that is, has very many zero coefficients, so that one would waste space in storing zeros, for instance, 9995 zeros per equation in a potential problem of 10^4 equations in 10^4 unknowns with typically only 5 nonzero terms per equation (more on this in Sec. 21.4).

Gauss–Seidel Iteration Method[1]

This is an iterative method of great practical importance, which we can simply explain in terms of an example.

EXAMPLE 1 **Gauss–Seidel Iteration**

We consider the linear system

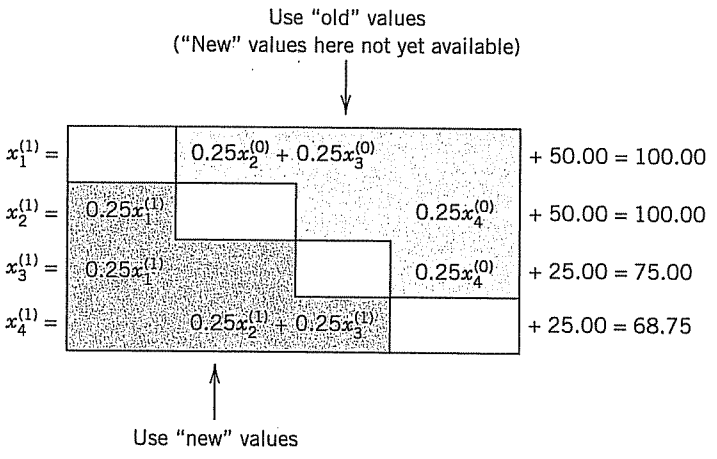
$$\text{(1)} \qquad \begin{aligned} x_1 - 0.25x_2 - 0.25x_3 &= 50 \\ -0.25x_1 + x_2 - 0.25x_4 &= 50 \\ -0.25x_1 + x_3 - 0.25x_4 &= 25 \\ -0.25x_2 - 0.25x_3 + x_4 &= 25. \end{aligned}$$

(Equations of this form arise in the numeric solution of PDEs and in spline interpolation.) We write the system in the form

$$\text{(2)} \qquad \begin{aligned} x_1 &= 0.25x_2 + 0.25x_3 + 50 \\ x_2 &= 0.25x_1 + 0.25x_4 + 50 \\ x_3 &= 0.25x_1 + 0.25x_4 + 25 \\ x_4 &= 0.25x_2 + 0.25x_3 + 25. \end{aligned}$$

These equations are now used for iteration; that is, we start from a (possibly poor) approximation to the solution, say $x_1^{(0)} = 100$, $x_2^{(0)} = 100$, $x_3^{(0)} = 100$, $x_4^{(0)} = 100$, and compute from (2) a perhaps better approximation

(3)



These equations (3) are obtained from (2) by substituting on the right the ***most recent*** approximation for each unknown. In fact, corresponding values replace previous ones as soon as they have been computed, so that in

[1]PHILIPP LUDWIG VON SEIDEL (1821–1896), German mathematician. For Gauss see footnote 5 in Sec. 5.4.

the second and third equations we use $x_1^{(1)}$ (not $x_1^{(0)}$), and in the last equation of (3) we use $x_2^{(1)}$ and $x_3^{(1)}$ (not $x_2^{(0)}$ and $x_3^{(0)}$). Using the same principle, we obtain in the next step

$$\begin{aligned}
x_1^{(2)} &= 0.25x_2^{(1)} + 0.25x_3^{(1)} + 50.00 = 93.750 \\
x_2^{(2)} &= 0.25x_1^{(2)} + 0.25x_4^{(1)} + 50.00 = 90.625 \\
x_3^{(2)} &= 0.25x_1^{(2)} + 0.25x_4^{(1)} + 25.00 = 65.625 \\
x_4^{(2)} &= 0.25x_2^{(2)} + 0.25x_3^{(2)} + 25.00 = 64.062
\end{aligned}$$

Further steps give the values

x_1	x_2	x_3	x_4
89.062	88.281	63.281	62.891
87.891	87.695	62.695	62.598
87.598	87.549	62.549	62.524
87.524	87.512	62.512	62.506
87.506	87.503	62.503	62.502

Hence convergence to the exact solution $x_1 = x_2 = 87.5$, $x_3 = x_4 = 62.5$ (verify!) seems rather fast. ▩

An algorithm for the Gauss–Seidel iteration is shown on the next page. To obtain the algorithm, let us derive the general formulas for this iteration.

We assume that $a_{jj} = 1$ for $j = 1, \cdots, n$. (Note that this can be achieved if we can rearrange the equations so that no diagonal coefficient is zero; then we may divide each equation by the corresponding diagonal coefficient.) We now write

$$\mathbf{A} = \mathbf{I} + \mathbf{L} + \mathbf{U} \qquad (a_{jj} = 1) \tag{4}$$

where $\mathbf{I}$ is the $n \times n$ unit matrix and $\mathbf{L}$ and $\mathbf{U}$ are respectively lower and upper triangular matrices with zero main diagonals. If we substitute (4) into $\mathbf{Ax} = \mathbf{b}$, we have

$$\mathbf{Ax} = (\mathbf{I} + \mathbf{L} + \mathbf{U})\mathbf{x} = \mathbf{b}.$$

Taking $\mathbf{Lx}$ and $\mathbf{Ux}$ to the right, we obtain, since $\mathbf{Ix} = \mathbf{x}$,

$$\mathbf{x} = \mathbf{b} - \mathbf{Lx} - \mathbf{Ux}. \tag{5}$$

Remembering from (3) in Example 1 that below the main diagonal we took "new" approximations and above the main diagonal "old" ones, we obtain from (5) the desired iteration formulas

$$\mathbf{x}^{(m+1)} = \mathbf{b} - \mathbf{L}\mathbf{x}^{(m+1)} - \mathbf{U}\mathbf{x}^{(m)} \qquad (a_{jj} = 1) \tag{6}$$

("New" $\mathbf{x}^{(m+1)}$, "Old" $\mathbf{x}^{(m)}$)

where $\mathbf{x}^{(m)} = \left[x_j^{(m)}\right]$ is the mth approximation and $\mathbf{x}^{(m+1)} = \left[x_j^{(m+1)}\right]$ is the $(m+1)$st approximation. In components this gives the formula in line 1 in Table 20.2. The matrix $\mathbf{A}$ must satisfy $a_{jj} \neq 0$ for all j. In Table 20.2 our assumption $a_{jj} = 1$ is no longer required, but is automatically taken care of by the factor $1/a_{jj}$ in line 1.

Table 20.2 Gauss–Seidel Iteration

ALGORITHM GAUSS–SEIDEL ($\mathbf{A}$, $\mathbf{b}$, $\mathbf{x}^{(0)}$, ϵ, N)

This algorithm computes a solution $\mathbf{x}$ of the system $\mathbf{Ax} = \mathbf{b}$ given an initial approximation $\mathbf{x}^{(0)}$, where $\mathbf{A} = [a_{jk}]$ is an $n \times n$ matrix with $a_{jj} \neq 0, j = 1, \cdots, n$.

INPUT: $\mathbf{A}$, $\mathbf{b}$, initial approximation $\mathbf{x}^{(0)}$, tolerance $\epsilon > 0$, maximum number of iterations N

OUTPUT: Approximate solution $\mathbf{x}^{(m)} = [x_j^{(m)}]$ or failure message that $\mathbf{x}^{(N)}$ does not satisfy the tolerance condition

For $m = 0, \cdots, N - 1$, do:

  For $j = 1, \cdots, n$, do:

1 $$x_j^{(m+1)} = \frac{1}{a_{jj}} \left(b_j - \sum_{k=1}^{j-1} a_{jk} x_k^{(m+1)} - \sum_{k=j+1}^{n} a_{jk} x_k^{(m)} \right)$$

  End

2 If $\max_j |x_j^{(m+1)} - x_j^{(m)}| < \epsilon$ then OUTPUT $\mathbf{x}^{(m+1)}$. Stop

  [*Procedure completed successfully*]

End

OUTPUT: "No solution satisfying the tolerance condition obtained after N iteration steps." Stop

[*Procedure completed unsuccessfully*]

End GAUSS–SEIDEL

Convergence and Matrix Norms

An iteration method for solving $\mathbf{Ax} = \mathbf{b}$ is said to **converge** for an initial $\mathbf{x}^{(0)}$ if the corresponding iterative sequence $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \cdots$ converges to a solution of the given system. Convergence depends on the relation between $\mathbf{x}^{(m)}$ and $\mathbf{x}^{(m+1)}$. To get this relation for the Gauss–Seidel method, we use (6). We first have

$$(\mathbf{I} + \mathbf{L})\mathbf{x}^{(m+1)} = \mathbf{b} - \mathbf{U}\mathbf{x}^{(m)}$$

and by multiplying by $(\mathbf{I} + \mathbf{L})^{-1}$ from the left,

(7) $$\mathbf{x}^{(m+1)} = \mathbf{C}\mathbf{x}^{(m)} + (\mathbf{I} + \mathbf{L})^{-1}\mathbf{b} \qquad \text{where} \qquad \mathbf{C} = -(\mathbf{I} + \mathbf{L})^{-1}\mathbf{U}.$$

The Gauss–Seidel iteration converges for every $\mathbf{x}^{(0)}$ if and only if all the eigenvalues (Sec. 8.1) of the "iteration matrix" $\mathbf{C} = [c_{jk}]$ have absolute value less than 1. (Proof in Ref. [E5], p. 191, listed in App. 1.)

CAUTION! If you want to get $\mathbf{C}$, first divide the rows of $\mathbf{A}$ by a_{jj} to have main diagonal $1, \cdots, 1$. If the **spectral radius** of $\mathbf{C}$ (= maximum of those absolute values) is small, then the convergence is rapid.

Sufficient Convergence Condition. A sufficient condition for convergence is

(8) $$\|\mathbf{C}\| < 1.$$

Here $\|\mathbf{C}\|$ is some **matrix norm,** such as

(9) $$\|\mathbf{C}\| = \sqrt{\sum_{j=1}^{n}\sum_{k=1}^{n} c_{jk}^2}$$ **(Frobenius norm)**

or the greatest of the sums of the $|c_{jk}|$ in a *column* of $\mathbf{C}$

(10) $$\|\mathbf{C}\| = \max_k \sum_{j=1}^{n} |c_{jk}|$$ **(Column "sum" norm)**

or the greatest of the sums of the $|c_{jk}|$ in a *row* of $\mathbf{C}$

(11) $$\|\mathbf{C}\| = \max_j \sum_{k=1}^{n} |c_{jk}|$$ **(Row "sum" norm).**

These are the most frequently used matrix norms in numerics.

In most cases the choice of one of these norms is a matter of computational convenience. However, the following example shows that sometimes one of these norms is preferable to the others.

EXAMPLE 2 Test of Convergence of the Gauss–Seidel Iteration

Test whether the Gauss–Seidel iteration converges for the system

$$\begin{aligned} 2x + y + z &= 4 \\ x + 2y + z &= 4 \\ x + y + 2z &= 4 \end{aligned} \qquad \text{written} \qquad \begin{aligned} x &= 2 - \tfrac{1}{2}y - \tfrac{1}{2}z \\ y &= 2 - \tfrac{1}{2}x - \tfrac{1}{2}z \\ z &= 2 - \tfrac{1}{2}x - \tfrac{1}{2}y. \end{aligned}$$

Solution. The decomposition (multiply the matrix by 1/2 − why?) is

$$\begin{bmatrix} 1 & 1/2 & 1/2 \\ 1/2 & 1 & 1/2 \\ 1/2 & 1/2 & 1 \end{bmatrix} = \mathbf{I} + \mathbf{L} + \mathbf{U} = \mathbf{I} + \begin{bmatrix} 0 & 0 & 0 \\ 1/2 & 0 & 0 \\ 1/2 & 1/2 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 1/2 & 1/2 \\ 0 & 0 & 1/2 \\ 0 & 0 & 0 \end{bmatrix}.$$

It shows that

$$\mathbf{C} = -(\mathbf{I} + \mathbf{L})^{-1}\mathbf{U} = -\begin{bmatrix} 1 & 0 & 0 \\ -1/2 & 1 & 0 \\ -1/4 & -1/2 & 1 \end{bmatrix}\begin{bmatrix} 0 & 1/2 & 1/2 \\ 0 & 0 & 1/2 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & -1/2 & -1/2 \\ 0 & 1/4 & -1/4 \\ 0 & 1/8 & 3/8 \end{bmatrix}.$$

We compute the Frobenius norm of $\mathbf{C}$

$$\|\mathbf{C}\| = \left(\frac{1}{4} + \frac{1}{4} + \frac{1}{16} + \frac{1}{16} + \frac{1}{64} + \frac{9}{64}\right)^{1/2} = \left(\frac{50}{64}\right)^{1/2} = 0.884 < 1$$

and conclude from (8) that this Gauss–Seidel iteration converges. It is interesting that the other two norms would permit no conclusion, as you should verify. Of course, this points to the fact that (8) is sufficient for convergence rather than necessary. ■

Residual. Given a system $\mathbf{Ax} = \mathbf{b}$, the **residual** $\mathbf{r}$ of $\mathbf{x}$ with respect to this system is defined by

(12) $$\mathbf{r} = \mathbf{b} - \mathbf{Ax}.$$

Clearly, $\mathbf{r} = \mathbf{0}$ if and only if $\mathbf{x}$ is a solution. Hence $\mathbf{r} \neq \mathbf{0}$ for an approximate solution. In the Gauss–Seidel iteration, at each stage we modify or *relax* a component of an approximate solution in order to reduce a component of $\mathbf{r}$ to zero. Hence the Gauss–Seidel iteration belongs to a class of methods often called **relaxation methods.** More about the residual follows in the next section.

Jacobi Iteration

The Gauss–Seidel iteration is a method of **successive corrections** because for each component we successively replace an approximation of a component by a corresponding new approximation as soon as the latter has been computed. An iteration method is called a method of **simultaneous corrections** if no component of an approximation $\mathbf{x}^{(m)}$ is used until ***all*** the components of $\mathbf{x}^{(m)}$ have been computed. A method of this type is the **Jacobi iteration,** which is similar to the Gauss–Seidel iteration but involves *not* using improved values until a step has been completed and then replacing $\mathbf{x}^{(m)}$ by $\mathbf{x}^{(m+1)}$ at once, directly before the beginning of the next step. Hence if we write $\mathbf{Ax} = \mathbf{b}$ (***with*** $a_{jj} = 1$ ***as before!***) in the form $\mathbf{x} = \mathbf{b} + (\mathbf{I} - \mathbf{A})\mathbf{x}$, the Jacobi iteration in matrix notation is

$$\mathbf{x}^{(m+1)} = \mathbf{b} + (\mathbf{I} - \mathbf{A})\mathbf{x}^{(m)} \qquad (a_{jj} = 1). \tag{13}$$

This method converges for every choice of $\mathbf{x}^{(0)}$ if and only if the spectral radius of $\mathbf{I} - \mathbf{A}$ is less than 1. It has recently gained greater practical interest since on parallel processors all n equations can be solved simultaneously at each iteration step.

For Jacobi, see Sec. 10.3. For exercises, see the problem set.

PROBLEM SET 20

1. Verify the claim at the end of Example 2.
2. Show that for the system in Example 2 the Jacobi iteration diverges. *Hint.* Use eigenvalues.

3–8 GAUSS–SEIDEL ITERATION

Do 5 steps, starting from $\mathbf{x}_0 = [1 \quad 1 \quad 1]^T$ and using 6S in the computation. *Hint.* Make sure that you solve each equation for the variable that has the largest coefficient (why?). Show the details.

3. $x_1 + x_2 + 6x_3 = -61.3$
   $x_1 + 9x_2 - 2x_3 = 49.1$
   $8x_1 + 2x_2 - x_3 = 185.8$

4. $x_2 + 7x_3 = 25.5$
   $5x_1 + x_2 = 0$
   $x_1 + 6x_2 + x_3 = -10.5$

5. $5x_1 + x_2 + 2x_3 = 19$
   $x_1 + 4x_2 - 2x_3 = -2$
   $2x_1 + 3x_2 + 8x_3 = 39$

6. $4x_1 - x_2 = 21$
   $-x_1 + 4x_2 - x_3 = -45$
   $- x_2 + 4x_3 = 33$

7. $10x_1 + x_2 + x_3 = 6$
   $x_1 + 10x_2 + x_3 = 6$
   $x_1 + x_2 + 10x_3 = 6$

8. $4x_1 + 5x_3 = 12.5$
   $x_1 + 6x_2 + 2x_3 = 18.5$
   $8x_1 + 2x_2 + x_3 = -11.5$

9. Apply the Gauss–Seidel iteration (3 steps) to the system in Prob. 7, starting from (a) 0, 0, 0, (b) 10, 10, 10. Compare and comment.
10. In Prob. 7, compute $\mathbf{C}$ (a) if you solve the first equation for x_1, the second for x_2, the third for x_3, proving convergence; (b) if you nonsensically solve the third equation for x_1, the first for x_2, the second for x_3, proving divergence.

11. CAS PROJECT. Gauss–Seidel Iteration. (a) Write a program for Gauss–Seidel iteration.

(b) Apply the program to $\mathbf{A}(t)\mathbf{x} = \mathbf{b}$, starting from $[0 \quad 0 \quad 0]^T$, where

$$\mathbf{A}(t) = \begin{bmatrix} 1 & t & t \\ t & 1 & t \\ t & t & 1 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix}.$$

For $t = 0.2, 0.5, 0.8, 0.9$ determine the number of steps to obtain the exact solution to 6S and the corresponding spectral radius of $\mathbf{C}$. Graph the number of steps and the spectral radius as functions of t and comment.

(c) Successive overrelaxation (SOR). Show that by adding and subtracting $\mathbf{x}^{(m)}$ on the right, formula (6) can be written

$$\mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} + \mathbf{b} - \mathbf{L}\mathbf{x}^{(m+1)} - (\mathbf{U} + \mathbf{I})\mathbf{x}^{(m)} \qquad (a_{jj} = 1).$$

Anticipation of further corrections motivates the introduction of an **overrelaxation factor** $\omega > 1$ to get the **SOR formula for Gauss–Seidel**

$$\mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} + \omega(\mathbf{b} - \mathbf{L}\mathbf{x}^{(m+1)} - (\mathbf{U} + \mathbf{I})\mathbf{x}^{(m)}) \qquad (a_{jj} = 1) \tag{14}$$

intended to give more rapid convergence. A recommended value is $\omega = 2/(1 + \sqrt{1 - \rho})$, where ρ is the spectral radius of $\mathbf{C}$ in (7). Apply SOR to the matrix in (b) for $t = 0.5$ and 0.8 and notice the improvement of convergence. (Spectacular gains are made with larger systems.)

12–15 JACOBI ITERATION

Do 5 steps, starting from $\mathbf{x}_0 = [1 \quad 1 \quad 1]^T$. Compare with the Gauss–Seidel iteration. Which of the two seems to converge faster? (Show the details of your work.)

12. The system in Prob. 6

13. The system in Prob. 5

14. The system in Prob. 8

15. Show convergence in Prob. 14 by verifying that $\mathbf{I} - \mathbf{A}$, where $\mathbf{A}$ is the matrix in Prob. 14 with the rows divided by the corresponding main diagonal entries, has the eigenvalues -0.519589 and $0.259795 \pm 0.246603i$.

16–20 NORMS

Compute the norms (9), (10), (11) for the following (square) matrices. Comment on the reasons for greater or smaller differences among the three numbers.

16. The matrix in Prob. 3

17. The matrix in Prob. 7

18. The matrix in Prob. 8

19. $\begin{bmatrix} 2k & -k & -k \\ k & -2k & k \\ -k & -k & 2k \end{bmatrix}$

20. $\begin{bmatrix} 4 & -3 & 5 \\ -7 & 1 & 0 \\ 17 & -12 & -2 \end{bmatrix}$

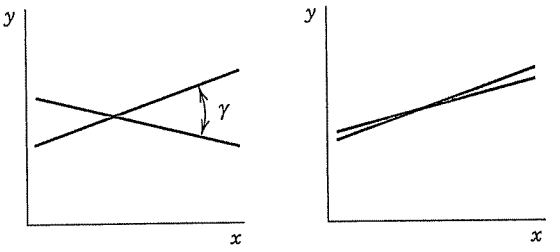
20.4 Linear Systems: Ill-Conditioning, Norms

One does not need much experience to observe that some systems $\mathbf{A}\mathbf{x} = \mathbf{b}$ are good, giving accurate solutions even under roundoff or coefficient inaccuracies, whereas others are bad, so that these inaccuracies affect the solution strongly. We want to see what is going on and whether or not we can "trust" a linear system. Let us first formulate the two relevant concepts (ill- and well-conditioned) for general numeric work and then turn to linear systems and matrices.

A computational problem is called **ill-conditioned** (or *ill-posed*) if "small" changes in the data (the input) cause "large" changes in the solution (the output). On the other hand, a problem is called **well-conditioned** (or *well-posed*) if "small" changes in the data cause only "small" changes in the solution.

These concepts are qualitative. We would certainly regard a magnification of inaccuracies by a factor 100 as "large," but could debate where to draw the line between "large" and "small," depending on the kind of problem and on our viewpoint. Double precision may sometimes help, but if data are measured inaccurately, one should attempt ***changing the mathematical setting*** of the problem to a well-conditioned one.

Let us now turn to linear systems. Figure 442 explains that ill-conditioning occurs if and only if the two equations give two nearly parallel lines, so that their intersection point (the solution of the system) moves substantially if we raise or lower a line just a little. For larger systems the situation is similar in principle, although geometry no longer helps. We shall see that we may regard ill-conditioning as an approach to singularity of the matrix.



(a) (b)

Fig. 442. (a) Well-conditioned and (b) ill-conditioned linear system of two equations in two unknowns

EXAMPLE 1 An Ill-Conditioned System

You may verify that the system

$$0.9999x - 1.0001y = 1$$
$$x - \quad y = 1$$

has the solution $x = 0.5$, $y = -0.5$, whereas the system

$$0.9999x - 1.0001y = 1$$
$$x - \quad y = 1 + \epsilon$$

has the solution $x = 0.5 + 5000.5\epsilon$, $y = -0.5 + 4999.5\epsilon$. This shows that the system is ill-conditioned because a change on the right of magnitude ϵ produces a change in the solution of magnitude 5000ϵ, approximately. We see that the lines given by the equations have nearly the same slope. ■

Well-conditioning can be asserted if the main diagonal entries of $\mathbf{A}$ have large absolute values compared to those of the other entries. Similarly if $\mathbf{A}^{-1}$ and $\mathbf{A}$ have maximum entries of about the same absolute value.

Ill-conditioning is indicated if $\mathbf{A}^{-1}$ has entries of large absolute value compared to those of the solution (about 5000 in Example 1) and if poor approximate solutions may still produce small residuals.

Residual. The *residual* $\mathbf{r}$ of an approximate solution $\tilde{\mathbf{x}}$ of $\mathbf{Ax} = \mathbf{b}$ is defined as

(1) $$\mathbf{r} = \mathbf{b} - \mathbf{A}\tilde{\mathbf{x}}.$$

Now $\mathbf{b} = \mathbf{Ax}$, so that

(2) $$\mathbf{r} = \mathbf{A}(\mathbf{x} - \tilde{\mathbf{x}}).$$

Hence $\mathbf{r}$ is small if $\tilde{\mathbf{x}}$ has high accuracy, but the converse may be false:

EXAMPLE 2 Inaccurate Approximate Solution with a Small Residual

The system

$$1.0001x_1 + \quad x_2 = 2.0001$$
$$x_1 + 1.0001x_2 = 2.0001$$

has the exact solution $x_1 = 1$, $x_2 = 1$. Can you see this by inspection? The very inaccurate approximation $\tilde{x}_1 = 2.0000$, $\tilde{x}_2 = 0.0001$ has the very small residual (to 4D)

$$\mathbf{r} = \begin{bmatrix} 2.0001 \\ 2.0001 \end{bmatrix} - \begin{bmatrix} 1.0001 & 1.0000 \\ 1.0000 & 1.0001 \end{bmatrix} \begin{bmatrix} 2.0000 \\ 0.0001 \end{bmatrix} = \begin{bmatrix} 2.0001 \\ 2.0001 \end{bmatrix} - \begin{bmatrix} 2.0003 \\ 2.0001 \end{bmatrix} = \begin{bmatrix} -0.0002 \\ 0.0000 \end{bmatrix}.$$

From this, a naive person might draw the false conclusion that the approximation should be accurate to 3 or 4 decimals.

Our result is probably unexpected, but we shall see that it has to do with the fact that the system is ill-conditioned. ■

Our goal is to show that ill-conditioning of a linear system and of its coefficient matrix $\mathbf{A}$ can be measured by a number, the *"condition number"* $\kappa(\mathbf{A})$. Other measures for ill-conditioning have also been proposed, but $\kappa(\mathbf{A})$ is probably the most widely used one. $\kappa(\mathbf{A})$ is defined in terms of norm, a concept of great general interest throughout numerics (and in modern mathematics in general!). We shall reach our goal in three steps, discussing

1. **Vector norms**
2. **Matrix norms**
3. **Condition number** κ of a square matrix.

Vector Norms

A **vector norm** for column vectors $\mathbf{x} = [x_j]$ with n components (n fixed) is a generalized length or distance. It is denoted by $\|\mathbf{x}\|$ and is defined by four properties of the usual length of vectors in three-dimensional space, namely,

(3)
(a) $\|\mathbf{x}\|$ is a nonnegative real number.
(b) $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = \mathbf{0}$.
(c) $\|k\mathbf{x}\| = |k|\,\|\mathbf{x}\|$ for all k.
(d) $\|\mathbf{x} + \mathbf{y}\| \leqq \|\mathbf{x}\| + \|\mathbf{y}\|$ (Triangle inequality).

If we use several norms, we label them by a subscript. Most important in connection with computations is the ***p-norm*** defined by

(4) $$\|\mathbf{x}\|_p = (|x_1|^p + |x_2|^p + \cdots + |x_n|^p)^{1/p}$$

where p is a fixed number and $p \geqq 1$. In practice, one usually takes $p = 1$ or 2 and, as a third norm, $\|\mathbf{x}\|_\infty$ (the latter as defined below), that is,

(5) $$\|\mathbf{x}\|_1 = |x_1| + \cdots + |x_n|$$ ("l_1-norm")

(6) $$\|\mathbf{x}\|_2 = \sqrt{x_1^2 + \cdots + x_n^2}$$ ("Euclidean" or "l_2-norm")

(7) $$\|\mathbf{x}\|_\infty = \max_j |x_j|$$ ("l_∞-norm").

For $n = 3$ the l_2-norm is the usual length of a vector in three-dimensional space. The l_1-norm and l_∞-norm are generally more convenient in computation. But all three norms are in common use.

EXAMPLE 3 Vector Norms

If $\mathbf{x}^T = [2 \quad -3 \quad 0 \quad 1 \quad -4]$, then $\|\mathbf{x}\|_1 = 10, \quad \|\mathbf{x}\|_2 = \sqrt{30}, \quad \|\mathbf{x}\|_\infty = 4.$ ∎

In three-dimensional space, two points with position vectors $\mathbf{x}$ and $\tilde{\mathbf{x}}$ have distance $|\mathbf{x} - \tilde{\mathbf{x}}|$ from each other. For a linear system $\mathbf{Ax} = \mathbf{b}$, this suggests that we take $\|\mathbf{x} - \tilde{\mathbf{x}}\|$ as a measure of inaccuraty and call it the **distance** between an exact and an approximate solution, or the **error** of $\tilde{\mathbf{x}}$.

Matrix Norm

If $\mathbf{A}$ is an $n \times n$ matrix and $\mathbf{x}$ any vector with n components, then $\mathbf{Ax}$ is a vector with n components. We now take a vector norm and consider $\|\mathbf{x}\|$ and $\|\mathbf{Ax}\|$. One can prove (see Ref. [E17]. p. 77, 92–93, listed in App. 1) that there is a number c (depending on $\mathbf{A}$) such that

$$\|\mathbf{Ax}\| \leqq c\|\mathbf{x}\| \qquad \text{for all } \mathbf{x}. \tag{8}$$

Let $\mathbf{x} \neq \mathbf{0}$. Then $\|\mathbf{x}\| > 0$ by (3b) and division gives $\|\mathbf{Ax}\|/\|\mathbf{x}\| \leqq c$. We obtain the smallest possible c valid for *all* $\mathbf{x}$ $(\neq \mathbf{0})$ by taking the maximum on the left. This smallest c is called the **matrix norm of A** *corresponding to the vector norm we picked* and is denoted by $\|\mathbf{A}\|$. Thus

$$\|\mathbf{A}\| = \max \frac{\|\mathbf{Ax}\|}{\|\mathbf{x}\|} \qquad (\mathbf{x} \neq \mathbf{0}), \tag{9}$$

the maximum being taken over all $\mathbf{x} \neq \mathbf{0}$. Alternatively [see (c) in Team Project 24],

$$\|\mathbf{A}\| = \max_{\|\mathbf{x}\| = 1} \|\mathbf{Ax}\|. \tag{10}$$

The maximum in (10) and thus also in (9) exists. And the name "matrix ***norm***" is justified because $\|\mathbf{A}\|$ satisfies (3) with $\mathbf{x}$ and $\mathbf{y}$ replaced by $\mathbf{A}$ and $\mathbf{B}$. (Proofs in Ref. [E17] pp. 77, 92–93.)

Note carefully that $\|\mathbf{A}\|$ depends on the vector norm that we selected. In particular, one can show that

for the l_1-norm (5) one gets the column "sum" norm (10), Sec. 20.3,

for the l_∞-norm (7) one gets the row "sum" norm (11), Sec. 20.3.

By taking our best possible (our smallest) $c = \|\mathbf{A}\|$ we have from (8)

$$\|\mathbf{Ax}\| \leqq \|\mathbf{A}\|\,\|\mathbf{x}\|. \tag{11}$$

This is the formula we shall need. Formula (9) also implies for two $n \times n$ matrices (see Ref. [E17], p. 98)

$$\|\mathbf{AB}\| \leqq \|\mathbf{A}\|\,\|\mathbf{B}\|, \qquad \text{thus} \qquad \|\mathbf{A}^n\| \leqq \|\mathbf{A}\|^n. \tag{12}$$

See Refs. [E9] and [E17] for other useful formulas on norms.

Before we go on, let us do a simple illustrative computation.

EXAMPLE 4 **Matrix Norms**

Compute the matrix norms of the coefficient matrix $\mathbf{A}$ in Example 1 and of its inverse $\mathbf{A}^{-1}$, assuming that we use (a) the l_1-vector norm, (b) the l_∞-vector norm.

Solution. We use (4*), Sec. 7.8, for the inverse and then (10) and (11) in Sec. 20.3. Thus

$$\mathbf{A} = \begin{bmatrix} 0.9999 & -1.0001 \\ 1.0000 & -1.0000 \end{bmatrix}, \qquad \mathbf{A}^{-1} = \begin{bmatrix} -5000.0 & 5000.5 \\ -5000.0 & 4999.5 \end{bmatrix}.$$

(a) The l_1-vector norm gives the column "sum" norm (10), Sec. 20.3; from Column 2 we thus obtain $\|\mathbf{A}\| = |-1.0001| + |-1.0000| = 2.0001$. Similarly, $\|\mathbf{A}^{-1}\| = 10000$.

(b) The l_∞-vector norm gives the row "sum" norm (11), Sec. 20.3; thus $\|\mathbf{A}\| = 2$, $\|\mathbf{A}^{-1}\| = 10000.5$ from Row 1. We notice that $\|\mathbf{A}^{-1}\|$ is surprisingly large, which makes the product $\|\mathbf{A}\|\ \|\mathbf{A}^{-1}\|$ large (20001). We shall see below that this is typical of an ill-conditioned system. ▩

Condition Number of a Matrix

We are now ready to introduce the key concept in our discussion of ill-conditioning, the **condition number** $\kappa(\mathbf{A})$ of a (nonsingular) square matrix $\mathbf{A}$, defined by

$$\kappa(\mathbf{A}) = \|\mathbf{A}\|\,\|\mathbf{A}^{-1}\|. \tag{13}$$

The role of the condition number is seen from the following theorem.

THEOREM 1 **Condition Number**

A linear system of equations $\mathbf{Ax} = \mathbf{b}$ *and its matrix* $\mathbf{A}$ *whose condition number* (13) *is small are well-conditioned. A large condition number indicates ill-conditioning.*

PROOF $\mathbf{b} = \mathbf{Ax}$ and (11) give $\|\mathbf{b}\| \leqq \|\mathbf{A}\|\|\mathbf{x}\|$. Let $\mathbf{b} \neq \mathbf{0}$ and $\mathbf{x} \neq \mathbf{0}$. Then division by $\|\mathbf{b}\|\,\|\mathbf{x}\|$ gives

$$\frac{1}{\|\mathbf{x}\|} \leqq \frac{\|\mathbf{A}\|}{\|\mathbf{b}\|}. \tag{14}$$

Multiplying (2) $\mathbf{r} = \mathbf{A}(\mathbf{x} - \tilde{\mathbf{x}})$ by $\mathbf{A}^{-1}$ from the left and interchanging sides, we have $\mathbf{x} - \tilde{\mathbf{x}} = \mathbf{A}^{-1}\mathbf{r}$. Now (11) with $\mathbf{A}^{-1}$ and $\mathbf{r}$ instead of $\mathbf{A}$ and $\mathbf{x}$ yields

$$\|\mathbf{x} - \tilde{\mathbf{x}}\| = \|\mathbf{A}^{-1}\mathbf{r}\| \leqq \|\mathbf{A}^{-1}\|\,\|\mathbf{r}\|.$$

Division by $\|\mathbf{x}\|$ [note that $\|\mathbf{x}\| \neq 0$ by (3b)] and use of (14) finally gives

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \leqq \frac{1}{\|\mathbf{x}\|}\,\|\mathbf{A}^{-1}\|\,\|\mathbf{r}\| \leqq \frac{\|\mathbf{A}\|}{\|\mathbf{b}\|}\,\|\mathbf{A}^{-1}\|\,\|\mathbf{r}\| = \kappa(\mathbf{A})\,\frac{\|\mathbf{r}\|}{\|\mathbf{b}\|}. \tag{15}$$

Hence if $\kappa(\mathbf{A})$ is small, a small $\|\mathbf{r}\|/\|\mathbf{b}\|$ implies a small relative error $\|\mathbf{x} - \tilde{\mathbf{x}}\|/\|\mathbf{x}\|$, so that the system is well-conditioned. However, this does not hold if $\kappa(\mathbf{A})$ is large; then a small $\|\mathbf{r}\|/\|\mathbf{b}\|$ does not necessarily imply a small relative error $\|\mathbf{x} - \tilde{\mathbf{x}}\|/\|\mathbf{x}\|$. ▩

EXAMPLE 5 Condition Numbers. Gauss–Seidel Iteration

$$\mathbf{A} = \begin{bmatrix} 5 & 1 & 1 \\ 1 & 4 & 2 \\ 1 & 2 & 4 \end{bmatrix} \quad \text{has the inverse} \quad \mathbf{A}^{-1} = \frac{1}{56}\begin{bmatrix} 12 & -2 & -2 \\ -2 & 19 & -9 \\ -2 & -9 & 19 \end{bmatrix}.$$

Since $\mathbf{A}$ is symmetric, (10) and (11) in Sec. 20.3 give the same condition number

$$\kappa(\mathbf{A}) = \|\mathbf{A}\|\,\|\mathbf{A}^{-1}\| = 7 \cdot \tfrac{1}{56} \cdot 30 = 3.75.$$

We see that a linear system $\mathbf{Ax} = \mathbf{b}$ with this $\mathbf{A}$ is well-conditioned.

For instance, if $\mathbf{b} = [14 \quad 0 \quad 28]^T$, the Gauss algorithm gives the solution $\mathbf{x} = [2 \quad -5 \quad 9]^T$ (confirm this). Since the main diagonal entries of $\mathbf{A}$ are relatively large, we can expect reasonably good convergence of the Gauss–Seidel iteration. Indeed, starting from, say, $\mathbf{x}_0 = [1 \quad 1 \quad 1]^T$, we obtain the first 8 steps (3D values)

x_1	x_2	x_3
1.000	1.000	1.000
2.400	−1.100	6.950
1.630	−3.882	8.534
1.870	−4.734	8.900
1.967	−4.942	8.979
1.993	−4.988	8.996
1.998	−4.997	8.999
2.000	−5.000	9.000
2.000	−5.000	9.000

EXAMPLE 6 Ill-Conditioned Linear System

Example 4 gives by (10) or (11), Sec. 20.3, for the matrix in Example 1 the very large condition number $\kappa(\mathbf{A}) = 2.0001 \cdot 10\,000 = 2 \cdot 10\,000.5 = 20\,0001$. This confirms that the system is very ill-conditioned.

Similarly in Example 2, where by (4*), Sec. 7.8 and 6D-computation,

$$\mathbf{A}^{-1} = \frac{1}{0.0002}\begin{bmatrix} 1.0001 & -1.0000 \\ -1.0000 & 1.0001 \end{bmatrix} = \begin{bmatrix} 5000.5 & -5.000.0 \\ -5000.0 & 5000.5 \end{bmatrix}$$

so that (10), Sec. 20.3, gives a very large $\kappa(\mathbf{A})$, explaining the surprising result in Example 2,

$$\kappa(\mathbf{A}) = (1.0001 + 1.0000)(5000.5 + 5000.0) \approx 20\,002.$$

In practice, $\mathbf{A}^{-1}$ will not be known, so that in computing the condition number $\kappa(\mathbf{A})$, one must estimate $\|\mathbf{A}^{-1}\|$. A method for this (proposed in 1979) is explained in Ref. [E9] listed in App. 1.

Inaccurate Matrix Entries. $\kappa(\mathbf{A})$ can be used for estimating the effect $\delta\mathbf{x}$ of an inaccuracy $\delta\mathbf{A}$ of $\mathbf{A}$ (errors of measurements of the a_{jk}, for instance). Instead of $\mathbf{Ax} = \mathbf{b}$ we then have

$$(\mathbf{A} + \delta\mathbf{A})(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b}.$$

Multiplying out and subtracting $\mathbf{Ax} = \mathbf{b}$ on both sides, we obtain

$$\mathbf{A}\delta\mathbf{x} + \delta\mathbf{A}(\mathbf{x} + \delta\mathbf{x}) = \mathbf{0}.$$

Multiplication by $\mathbf{A}^{-1}$ from the left and taking the second term to the right gives

$$\delta\mathbf{x} = -\mathbf{A}^{-1}\delta\mathbf{A}(\mathbf{x} + \delta\mathbf{x}).$$

Applying (11) with $\mathbf{A}^{-1}$ and vector $\delta\mathbf{A}(\mathbf{x} + \delta\mathbf{x})$ instead of $\mathbf{A}$ and $\mathbf{x}$, we get

$$\|\delta\mathbf{x}\| = \|\mathbf{A}^{-1}\delta\mathbf{A}(\mathbf{x} + \delta\mathbf{x})\| \leqq \|\mathbf{A}^{-1}\| \, \|\delta\mathbf{A}(\mathbf{x} + \delta\mathbf{x})\| .$$

Applying (11) on the right, with $\delta\mathbf{A}$ and $\mathbf{x} - \delta\mathbf{x}$ instead of $\mathbf{A}$ and $\mathbf{x}$, we obtain

$$\|\delta\mathbf{x}\| \leqq \|\mathbf{A}^{-1}\| \, \|\delta\mathbf{A}\| \, \|\mathbf{x} + \delta\mathbf{x}\| .$$

Now $\|\mathbf{A}^{-1}\| = \kappa(\mathbf{A})/\|\mathbf{A}\|$ by the definition of $\kappa(\mathbf{A})$, so that division by $\|\mathbf{x} + \delta\mathbf{x}\|$ shows that the relative inaccuracy of $\mathbf{x}$ is related to that of $\mathbf{A}$ via the condition number by the inequality

(16)
$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \approx \frac{\|\delta\mathbf{x}\|}{\|\mathbf{x} + \delta\mathbf{x}\|} \leqq \|\mathbf{A}^{-1}\| \, \|\delta\mathbf{A}\| = \kappa(\mathbf{A}) \frac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|} .$$

Conclusion. If the system is well-conditioned, small inaccuracies $\|\delta\mathbf{A}\|/\|\mathbf{A}\|$ can have only a small effect on the solution. However, in the case of ill-conditioning, if $\|\delta\mathbf{A}\|/\|\mathbf{A}\|$ is small, $\|\delta\mathbf{x}\|/\|\mathbf{x}\|$ ***may*** be large.

Inaccurate Right Side. You may show that, similarly, when $\mathbf{A}$ is accurate, an inaccuracy $\delta\mathbf{b}$ of $\mathbf{b}$ causes an inaccuracy $\delta\mathbf{x}$ satisfying

(17)
$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leqq \kappa(\mathbf{A}) \frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} .$$

Hence $\|\delta\mathbf{x}\|/\|\mathbf{x}\|$ must remain relatively small whenever $\kappa(\mathbf{A})$ is small.

EXAMPLE 7 **Inaccuracies. Bounds (16) and (17)**

If each of the nine entries of $\mathbf{A}$ in Example 5 is measured with an inaccuracy of 0.1, then $\|\delta\mathbf{A}\| = 9 \cdot 0.1$ and (16) gives

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leqq 7.5 \cdot \frac{3 \cdot 0.1}{7} = 0.321 \qquad \text{thus} \qquad \|\delta\mathbf{x}\| \leqq 0.321 \, \|\mathbf{x}\| = 0.321 \cdot 16 = 5.14.$$

By experimentation you will find that the actual inaccuracy $\|\delta\mathbf{x}\|$ is only about 30% of the bound 5.14. This is typical.

Similarly, if $\delta\mathbf{b} = [0.1 \quad 0.1 \quad 0.1]^T$, then $\|\delta\mathbf{b}\| = 0.3$ and $\|\mathbf{b}\| = 42$ in Example 5, so that (17) gives

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leqq 7.5 \cdot \frac{0.3}{42} = 0.0536, \qquad \text{hence} \qquad \|\delta\mathbf{x}\| \leqq 0.0536 \cdot 16 = 0.857$$

but this bound is again much greater than the actual inaccuracy, which is about 0.15. ■

Further Comments on Condition Numbers. The following additional explainations may be helpful.

1. There is no sharp dividing line between "well-conditioned" and "ill-conditioned," but generally the situation will get worse as we go from systems with small $\kappa(\mathbf{A})$ to systems with larger $\kappa(\mathbf{A})$. Now always $\kappa(\mathbf{A}) \geqq 1$, so that values of 10 or 20 or so give no reason for concern, whereas $\kappa(\mathbf{A}) = 100$, say, calls for caution, and systems such as those in Examples 1 and 2 are extremely ill-conditioned.

2. If $\kappa(\mathbf{A})$ is large (or small) in one norm, it will be large (or small, respectively) in any other norm. See Example 5.

3. The literature on ill-conditioning is extensive. For an introduction to it, see [E9].

This is the end of our discussion of numerics for solving linear systems. In the next section we consider curve fitting, an important area in which solutions are obtained from linear systems.

PROBLEM SET 20.4

1–8 VECTOR NORMS

Compute (5), (6), (7). Compute a corresponding **unit vector** (vector of norm 1) with respect to the l_∞-norm.

1. $[1 \quad -6 \quad 5]$

2. $[0.4 \quad -1.2 \quad 0 \quad 8.0]$

3. $[-4 \quad 4 \quad 3 \quad -3]$

4. $[0 \quad 0 \quad 1 \quad 0 \quad 0]$

5. $[0.3 \quad -0.1 \quad 0.5 \quad 1.0]$

6. $[16 \quad 21 \quad 54 \quad -119]$

7. $[1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1]$

8. $[3 \quad 0 \quad 0 \quad -3 \quad 0]$

9. Show that $\|\mathbf{x}\|_\infty \leqq \|\mathbf{x}\|_2 \leqq \|\mathbf{x}\|_1$.

10–15 MATRIX NORMS, CONDITION NUMBERS

Compute the matrix norm and the condition number corresponding to the l_1-vector norm.

10. $\begin{bmatrix} -3 & 4 \\ 1 & 2 \end{bmatrix}$ **11.** $\begin{bmatrix} 5 & 7 \\ 7 & 10 \end{bmatrix}$

12. $\begin{bmatrix} \sqrt{3} & 3 \\ 0 & -\sqrt{3} \end{bmatrix}$ **13.** $\begin{bmatrix} 0 & 0 & 100 \\ 0 & 100 & 0 \\ 0.01 & 0 & 0 \end{bmatrix}$

14. $\begin{bmatrix} 21 & 10.5 & 7 & 5.25 \\ 10.5 & 7 & 5.25 & 4.2 \\ 7 & 5.25 & 4.2 & 3.5 \\ 5.25 & 4.2 & 3.5 & 3 \end{bmatrix}$

15. $\begin{bmatrix} 1 & 0.1 & 0 \\ 0.1 & 1 & 0.1 \\ 0 & 0.1 & 1 \end{bmatrix}$

16. Verify (11) for $\mathbf{x} = [4 \quad -5 \quad 2]^T$ taken with the l_∞-norm and the matrix in Prob. 15.

17. Verify (12) for the matrices in Probs. 10 and 11.

18. Verify the calculations in Examples 5 and 6 of the text.

19–20 ILL-CONDITIONED SYSTEMS

Solve $\mathbf{Ax} = \mathbf{b}_1$, $\mathbf{Ax} = \mathbf{b}_2$, compare the solutions, and comment. Compute the condition number of $\mathbf{A}$.

19. $\mathbf{A} = \begin{bmatrix} 2 & 1.4 \\ 1.4 & 1 \end{bmatrix}$, $\mathbf{b}_1 = \begin{bmatrix} 1.4 \\ 1 \end{bmatrix}$, $\mathbf{b}_2 = \begin{bmatrix} 1.44 \\ 1 \end{bmatrix}$

20. $\mathbf{A} = \begin{bmatrix} 5 & -7 \\ -7 & 10 \end{bmatrix}$, $\mathbf{b}_1 = \begin{bmatrix} -2 \\ 3 \end{bmatrix}$, $\mathbf{b}_2 = \begin{bmatrix} -2 \\ 3.1 \end{bmatrix}$

21. (Residual) For $\mathbf{Ax} = \mathbf{b}_1$ in Prob. 19 guess what the residual of $\tilde{\mathbf{x}} = [113 \quad -160]^T$ might be (the solution being $\mathbf{x} = [0 \quad 1]^T$). Then calculate and comment.

22. Show that $\kappa(\mathbf{A}) \geqq 1$ for the matrix norms (10), (11), Sec. 20.3, and $\kappa(\mathbf{A}) \geqq \sqrt{n}$ for the Frobenius norm (9), Sec. 20.3.

23. CAS EXPERIMENT. Hilbert Matrices. The 3×3 Hilbert matrix is

$$\mathbf{H}_3 = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \end{bmatrix}.$$

The $n \times n$ Hilbert matrix is $\mathbf{H}_n = [h_{jk}]$, where $h_{jk} = 1/(j + k - 1)$. (Similar matrices occur in curve fitting by least squares.) Compute the condition number $\kappa(\mathbf{H}_n)$ for the matrix norm corresponding to the l_∞- (or l_1-) vector norm, for $n = 2, 3, \cdots, 6$ (or further if you wish). Try to find a formula that gives reasonable approximate values of these rapidly growing numbers. Solve a few linear systems involving an $\mathbf{H}_n$ of your choice.

24. TEAM PROJECT. Norms. (a) Vector norms in our text are **equivalent,** that is, they are related by double inequalities; for instance,

$$(18)\quad \text{(a)}\quad \|\mathbf{x}\|_\infty \leqq \|\mathbf{x}\|_1 \leqq n\|\mathbf{x}\|_\infty$$

$$\text{(b)}\quad \frac{1}{n}\|\mathbf{x}\|_1 \leqq \|\mathbf{x}\|_\infty \leqq \|\mathbf{x}\|_1.$$

Hence if for some $\mathbf{x}$, one norm is large (or small), the other norm must also be large (or small). Thus in many investigations the particular choice of a norm is not essential. Prove (18).

(b) The **Cauchy–Schwarz inequality** is

$$|\mathbf{x}^T\mathbf{y}| \leqq \|\mathbf{x}\|_2\,\|\mathbf{y}\|_2.$$

It is very important. (Proof in Ref. [GR7] listed in App. 1.) Use it to prove

$$(19a)\qquad \|\mathbf{x}\|_2 \leqq \|\mathbf{x}\|_1 \leqq \sqrt{n}\,\|\mathbf{x}\|_2$$

$$(19b)\qquad \frac{1}{\sqrt{n}}\,\|\mathbf{x}\|_1 \leqq \|\mathbf{x}\|_2 \leqq \|\mathbf{x}\|_1.$$

(c) Formula (10) is often more practical than (9). Derive (10) from (9).

(d) Matrix norms. Illustrate (11) with examples. Give examples of (12) with equality as well as with strict inequality. Prove that the matrix norms (10), (11) in Sec. 20.3 satisfy the *axioms of a norm*

$$\|\mathbf{A}\| \geqq 0.$$

$$\|\mathbf{A}\| = 0 \text{ if and only if } \mathbf{A} = \mathbf{0},$$

$$\|k\mathbf{A}\| = |k|\,\|\mathbf{A}\|,$$

$$\|\mathbf{A} + \mathbf{B}\| \leqq \|\mathbf{A}\| + \|\mathbf{B}\|.$$

25. WRITING PROJECT. Norms and Their Use in This Section. Make a list of the most important of the many ideas covered in this section and write a two-page report on them.

20.5 Least Squares Method

Having discussed numerics for linear systems, we now turn to an important application, curve fitting, in which the solutions are obtained from linear systems.

In **curve fitting** we are given n points (pairs of numbers) $(x_1, y_1), \cdots, (x_n, y_n)$ and we want to determine a function $f(x)$ such that

$$f(x_1) \approx y_1, \cdots, f(x_n) \approx y_n,$$

approximately. The type of function (for example, polynomials, exponential functions, sine and cosine functions) may be suggested by the nature of the problem (the underlying physical law, for instance), and in many cases a polynomial of a certain degree will be appropriate.

Let us begin with a motivation.

If we require strict equality $f(x_1) = y_1, \cdots, f(x_n) = y_n$ and use polynomials of sufficiently high degree, we may apply one of the methods discussed in Sec. 19.3 in connection with interpolation. However, in certain situations this would not be the appropriate solution of the actual problem. For instance, to the four points

$$(1)\quad (-1.3,\ 0.103),\qquad (-0.1,\ 1.099),\qquad (0.2,\ 0.808),\qquad (1.3,\ 1.897)$$

there corresponds the interpolation polynomial $f(x) = x^3 - x + 1$ (Fig. 443), but if we

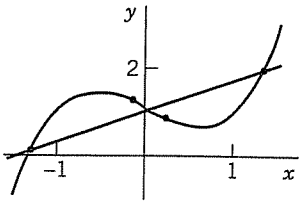


Fig. 443. Approximate fitting of a straight line

graph the points, we see that they lie nearly on a straight line. Hence if these values are obtained in an experiment and thus involve an experimental error, and if the nature of the experiment suggests a linear relation, we better fit a straight line through the points (Fig. 443). Such a line may be useful for predicting values to be expected for other values of x. A widely used principle for fitting straight lines is the **method of least squares** by Gauss and Legendre. In the present situation it may be formulated as follows.

Method of Least Squares. *The straight line*

$$y = a + bx \tag{2}$$

should be fitted through the given points $(x_1, y_1), \cdots, (x_n, y_n)$ *so that the sum of the squares of the distances of those points from the straight line is minimum, where the distance is measured in the vertical direction (the y-direction).*

The point on the line with abscissa x_j has the ordinate $a + bx_j$. Hence its distance from (x_j, y_j) is $|y_j - a - bx_j|$ (Fig. 444) and that sum of squares is

$$q = \sum_{j=1}^{n} (y_j - a - bx_j)^2.$$

q depends on a and b. A necessary condition for q to be minimum is

$$\frac{\partial q}{\partial a} = -2 \sum (y_j - a - bx_j) = 0$$

$$\frac{\partial q}{\partial b} = -2 \sum x_j (y_j - a - bx_j) = 0 \tag{3}$$

(where we sum over j from 1 to n). Dividing by 2, writing each sum as three sums, and taking one of them to the right, we obtain the result

$$\begin{aligned} an \quad + b \sum x_j &= \sum y_j \\ a \sum x_j + b \sum x_j^2 &= \sum x_j y_j. \end{aligned} \tag{4}$$

These equations are called the **normal equations** of our problem.

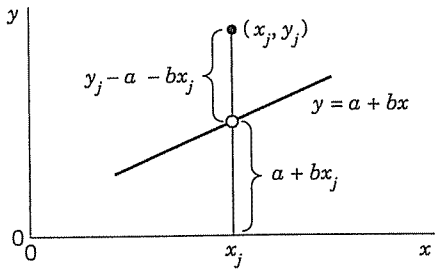


Fig. 444. Vertical distance of a point (x_j, y_j) from a straight line $y = a + bx$

EXAMPLE 1 Straight Line

Using the method of least squares, fit a straight line to the four points given in formula (1).

Solution. We obtain

$$n = 4, \qquad \sum x_j = 0.1, \qquad \sum x_j^2 = 3.43, \qquad \sum y_j = 3.907, \qquad \sum x_j y_j = 2.3839.$$

Hence the normal equations are

$$4a + 0.10b = 3.9070$$
$$0.1a + 3.43b = 2.3839.$$

The solution (rounded to 4D) is $a = 0.9601$, $b = 0.6670$, and we obtain the straight line (Fig. 443)

$$y = 0.9601 + 0.6670x.$$

Curve Fitting by Polynomials of Degree m

Our method of curve fitting can be generalized from a polynomial $y = a + bx$ to a polynomial of degree m

$$p(x) = b_0 + b_1 x + \cdots + b_m x^m \tag{5}$$

where $m \leqq n - 1$. Then q takes the form

$$q = \sum_{j=1}^{n} (y_j - p(x_j))^2$$

and depends on $m + 1$ parameters $b_0, \cdots, b_m$. Instead of (3) we then have $m + 1$ conditions

$$\frac{\partial q}{\partial b_0} = 0, \qquad \cdots, \qquad \frac{\partial q}{\partial b_m} = 0 \tag{6}$$

which give a system of $m + 1$ normal equations.

In the case of a quadratic polynomial

$$p(x) = b_0 + b_1 x + b_2 x^2 \tag{7}$$

the normal equations are (summation from 1 to n)

$$\begin{aligned} b_0 n \quad &+ b_1 \sum x_j \quad + b_2 \sum x_j^2 = \sum y_j \\ b_0 \sum x_j \quad &+ b_1 \sum x_j^2 + b_2 \sum x_j^3 = \sum x_j y_j \\ b_0 \sum x_j^2 \quad &+ b_1 \sum x_j^3 + b_2 \sum x_j^4 = \sum x_j^2 y_j. \end{aligned} \tag{8}$$

The derivation of (8) is left to the reader.

EXAMPLE 2 Quadratic Parabola by Least Squares

Fit a parabola through the data (0, 5), (2, 4), (4, 1), (6, 6), (8, 7).

Solution. For the normal equations we need $n = 5$, $\Sigma x_j = 20$, $\Sigma x_j^2 = 120$, $\Sigma x_j^3 = 800$, $\Sigma x_j^4 = 5664$, $\Sigma y_j = 23$, $\Sigma x_j y_j = 104$, $\Sigma x_j^2 y_j = 696$. Hence these equations are

$$\begin{aligned} 5b_0 + 20b_1 + 120b_2 &= 23 \\ 20b_0 + 120b_1 + 800b_2 &= 104 \\ 120b_0 + 800b_1 + 5664b_2 &= 696. \end{aligned}$$

Solving them we obtain the quadratic least squares parabola (Fig. 445)

$$y = 5.11429 - 1.41429x + 0.21429x^2.$$

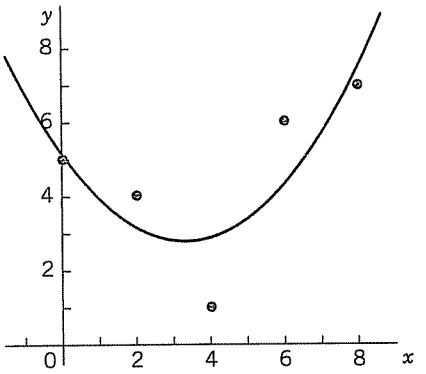


Fig. 445. Least squares parabola in Example 2

For a general polynomial (5) the normal equations form a linear system of equations in the unknowns $b_0, \cdots, b_m$. When its matrix $\mathbf{M}$ is nonsingular, we can solve the system by Cholesky's method (Sec. 20.2) because then $\mathbf{M}$ is positive definite (and symmetric). When the equations are nearly linearly dependent, the normal equations may become ill-conditioned and should be replaced by other methods; see [E5], Sec. 5.7, listed in App. 1.

The least squares method also plays a role in statistics (see Sec. 25.9).

PROBLEM SET 20.5

1–6 FITTING A STRAIGHT LINE

Fit a straight line to the given points (x, y) by least squares. Show the details. Check your result by sketching the points and the line. Judge the goodness of fit.

1. (2, 0), (3, 4), (4, 10), (5, 16)
2. How does the line in Prob. 1 change if you add a point far above it, say, (3, 20)?
3. (2.5, 8.0), (5.0, 6.9), (7.5, 6.2), (10.0, 5.0)
4. **(Ohm's law $U = Ri$)** Estimate the resistance R from the least squares line that fits $(i, U) = (2.0, 104)$, (4.0, 206), (6.0, 314), (10.0, 530).
5. **(Average speed)** Estimate the average speed v_{av} of a car traveling according to $s = v \cdot t$ [km] (s = distance traveled, t [h] = time) from $(t, s) = (9, 140)$, (10, 220), (11, 310), (12, 410).
6. **(Hooke's law $F = ks$)** Estimate the spring modulus k from the force F [lb] and extension s [cm], where $(F, s) = (1, 0.50)$, (2, 1.02), (4, 1.99), (6, 3.01), (10, 4.98), (20, 10.03).
7. Derive the normal equations (8).

8–10 FITTING A QUADRATIC PARABOLA

Fit a parabola (7) to the given points (x, y) by least squares. Check by sketching.

8. $(-1, 3)$, (0, 0), (1, 2), (2, 8)
9. (0, 4), (2, 2), (4, -1), (6, -5)

10.

Worker's time on duty x [h]	1	2	3	4	5
Worker's reaction time [sec]	1.50	1.28	1.40	1.85	2.20

11. Fit (2) and (7) by least squares to $(-1.0, 5.4)$, $(-0.5, 4.1)$, $(0, 3.9)$, $(0.5, 4.8)$, $(1.0, 6.3)$, $(1.5, 9.3)$. Graph the data and the curves on common axes and comment.

12. **(Cubic parabola)** Derive the formula for the normal equations of a cubic least squares parabola.

13. Fit curves (2) and (7) and a cubic parabola by least squares to $(-2, -35)$, $(-1, -9)$, $(0, -1)$, $(1, -1)$, $(2, 17)$, $(3, 63)$. Graph the three curves and the points on common axes. Comment on the goodness of fit.

14. **CAS PROJECT. Least Squares.** Write programs for calculating and solving the normal equations (4) and (8). Apply the programs to Probs. 3, 5, 9, 11. If your CAS has a command for fitting (Maple and Mathematica do), compare your results with those by your CAS commands.

15. **CAS EXPERIMENT. Least Squares versus Interpolation.** For the given data and for data of your choice find the interpolation polynomial and the least squares approximations (linear, quadratic, etc.). Compare and comment.

(a) $(-2, 0)$, $(-1, 0)$, $(0, 1)$, $(1, 0)$, $(2, 0)$

(b) $(-4, 0)$, $(-3, 0)$, $(-2, 0)$, $(-1, 0)$, $(0, 1)$, $(1, 0)$, $(2, 0)$, $(3, 0)$, $(4, 0)$

(c) Choose five points on a straight line, e.g., $(0, 0)$, $(1, 1)$, $\cdots$, $(4, 4)$. Move one point 1 unit upward and find the quadratic least squares polynomial. Do this for each point. Graph the five polynomials on common axes. Which of the five motions has the greatest effect?

16. **TEAM PROJECT. The least squares approximation of a function** $f(x)$ on an interval $a \leqq x \leqq b$ by a function

$$F_m(x) = a_0 y_0(x) + a_1 y_1(x) + \cdots + a_m y_m(x)$$

where $y_0(x), \cdots, y_m(x)$ are given functions, requires the determination of the coefficients $a_0, \cdots, a_m$ such that

$$(9) \qquad \int_a^b [f(x) - F_m(x)]^2 \, dx$$

becomes minimum. This integral is denoted by $\|f - F_m\|^2$, and $\|f - F_m\|$ is called the L_2**-norm** of $f - F_m$ (L suggesting Lebesgue[2]). A necessary condition for that minimum is given by $\partial \|f - F_m\|^2 / \partial a_j = 0$, $j = 0, \cdots, m$ [the analog of (6)]. **(a)** Show that this leads to $m + 1$ normal equations ($j = 0, \cdots, m$)

$$\sum_{k=0}^{m} h_{jk} a_k = b_j \qquad \text{where}$$

$$(10) \qquad h_{jk} = \int_a^b y_j(x) y_k(x) \, dx,$$

$$b_j = \int_a^b f(x) y_j(x) \, dx.$$

(b) Polynomial. What form does (10) take if $F_m(x) = a_0 + a_1 x + \cdots + a_m x^m$? What is the coefficient matrix of (10) in this case when the interval is $0 \leqq x \leqq 1$?

(c) Orthogonal functions. What are the solutions of (10) if $y_0(x), \cdots, y_m(x)$ are orthogonal on the interval $a \leqq x \leqq b$? (For the definition, see Sec. 5.7. See also Sec. 5.8.)

20.6 Matrix Eigenvalue Problems: Introduction

In the remaining sections of this chapter we discuss some of the most important ideas and numeric methods for matrix eigenvalue problems. This very extensive part of numeric linear algebra is of great practical importance, with much research going on, and hundreds, if not thousands of papers published in various mathematical journals (see the references in [E8], [E9], [E11], [E29]). We begin with the concepts and general results we shall need in explaining and applying numeric methods for eigenvalue problems. (For typical models of eigenvalue problems see Chap. 8.)

[2]HENRI LEBESGUE (1875–1941), great French mathematician, creator of a modern theory of measure and integration in his famous doctoral thesis of 1902.

An **eigenvalue** or **characteristic value** (or *latent root*) of a given $n \times n$ matrix $\mathbf{A} = [a_{jk}]$ is a real or complex number λ such that the vector equation

$$\mathbf{Ax} = \lambda\mathbf{x} \tag{1}$$

has a nontrivial solution, that is, a solution $\mathbf{x} \neq \mathbf{0}$, which is then called an **eigenvector** or **characteristic vector** of $\mathbf{A}$ corresponding to that eigenvalue λ. The set of all eigenvalues of $\mathbf{A}$ is called the **spectrum** of $\mathbf{A}$. Equation (1) can be written

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0} \tag{2}$$

where $\mathbf{I}$ is the $n \times n$ unit matrix. This homogeneous system has a nontrivial solution if and only if the **characteristic determinant** $\det(\mathbf{A} - \lambda\mathbf{I})$ is 0 (see Theorem 2 in Sec. 7.5). This gives (see Sec. 8.1)

THEOREM 1

Eigenvalues

The eigenvalues of $\mathbf{A}$ *are the solutions* λ *of the* **characteristic equation**

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} a_{11} - \lambda & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} - \lambda & \cdots & a_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} - \lambda \end{vmatrix} = 0. \tag{3}$$

Developing the characteristic determinant, we obtain the **characteristic polynomial** of $\mathbf{A}$, which is of degree n in λ. Hence $\mathbf{A}$ has at least one and at most n numerically different eigenvalues. If $\mathbf{A}$ is real, so are the coefficients of the characteristic polynomial. By familiar algebra it follows that then the roots (the eigenvalues of $\mathbf{A}$) are ***real or complex conjugates*** in pairs.

We shall usually denote the eigenvalues of $\mathbf{A}$ by

$$\lambda_1, \lambda_2, \cdots, \lambda_n$$

with the understanding that some (or all) of them may be equal.

The sum of these n eigenvalues equals the sum of the entries on the main diagonal of $\mathbf{A}$, called the **trace** of $\mathbf{A}$; thus

$$\text{trace } \mathbf{A} = \sum_{j=1}^{n} a_{jj} = \sum_{k=1}^{n} \lambda_k. \tag{4}$$

Also, the product of the eigenvalues equals the determinant of $\mathbf{A}$,

$$\det \mathbf{A} = \lambda_1\lambda_2 \cdots \lambda_n. \tag{5}$$

Both formulas follow from the product representation of the characteristic polynomial, which we denote by $f(\lambda)$,

$$f(\lambda) = (-1)^n(\lambda - \lambda_1)(\lambda - \lambda_2) \cdots (\lambda - \lambda_n).$$

If we take equal factors together and denote the *numerically distinct* eigenvalues of $\mathbf{A}$ by $\lambda_1, \cdots, \lambda_r$ $(r \leqq n)$, then the product becomes

$$f(\lambda) = (-1)^n(\lambda - \lambda_1)^{m_1}(\lambda - \lambda_2)^{m_2} \cdots (\lambda - \lambda_r)^{m_r}. \tag{6}$$

The exponent m_j is called the **algebraic multiplicity** of λ_j. The maximum number of linearly independent eigenvectors corresponding to λ_j is called the **geometric multiplicity** of λ_j. It is equal to or smaller than m_j.

A subspace S of R^n or C^n (if $\mathbf{A}$ is complex) is called an **invariant subspace** of $\mathbf{A}$ if for every $\mathbf{v}$ in S the vector $\mathbf{Av}$ is also in S. **Eigenspaces** of $\mathbf{A}$ (spaces of eigenvectors; Sec. 8.1) are important invariant subspaces of $\mathbf{A}$.

An $n \times n$ matrix $\mathbf{B}$ is called **similar** to $\mathbf{A}$ if there is a nonsingular $n \times n$ matrix $\mathbf{T}$ such that

$$\mathbf{B} = \mathbf{T}^{-1}\mathbf{AT}. \tag{7}$$

Similarity is important for the following reason.

THEOREM 2

Similar Matrices

Similar matrices have the same eigenvalues. If $\mathbf{x}$ *is an eigenvector of* $\mathbf{A}$*, then* $\mathbf{y} = \mathbf{T}^{-1}\mathbf{x}$ *is an eigenvector of* $\mathbf{B}$ *in* (7) *corresponding to the same eigenvalue.* (Proof in Sec. 8.4.)

Another theorem that has various applications in numerics is as follows.

THEOREM 3

Spectral Shift

If $\mathbf{A}$ *has the eigenvalues* $\lambda_1, \cdots, \lambda_n$*, then* $\mathbf{A} - k\mathbf{I}$ *with arbitrary* k *has the eigenvalues* $\lambda_1 - k, \cdots, \lambda_n - k$.

This theorem is a special case of the following **spectral mapping theorem.**

THEOREM 4

Polynomial Matrices

If λ *is an eigenvalue of* $\mathbf{A}$*, then*

$$q(\lambda) = \alpha_s\lambda^s + \alpha_{s-1}\lambda^{s-1} + \cdots + \alpha_1\lambda + \alpha_0$$

is an eigenvalue of the **polynomial matrix**

$$q(\mathbf{A}) = \alpha_s\mathbf{A}^s + \alpha_{s-1}\mathbf{A}^{s-1} + \cdots + \alpha_1\mathbf{A} + \alpha_0\mathbf{I}.$$

PROOF $\mathbf{Ax} = \lambda\mathbf{x}$ implies $\mathbf{A}^2\mathbf{x} = \mathbf{A}\lambda\mathbf{x} = \lambda\mathbf{Ax} = \lambda^2\mathbf{x}$, $\mathbf{A}^3\mathbf{x} = \lambda^3\mathbf{x}$, etc. Thus

$$\begin{aligned} q(\mathbf{A})\mathbf{x} &= (\alpha_s\mathbf{A}^s + \alpha_{s-1}\mathbf{A}^{s-1} + \cdots)\,\mathbf{x} \\ &= \alpha_s\mathbf{A}^s\mathbf{x} + \alpha_{s-1}\mathbf{A}^{s-1}\mathbf{x} + \cdots \\ &= \alpha_s\lambda^s\mathbf{x} + \alpha_{s-1}\lambda^{s-1}\mathbf{x} + \cdots = q(\lambda)\,\mathbf{x}. \end{aligned}$$

■

The eigenvalues of important special matrices can be characterized as follows.

THEOREM 5

Special Matrices

The eigenvalues of Hermitian matrices (i.e., $\bar{\mathbf{A}}^T = \mathbf{A}$), *hence of real symmetric matrices* (i.e., $\mathbf{A}^T = \mathbf{A}$), *are real. The eigenvalues of skew-Hermitian matrices* (i.e., $\bar{\mathbf{A}}^T = -\mathbf{A}$), *hence of real skew-symmetric matrices* (i.e., $\mathbf{A}^T = -\mathbf{A}$) *are pure imaginary or* 0. *The eigenvalues of unitary matrices* (i.e., $\bar{\mathbf{A}}^T = \mathbf{A}^{-1}$), *hence of orthogonal matrices* (i.e., $\mathbf{A}^T = \mathbf{A}^{-1}$), *have absolute value* 1. (Proofs in Secs. 8.3 and 8.5.)

The **choice of a numeric method** for matrix eigenvalue problems depends essentially on two circumstances, on the kind of matrix (real symmetric, real general, complex, sparse, or full) and on the kind of information to be obtained, that is, whether one wants to know all eigenvalues or merely specific ones, for instance, the largest eigenvalue, whether eigenvalues *and* eigenvectors are wanted, and so on. It is clear that we cannot enter into a systematic discussion of all these and further possibilities that arise in practice, but we shall concentrate on some basic aspects and methods that will give us a general understanding of this fascinating field.

20.7 Inclusion of Matrix Eigenvalues

The whole numerics for matrix eigenvalues is motivated by the fact that except for a few trivial cases we cannot determine eigenvalues *exactly* by a finite process because these values are the roots of a polynomial of nth degree. Hence we must mainly use iteration.

In this section we state a few general theorems that give approximations and error bounds for eigenvalues. Our matrices will continue to be real (except in formula (5) below), but since (nonsymmetric) matrices may have complex eigenvalues, complex numbers will play a (very modest) role in this section.

The important theorem by Gerschgorin gives a region consisting of closed circular disks in the complex plane and including all the eigenvalues of a given matrix. Indeed, for each $j = 1, \cdots, n$ the inequality (1) in the theorem determines a closed circular disk in the complex λ-plane with center a_{jj} and radius given by the right side of (1); and Theorem 1 states that each of the eigenvalues of $\mathbf{A}$ lies in one of these n disks.

THEOREM 1

Gerschgorin's Theorem

Let λ be an eigenvalue of an arbitrary $n \times n$ matrix $\mathbf{A} = [a_{jk}]$. Then for some integer j $(1 \leqq j \leqq n)$ we have

$$\textbf{(1)} \qquad |a_{jj} - \lambda| \leqq |a_{j1}| + |a_{j2}| + \cdots + |a_{j,j-1}| + |a_{j,j+1}| + \cdots + |a_{jn}|.$$

PROOF Let $\mathbf{x}$ be an eigenvector corresponding to an eigenvalue λ of $\mathbf{A}$. Then

$$(2) \qquad \mathbf{Ax} = \lambda\mathbf{x} \qquad \text{or} \qquad (\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}.$$

Let x_j be a component of $\mathbf{x}$ that is largest in absolute value. Then we have $|x_m/x_j| \leqq 1$ for $m = 1, \cdots, n$. The vector equation (2) is equivalent to a system of n equations for the

n components of the vectors on both sides. The jth of these n equations with j as just indicated is

$$a_{j1}x_1 + \cdots + a_{j,j-1}x_{j-1} + (a_{jj} - \lambda)x_j + a_{j,j+1}x_{j+1} + \cdots + a_{jn}x_n = 0.$$

Division by x_j (which cannot be zero; why?) and reshuffling terms gives

$$a_{jj} - \lambda = -a_{j1}\frac{x_1}{x_j} - \cdots - a_{j,j-1}\frac{x_{j-1}}{x_j} - a_{j,j+1}\frac{x_{j+1}}{x_j} - \cdots - a_{jn}\frac{x_n}{x_j}.$$

By taking absolute values on both sides of this equation, applying the triangle inequality $|a + b| \leqq |a| + |b|$ (where a and b are any complex numbers), and observing that because of the choice of j (which is crucial!), $|x_1/x_j| \leqq 1, \cdots, |x_n/x_j| \leqq 1$, we obtain (1), and the theorem is proved. ∎

EXAMPLE 1 **Gerschgorin's Theorem**

For the eigenvalues of the matrix

$$\mathbf{A} = \begin{bmatrix} 0 & 1/2 & 1/2 \\ 1/2 & 5 & 1 \\ 1/2 & 1 & 1 \end{bmatrix}$$

we get the Gerschgorin disks (Fig. 446)

D_1: Center 0, radius 1, D_2: Center 5, radius 1.5, D_3: Center 1, radius 1.5.

The centers are the main diagonal entries of **A**. These would be the eigenvalues of **A** if **A** were diagonal. We can take these values as crude approximations of the unknown eigenvalues (3D values) $\lambda_1 = -0.209$, $\lambda_2 = 5.305$, $\lambda_3 = 0.904$ (verify this); then the radii of the disks are corresponding error bounds.

Since **A** is symmetric, it follows from Theorem 5, Sec. 20.6, that the spectrum of **A** must actually lie in the intervals $[-1, 2.5]$ and $[3.5, 6.5]$.

It is interesting that here the Gerschgorin disks form two disjoint sets, namely, $D_1 \cup D_3$, which contains two eigenvalues, and D_2, which contains one eigenvalue. This is typical, as the following theorem shows. ∎

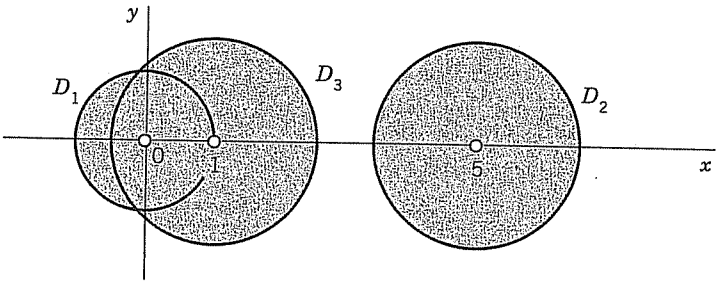


Fig. 446. Gerschgorin disks in Example 1

THEOREM 2 **Extension of Gerschgorin's Theorem**

If p Gerschgorin disks form a set S that is disjoint from the $n - p$ other disks of a given matrix **A**, *then S contains precisely p eigenvalues of* **A** *(each counted with its algebraic multiplicity, as defined in Sec.* 20.6).

Idea of Proof. Set $\mathbf{A} = \mathbf{B} + \mathbf{C}$, where **B** is the diagonal matrix with entries a_{jj}, and apply Theorem 1 to $\mathbf{A}_t = \mathbf{B} + t\mathbf{C}$ with real t growing from 0 to 1. ∎

EXAMPLE 2 **Another Application of Gerschgorin's Theorem. Similarity**

Suppose that we have diagonalized a matrix by some numeric method that left us with some off-diagonal entries of size 10^{-5}, say,

$$\mathbf{A} = \begin{bmatrix} 2 & 10^{-5} & 10^{-5} \\ 10^{-5} & 2 & 10^{-5} \\ 10^{-5} & 10^{-5} & 4 \end{bmatrix}.$$

What can we conclude about deviations of the eigenvalues from the main diagonal entries?

Solution. By Theorem 2, one eigenvalue must lie in the disk of radius $2 \cdot 10^{-5}$ centered at 4 and two eigenvalues (or an eigenvalue of algebraic multiplicity 2) in the disk of radius $2 \cdot 10^{-5}$ centered at 2. Actually, since the matrix is symmetric, these eigenvalues must lie in the intersections of these disks and the real axis, by Theorem 5 in Sec. 20.6.

We show how an isolated disk can always be reduced in size by a similarity transformation. The matrix

$$\mathbf{B} = \mathbf{T}^{-1}\mathbf{A}\mathbf{T} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 10^{-5} \end{bmatrix} \begin{bmatrix} 2 & 10^{-5} & 10^{-5} \\ 10^{-5} & 2 & 10^{-5} \\ 10^{-5} & 10^{-5} & 4 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 10^{5} \end{bmatrix}$$

$$= \begin{bmatrix} 2 & 10^{-5} & 1 \\ 10^{-5} & 2 & 1 \\ 10^{-10} & 10^{-10} & 4 \end{bmatrix}$$

is similar to **A.** Hence by Theorem 2, Sec. 20.6, it has the same eigenvalues as **A.** From Row 3 we get the smaller disk of radius $2 \cdot 10^{-10}$. Note that the other disks got bigger, approximately by a factor of 10^5. And in choosing **T** we have to watch that the new disks do not overlap with the disk whose size we want to decrease.

For further interesting facts, see the new book [E28]. ▩

By definition, a **diagonally dominant** matrix $\mathbf{A} = [a_{jk}]$ is an $n \times n$ matrix such that

$$|a_{jj}| \geqq \sum_{k \neq j} |a_{jk}| \qquad j = 1, \cdots, n \tag{3}$$

where we sum over all off-diagonal entries in Row j. The matrix is said to be **strictly diagonally dominant** if $>$ in (3) for all j. Use Theorem 1 to prove the following basic property.

THEOREM 3

Strict Diagonal Dominance

Strictly diagonally dominant matrices are nonsingular.

Further Inclusion Theorems

An **inclusion theorem** is a theorem that specifies a set which contains at least one eigenvalue of a given matrix. Thus, Theorems 1 and 2 are inclusion theorems; they even include the whole spectrum. We now discuss some famous theorems that yield further inclusions of eigenvalues. We state the first two of them without proofs (which would exceed the level of this book).

THEOREM 4

Schur's Theorem[3]

Let $\mathbf{A} = [a_{jk}]$ *be an* $n \times n$ *matrix. Then for each of its eigenvalues* $\lambda_1, \cdots, \lambda_n$,

$$|\lambda_m|^2 \leqq \sum_{i=1}^{n} |\lambda_i|^2 \leqq \sum_{j=1}^{n} \sum_{k=1}^{n} |a_{jk}|^2 \quad \textbf{(Schur's inequality).} \tag{4}$$

In (4) *the second equality sign holds if and only if* $\mathbf{A}$ *is such that*

$$\bar{\mathbf{A}}^{\mathsf{T}}\mathbf{A} = \mathbf{A}\bar{\mathbf{A}}^{\mathsf{T}}. \tag{5}$$

Matrices that satisfy (5) are called **normal matrices.** It is not difficult to see that Hermitian, skew-Hermitian, and unitary matrices are normal, and so are real symmetric, skew-symmetric, and orthogonal matrices.

EXAMPLE 3 **Bounds for Eigenvalues Obtained from Schur's Inequality**

For the matrix

$$\mathbf{A} = \begin{bmatrix} 26 & -2 & 2 \\ 2 & 21 & 4 \\ 4 & 2 & 28 \end{bmatrix}$$

we obtain from Schur's inequality $|\lambda| \leqq \sqrt{1949} = 44.1475$. You may verify that the eigenvalues are 30, 25, and 20. Thus $30^2 + 25^2 + 20^2 = 1925 < 1949$; in fact, $\mathbf{A}$ is not normal. ▩

The preceding theorems are valid for every real or complex square matrix. Other theorems hold for special classes of matrices only. Famous is the following.

THEOREM 5

Perron's Theorem[4]

Let $\mathbf{A}$ *be a real* $n \times n$ *matrix whose entries are all positive. Then* $\mathbf{A}$ *has a positive real eigenvalue* $\lambda = \rho$ *of multiplicity* 1. *The corresponding eigenvector can be chosen with all components positive. (The other eigenvalues are less than* ρ *in absolute value.)*

For a proof see Ref. [B3], vol. II, pp. 53–62. The theorem also holds for matrices with *nonnegative* real entries (**"Perron–Frobenius Theorem"**[4]) provided $\mathbf{A}$ is **irreducible,** that is, it cannot be brought to the following form by interchanging rows and columns; here $\mathbf{B}$ and $\mathbf{F}$ are square and $\mathbf{0}$ is a zero matrix.

$$\begin{bmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{0} & \mathbf{F} \end{bmatrix}$$

[3]ISSAI SCHUR (1875–1941), German mathematician, also known by his important work in group theory.

[4]OSKAR PERRON (1880–1975), GEORG FROBENIUS (1849–1917), LOTHAR COLLATZ (1910–1990), German mathematicians, known for their work in potential theory, ODEs (Sec. 5.4) and group theory, and numerics, respectively.

Perron's theorem has various applications, for instance, in economics. It is interesting that one can obtain from it a theorem that gives a numeric algorithm:

THEOREM 6

Collatz Inclusion Theorem[4]

Let $\mathbf{A} = [a_{jk}]$ *be a real* $n \times n$ *matrix whose elements are all positive. Let* $\mathbf{x}$ *be any real vector whose components* $x_1, \cdots, x_n$ *are positive, and let* $y_1, \cdots, y_n$ *be the components of the vector* $\mathbf{y} = \mathbf{Ax}$. *Then the closed interval on the real axis bounded by the smallest and the largest of the* n *quotients* $q_j = y_j/x_j$ *contains at least one eigenvalue of* $\mathbf{A}$.

PROOF We have $\mathbf{Ax} = \mathbf{y}$ or

$$\mathbf{y} - \mathbf{Ax} = \mathbf{0}. \tag{6}$$

The transpose $\mathbf{A}^T$ satisfies the conditions of Theorem 5. Hence $\mathbf{A}^T$ has a positive eigenvalue λ and, corresponding to this eigenvalue, an eigenvector $\mathbf{u}$ whose components u_j are all positive. Thus $\mathbf{A}^T\mathbf{u} = \lambda\mathbf{u}$, and by taking the transpose we obtain $\mathbf{u}^T\mathbf{A} = \lambda\mathbf{u}^T$. From this and (6) we have

$$\mathbf{u}^T(\mathbf{y} - \mathbf{Ax}) = \mathbf{u}^T\mathbf{y} - \mathbf{u}^T\mathbf{Ax} = \mathbf{u}^T\mathbf{y} - \lambda\mathbf{u}^T\mathbf{x} = \mathbf{u}^T(\mathbf{y} - \lambda\mathbf{x}) = 0$$

or written out

$$\sum_{j=1}^{n} u_j(y_j - \lambda x_j) = 0.$$

Since all the components u_j are positive, it follows that

$$\text{(7)} \qquad \begin{array}{llll} y_j - \lambda x_j \geqq 0, & \text{that is,} & q_j \geqq \lambda & \text{for at least one } j, \\ y_j - \lambda x_j \leqq 0, & \text{that is,} & q_j \leqq \lambda & \text{for at least one } j. \end{array} \qquad \text{and}$$

Since $\mathbf{A}$ and $\mathbf{A}^T$ have the same eigenvalues, λ is an eigenvalue of $\mathbf{A}$, and from (7) the statement of the theorem follows. ∎

EXAMPLE 4 **Bounds for Eigenvalues from Collatz's Theorem. Iteration**

For a given matrix $\mathbf{A}$ with positive entries we choose an $\mathbf{x} = \mathbf{x}_0$ and **iterate,** that is, we compute $\mathbf{x}_1 = \mathbf{Ax}_0$. $\mathbf{x}_2 = \mathbf{Ax}_1, \cdots, \mathbf{x}_{20} = \mathbf{Ax}_{19}$. In each step, taking $\mathbf{x} = \mathbf{x}_j$ and $\mathbf{y} = \mathbf{Ax}_j = \mathbf{x}_{j+1}$ we compute an inclusion interval by Collatz's theorem. This gives (6S)

$$\mathbf{A} = \begin{bmatrix} 0.49 & 0.02 & 0.22 \\ 0.02 & 0.28 & 0.20 \\ 0.22 & 0.20 & 0.40 \end{bmatrix}, \mathbf{x}_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \mathbf{x}_1 = \begin{bmatrix} 0.73 \\ 0.50 \\ 0.82 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} 0.5481 \\ 0.3186 \\ 0.5886 \end{bmatrix},$$

$$\cdots, \mathbf{x}_{19} = \begin{bmatrix} 0.00216309 \\ 0.00108155 \\ 0.00216309 \end{bmatrix}, \mathbf{x}_{20} = \begin{bmatrix} 0.00155743 \\ 0.000778713 \\ 0.00155743 \end{bmatrix}$$

and the intervals $0.5 \leqq \lambda \leqq 0.82$, $0.3186/0.50 = 0.6372 \leqq \lambda \leqq 0.5481/0.73 = 0.750822$, etc. These intervals have length

j	1	2	3	10	15	20
Length	0.32	0.113622	0.0539835	0.0004217	0.0000132	0.0000004

Using the characteristic polynomial, you may verify that the eigenvalues of $\mathbf{A}$ are 0.72, 0.36, 0.09, so that those intervals include the largest eigenvalue, 0.72. Their lengths decreased with j, so that the iteration was worthwhile. The reason will appear in the next section, where we discuss an iteration method for eigenvalues. ■

PROBLEM SET 20.7

1–6 GERSCHGORIN DISKS

Find and sketch disks or intervals that contain the eigenvalues. If you have a CAS, find the spectrum and compare.

1. $\begin{bmatrix} 6 & -1 & 0 \\ -1 & 7 & 1 \\ 0 & 1 & 8 \end{bmatrix}$

2. $\begin{bmatrix} 5 & 10^{-2} & 10^{-2} \\ 10^{-2} & 8 & 10^{-2} \\ 10^{-2} & 10^{-2} & 9 \end{bmatrix}$

3. $\begin{bmatrix} 5 & -2 & 2 \\ 2 & 0 & 4 \\ 4 & 2 & 7 \end{bmatrix}$

4. $\begin{bmatrix} 1+i & 0.3 & 0.5i \\ 0.3i & -3+2i & 0.1 \\ 0.1i & 0.2 & 4-i \end{bmatrix}$

5. $\begin{bmatrix} 4i & 0.1i & 1+i \\ 0.1i & 0 & 0 \\ -1+i & 0 & 9i \end{bmatrix}$

6. $\begin{bmatrix} 10 & 0.1 & -0.2 \\ 0.1 & 6 & 0 \\ -0.2 & 0 & 3 \end{bmatrix}$

7. (Similarity) Find $\mathbf{T}^{-\mathrm{T}}\mathbf{A}\mathbf{T}$ such that in Prob. 2 the radius of the Gerschgorin circle with center 5 is reduced by a factor 1/100.

8. By what integer factor can you at most reduce the Gerschgorin circle with center 3 in Prob. 6?

9. If a symmetric $n \times n$ matrix $\mathbf{A} = [a_{jk}]$ has been diagonalized except for small off-diagonal entries of size 10^{-6}, what can you say about the eigenvalues?

10. (Extended Gerschgorin theorem) Prove Theorem 2.

11. Prove Theorem 3.

12. (Normal matrices) Show that Hermitian, skew-Hermitian, and unitary matrices (hence real symmetric, skew-symmetric, and orthogonal matrices) are normal. Why is this of practical interest?

13. (Spectral radius $\rho(\mathbf{A})$) Show that $\rho(\mathbf{A})$ cannot be greater than the row sum norm of $\mathbf{A}$.

14. (Eigenvalues on the circle) Illustrate with a 2×2 matrix that an eigenvalue may very well lie on a Gerschgorin circle (so that Gerschgorin disks can generally not be replaced with smaller disks without losing the inclusion property).

15–17 SCHUR'S INEQUALITY

Use (4) to obtain an upper bound for the spectral radius:

15. In Prob. 1

16. In Prob. 6

17. In Prob. 3

18–19 COLLATZ'S THEOREM

Apply Theorem 6, choosing the given vectors as vectors $\mathbf{x}$.

18. $\begin{bmatrix} 10 & 1 & 1 \\ 1 & 9 & 3 \\ 1 & 3 & 11 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 3 \\ 2 \end{bmatrix}$

19. $\begin{bmatrix} 4 & 2 & 2 \\ 2 & 4 & 2 \\ 2 & 2 & 4 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$

20. CAS EXPERIMENT. Collatz Iteration. (a) Write a program for the iteration in Example 4 (with any $\mathbf{A}$ and $\mathbf{x}_0$) that at each step prints the midpoint (why?), the endpoints, and the length of the inclusion interval.

(b) Apply the program to symmetric matrices of your choice. Explore how convergence depends on the choice of initial vectors. Can you construct cases in which the lengths of the inclusion intervals are not monotone decreasing? Can you explain the reason? Can you experiment on the effect of rounding?

20.8 Power Method for Eigenvalues

A simple standard procedure for computing approximate values of the eigenvalues of an $n \times n$ matrix $\mathbf{A} = [a_{jk}]$ is the **power method.** In this method we start from any vector $\mathbf{x}_0$ ($\neq \mathbf{0}$) with n components and compute successively

$$\mathbf{x}_1 = \mathbf{A}\mathbf{x}_0, \qquad \mathbf{x}_2 = \mathbf{A}\mathbf{x}_1, \qquad \cdots, \qquad \mathbf{x}_s = \mathbf{A}\mathbf{x}_{s-1}.$$

For simplifying notation, we denote $\mathbf{x}_{s-1}$ by $\mathbf{x}$ and $\mathbf{x}_s$ by $\mathbf{y}$, so that $\mathbf{y} = \mathbf{A}\mathbf{x}$.

The method applies to any $n \times n$ matrix $\mathbf{A}$ that has a **dominant eigenvalue** (a λ such that $|\lambda|$ is greater than the absolute values of the other eigenvalues). If $\mathbf{A}$ is *symmetric,* it also gives the error bound (2), in addition to the approximation (1).

THEOREM 1

Power Method, Error Bounds

Let $\mathbf{A}$ be an $n \times n$ real symmetric matrix. Let $\mathbf{x}$ ($\neq \mathbf{0}$) be any real vector with n components. Furthermore, let

$$\mathbf{y} = \mathbf{A}\mathbf{x}, \qquad m_0 = \mathbf{x}^{\mathsf{T}}\mathbf{x}, \qquad m_1 = \mathbf{x}^{\mathsf{T}}\mathbf{y}, \qquad m_2 = \mathbf{y}^{\mathsf{T}}\mathbf{y}.$$

Then the quotient

(1) $$q = \frac{m_1}{m_0}$$ **(Rayleigh[5] quotient)**

is an approximation for an eigenvalue λ of $\mathbf{A}$ (usually that which is greatest in absolute value, but no general statements are possible).

Furthermore, if we set $q = \lambda - \epsilon$, so that ϵ is the error of q, then

(2) $$|\epsilon| \leqq \delta = \sqrt{\frac{m_2}{m_0} - q^2}.$$

[5]LORD RAYLEIGH (JOHN WILLIAM STRUTT) (1842–1919), great English physicist and mathematician, professor at Cambridge and London, known for his important contributions to various branches of applied mathematics and theoretical physics, in particular, the theory of waves, elasticity, and hydrodynamics. In 1904 he received a Nobel Prize in physics.

PROOF δ^2 denotes the radicand in (2). Since $m_1 = qm_0$ by (1), we have

$$(3) \qquad (\mathbf{y} - q\mathbf{x})^\mathsf{T}(\mathbf{y} - q\mathbf{x}) = m_2 - 2qm_1 + q^2m_0 = m_2 - q^2m_0 = \delta^2 m_0.$$

Since $\mathbf{A}$ is real symmetric, it has an orthogonal set of n real unit eigenvectors $\mathbf{z}_1, \cdots, \mathbf{z}_n$ corresponding to the eigenvalues $\lambda_1, \cdots, \lambda_n$, respectively (some of which may be equal). (Proof in Ref. [B3], vol. 1, pp. 270–272, listed in App. 1.) Then $\mathbf{x}$ has a representation of the form

$$\mathbf{x} = a_1\mathbf{z}_1 + \cdots + a_n\mathbf{z}_n.$$

Now $\mathbf{A}\mathbf{z}_1 = \lambda_1\mathbf{z}_1$, etc., and we obtain

$$\mathbf{y} = \mathbf{A}\mathbf{x} = a_1\lambda_1\mathbf{z}_1 + \cdots + a_n\lambda_n\mathbf{z}_n$$

and, since the $\mathbf{z}_j$ are orthogonal unit vectors,

$$(4) \qquad m_0 = \mathbf{x}^\mathsf{T}\mathbf{x} = a_1^2 + \cdots + a_n^2.$$

It follows that in (3),

$$\mathbf{y} - q\mathbf{x} = a_1(\lambda_1 - q)\mathbf{z}_1 + \cdots + a_n(\lambda_n - q)\mathbf{z}_n.$$

Since the $\mathbf{z}_j$ are orthogonal unit vectors, we thus obtain from (3)

$$(5) \qquad \delta^2 m_0 = (\mathbf{y} - q\mathbf{x})^\mathsf{T}(\mathbf{y} - q\mathbf{x}) = a_1^2(\lambda_1 - q)^2 + \cdots + a_n^2(\lambda_n - q)^2.$$

Now let λ_c be an eigenvalue of $\mathbf{A}$ to which q is closest, where c suggests "closest". Then $(\lambda_c - q)^2 \leqq (\lambda_j - q)^2$ for $j = 1, \cdots, n$. From this and (5) we obtain the inequality

$$\delta^2 m_0 \geqq (\lambda_c - q)^2(a_1^2 + \cdots + a_n^2) = (\lambda_c - q)^2 m_0.$$

Dividing by m_0, taking square roots, and recalling the meaning of δ^2 gives

$$\delta = \sqrt{\frac{m_2}{m_0} - q^2} \geqq |\lambda_c - q|.$$

This shows that δ is a bound for the error ϵ of the approximation q of an eigenvalue of $\mathbf{A}$ and completes the proof. ∎

The main advantage of the method is its simplicity. And it can handle *sparse matrices* too large to store as a full square array. Its disadvantage is its possibly slow convergence. From the proof of Theorem 1 we see that the speed of convergence depends on the ratio of the dominant eigenvalue to the next in absolute value (2:1 in Example 1, below).

If we want a convergent sequence of **eigenvectors,** then at the beginning of each step we **scale** the vector, say, by dividing its components by an absolutely largest one, as in Example 1, as follows.

EXAMPLE 1 **Application of Theorem 1. Scaling**

For the symmetric matrix $\mathbf{A}$ in Example 4, Sec. 20.7, and $\mathbf{x}_0 = [1 \quad 1 \quad 1]^T$ we obtain from (1) and (2) and the indicated scaling

$$\mathbf{A} = \begin{bmatrix} 0.49 & 0.02 & 0.22 \\ 0.02 & 0.28 & 0.20 \\ 0.22 & 0.20 & 0.40 \end{bmatrix}, \quad \mathbf{x}_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{x}_1 = \begin{bmatrix} 0.890244 \\ 0.609756 \\ 1 \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} 0.931193 \\ 0.541284 \\ 1 \end{bmatrix}$$

$$\mathbf{x}_5 = \begin{bmatrix} 0.990663 \\ 0.504682 \\ 1 \end{bmatrix}, \qquad \mathbf{x}_{10} = \begin{bmatrix} 0.999707 \\ 0.500146 \\ 1 \end{bmatrix}, \qquad \mathbf{x}_{15} = \begin{bmatrix} 0.999991 \\ 0.500005 \\ 1 \end{bmatrix}.$$

Here $\mathbf{A}\mathbf{x}_0 = [0.73 \quad 0.5 \quad 0.82]^T$, scaled to $\mathbf{x}_1 = [0.73/0.82 \quad 0.5/0.82 \quad 1]^T$, etc. The dominant eigenvalue is 0.72, an eigenvector $[1 \quad 0.5 \quad 1]^T$. The corresponding q and δ are computed each time before the next scaling. Thus in the first step,

$$q = \frac{m_1}{m_0} = \frac{\mathbf{x}_0^T \mathbf{A}\mathbf{x}_0}{\mathbf{x}_0^T \mathbf{x}_0} = \frac{2.05}{3} = 0.683333$$

$$\delta = \left(\frac{m_2}{m_0} - q^2 \right)^{1/2} = \left(\frac{(\mathbf{A}\mathbf{x}_0)^T \mathbf{A}\mathbf{x}_0}{\mathbf{x}_0^T \mathbf{x}_0} - q^2 \right)^{1/2} = \left(\frac{1.4553}{3} - q^2 \right)^{1/2} = 0.134743.$$

This gives the following values of q, δ, and the error $\epsilon = 0.72 - q$ (calculations with 10D, rounded to 6D):

j	1	2	5	10
q	0.683333	0.716048	0.719944	0.720000
δ	0.134743	0.038887	0.004499	0.000141
ϵ	0.036667	0.003952	0.000056	$5 \cdot 10^{-8}$

The error bounds are much larger than the actual errors. This is typical, although the bounds cannot be improved; that is, for special symmetric matrices they agree with the errors.

Our present results are somewhat better than those of Collatz's method in Example 4 of Sec. 20.7, at the expense of more operations. ■

Spectral shift, the transition from $\mathbf{A}$ to $\mathbf{A} - k\mathbf{I}$, shifts every eigenvalue by $-k$. Although finding a good k can hardly be made automatic, it may be helped by some other method or small preliminary computational experiments. In Example 1, Gerschgorin's theorem gives $-0.02 \leqq \lambda \leqq 0.82$ for the whole spectrum (verify!). Shifting by -0.4 might be too much (then $-0.42 \leqq \lambda \leqq 0.42$), so let us try -0.2.

EXAMPLE 2 **Power Method with Spectral Shift**

For $\mathbf{A} - 0.2\mathbf{I}$ with $\mathbf{A}$ as in Example 1 we obtain the following substantial improvements (where the index 1 refers to Example 1 and the index 2 to the present example).

j	1	2	5	10
δ_1	0.134743	0.038887	0.004499	0.000141
δ_2	0.134743	0.034474	0.000693	$1.8 \cdot 10^{-6}$
ϵ_1	0.036667	0.003952	0.000056	$5 \cdot 10^{-8}$
ϵ_2	0.036667	0.002477	$1.3 \cdot 10^{-6}$	$9 \cdot 10^{-12}$

■

PROBLEM SET 20.8

1–7 POWER METHOD WITH SCALING

Apply the power method (3 steps) with scaling, using $\mathbf{x}_0 = [1 \quad 1]^T$ or $[1 \quad 1 \quad 1]^T$ or $[1 \quad 1 \quad 1 \quad 1]^T$, as applicable. Give Rayleigh quotients and error bounds. Show the details of your work.

1. $\begin{bmatrix} 3.5 & 2.0 \\ 2.0 & 0.5 \end{bmatrix}$ **2.** $\begin{bmatrix} 0.6 & 0.8 \\ 0.8 & -0.6 \end{bmatrix}$

3. $\begin{bmatrix} 5 & 2 \\ 2 & 7 \end{bmatrix}$

4. $\begin{bmatrix} -2 & 2 & 3 \\ 2 & 1 & 6 \\ 3 & 6 & -2 \end{bmatrix}$ **5.** $\begin{bmatrix} 2 & -1 & 1 \\ -1 & 3 & 2 \\ 1 & 2 & 3 \end{bmatrix}$

6. $\begin{bmatrix} 0 & 4 & 0 & 1 \\ 4 & -1 & 2 & 8 \\ 0 & 2 & 3 & 2 \\ 1 & 8 & 2 & -2 \end{bmatrix}$ **7.** $\begin{bmatrix} 5 & 1 & 0 & 0 \\ 1 & 3 & 1 & 0 \\ 0 & 1 & 3 & 1 \\ 0 & 0 & 1 & 5 \end{bmatrix}$

8. (Optimality of δ) In Prob. 2 choose $\mathbf{x}_0 = [3 \quad -1]^T$ and show that $q = 0$ and $\delta = 1$ for all steps and that the eigenvalues are ± 1, so that the interval $[q - \delta, q + \delta]$ cannot be shortened in general! Experiment with other $\mathbf{x}_0$.

9. Prove that if $\mathbf{x}$ is an eigenvector, then $\delta = 0$ in (2). Give two examples.

10. (Rayleigh quotient) Why does q generally approximate the eigenvalue of greatest absolute value? When will q be a good approximation?

11. (Spectral shift, smallest eigenvalue) In Prob. 5 set $\mathbf{B} = \mathbf{A} - 3\mathbf{I}$ (as perhaps suggested by the diagonal entries) and try whether you may get a sequence of q's converging to an eigenvalue of $\mathbf{A}$ that *is smallest* (not largest) in absolute value. Use $\mathbf{x}_0 = [1 \quad 1 \quad 1]^T$. Do 8 steps. Verify that $\mathbf{A}$ has the spectrum $\{0, 3, 5\}$.

12. CAS EXPERIMENT. Power Method with Scaling. Shifting. **(a)** Write a program for $n \times n$ matrices that prints every step. Apply it to the (nonsymmetric!) matrix (20 steps), starting from $[1 \quad 1 \quad 1]^T$.

$$\mathbf{A} = \begin{bmatrix} 15 & 12 & 3 \\ 18 & 44 & 18 \\ -19 & -36 & -7 \end{bmatrix}.$$

(b) Experiment in (a) with shifting. Which shift do you find optimal?

(c) Write a program as in (a) but for symmetric matrices that prints vectors, scaled vectors, q, and δ. Apply it to the matrix in Prob. 6.

(d) Find a (nonsymmetric) matrix for which δ in (2) is no longer an error bound.

(e) Experiment systematically with speed of convergence by choosing matrices with the second greatest eigenvalue (i) almost equal to the greatest, (ii) somewhat different, (iii) much different.

20.9 Tridiagonalization and QR-Factorization

We consider the problem of computing *all* the eigenvalues of a ***real symmetric*** matrix $\mathbf{A} = [a_{jk}]$, discussing a method widely used in practice. In the ***first stage*** we reduce the given matrix stepwise to a **tridiagonal matrix,** that is, a matrix having all its nonzero entries on the main diagonal and in the positions immediately adjacent to the main diagonal (such as $\mathbf{A}_3$ in Fig. 447, Third Step). This reduction was invented by A. S. Householder (*J. Assn. Comput. Machinery* **5** (1958), 335–342). See also Ref. [E29] in App. 1.

This Householder tridiagonalization will simplify the matrix without changing its eigenvalues. The latter will then be determined (approximately) by factoring the tridiagonalized matrix, as discussed later in this section.

Householder's Tridiagonalization Method

An $n \times n$ real symmetric matrix $\mathbf{A} = [a_{jk}]$ being given, we reduce it by $n - 2$ successive similarity transformations (see Sec. 20.6) involving matrices $\mathbf{P}_1, \cdots, \mathbf{P}_{n-2}$ to tridiagonal form. These matrices are orthogonal and symmetric. Thus $\mathbf{P}_1^{-1} = \mathbf{P}_1^{\mathsf{T}} = \mathbf{P}_1$ and similarly for the others. These transformations produce from the given $\mathbf{A}_0 = \mathbf{A} = [a_{jk}]$ the matrices $\mathbf{A}_1 = [a_{jk}^{(1)}]$, $\mathbf{A}_2 = [a_{jk}^{(2)}]$, $\cdots$, $\mathbf{A}_{n-2} = [a_{jk}^{(n-2)}]$ in the form

$$\begin{aligned} \mathbf{A}_1 &= \mathbf{P}_1\mathbf{A}_0\mathbf{P}_1 \\ \mathbf{A}_2 &= \mathbf{P}_2\mathbf{A}_1\mathbf{P}_2 \\ &\cdots\cdots\cdots\cdots \\ \mathbf{B} = \mathbf{A}_{n-2} &= \mathbf{P}_{n-2}\mathbf{A}_{n-3}\mathbf{P}_{n-2}. \end{aligned} \tag{1}$$

The transformations (1) create the necessary zeros, in the first step in Row 1 and Column 1, in the second step in Row 2 and Column 2, etc., as Fig. 447 illustrates for a 5×5 matrix. $\mathbf{B}$ is tridiagonal.

How do we determine $\mathbf{P}_1, \mathbf{P}_2, \cdots \mathbf{P}_{n-2}$? Now, all these $\mathbf{P}_r$ are of the form

$$\mathbf{P}_r = \mathbf{I} - 2\mathbf{v}_r\mathbf{v}_r^{\mathsf{T}} \qquad (r = 1, \cdots, n-2) \tag{2}$$

where $\mathbf{I}$ is the $n \times n$ unit matrix and $\mathbf{v}_r = [v_{jr}]$ is a unit vector with its first r components 0; thus

$$\mathbf{v}_1 = \begin{bmatrix} 0 \\ * \\ * \\ \vdots \\ * \end{bmatrix}, \qquad \mathbf{v}_2 = \begin{bmatrix} 0 \\ 0 \\ * \\ \vdots \\ * \end{bmatrix}, \qquad \cdots, \qquad \mathbf{v}_{n-2} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ * \\ * \end{bmatrix} \tag{3}$$

where the asterisks denote the other components (which will be nonzero in general).

Step 1. $\mathbf{v}_1$ has the components

$$v_{11} = 0$$

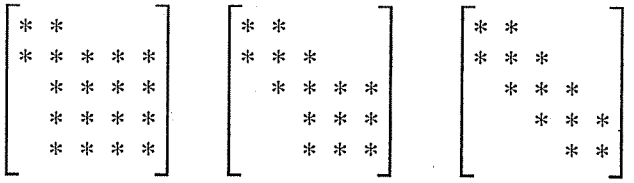
$$\text{(a)} \qquad v_{21} = \sqrt{\frac{1}{2}\left(1 + \frac{|a_{21}|}{S_1}\right)}$$

$$\text{(b)} \qquad v_{j1} = \frac{a_{j1} \operatorname{sgn} a_{21}}{2v_{21}S_1} \qquad j = 3, 4, \cdots, n \tag{4}$$

where

$$\text{(c)} \qquad S_1 = \sqrt{a_{21}^2 + a_{31}^2 + \cdots + a_{n1}^2}$$

where $S_1 > 0$, and $\operatorname{sgn} a_{21} = +1$ if $a_{21} \geqq 0$ and $\operatorname{sgn} a_{21} = -1$ if $a_{21} < 0$. With this we compute $\mathbf{P}_1$ by (2) and then $\mathbf{A}_1$ by (1). This was the first step.



First Step $\mathbf{A}_1 = \mathbf{P}_1\mathbf{A}\mathbf{P}_1$ Second Step $\mathbf{A}_2 = \mathbf{P}_2\mathbf{A}_1\mathbf{P}_2$ Third Step $\mathbf{A}_3 = \mathbf{P}_3\mathbf{A}_2\mathbf{P}_3$

Fig. 447. Householder's method for a 5 × 5 matrix. Positions left blank are zeros created by the method.

Step 2. We compute $\mathbf{v}_2$ by (4) with all subscripts increased by 1 and the a_{jk} replaced by $a_{jk}^{(1)}$, the entries of $\mathbf{A}_1$ just computed. Thus [see also (3)]

$$v_{12} = v_{22} = 0$$

$$v_{32} = \sqrt{\frac{1}{2}\left(1 + \frac{|a_{32}^{(1)}|}{S_2}\right)} \tag{4*}$$

$$v_{j2} = \frac{a_{j2}^{(1)} \operatorname{sgn} a_{32}^{(1)}}{2v_{32}S_2} \qquad j = 4, 5, \cdots, n$$

where

$$S_2 = \sqrt{a_{32}^{(1)2} + a_{42}^{(1)2} + \cdots + a_{n2}^{(1)2}}.$$

With this we compute $\mathbf{P}_2$ by (2) and then $\mathbf{A}_2$ by (1).

Step 3. We compute $\mathbf{v}_3$ by (4*) with all subscripts increased by 1 and the $a_{jk}^{(1)}$ replaced by the entries $a_{jk}^{(2)}$ of $\mathbf{A}_2$, and so on.

EXAMPLE 1 Householder Tridiagonalization

Tridiagonalize the real symmetric matrix

$$\mathbf{A} = \mathbf{A}_0 = \begin{bmatrix} 6 & 4 & 1 & 1 \\ 4 & 6 & 1 & 1 \\ 1 & 1 & 5 & 2 \\ 1 & 1 & 2 & 5 \end{bmatrix}.$$

Solution. ***Step 1.*** We compute $s_1^2 = 4^2 + 1^2 + 1^2 = 18$ from (4c). Since $a_{21} = 4 > 0$, we have $\operatorname{sgn} a_{21} = +1$ in (4b) and get from (4) by straightforward computation

$$\mathbf{v}_1 = \begin{bmatrix} 0 \\ v_{21} \\ v_{31} \\ v_{41} \end{bmatrix} = \begin{bmatrix} 0 \\ 0.985\,598\,56 \\ 0.119\,573\,16 \\ 0.119\,573\,16 \end{bmatrix}.$$

From this and (2),

$$\mathbf{P}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -0.942\,809\,04 & -0.235\,702\,27 & -0.235\,702\,27 \\ 0 & -0.235\,702\,27 & 0.971\,404\,52 & -0.028\,595\,48 \\ 0 & -0.235\,702\,27 & -0.028\,595\,48 & 0.971\,404\,52 \end{bmatrix}.$$

From the first line in (1) we now get

$$\mathbf{A}_1 = \mathbf{P}_1\mathbf{A}_0\mathbf{P}_1 = \begin{bmatrix} 6 & -\sqrt{18} & 0 & 0 \\ -\sqrt{18} & 7 & -1 & -1 \\ 0 & -1 & 9/2 & 3/2 \\ 0 & -1 & 3/2 & 9/2 \end{bmatrix}.$$

Step 2. From (4*) we compute $s_2^2 = 2$ and

$$\mathbf{v}_2 = \begin{bmatrix} 0 \\ 0 \\ v_{32} \\ v_{42} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0.923\,879\,53 \\ 0.382\,683\,43 \end{bmatrix}.$$

From this and (2),

$$\mathbf{P}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1/\sqrt{2} & -1/\sqrt{2} \\ 0 & 0 & -1\sqrt{2} & 1/\sqrt{2} \end{bmatrix}.$$

The second line in (1) now gives

$$\mathbf{B} = \mathbf{A}_2 = \mathbf{P}_2\mathbf{A}_1\mathbf{P}_2 = \begin{bmatrix} 6 & -\sqrt{18} & 0 & 0 \\ -\sqrt{18} & 7 & \sqrt{2} & 0 \\ 0 & \sqrt{2} & 6 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}.$$

This matrix $\mathbf{B}$ is tridiagonal. Since our given matrix has order $n = 4$, we needed $n - 2 = 2$ steps to accomplish this reduction, as claimed. (Do you see that we got more zeros than we can expect in general?)

$\mathbf{B}$ is similar to $\mathbf{A}$, as we now show in general. This is essential because $\mathbf{B}$ thus has the same spectrum as $\mathbf{A}$, by Theorem 2 in Sec. 20.6. ▩

B *Similar to* A. We assert that $\mathbf{B}$ in (1) is similar to $\mathbf{A} = \mathbf{A}_0$. The matrix $\mathbf{P}_r$ is symmetric; indeed,

$$\mathbf{P}_r^{\mathsf{T}} = (\mathbf{I} - 2\mathbf{v}_r\mathbf{v}_r^{\mathsf{T}})^{\mathsf{T}} = \mathbf{I}^{\mathsf{T}} - 2(\mathbf{v}_r\mathbf{v}_r^{\mathsf{T}})^{\mathsf{T}} = \mathbf{I} - 2\mathbf{v}_r\mathbf{v}_r^{\mathsf{T}} = \mathbf{P}_r.$$

Also, $\mathbf{P}_r$ is orthogonal because $\mathbf{v}_r$ is a unit vector, so that $\mathbf{v}_r^{\mathsf{T}}\mathbf{v}_r = 1$ and thus

$$\begin{aligned} \mathbf{P}_r\mathbf{P}_r^{\mathsf{T}} = \mathbf{P}_r^2 = (\mathbf{I} - 2\mathbf{v}_r\mathbf{v}_r^{\mathsf{T}})^2 &= \mathbf{I} - 4\mathbf{v}_r\mathbf{v}_r^{\mathsf{T}} + 4\mathbf{v}_r\mathbf{v}_r^{\mathsf{T}}\mathbf{v}_r\mathbf{v}_r^{\mathsf{T}} \\ &= \mathbf{I} - 4\mathbf{v}_r\mathbf{v}_r^{\mathsf{T}} + 4\mathbf{v}_r(\mathbf{v}_r^{\mathsf{T}}\mathbf{v}_r)\mathbf{v}_r^{\mathsf{T}} = \mathbf{I}. \end{aligned}$$

Hence $\mathbf{P}_r^{-1} = \mathbf{P}_r^{\mathsf{T}} = \mathbf{P}_r$ and from (1) we now obtain

$$\begin{aligned} \mathbf{B} &= \mathbf{P}_{n-2}\mathbf{A}_{n-3}\mathbf{P}_{n-2} = \cdots \\ \cdots &= \mathbf{P}_{n-2}\mathbf{P}_{n-3}\cdots\mathbf{P}_1\mathbf{A}\mathbf{P}_1\cdots\mathbf{P}_{n-3}\mathbf{P}_{n-2} \\ &= \mathbf{P}_{n-2}^{-1}\mathbf{P}_{n-3}^{-1}\cdots\mathbf{P}_1^{-1}\mathbf{A}\mathbf{P}_1\cdots\mathbf{P}_{n-3}\mathbf{P}_{n-2} \\ &= \mathbf{P}^{-1}\mathbf{A}\mathbf{P} \end{aligned}$$

where $\mathbf{P} = \mathbf{P}_1\mathbf{P}_2\cdots\mathbf{P}_{n-2}$. This proves our assertion. ▩

QR-Factorization Method

In 1958 H. Rutishauser of Switzerland proposed the idea of using the LU-factorization (Sec. 20.2; he called it LR-factorization) in solving eigenvalue problems. An improved version of Rutishauser's method (avoiding breakdown if certain submatrices become singular, etc.; see Ref. [E29]) is the QR-method, independently proposed by the American J. G. F. Francis (*Computer J.* **4** (1961–62), 265–271, 332–345) and the Russian V. N. Kublanovskaya (*Zhurnal Vych. Mat. i Mat. Fiz.* **1** (1961), 555–570). The QR-method uses the factorization **QR** with orthogonal **Q** and upper triangular **R.** We discuss the QR-method for a real ***symmetric*** matrix. (For extensions to general matrices see Ref. [E29] in App. 1.)

In this method we first transform a given real symmetric $n \times n$ matrix $\mathbf{A}$ into a tridiagonal matrix $\mathbf{B_0} = \mathbf{B}$ by Householder's method. This creates many zeros and thus reduces the amount of further work. Then we compute $\mathbf{B_1}, \mathbf{B_2}, \cdots$ stepwise according to the following iteration method.

Step 1. Factor $\mathbf{B_0} = \mathbf{Q_0R_0}$ with orthogonal $\mathbf{R_0}$ and upper triangular $\mathbf{R_0}$. Then compute $\mathbf{B_1} = \mathbf{R_0Q_0}$.

Step 2. Factor $\mathbf{B_1} = \mathbf{Q_1R_1}$. Then compute $\mathbf{B_2} = \mathbf{R_1Q_1}$.

General Step s + 1.

(5)
$$\text{(a)} \quad \text{Factor } \mathbf{B}_s = \mathbf{Q}_s\mathbf{R}_s.$$
$$\text{(b)} \quad \text{Compute } \mathbf{B}_{s+1} = \mathbf{R}_s\mathbf{Q}_s.$$

Here $\mathbf{Q}_s$ is orthogonal and $\mathbf{R}_s$ upper triangular. The factorization (5a) will be explained below.

$\mathbf{B}_{s+1}$ Similar to B. Convergence to a Diagonal Matrix. From (5a) we have $\mathbf{R}_s = \mathbf{Q}_s^{-1}\mathbf{B}_s$. Substitution into (5b) gives

(6)
$$\mathbf{B}_{s+1} = \mathbf{R}_s\mathbf{Q}_s = \mathbf{Q}_s^{-1}\mathbf{B}_s\mathbf{Q}_s.$$

Thus $\mathbf{B}_{s+1}$ is similar to $\mathbf{B}_s$. Hence $\mathbf{B}_{s+1}$ is similar to $\mathbf{B_0} = \mathbf{B}$ for all s. By Theorem 2, Sec. 20.6, this implies that $\mathbf{B}_{s+1}$ has the same eigenvalues as **B.**

Also, $\mathbf{B}_{s+1}$ is symmetric. This follows by induction. Indeed, $\mathbf{B_0} = \mathbf{B}$ is symmetric. Assuming $\mathbf{B}_s$ to be symmetric, that is, $\mathbf{B}_s^{\mathsf{T}} = \mathbf{B}_s$, and using $\mathbf{Q}_s^{-1} = \mathbf{Q}_s^{\mathsf{T}}$ (since $\mathbf{Q}_s$ is orthogonal), we get from (6) the symmetry,

$$\mathbf{B}_{s+1}^{\mathsf{T}} = (\mathbf{Q}_s^{\mathsf{T}}\mathbf{B}_s\mathbf{Q}_s)^{\mathsf{T}} = \mathbf{Q}_s^{\mathsf{T}}\mathbf{B}_s^{\mathsf{T}}\mathbf{Q}_s = \mathbf{Q}_s^{\mathsf{T}}\mathbf{B}_s\mathbf{Q}_s = \mathbf{B}_{s+1}.$$

If the eigenvalues of **B** are different in absolute value, say, $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|$, then

$$\lim_{s\to\infty} \mathbf{B}_s = \mathbf{D}$$

where **D** is diagonal, with main diagonal entries $\lambda_1, \lambda_2, \cdots, \lambda_n$. (Proof in Ref. [E29] listed in App. 1.)

How to Get the QR-Factorization, say, $\mathbf{B} = \mathbf{B}_0 = [b_{jk}] = \mathbf{Q}_0\mathbf{R}_0$. The tridiagonal matrix $\mathbf{B}$ has $n - 1$ generally nonzero entries below the main diagonal. These are $b_{21}, b_{32}, \cdots, b_{n,n-1}$. We multiply $\mathbf{B}$ from the left by a matrix $\mathbf{C}_2$ such that $\mathbf{C}_2\mathbf{B} = [b_{jk}^{(2)}]$ has $b_{21}^{(2)} = 0$. We multiply this by a matrix $\mathbf{C}_3$ such that $\mathbf{C}_3\mathbf{C}_2\mathbf{B} = [b_{jk}^{(3)}]$ has $b_{32}^{(3)} = 0$, etc. After $n - 1$ such multiplications we are left with an upper triangular matrix $\mathbf{R}_0$, namely,

$$\mathbf{C}_n\mathbf{C}_{n-1} \cdots \mathbf{C}_3\mathbf{C}_2\mathbf{B}_0 = \mathbf{R}_0. \tag{7}$$

These $n \times n$ matrices $\mathbf{C}_j$ are very simple. $\mathbf{C}_j$ has the 2×2 submatrix

$$\begin{bmatrix} \cos\theta_j & \sin\theta_j \\ -\sin\theta_j & \cos\theta_j \end{bmatrix} \qquad (\theta_j \text{ suitable})$$

in Rows $j - 1$ and j and Columns $j - 1$ and j; everywhere else on the main diagonal the matrix $\mathbf{C}_j$ has entries 1; and all its other entries are 0. (This submatrix is the matrix of a plane rotation through the angle θ_j; see Team Project 28, Sec. 7.2.) For instance, if $n = 4$, writing $c_j = \cos\theta_j$, $s_j = \sin\theta_j$, we have

$$\mathbf{C}_2 = \begin{bmatrix} c_2 & s_2 & 0 & 0 \\ -s_2 & c_2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{C}_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & c_3 & s_3 & 0 \\ 0 & -s_3 & c_3 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{C}_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & c_4 & s_4 \\ 0 & 0 & -s_4 & c_4 \end{bmatrix}.$$

These $\mathbf{C}_j$ are orthogonal. Hence their product in (7) is orthogonal, and so is the inverse of this product. We call this inverse $\mathbf{Q}_0$. Then from (7),

$$\mathbf{B}_0 = \mathbf{Q}_0\mathbf{R}_0 \tag{8}$$

where, with $\mathbf{C}_j^{-1} = \mathbf{C}_j^{\mathsf{T}}$,

$$\mathbf{Q}_0 = (\mathbf{C}_n\mathbf{C}_{n-1} \cdots \mathbf{C}_3\mathbf{C}_2)^{-1} = \mathbf{C}_2^{\mathsf{T}}\mathbf{C}_3^{\mathsf{T}} \cdots \mathbf{C}_{n-1}^{\mathsf{T}}\mathbf{C}_n^{\mathsf{T}}. \tag{9}$$

This is our QR-factorization of $\mathbf{B}_0$. From it we have by (5b) with $s = 0$

$$\mathbf{B}_1 = \mathbf{R}_0\mathbf{Q}_0 = \mathbf{R}_0\mathbf{C}_2^{\mathsf{T}}\mathbf{C}_3^{\mathsf{T}} \cdots \mathbf{C}_{n-1}^{\mathsf{T}}\mathbf{C}_n^{\mathsf{T}}. \tag{10}$$

We do not need $\mathbf{Q}_0$ explicitly, but to get $\mathbf{B}_1$ from (10), we first compute $\mathbf{R}_0\mathbf{C}_2^{\mathsf{T}}$, then $(\mathbf{R}_0\mathbf{C}_2^{\mathsf{T}})\mathbf{C}_3^{\mathsf{T}}$, etc. Similarly in the further steps that produce $\mathbf{B}_2, \mathbf{B}_3, \cdots$.

Determination of $\cos\theta_j$ and $\sin\theta_j$. We finally show how to find the angles of rotation. $\cos\theta_2$ and $\sin\theta_2$ in $\mathbf{C}_2$ must be such that $b_{21}^{(2)} = 0$ in the product

$$\mathbf{C}_2\mathbf{B} = \begin{bmatrix} c_2 & s_2 & 0 & \cdots \\ -s_2 & c_2 & 0 & \cdots \\ \cdot & \cdot & \cdot & \cdots \\ \cdot & \cdot & \cdot & \cdots \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & b_{13} & \cdots \\ b_{21} & b_{22} & b_{23} & \cdots \\ \cdot & \cdot & \cdot & \cdots \\ \cdot & \cdot & \cdot & \cdots \end{bmatrix}.$$

Now $b_{21}^{(2)}$ is obtained by multiplying the second row of $\mathbf{C}_2$ by the first column of $\mathbf{B}$,

$$b_{21}^{(2)} = -s_2 b_{11} + c_2 b_{21} = -(\sin\theta_2) b_{11} + (\cos\theta_2) b_{21} = 0.$$

Hence $\tan\theta_2 = s_2/c_2 = b_{21}/b_{11}$, and

$$\cos\theta_2 = \frac{1}{\sqrt{1+\tan^2\theta_2}} = \frac{1}{\sqrt{1+(b_{21}/b_{11})^2}}$$

(11)

$$\sin\theta_2 = \frac{\tan\theta_2}{\sqrt{1+\tan^2\theta_2}} = \frac{b_{21}/b_{11}}{\sqrt{1+(b_{21}/b_{11})^2}}.$$

Similarly for $\theta_3, \theta_4, \cdots$. The next example illustrates all this.

EXAMPLE 2 QR-Factorization Method

Compute all the eigenvalues of the matrix

$$\mathbf{A} = \begin{bmatrix} 6 & 4 & 1 & 1 \\ 4 & 6 & 1 & 1 \\ 1 & 1 & 5 & 2 \\ 1 & 1 & 2 & 5 \end{bmatrix}.$$

Solution. We first reduce $\mathbf{A}$ to tridiagonal form. Applying Householder's method, we obtain (see Example 1)

$$\mathbf{A}_2 = \begin{bmatrix} 6 & -\sqrt{18} & 0 & 0 \\ -\sqrt{18} & 7 & \sqrt{2} & 0 \\ 0 & \sqrt{2} & 6 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}.$$

From the characteristic determinant we see that $\mathbf{A}_2$, hence $\mathbf{A}$, has the eigenvalue 3. (Can you see this directly from $\mathbf{A}_2$?) Hence it suffices to apply the QR-method to the tridiagonal 3×3 matrix

$$\mathbf{B}_0 = \mathbf{B} = \begin{bmatrix} 6 & -\sqrt{18} & 0 \\ -\sqrt{18} & 7 & \sqrt{2} \\ 0 & \sqrt{2} & 6 \end{bmatrix}.$$

Step 1. We multiply $\mathbf{B}$ from the left by

$$\mathbf{C}_2 = \begin{bmatrix} \cos\theta_2 & \sin\theta_2 & 0 \\ -\sin\theta_2 & \cos\theta_2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and then } \mathbf{C}_2\mathbf{B} \text{ by} \quad \mathbf{C}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_3 & \sin\theta_3 \\ 0 & -\sin\theta_3 & \cos\theta_3 \end{bmatrix}.$$

Here $(-\sin\theta_2)\cdot 6 + (\cos\theta_2)(-\sqrt{18}) = 0$ gives (11) $\cos\theta_2 = 0.816\,496\,58$ and $\sin\theta_2 = -0.577\,350\,27$. With these values we compute

$$\mathbf{C}_2\mathbf{B} = \begin{bmatrix} 7.348\,469\,23 & -7.505\,553\,50 & -0.816\,496\,58 \\ 0 & 3.265\,986\,32 & 1.154\,700\,54 \\ 0 & 1.414\,213\,56 & 6.000\,000\,00 \end{bmatrix}.$$

In $\mathbf{C}_3$ we get from $(-\sin\theta_3)\cdot 3.265\,986\,32 + (\cos\theta_3)\cdot 1.414\,213\,56 = 0$ the values $\cos\theta_3 = 0.917\,662\,94$ and $\sin\theta_3 = 0.397\,359\,71$. This gives

$$\mathbf{R}_0 = \mathbf{C}_3\mathbf{C}_2\mathbf{B} = \begin{bmatrix} 7.348\,469\,23 & -7.505\,553\,50 & -0.816\,496\,58 \\ 0 & 3.559\,026\,08 & 3.443\,784\,13 \\ 0 & 0 & 5.047\,146\,15 \end{bmatrix}.$$

From this we compute

$$\mathbf{B}_1 = \mathbf{R}_0\mathbf{C}_2{}^{\mathsf{T}}\mathbf{C}_3{}^{\mathsf{T}} = \begin{bmatrix} 10.333\,333\,33 & -2.054\,804\,67 & 0 \\ -2.054\,804\,67 & 4.035\,087\,72 & 2.005\,532\,51 \\ 0 & 2.005\,532\,51 & 4.631\,578\,95 \end{bmatrix}$$

which is symmetric and tridiagonal. The off-diagonal entries in $\mathbf{B}_1$ are still large in absolute value. Hence we have to go on.

Step 2. We do the same computations as in the first step, with $\mathbf{B}_0 = \mathbf{B}$ replaced by $\mathbf{B}_1$ and $\mathbf{C}_2$ and $\mathbf{C}_3$ changed accordingly, the new angles being $\theta_2 = -0.196\,291\,533$ and $\theta_3 = 0.513\,415\,589$. We obtain

$$\mathbf{R}_1 = \begin{bmatrix} 10.535\,653\,75 & -2.802\,322\,41 & -0.391\,145\,88 \\ 0 & 4.083\,295\,84 & 3.988\,240\,28 \\ 0 & 0 & 3.068\,326\,68 \end{bmatrix}$$

and from this

$$\mathbf{B}_2 = \begin{bmatrix} 10.879\,879\,88 & -0.796\,379\,18 & 0 \\ -0.796\,379\,18 & 5.447\,386\,64 & 1.507\,025\,00 \\ 0 & 1.507\,025\,00 & 2.672\,733\,48 \end{bmatrix}.$$

We see that the off-diagonal entries are somewhat smaller in absolute value than those of $\mathbf{B}_1$, but still much too large for the diagonal entries to be good approximations of the eigenvalues of $\mathbf{B}$.

Further Steps. We list the main diagonal entries and the absolutely largest off-diagonal entry, which is $|b_{12}^{(j)}| = |b_{21}^{(j)}|$ in all steps. You may show that the given matrix $\mathbf{A}$ has the spectrum 11, 6, 3, 2.

Step j	$b_{11}^{(j)}$	$b_{22}^{(j)}$	$b_{33}^{(j)}$	$\max_{j\neq k} \lvert b_{jk}^{(j)}\rvert$
3	10.966 892 9	5.945 898 56	2.087 208 51	0.585 235 82
5	10.997 087 2	6.001 815 41	2.001 097 38	0.120 653 34
7	10.999 742 1	6.000 244 39	2.000 013 55	0.035 911 07
9	10.999 977 2	6.000 022 67	2.000 000 17	0.010 684 77

Looking back at our discussion, we recognize that the purpose of applying Householder's tridiagonalization before the QR-factorization method is a substantial reduction of cost in each QR-factorization, in particular if $\mathbf{A}$ is large.

Convergence acceleration and thus further reduction of cost can be achieved by a **spectral shift,** that is, by taking $\mathbf{B}_s - k_s\mathbf{I}$ instead of $\mathbf{B}_s$ with a suitable k_s. Possible choices of k_s are discussed in Ref. [E29], p. 510.

PROBLEM SET 20.9

1–4 HOUSEHOLDER TRIDIAGONALIZATION

Tridiagonalize, showing the details:

1. $\begin{bmatrix} 3.5 & 1.0 & 1.5 \\ 1.0 & 5.0 & 3.0 \\ 1.5 & 3.0 & 3.5 \end{bmatrix}$ **2.** $\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$

3. $\begin{bmatrix} 0.98 & 0.04 & 0.44 \\ 0.04 & 0.56 & 0.40 \\ 0.44 & 0.40 & 0.80 \end{bmatrix}$

4. $\begin{bmatrix} 8 & 8 & 2 & 2 \\ 8 & 8 & 2 & 2 \\ 2 & 2 & 6 & 4 \\ 2 & 2 & 4 & 6 \end{bmatrix}$

5–9 QR-FACTORIZATION

Do three QR-steps to find approximations of the eigenvalues of:

5. The matrix in the answer to Prob. 1

6. The matrix in the answer to Prob. 3

7. $\begin{bmatrix} 18 & 2 & 0 \\ 2 & 8 & 2 \\ 0 & 2 & 2 \end{bmatrix}$

8. $\begin{bmatrix} 16.2 & -0.1 & 0 \\ -0.1 & -4.3 & 0.2 \\ 0 & 0.2 & 4.1 \end{bmatrix}$

9. $\begin{bmatrix} 7.0 & 0.1 & 0 \\ 0.1 & 4.0 & 0.1 \\ 0 & 0.1 & 1.0 \end{bmatrix}$

10. **CAS EXPERIMENT. QR-Method.** Try to find out experimentally on what properties of a matrix the speed of decrease of off-diagonal entries in the QR-method depends. For this purpose write a program that first tridiagonalizes and then does QR-steps. Try the program out on the matrices in Probs. 1, 3, and 4. Summarize your findings in a short report.

CHAPTER 20 REVIEW QUESTIONS AND PROBLEMS

1. What are the main problem areas in numeric linear algebra?
2. What is pivoting? When and how would you apply it?
3. What happens if you apply Gauss elimination to a system that has no solutions?
4. What is Doolittle's method? Its connection to Gauss elimination?
5. What is Cholesky's method? When would you apply it?
6. What do you know about the convergence of the Gauss–Seidel method?
7. What is ill-conditioning? What is the condition number and its significance?
8. What is least squares approximation? What are the normal equations?
9. What is an eigenvalue of a matrix? Why are eigenvalue problems important? Give typical examples.
10. Why are similarity transformations of matrices important in designing numeric methods? Give examples.
11. What is the power method for eigenvalues? What are its advantages and disadvantages?
12. State Gerschgorin's theorem from memory. Can you remember its proof?
13. State Schur's inequality and give some applications of it.
14. What is tridiagonalization? When would you apply it?
15. What is the idea of the QR-method? When would you apply the method?

16–19 GAUSS ELIMINATION

Solve:

16. $$\begin{aligned} 4x_2 - 3x_3 &= 11.8 \\ 5x_1 + 3x_2 + x_3 &= 34.2 \\ 6x_1 - 7x_2 + 2x_3 &= -3.1 \end{aligned}$$

17. $$\begin{aligned} x_1 + x_2 + x_3 &= 5 \\ x_1 + 2x_2 + 2x_3 &= 6 \\ x_1 + 2x_2 + 3x_3 &= 8 \end{aligned}$$

18. $$\begin{aligned} 5x_1 + x_2 - 3x_3 &= 17 \\ -5x_2 + 15x_3 &= -10 \\ 2x_1 - 3x_2 + 9x_3 &= 0 \end{aligned}$$

19. $$\begin{aligned} 2x_1 + 3x_3 &= 15 \\ 4x_2 - x_3 &= -13 \\ 3x_1 - x_2 + 5x_3 &= 26 \end{aligned}$$

20. Solve Prob. 17 by Doolittle's method.
21. Solve Prob. 17 by Cholesky's method.

22–24 INVERSE MATRIX

Compute the inverse of:

22. $\begin{bmatrix} 1.0 & 2.0 & 0.5 \\ 0.5 & 1.0 & 0.5 \\ 1.5 & 2.0 & 1.0 \end{bmatrix}$

23. $\begin{bmatrix} 1.5 & 2.0 & 1.0 \\ 2.0 & 3.5 & 1.5 \\ 1.0 & 1.5 & 9.0 \end{bmatrix}$

24. $\begin{bmatrix} 5 & 1 & 1 \\ 1 & 6 & 0 \\ 1 & 0 & 8 \end{bmatrix}$

25–26 GAUSS–SEIDEL METHOD

Do 3 steps without scaling, starting from $[1 \quad 1 \quad 1]^T$:

25.
$$\begin{aligned} x_1 + 15x_2 - x_3 &= 11 \\ 10x_1 + 3x_2 &= -17 \\ 2x_1 - x_2 + 5x_3 &= 5 \end{aligned}$$

26.
$$\begin{aligned} 4x_1 - x_2 &= 5.5 \\ 4x_2 - x_3 &= 0.4 \\ -x_1 + 4x_3 &= 11.2 \end{aligned}$$

27–32 VECTOR NORMS

Compute the ℓ_1-, ℓ_2-, and ℓ_∞-norms of the vectors

27. $[0 \quad 4 \quad -8 \quad 3]^T$

28. $[3 \quad 8 \quad -11]^T$

29. $[-4 \quad 1 \quad 0 \quad 2]^T$

30. $[0 \quad 0 \quad 1 \quad 0]^T$

31. $[-5 \quad -2 \quad 7 \quad 0 \quad 0]^T$

32. $[0.3 \quad 1.4 \quad 0.2 \quad -0.6]^T$

33–35 MATRIX NORM

Compute the matrix norm corresponding to the ℓ_∞-vector norm for the coefficient matrix:

33. In Prob. 17

34. In Prob. 18

35. In Prob. 19

36–38 CONDITION NUMBER

Compute the condition number (corresponding to the ℓ_∞-vector norm) of the coefficient matrix:

36. In Prob. 22

37. In Prob. 23

38. In Prob. 24

39–40 FITTING BY LEAST SQUARES

Fit:

39. A straight line to $(-2, 0.1)$, $(0, 1.9)$, $(2, 3.8)$, $(4, 6.1)$, $(6, 7.8)$

40. A quadratic parabola to $(1, 9)$, $(2, 5)$, $(3, 4)$, $(4, 5)$, $(5, 7)$

41–43 EIGENVALUES

Find three circular disks that must contain all the eigenvalues of the matrix:

41. In Prob. 22

42. In Prob. 23

43. In Prob. 24

44. (Power method) Do 4 steps of the power method for the matrix in Prob. 24, starting from $[1 \quad 1 \quad 1]^T$ and computing the Rayleigh quotients and error bounds.

45. (Householder and QR) Tridiagonalize the matrix in Prob. 23. Then apply 3 QR steps. (Spectrum (6S): 9.65971, 4.07684, 0.263451)

SUMMARY OF CHAPTER 20
Numeric Linear Algebra

Main tasks are the numeric solution of linear systems (Secs. 20.1–20.4), curve fitting (Sec. 20.5), and eigenvalue problems (Secs. 20.6–20.9).

Linear systems $\mathbf{Ax} = \mathbf{b}$ with $\mathbf{A} = [a_{jk}]$, written out

$$(1) \qquad \begin{aligned} &E_1: \quad a_{11}x_1 + \cdots + a_{1n}x_n = b_1 \\ &E_2: \quad a_{21}x_1 + \cdots + a_{2n}x_n = b_2 \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ &E_n: \quad a_{n1}x_1 + \cdots + a_{nn}x_n = b_n \end{aligned}$$

can be solved by a **direct method** (one in which the number of numeric operations can be specified in advance, e.g., Gauss's elimination) or by an **indirect** or **iterative method** (in which an initial approximation is improved stepwise).

The **Gauss elimination** (Sec. 20.1) is direct, namely, a systematic elimination process that reduces (1) stepwise to triangular form. In Step 1 we eliminate x_1 from equations E_2 to E_n by subtracting (a_{21}/a_{11}) E_1 from E_2, then (a_{31}/a_{11}) E_1 from E_3, etc. Equation E_1 is called the **pivot equation** in this step and a_{11} the **pivot.** In Step 2 we take the new second equation as pivot equation and eliminate x_2, etc. If the triangular form is reached, we get x_n from the last equation, then x_{n-1} from the second last, etc. **Partial pivoting** (= interchange of equations) is *necessary* if candidates for pivots are zero, and *advisable* if they are small in absolute value.

Doolittle's, Crout's, and **Cholesky's methods** in Sec. 20.2 are variants of the Gauss elimination. They factor $\mathbf{A} = \mathbf{LU}$ ($\mathbf{L}$ lower triangular, $\mathbf{U}$ upper triangular) and solve $\mathbf{Ax} = \mathbf{LUx} = \mathbf{b}$ by solving $\mathbf{Ly} = \mathbf{b}$ for $\mathbf{y}$ and then $\mathbf{Ux} = \mathbf{y}$ for $\mathbf{x}$.

In the **Gauss–Seidel iteration** (Sec. 20.3) we make $a_{11} = a_{22} = \cdots = a_{nn} = 1$ (by division) and write $\mathbf{Ax} = (\mathbf{I} + \mathbf{L} + \mathbf{U})\mathbf{x} = \mathbf{b}$; thus $\mathbf{x} = \mathbf{b} - (\mathbf{L} + \mathbf{U})\mathbf{x}$, which suggests the iteration formula

$$\mathbf{x}^{(m+1)} = \mathbf{b} - \mathbf{L}\mathbf{x}^{(m+1)} - \mathbf{U}\mathbf{x}^{(m)} \tag{2}$$

in which we always take the most recent approximate x_j's on the right. If $\|\mathbf{C}\| < 1$, where $\mathbf{C} = -(\mathbf{I} + \mathbf{L})^{-1}\mathbf{U}$, then this process converges. Here, $\|\mathbf{C}\|$ denotes any matrix norm (Sec. 20.3).

If the **condition number** $\kappa(\mathbf{A}) = \|\mathbf{A}\|\ \|\mathbf{A}^{-1}\|$ of $\mathbf{A}$ is large, then the system $\mathbf{Ax} = \mathbf{b}$ is **ill-conditioned** (Sec. 20.4), and a small **residual** $\mathbf{r} = \mathbf{b} - \mathbf{A}\tilde{\mathbf{x}}$ does ***not*** imply that $\tilde{\mathbf{x}}$ is close to the exact solution.

The fitting of a polynomial $p(x) = b_0 + b_1x + \cdots + b_mx^m$ through given data (points in the xy-plane) $(x_1, y_1), \cdots, (x_n, y_n)$ by the method of **least squares** is discussed in Sec. 20.5 (and in statistics in Sec. 25.9).

Eigenvalues λ (values λ for which $\mathbf{Ax} = \lambda\mathbf{x}$ has a solution $\mathbf{x} \neq \mathbf{0}$, called an **eigenvector**) can be characterized by inequalities (Sec. 20.7), e.g. in **Gerschgorin's theorem,** which gives n circular disks which contain the whole spectrum (all eigenvalues) of $\mathbf{A}$, of centers a_{jj} and radii $\Sigma|a_{jk}|$ (sum over k from 1 to n, $k \neq j$).

Approximations of eigenvalues can be obtained by iteration, starting from an $\mathbf{x}_0 \neq \mathbf{0}$ and computing $\mathbf{x}_1 = \mathbf{Ax}_0$, $\mathbf{x}_2 = \mathbf{Ax}_1$, $\cdots$, $\mathbf{x}_n = \mathbf{Ax}_{n-1}$. In this **power method** (Sec. 20.8) the **Rayleigh quotient**

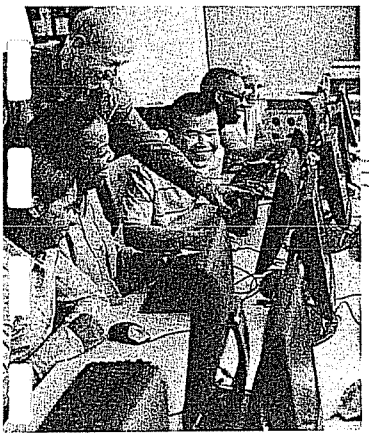
$$q = \frac{(\mathbf{Ax})^{\mathsf{T}}\mathbf{x}}{\mathbf{x}^{\mathsf{T}}\mathbf{x}} \qquad (\mathbf{x} = \mathbf{x}_n) \tag{3}$$

gives an approximation of an eigenvalue (usually that of the greatest absolute value) and, if $\mathbf{A}$ is symmetric, an error bound is

$$|\epsilon| \leqq \sqrt{\frac{(\mathbf{Ax})^{\mathsf{T}}\mathbf{Ax}}{\mathbf{x}^{\mathsf{T}}\mathbf{x}} - q^2}\,. \tag{4}$$

Convergence may be slow but can be improved by a *spectral shift.*

For determining all the eigenvalues of a symmetric matrix $\mathbf{A}$ it is best to first tridiagonalize $\mathbf{A}$ and then to apply the QR-method (Sec. 20.9), which is based on a factorization $\mathbf{A} = \mathbf{QR}$ with orthogonal $\mathbf{Q}$ and upper triangular $\mathbf{R}$ and uses similarity transformations.



CHAPTER 21

Numerics for ODEs and PDEs

Numeric methods for differential equations are of great practical importance to the engineer and physicist because practical problems often lead to differential equations that cannot be solved by one of the methods in Chaps. 1–6 or 12 or by similar methods. Also, sometimes an ODE does have a solution formula (as the ODEs in Secs. 1.3–1.5 do), which, however, in some specific cases may become so complicated that one prefers to apply a numeric method instead.

This chapter explains and applies basic methods for the numeric solution of ODEs (Secs. 21.1–21.3) and PDEs (Secs. 21.4–21.7).

***Sections* 21.1 *and* 21.2 *may be studied immediately after Chap.* 1 *and Sec.* 21.3 *immediately after Chap.* 2,** because these sections are independent of Chaps. 19 and 20.

***Sections* 21.4–21.7 *on PDEs may be studied immediately after Chap.* 12** if students have some knowledge of linear systems of algebraic equations.

Prerequisite: Secs. 1.1–1.5 for ODEs, Secs. 12.1–12.3, 12.5, 12.10 for PDEs.
References and Answers to Problems App. 1 Part E (see also Parts A and C), App. 2.

21.1 Methods for First-Order ODEs

From Chap. 1 we know that an ODE of the first order is of the form $F(x, y, y') = 0$ and can often be written in the explicit form $y' = f(x, y)$. An **initial value problem** for this equation is of the form

$$y' = f(x, y), \qquad y(x_0) = y_0 \tag{1}$$

where x_0 and y_0 are given and we assume that the problem has a unique solution on some open interval $a < x < b$ containing x_0.

In this section we shall discuss methods of computing approximate numeric values of the solution $y(x)$ of (1) at the equidistant points on the x-axis

$$x_1 = x_0 + h, \qquad x_2 = x_0 + 2h, \qquad x_3 = x_0 + 3h, \qquad \cdots$$

where the **step size** h is a fixed number, for instance, 0.2 or 0.1 or 0.01, whose choice we discuss later in this section. Those methods are **step-by-step methods,** using the same formula in each step. Such formulas are suggested by the Taylor series

$$y(x + h) = y(x) + hy'(x) + \frac{h^2}{2}y''(x) + \cdots. \tag{2}$$

For a small h the higher powers h^2, h^3, $\cdots$ are very small. This suggests the crude approximation

$$y(x + h) \approx y(x) + hy'(x)$$
$$= y(x) + hf(x, y)$$

(with the second line obtained from the given ODE) and the following iteration process. In the first step we compute

$$y_1 = y_0 + hf(x_0, y_0)$$

which approximates $y(x_1) = y(x_0 + h)$. In the second step we compute

$$y_2 = y_1 + hf(x_1, y_1)$$

which approximates $y(x_2) = y(x_0 + 2h)$, etc., and in general

$$y_{n+1} = y_n + hf(x_n, y_n) \qquad (n = 0, 1, \cdots). \tag{3}$$

This is called the **Euler method** or the **Euler–Cauchy method.** Geometrically it is an approximation of the curve of $y(x)$ by a polygon whose first side is tangent to this curve at x_0 (see Fig. 448).

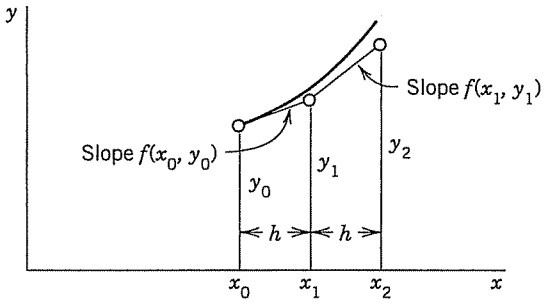


Fig. 448. Euler method

This crude method is hardly ever used in practice, but since it is simple, it nicely explains the principle of methods based on the Taylor series.

Taylor's formula with remainder has the form

$$y(x + h) = y(x) + hy'(x) + \tfrac{1}{2}h^2y''(\xi)$$

(where $x \leqq \xi \leqq x + h$). It shows that in the Euler method the *truncation error in each step* or **local truncation error** is proportional to h^2, written $O(h^2)$, where O suggests *order* (see also Sec. 20.1). Now over a fixed x-interval in which we want to solve an ODE the number of steps is proportional to $1/h$. Hence the *total error* or **global error** is proportional to $h^2(1/h) = h^1$. For this reason, the Euler method is called a **first-order method.** In addition, there are **roundoff errors** in this and other methods, which may affect the accuracy of the values $y_1, y_2, \cdots$ more and more as n increases, as we shall see.

Table 21.1 Euler Method Applied to (4) in Example 1 and Error

n	x_n	y_n	$0.2(x_n + y_n)$	Exact Values	Error ϵ_n
0	0.0	0.000	0.000	0.000	0.000
1	0.2	0.000	0.040	0.021	0.021
2	0.4	0.040	0.088	0.092	0.052
3	0.6	0.128	0.146	0.222	0.094
4	0.8	0.274	0.215	0.426	0.152
5	1.0	0.489		0.718	0.229

EXAMPLE 1 Euler Method

Apply the Euler method to the following initial value problem, choosing $h = 0.2$ and computing $y_1, \cdots, y_5$:

$$y' = x + y, \qquad y(0) = 0. \tag{4}$$

Solution. Here $f(x, y) = x + y$; hence $f(x_n, y_n) = x_n + y_n$, and we see that (3) becomes

$$y_{n+1} = y_n + 0.2(x_n + y_n).$$

Table 21.1 shows the computations, the values of the exact solution

$$y(x) = e^x - x - 1$$

obtained from (4) in Sec. 1.5, and the error. In practice the exact solution is unknown, but an indication of the accuracy of the values can be obtained by applying the Euler method once more with step $2h = 0.4$, letting y_n^* denote the approximation now obtained, and comparing corresponding approximations. This computation is:

x_n	y_n^*	$0.4(x_n + y_n)$	y_n in Table 21.1	Difference $y_n - y_n^*$
0.0	0.000	0.000	0.000	0.000
0.4	0.000	0.160	0.040	0.040
0.8	0.160		0.274	0.114

Let ϵ_n and ϵ_n^* be the errors of the computations with h and $2h$, respectively. Since the error is of order h^2, in a switch from h to $2h$ it is multiplied by $2^2 = 4$, but since we need only half as many steps as before, it will be multiplied only by $4/2 = 2$. Hence $\epsilon_n^* \approx 2\epsilon_n$ so that the difference is $\epsilon_n^* - \epsilon_n \approx 2\epsilon_n - \epsilon_n = \epsilon_n$. Now $y = y_n + \epsilon_n = y_n^* + \epsilon_n^*$ by the definition of error; hence $\epsilon_n^* - \epsilon_n = y_n - y_n^*$ indicates ϵ_n qualitatively. In our computations, $y_2 - y_2^* = 0.04 - 0 = 0.04$ (actual error 0.052, see Table 21.1) and $y_4 - y_4^* = 0.274 - 0.160 = 0.114$ (actually 0.152). ■

EXAMPLE 2 Euler Method for a Nonlinear ODE

Figure 449 concerns the initial value problem

$$y' = (y - 0.01x^2)^2 \sin(x^2) + 0.02x, \qquad y(0) = 0.4 \tag{5}$$

and shows the curve of the solution $y = 1/[2.5 - S(x)] + 0.01x^2$ where $S(x)$ is the Fresnel integral (38) in App. 3.1. It also shows 80 approximate values for $0 \leqq x \leqq 4$ obtained by the Euler method from (3),

$$y_{n+1} = y_n + 0.05\left[(y_n - 0.01x_n^2)^2 \sin(x_n^2) + 0.02x_n\right].$$

Although $h = 0.05$ is smaller than h in Example 1, the accuracy is still not good. It is interesting that the error is not monotone increasing, obviously since the solution is not monotone. We shall return to this ODE in the problem set. ■

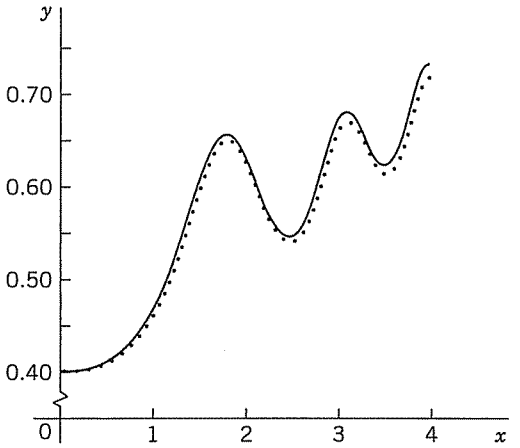


Fig. 449. Solution curve and Euler approximation in Example 2

Automatic Variable Step Size Selection in Modern Numeric Software

The idea of adaptive integration as motivated and explained in Sec. 19.5 applies equally well to the numeric solution of ODEs. It now concerns automatically changing the step size h depending on the variability of $y' = f$ determined by

$$y'' = f' = f_x + f_y y' = f_x + f_y f. \tag{6*}$$

Accordingly, modern software automatically selects variable step sizes h_n so that the error of the solution will not exceed a given maximum size TOL (suggesting *tolerance*). Now for the Euler method, when the step size is $h = h_n$, the local error at x_n is about $\frac{1}{2}h_n^2\,|y''(x_n)|$. We require that this be equal to a given tolerance TOL,

$$\text{(6)} \qquad \text{(a)} \quad \tfrac{1}{2}h_n^2|y''(x_n)| = \text{TOL}, \qquad \text{thus} \qquad \text{(b)} \quad h_n = \sqrt{\frac{2\ \text{TOL}}{|y''(x_n)|}}\,.$$

$y''(x)$ must not be zero on the interval J: $x_0 \leqq x = x_N$ on which the solution is wanted. Let K be the minimum of $|y''(x)|$ on J and assume that $K > 0$. Minimum $|y''(x)|$ corresponds to maximum $h = H = \sqrt{2\ \text{TOL}/K}$ by (6). Thus, $\sqrt{2\ \text{TOL}} = H\sqrt{K}$. We can insert this into (6b), obtaining by straightforward algebra

$$\text{(7)} \qquad h_n = \varphi(x_n)H \qquad \text{where} \qquad \varphi(x_n) = \sqrt{\frac{K}{|y''(x_n)|}}\,.$$

For other methods, automatic step size selection is based on the same principle.

Improved Euler Method

By taking more terms in (2) into account we obtain numeric methods of higher order and precision. But there is a practical problem. If we substitute $y' = f(x, y(x))$ into (2), we have

$$y(x + h) = y(x) + hf + \tfrac{1}{2}h^2 f' + \tfrac{1}{6}h^3 f'' + \cdots. \tag{2*}$$

Now y in f depends on x, so that we have f' as shown in (6*) and f'', f''' even much more cumbersome. The ***general strategy*** now is to avoid the computation of these derivatives and to replace it by computing f for one or several suitably chosen auxiliary values of (x, y). "Suitably" means that these values are chosen to make the order of the method as high as possible (to have high accuracy). Let us discuss two such methods that are of practical importance, namely, the improved Euler method and the (classical) Runge-Kutta method.

In the **improved Euler method** or **improved Euler–Cauchy method** (sometimes also called **Heun method**), in each step we compute first the auxiliary value

(8a) $$y_{n+1}^* = y_n + hf(x_n, y_n)$$

and then the new value

(8b) $$y_{n+1} = y_n + \tfrac{1}{2}h[f(x_n, y_n) + f(x_{n+1}, y_{n+1}^*)].$$

This method has a simple geometric interpretation. In fact, we may say that in the interval from x_n to $x_n + \frac{1}{2}h$ we approximate the solution y by the straight line through (x_n, y_n) with slope $f(x_n, y_n)$, and then we continue along the straight line with slope $f(x_{n+1}, y_{n+1}^*)$ until x reaches x_{n+1}.

The improved Euler–Cauchy method is a **predictor–corrector method,** because in each step we first ***predict*** a value by (8a) and then ***correct*** it by (8b).

In algorithmic form, using the notations $k_1 = hf(x_n, y_n)$ in (8a) and $k_2 = hf(x_{n+1}, y_{n+1}^*)$ in (8b), we can write this method as shown in Table 21.2.

Table 21.2 Improved Euler Method (Heun's Method)

ALGORITHM EULER (f, x_0, y_0, h, N)

This algorithm computes the solution of the initial value problem $y' = f(x, y)$, $y(x_0) = y_0$ at equidistant points $x_1 = x_0 + h$, $x_2 = x_0 + 2h, \cdots, x_N = x_0 + Nh$; here f is such that this problem has a unique solution on the interval $[x_0, x_N]$ (see Sec. 1.7).

INPUT: Initial values x_0, y_0, step size h, number of steps N

OUTPUT: Approximation y_{n+1} to the solution $y(x_{n+1})$ at $x_{n+1} = x_0 + (n + 1)h$, where $n = 0, \cdots, N - 1$

For $n = 0, 1, \cdots, N - 1$ do:

$x_{n+1} = x_n + h$

$k_1 = hf(x_n, y_n)$

$k_2 = hf(x_{n+1}, y_n + k_1)$

$y_{n+1} = y_n + \frac{1}{2}(k_1 + k_2)$

OUTPUT x_{n+1}, y_{n+1}

End

Stop

End EULER

EXAMPLE 3 Improved Euler Method

Apply the improved Euler method to the initial value problem (4), choosing $h = 0.2$, as before.

Solution. For the present problem we have in Table 21.2

$$k_1 = 0.2(x_n + y_n)$$

$$k_2 = 0.2(x_n + 0.2 + y_n + 0.2(x_n + y_n))$$

$$y_{n+1} = y_n + \frac{0.2}{2}(2.2x_n + 2.2y_n + 0.2) = y_n + 0.22(x_n + y_n) + 0.02.$$

Table 21.3 shows that our present results are more accurate than those in Example 1; see also Table 21.6. ■

Table 21.3 Improved Euler Method Applied to (4) and Error

n	x_n	y_n	$0.22(x_n + y_n) + 0.02$	Exact Values (4D)	Error
0	0.0	0.0000	0.0200	0.0000	0.0000
1	0.2	0.0200	0.0684	0.0214	0.0014
2	0.4	0.0884	0.1274	0.0918	0.0034
3	0.6	0.2158	0.1995	0.2221	0.0063
4	0.8	0.4153	0.2874	0.4255	0.0102
5	1.0	0.7027		0.7183	0.0156

Error of the Improved Euler Method. *The local error is of order h^3 and the global error of order h^2, so that the method is a* **second-order method.**

PROOF Setting $\tilde{f}_n = f(x_n, y(x_n))$ and using (2*), we have

$$y(x_n + h) - y(x_n) = h\tilde{f}_n + \tfrac{1}{2}h^2\tilde{f}_n' + \tfrac{1}{6}h^3\tilde{f}_n'' + \cdots. \tag{9a}$$

Approximating the expression in the brackets in (8b) by $\tilde{f}_n + \tilde{f}_{n+1}$ and again using the Taylor expansion, we obtain from (8b)

$$\begin{aligned} y_{n+1} - y_n &\approx \tfrac{1}{2}h\left[\tilde{f}_n + \tilde{f}_{n+1}\right] \\ &= \tfrac{1}{2}h\left[\tilde{f}_n + (\tilde{f}_n + h\tilde{f}_n' + \tfrac{1}{2}h^2\tilde{f}_n'' + \cdots)\right] \\ &= h\tilde{f}_n + \tfrac{1}{2}h^2\tilde{f}_n' + \tfrac{1}{4}h^3\tilde{f}_n'' + \cdots \end{aligned} \tag{9b}$$

(where $' = d/dx_n$, etc.). Subtraction of (9b) from (9a) gives the local error

$$\frac{h^3}{6}\tilde{f}_n'' - \frac{h^3}{4}\tilde{f}_n'' + \cdots = -\frac{h^3}{12}\tilde{f}_n'' + \cdots.$$

Since the number of steps over a fixed x-interval is proportional to $1/h$, the global error is of order $h^3/h = h^2$, so that the method is of second order. ■

Runge–Kutta Methods (RK Methods)

A method of great practical importance and much greater accuracy than that of the improved Euler method is the *classical Runge–Kutta method of fourth order,* which we call briefly the **Runge–Kutta method.**[1] It is shown in Table 21.4. We see that in each step we first compute four auxiliary quantities k_1, k_2, k_3, k_4 and then the new value y_{n+1}. The method is well suited to the computer because it needs no special starting procedure, makes light demand on storage, and repeatedly uses the same straightforward computational procedure. It is numerically stable.

Note that if f depends only on x, this method reduces to Simpson's rule of integration (Sec. 19.5). Note further that $k_1, \cdots, k_4$ depend on n and generally change from step to step.

Table 21.4 Classical Runge–Kutta Method of Fourth Order

ALGORITHM RUNGE–KUTTA (f, x_0, y_0, h, N).

This algorithm computes the solution of the initial value problem $y' = f(x, y)$, $y(x_0) = y_0$ at equidistant points

$$x_1 = x_0 + h,\ x_2 = x_0 + 2h, \cdots, x_N = x_0 + Nh;$$

here f is such that this problem has a unique solution on the interval $[x_0, x_N]$ (see Sec. 1.7).

INPUT: Function f, initial values x_0, y_0, step size h, number of steps N

OUTPUT: Approximation y_{n+1} to the solution $y(x_{n+1})$ at $x_{n+1} = x_0 + (n + 1)h$, where $n = 0, 1, \cdots, N - 1$

For $n = 0, 1, \cdots, N - 1$ do:

$$k_1 = hf(x_n, y_n)$$
$$k_2 = hf(x_n + \tfrac{1}{2}h, y_n + \tfrac{1}{2}k_1)$$
$$k_3 = hf(x_n + \tfrac{1}{2}h, y_n + \tfrac{1}{2}k_2)$$
$$k_4 = hf(x_n + h, y_n + k_3)$$
$$x_{n+1} = x_n + h$$
$$y_{n+1} = y_n + \tfrac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)$$
OUTPUT x_{n+1}, y_{n+1}

End

Stop

End RUNGE–KUTTA

[1]Named after the German mathematicians KARL RUNGE (Sec. 19.4) and WILHELM KUTTA (1867–1944). Runge [*Math. Annalen* **46** (1895), 167–178], KARL HEUN [*Zeitschr. Math. Phys.* **45** (1900), 23–38], and Kutta [*Zeitschr. Math. Phys.* **46** 1901), 435–453] developed various such methods. Theoretically, there are infinitely many fourth-order methods using four function values per step. The method in Table 21.4 is most popular from a practical viewpoint because of its "symmetrical" form and its simple coefficients. It was given by Kutta.

EXAMPLE 4 Classical Runge–Kutta Method

Apply the Runge–Kutta method to the initial value problem (4) in Example 1, choosing $h = 0.2$, as before, and computing five steps.

Solution. For the present problem we have $f(x, y) = x + y$. Hence

$$k_1 = 0.2(x_n + y_n), \qquad k_2 = 0.2(x_n + 0.1 + y_n + 0.5k_1),$$
$$k_3 = 0.2(x_n + 0.1 + y_n + 0.5k_2), \qquad k_4 = 0.2(x_n + 0.2 + y_n + k_3).$$

Table 21.5 shows the results and their errors, which are smaller by factors 10^3 and 10^4 than those for the two Euler methods. See also Table 21.6. We mention in passing that since the present $k_1, \cdots, k_4$ are simple, operations were saved by substituting k_1 into k_2, then k_2 into k_3, etc.; the resulting formula is shown in Column 4 of Table 21.5. ■

Table 21.5 Runge–Kutta Method Applied to (4)

n	x_n	y_n	$0.2214(x_n + y_n) + 0.0214$	Exact Values (6D) $y = e^x - x - 1$	$10^6 \times$ Error of y_n
0	0.0	0	0.021 400	0.000 000	0
1	0.2	0.021 400	0.070 418	0.021 403	3
2	0.4	0.091 818	0.130 289	0.091 825	7
3	0.6	0.222 107	0.203 414	0.222 119	12
4	0.8	0.425 521	0.292 730	0.425 541	20
5	1.0	0.718 251		0.718 282	31

Table 21.6 Comparison of the Accuracy of the Three Methods Under Consideration in the Case of the Initial Value Problem (4), with $h = 0.2$

x	$y = e^x - x - 1$	Error		
		Euler (Table 21.1)	Improved Euler (Table 21.3)	Runge–Kutta (Table 21.5)
0.2	0.021 403	0.021	0.0014	0.000 003
0.4	0.091 825	0.052	0.0034	0.000 007
0.6	0.222 119	0.094	0.0063	0.000 011
0.8	0.425 541	0.152	0.0102	0.000 020
1.0	0.718 282	0.229	0.0156	0.000 031

Error and Step Size Control. RKF (Runge–Kutta–Fehlberg)

The idea of adaptive integration (Sec. 19.5) has analogs for Runge–Kutta (and other) methods. In Table 21.4 for RK (Runge–Kutta), if we compute in each step approximations $\tilde{y}$ and $\tilde{\tilde{y}}$ with step sizes h and $2h$, respectively, the latter has error per step equal to $2^5 = 32$ times that of the former; however, since we have only half as many steps for $2h$, the actual factor is $2^5/2 = 16$, so that, say,

$$\epsilon^{(2h)} \approx 16\epsilon^{(h)} \qquad \text{and thus} \qquad y^{(h)} - y^{(2h)} = \epsilon^{(2h)} - \epsilon^{(h)} \approx (16 - 1)\epsilon^{(h)}.$$

Hence the error $\epsilon = \epsilon^{(h)}$ for step size h is about

(10) $$\epsilon = \tfrac{1}{15}(\tilde{y} - \tilde{\tilde{y}})$$

where $\tilde{y} - \tilde{\tilde{y}} = y^{(h)} - y^{(2h)}$, as said before. Table 21.7 illustrates (10) for the initial value problem

(11) $$y' = (y - x - 1)^2 + 2, \qquad y(0) = 1,$$

the step size $h = 0.1$ and $0 \leqq x \leqq 0.4$. We see that the estimate is close to the actual error. This method of error estimation is simple but may be unstable.

Table 21.7 Runge–Kutta Method Applied to the Initial Value Problem (11) and Error Estimate (10). Exact Solution $y = \tan x + x + 1$

x	$\tilde{y}$ (Step size h)	$\tilde{\tilde{y}}$ (Step size $2h$)	Error Estimate (10)	Actual Error	Exact Solution (9D)
0.0	1.000 000 000	1.000 000 000	0.000 000 000	0.000 000 000	1.000 000 000
0.1	1.200 334 589			0.000 000 083	1.200 334 672
0.2	1.402 709 878	1.402 707 408	0.000 000 165	0.000 000 157	1.402 710 036
0.3	1.609 336 039			0.000 000 210	1.609 336 250
0.4	1.822 792 993	1.822 788 993	0.000 000 267	0.000 000 226	1.822 793 219

RKF. E. Fehlberg [*Computing* **6** (1970), 61–71] proposed and developed error control by using two RK methods of different orders to go from (x_n, y_n) to (x_{n+1}, y_{n+1}). The difference of the computed y-values at x_{n+1} gives an error estimate to be used for step size control. Fehlberg discovered two RK formulas that together need only 6 function evaluations per step. We present these formulas here because RKF has become quite popular. For instance, Maple uses it (also for systems of ODEs).

Fehlberg's fifth-order RK method is

(12a) $$y_{n+1} = y_n + \gamma_1 k_1 + \cdots + \gamma_6 k_6$$

with coefficient vector $\gamma = [\gamma_1 \cdots \gamma_6]$,

(12b) $$\gamma = \left[\tfrac{16}{135} \quad 0 \quad \tfrac{6656}{12825} \quad \tfrac{28561}{56430} \quad -\tfrac{9}{50} \quad \tfrac{2}{55}\right].$$

His **fourth-order RK method** is

(13a) $$y^*_{n+1} = y_n + \gamma^*_1 k_1 + \cdots + \gamma^*_5 k_5$$

with coefficient vector

(13b) $$\gamma^* = \left[\tfrac{25}{216} \quad 0 \quad \tfrac{1408}{2565} \quad \tfrac{2197}{4104} \quad -\tfrac{1}{5}\right].$$

In both formulas we use only 6 different function evaluations altogether, namely,

$$
\begin{aligned}
k_1 &= hf(x_n, y_n) \\
k_2 &= hf(x_n + \tfrac{1}{4}h, \quad y_n + \tfrac{1}{4}k_1) \\
k_3 &= hf(x_n + \tfrac{3}{8}h, \quad y_n + \tfrac{3}{32}k_1 + \tfrac{9}{32}k_2) \\
k_4 &= hf(x_n + \tfrac{12}{13}h, \quad y_n + \tfrac{1932}{2197}k_1 - \tfrac{7200}{2197}k_2 + \tfrac{7296}{2197}k_3) \\
k_5 &= hf(x_n + h, \quad y_n + \tfrac{439}{216}k_1 - 8k_2 + \tfrac{3680}{513}k_3 - \tfrac{845}{4104}k_4) \\
k_6 &= hf(x_n + \tfrac{1}{2}h, \quad y_n - \tfrac{8}{27}k_1 + 2k_2 - \tfrac{3544}{2565}k_3 + \tfrac{1859}{4104}k_4 - \tfrac{11}{40}k_5).
\end{aligned} \tag{14}
$$

The difference of (12) and (13) gives the **error estimate**

$$
\epsilon_{n+1} \approx y_{n+1} - y_{n+1}^* = \tfrac{1}{360}k_1 - \tfrac{128}{4275}k_3 - \tfrac{2197}{75240}k_4 + \tfrac{1}{50}k_5 + \tfrac{2}{55}k_6. \tag{15}
$$

EXAMPLE 5 Runge–Kutta–Fehlberg

For the initial value problem (11) we obtain from (12)–(14) with $h = 0.1$ in the first step the 12S-values

$$
\begin{aligned}
k_1 &= 0.200000\,000000 & k_2 &= 0.200062\,500000 \\
k_3 &= 0.200140\,756867 & k_4 &= 0.200856\,926154 \\
k_5 &= 0.201006\,676700 & k_6 &= 0.200250\,418651
\end{aligned}
$$

$$
y_1^* = 1.200334\,66949
$$

$$
y_1 = 1.200334\,67253
$$

and the error estimate

$$
\epsilon_1 \approx y_1 - y_1^* = 0.000000\,00304.
$$

The exact 12S-value is $y(0.1) = 1.200334\,67209$. Hence the actual error of y_1 is $-4.4 \cdot 10^{-10}$, smaller than that in Table 21.7 by a factor 200. ■

Table 21.8 summarizes essential features of the methods in this section. It can be shown that these methods are *numerically stable* (definition in Sec. 19.1). They are **one-step methods** because in each step we use the data of just *one* preceding step, in contrast to **multistep methods** where in each step we use data from *several* preceding steps, as we shall see in the next section.

Table 21.8 Methods Considered and Their Order (= Their Global Error)

Method	Function Evaluation per Step	Global Error	Local Error
Euler	1	$O(h)$	$O(h^2)$
Improved Euler	2	$O(h^2)$	$O(h^3)$
RK (fourth order)	4	$O(h^4)$	$O(h^5)$
RKF	6	$O(h^5)$	$O(h^6)$

Backward Euler Method. Stiff ODEs

The **backward Euler formula** for numerically solving (1) is

$$\textbf{(16)} \qquad y_{n+1} = y_n + hf(x_{n+1}, y_{n+1}) \qquad (n = 0, 1, \cdots).$$

This formula is obtained by evaluating the right side at the *new* location (x_{n+1}, y_{n+1}); this is called the **backward Euler scheme.** For known y_n it gives y_{n+1} *implicitly,* so it defines an **implicit method,** in contrast to the Euler method (3), which gives y_{n+1} explicitly. Hence (16) must be solved for y_{n+1}. How difficult this is depends on f in (1). For a linear ODE this provides no problem, as Example 6 (below) illustrates. The method is particularly useful for "stiff" ODEs, as they occur quite frequently in the study of vibrations, electric circuits, chemical reactions, etc. The situation of stiffness is roughly as follows; for details, see, for example, [E5], [E25], [E26] in App. 1.

Error terms of the methods considered so far involve a higher derivative. And we ask what happens if we let h *increase.* Now if the error (the derivative) grows fast but the desired solution also grows fast, nothing will happen. However, if that solution does not grow fast, then with growing h the error term can take over to an extent that the numeric result becomes completely nonsensical, as in Fig. 450. Such an ODE for which h must thus be restricted to small values, and the physical system the ODE models, are called **stiff.** This term is suggested by a mass–spring system with a stiff spring (spring with a large k; see Sec. 2.4). Example 6 illustrates that implicit methods remove the difficulty of increasing h in the case of stiffness: it can be shown that in the application of an implicit method the solution remains stable under any increase of h, although the accuracy decreases with increasing h.

EXAMPLE 6 Backward Euler Method. Stiff ODE

The initial value problem

$$y' = f(x, y) = -20y + 20x^2 + 2x, \qquad y(0) = 1$$

has the solution (verify!)

$$y = e^{-20x} + x^2.$$

The backward Euler formula (16) is

$$y_{n+1} = y_n + hf(x_{n+1}, y_{n+1}) = y_n + h(-20y_{n+1} + 20x_{n+1}^2 + 2x_{n+1}).$$

Noting that $x_{n+1} = x_n + h$, taking the term $-20hy_{n+1}$ to the left, and dividing, we obtain

$$(16^*) \qquad y_{n+1} = \frac{y_n + h[20(x_n + h)^2 + 2(x_n + h)]}{1 + 20h}.$$

The numeric results in Table 21.9 show the following.

Stability of the backward Euler method for $h = 0.05$ and also for $h = 0.2$ with an error increase by about a factor 4 for $h = 0.2$,

Stability of the Euler method for $h = 0.05$ but instability for $h = 0.1$ (Fig. 450),

Stability of RK for $h = 0.1$ but instability for $h = 0.2$.

This illustrates that the ODE is stiff. Note that even in the case of stability the approximation of the solution near $x = 0$ is poor.

Stiffness will be considered further in Sec. 21.3 in connection with systems of ODEs.

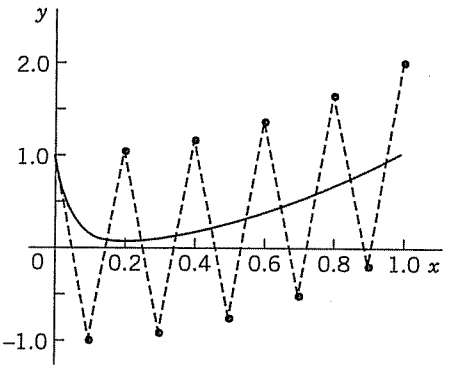


Fig. 450. Euler method with $h = 0.1$ for the stiff ODE in Example 6 and exact solution

Table 21.9 Backward Euler Method (BEM) for Example 6. Comparison with Euler and RK

x	BEM $h = 0.05$	BEM $h = 0.2$	Euler $h = 0.05$	Euler $h = 0.1$	RK $h = 0.1$	RK $h = 0.2$	Exact
0.0	1.00000	1.00000	1.00000	1.00000	1.00000	1.000	1.00000
0.1	0.26188		0.00750	−1.00000	0.34500		0.14534
0.2	0.10484	0.24800	0.03750	1.04000	0.15333	5.093	0.05832
0.3	0.10809		0.08750	−0.92000	0.12944		0.09248
0.4	0.16640	0.20960	0.15750	1.16000	0.17482	25.48	0.16034
0.5	0.25347		0.24750	−0.76000	0.25660		0.25004
0.6	0.36274	0.37792	0.35750	1.36000	0.36387	127.0	0.36001
0.7	0.49256		0.48750	−0.52000	0.49296		0.49001
0.8	0.64252	0.65158	0.63750	1.64000	0.64265	634.0	0.64000
0.9	0.81250		0.80750	−0.20000	0.81255		0.81000
1.0	1.00250	1.01032	0.99750	2.00000	1.00252	3168	1.00000

PROBLEM SET 21.1

1–4 EULER METHOD

Do 10 steps. Solve the problem exactly. Compute the error. (Show the details.)

1. $y' = y$, $y(0) = 1$, $h = 0.1$
2. $y' = y$, $y(0) = 1$, $h = 0.01$
3. $y' = (y - x)^2$, $y(0) = 0$, $h = 0.1$
4. $y' = (y + x)^2$, $y(0) = 0$, $h = 0.1$

5–10 IMPROVED EULER METHOD

Do 10 steps. Solve exactly. Compute the error. (Show the details.)

5. $y' = y$, $y(0) = 1$, $h = 0.1$. Compare with Prob. 1 and comment.
6. **(Logistic population)** $y' = y - y^2$, $y(0) = 0.2$, $h = 0.1$
7. $y' - xy^2 = 0$, $y(0) = 1$, $h = 0.1$
8. $y' + y \tan x = \sin 2x$, $y(0) = 1$, $h = 0.1$
9. Do Prob. 7 using the Euler method with $h = 0.1$ and compare the accuracy.
10. $y' = 1 - \frac{1}{4}y^2$, $y(0) = 0$, $h = 0.1$

11–17 CLASSICAL RUNGE–KUTTA METHOD OF FOURTH ORDER

Do 10 steps. Compare as indicated. Comment. (Show the details.)

11. $y' - xy^2 = 0$, $y(0) = 1$, $h = 0.1$. Compare with Prob. 7. Apply (10) to y_{10}.
12. $y' = y - y^2$, $y(0) = 0.2$, $h = 0.1$. Compare with Prob. 6. Apply (10) to y_{10}.

13. $y' = (1 + x^{-1})y$, $y(1) = e$, $h = 0.2$
14. $y' = \frac{1}{2}(y/x - x/y)$, $y(2) = 2$, $h = 0.2$
15. $y' + y \tan x = \sin 2x$, $y(0) = 1$, $h = 0.1$
16. In Prob. 15 use $h = 0.2$ (5 steps) and compare the error.
17. $y' + 5x^4y^2 = 0$, $y(0) = 1$, $h = 0.2$
18. **Kutta's third-order method** is defined by $y_{n+1} = y_n + \frac{1}{6}(k_1 + 4k_2 + k_3^*)$ with k_1 and k_2 as in RK (Table 21.4) and $k_3^* = hf(x_{n+1}, y_n - k_1 + 2k_2)$. Apply this method to (4) in Example 1. Choose $h = 0.2$ and do 5 steps. Compare with Table 21.6.
19. **CAS EXPERIMENT. Euler–Cauchy vs. RK.** (a) Solve (5) in Example 2 by Euler, Improved Euler, and RK for $0 \leqq x \leqq 5$ with step $h = 0.2$. Compare the errors for $x = 1, 3, 5$ and comment.

(b) Graph solution curves of the ODE in (5) for various positive and negative initial values.

(c) Do a similar experiment as in (a) for an initial value problem that has a monotone increasing or monotone decreasing solution. Compare the behavior of the error with that in (a). Comment.

20. **CAS EXPERIMENT. RKF.** (a) Write a program for RKF that gives x_n, y_n, the estimate (10), and if the solution is known, the actual error ϵ_n.

(b) Apply the program to Example 5 in the text (10 steps, $h = 0.1$).

(c) ϵ_n in (b) gives a relatively good idea of the size of the actual error. Is this typical or accidental? Find out by experimentation with other problems on what properties of the ODE or solution this might depend.

21.2 Multistep Methods

In a **one-step method** we compute y_{n+1} using only a single step, namely, the previous value y_n. *One-step methods are* **"self-starting,"** they need no help to get going because they obtain y_1 from the initial value y_0, etc. All methods in Sec. 21.1 are one-step.

In contrast, a **multistep method** uses in each step values from two or more previous steps. These methods are motivated by the expectation that the additional information will increase accuracy and stability. But to get started, one needs values, say, y_0, y_1, y_2, y_3 in a 4-step method, obtained by Runge–Kutta or another accurate method. Thus, multistep methods are not self-starting. Such methods are obtained as follows.

Adams–Bashforth Methods

We consider an initial value problem

$$y' = f(x, y), \qquad y(x_0) = y_0 \tag{1}$$

as before, with f such that the problem has a unique solution on some open interval containing x_0. We integrate $y' = f(x, y)$ from x_n to $x_{n+1} = x_n + h$. This gives

$$\int_{x_n}^{x_{n+1}} y'(x)\,dx = y(x_{n+1}) - y(x_n) = \int_{x_n}^{x_{n+1}} f(x, y(x))\,dx.$$

Now comes the main idea. We replace $f(x, y(x))$ by an interpolation polynomial $p(x)$ (see Sec. 19.3), so that we can later integrate. This gives approximations y_{n+1} of $y(x_{n+1})$ and y_n of $y(x_n)$,

$$y_{n+1} = y_n + \int_{x_n}^{x_{n+1}} p(x)\,dx. \tag{2}$$

Different choices of $p(x)$ will now produce different methods. We explain the principle by taking a cubic polynomial, namely, the polynomial $p_3(x)$ that at (equidistant)

$$x_n, \qquad x_{n-1}, \qquad x_{n-2}, \qquad x_{n-3}$$

has the respective values

$$\begin{aligned} f_n &= f(x_n, y_n) \\ f_{n-1} &= f(x_{n-1}, y_{n-1}) \\ f_{n-2} &= f(x_{n-2}, y_{n-2}) \\ f_{n-3} &= f(x_{n-3}, y_{n-3}). \end{aligned} \tag{3}$$

This will lead to a practically useful formula. We can obtain $p_3(x)$ from Newton's backward difference formula (18), Sec. 19.3:

$$p_3(x) = f_n + r\nabla f_n + \tfrac{1}{2}r(r+1)\nabla^2 f_n + \tfrac{1}{6}r(r+1)(r+2)\nabla^3 f_n$$

where

$$r = \frac{x - x_n}{h}.$$

We integrate $p_3(x)$ over x from x_n to $x_{n+1} = x_n + h$, thus over r from 0 to 1. Since

$$x = x_n + hr, \qquad \text{we have} \qquad dx = h\,dr.$$

The integral of $\frac{1}{2}r(r+1)$ is 5/12 and that of $\frac{1}{6}r(r+1)(r+2)$ is 3/8. We thus obtain

$$\int_{x_n}^{x_{n+1}} p_3\,dx = h\int_0^1 p_3\,dr = h\left(f_n + \frac{1}{2}\nabla f_n + \frac{5}{12}\nabla^2 f_n + \frac{3}{8}\nabla^3 f_n\right). \tag{4}$$

It is practical to replace these differences by their expressions in terms of f:

$$\begin{aligned} \nabla f_n &= f_n - f_{n-1} \\ \nabla^2 f_n &= f_n - 2f_{n-1} + f_{n-2} \\ \nabla^3 f_n &= f_n - 3f_{n-1} + 3f_{n-2} - f_{n-3}. \end{aligned}$$

We substitute this into (4) and collect terms. This gives the multistep formula of the **Adams–Bashforth method** *of fourth order*

$$y_{n+1} = y_n + \frac{h}{24}(55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3}). \tag{5}$$

It expresses the new value y_{n+1} [approximation of the solution y of (1) at x_{n+1}] in terms of 4 values of f computed from the y-values obtained in the preceding 4 steps. The local truncation error is of order h^5, as can be shown, so that the global error is of order h^4; hence (5) does define a fourth-order method.

Adams–Moulton Methods

Adams–Moulton methods are obtained if for $p(x)$ in (2) we choose a polynomial that interpolates $f(x, y(x))$ at $x_{n+1}, x_n, x_{n-1}, \cdots$ (as opposed to $x_n, x_{n-1}, \cdots$ used before; this is the main point). We explain the principle for the cubic polynomial $\tilde{p}_3(x)$ that interpolates at $x_{n+1}, x_n, x_{n-1}, x_{n-2}$. (Before we had $x_n, x_{n-1}, x_{n-2}, x_{n-3}$.) Again using (18) in Sec. 19.3 but now setting $r = (x - x_{n+1})/h$, we have

$$\tilde{p}_3(x) = f_{n+1} + r\nabla f_{n+1} + \tfrac{1}{2}r(r+1)\nabla^2 f_{n+1} + \tfrac{1}{6}r(r+1)(r+2)\nabla^3 f_{n+1}.$$

We now integrate over x from x_n to x_{n+1} as before. This corresponds to integrating over r from -1 to 0. We obtain

$$\int_{x_n}^{x_{n+1}} \tilde{p}_3(x)\,dx = h\left(f_{n+1} - \frac{1}{2}\nabla f_{n+1} - \frac{1}{12}\nabla^2 f_{n+1} - \frac{1}{24}\nabla^3 f_{n+1}\right).$$

Replacing the differences as before gives

$$(6) \qquad y_{n+1} = y_n + \int_{x_n}^{x_{n+1}} \tilde{p}_3(x)\,dx = y_n + \frac{h}{24}(9f_{n+1} + 19f_n - 5f_{n-1} + f_{n-2}).$$

This is usually called an **Adams–Moulton formula.** It is an **implicit formula** because $f_{n+1} = f(x_{n+1}, y_{n+1})$ appears on the right, so that it defines y_{n+1} only *implicitly,* in contrast to (5), which is an **explicit formula,** not involving y_{n+1} on the right. To use (6) we must ***predict*** a value y_{n+1}^*, for instance, by using (5), that is,

$$\textbf{(7a)} \qquad y_{n+1}^* = y_n + \frac{h}{24}(55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3}).$$

The ***corrected*** new value y_{n+1} is then obtained from (6) with f_{n+1} replaced by $f_{n+1}^* = f(x_{n+1}, y_{n+1}^*)$ and the other f's as in (6); thus,

$$\textbf{(7b)} \qquad y_{n+1} = y_n + \frac{h}{24}(9f_{n+1}^* + 19f_n - 5f_{n-1} + f_{n-2}).$$

This **predictor–corrector method** (7a), (7b) is usually called the **Adams–Moulton method** *of fourth order.* It has the advantage over RK that (7) gives the error estimate

$$\epsilon_{n+1} \approx \tfrac{1}{15}(y_{n+1} - y_{n+1}^*),$$

as can be shown. This is the analog of (10) in Sec. 21.1.

Sometimes the name 'Adams–Moulton method' is reserved for the method with *several* corrections per step by (7b) until a specific accuracy is reached. Popular codes exist for both versions of the method.

Getting Started. In (5) we need f_0, f_1, f_2, f_3. Hence from (3) we see that we must first compute y_1, y_2, y_3 by some other method of comparable accuracy, for instance, by RK or by RKF. For other choices see Ref. [E26] listed in App. 1.

EXAMPLE 1 Adams–Bashforth Prediction (7a), Adams–Moulton Correction (7b)

Solve the initial value problem

$$y' = x + y, \qquad y(0) = 0 \tag{8}$$

by (7a), (7b) on the interval $0 \leqq x \leqq 2$, choosing $h = 0.2$.

Solution. The problem is the same as in Examples 1–3, Sec. 21.1, so that we can compare the results. We compute starting values y_1, y_2, y_3 by the classical Runge–Kutta method. Then in each step we predict by (7a) and make one correction by (7b) before we execute the next step. The results are shown and compared with the exact values in Table 21.10. We see that the corrections improve the accuracy considerably. This is typical. ■

Table 21.10 Adams–Moulton Method Applied to the Initial Value Problem (8); Predicted Values Computed by (7a) and Corrected Values by (7b)

n	x_n	Starting y_n	Predicted y_n^*	Corrected y_n	Exact Values	$10^6 \cdot$ Error of y_n
0	0.0	0.000 000			0.000 000	0
1	0.2	0.021 400			0.021 403	3
2	0.4	0.091 818			0.091 825	7
3	0.6	0.222 107			0.222 119	12
4	0.8		0.425 361	0.425 529	0.425 541	12
5	1.0		0.718 066	0.718 270	0.718 282	12
6	1.2		1.119 855	1.120 106	1.120 117	11
7	1.4		1.654 885	1.655 191	1.655 200	9
8	1.6		2.352 653	2.353 026	2.353 032	6
9	1.8		3.249 190	3.249 646	3.249 647	1
10	2.0		4.388 505	4.389 062	4.389 056	−6

Comments on Comparison of Methods. An Adams–Moulton formula is generally much more accurate than an Adams–Bashforth formula of the same order. This justifies the greater complication and expense in using the former. The method (7a), (7b) is ***numerically stable,*** whereas the exclusive use of (7a) might cause instability. Step size control is relatively simple. If |Corrector − Predictor| > TOL, use interpolation to generate "old" results at half the current step size and then try $h/2$ as the new step.

Whereas the Adams–Moulton formula (7a), (7b) needs only 2 evaluations per step, Runge–Kutta needs 4; however, with Runge–Kutta one may be able to take a step size more than twice as large, so that a comparison of this kind (widespread in the literature) is meaningless.

For more details, see Refs. [E25], [E26] listed in App. 1.

PROBLEM SET 21.2

1. Carry out and show the details of the calculations leading to (4)–(7) in the text.

2–11 ADAMS–MOULTON METHOD (7a), (7b)

Solve the initial value problems by Adams–Moulton, 10 steps with 1 correction per step. Solve exactly and compute the error. (Use RK where no starting values are given.)

2. $y' = y$, $y(0) = 1$, $h = 0.1$ (1.105171, 1.221403, 1.349859)

3. $y' = -0.2xy$, $y(0) = 1$, $h = 0.2$

4. $y' = 2xy$, $y(0) = 1$, $h = 0.1$

5. $y' = 1 + y^2$, $y(0) = 0$, $h = 0.1$

6. Do Prob. 4 by RK, 5 steps, $h = 0.2$. Compare the errors.

7. Do Prob. 5 by RK, 5 steps, $h = 0.2$. Compare the errors.

8. $y' = x/y$, $y(1) = 3$, $h = 0.2$

9. $y' = (x + y - 4)^2$, $y(0) = 4$, $h = 0.2$, only 7 steps (why?)

10. $y' = 1 - 4y^2$, $y(0) = 0$, $h = 0.1$

11. $y' = x + y$, $y(0) = 0$, $h = 0.1$ (0.00517083, 0.0214026, 0.0498585)

12. Show that by applying the method in the text to a polynomial of second degree we obtain the predictor and corrector formulas

$$y_{n+1}^* = y_n + \frac{h}{12}(23f_n - 16f_{n-1} + 5f_{n-2}),$$

$$y_{n+1} = y_n + \frac{h}{12}(5f_{n+1} + 8f_n - f_{n-1}).$$

13. Use Prob. 12 to solve $y' = 2xy$, $y(0) = 1$ (10 steps, $h = 0.1$, RK starting values). Compare with the exact solution and comment.

14. How much can you reduce the error in Prob. 13 by halving h (20 steps, $h = 0.05$)? First guess, then compute.

15. **CAS PROJECT. Adams–Moulton.** **(a) Accurate starting** is important in (7a), (7b). Illustrate this in Example 1 of the text by using starting values from the improved Euler–Cauchy method and compare the results with those in Table 21.9.

(b) How much does the error in Prob. 11 decrease if you use exact starting values (instead of RK-values)?

(c) Experiment to find out for what ODEs poor starting is very damaging and for what ODEs it is not.

(d) The classical **RK method** often gives the same accuracy with step $2h$ as Adams–Moulton with step h, so that the total number of function evaluations is the same in both cases. Illustrate this with Prob. 8. (Hence corresponding comparisons in the literature in favor of Adams–Moulton are not valid. See also Probs. 6 and 7.)

21.3 Methods for Systems and Higher Order ODEs

Initial value problems for first-order systems of ODEs are of the form

$$\mathbf{y}' = \mathbf{f}(x, \mathbf{y}), \qquad \mathbf{y}(x_0) = \mathbf{y}_0, \tag{1}$$

in components

$$\begin{aligned} y_1' &= f_1(x, y_1, \cdots, y_m), & y_1(x_0) &= y_{10} \\ y_2' &= f_2(x, y_1, \cdots, y_m), & y_2(x_0) &= y_{20} \\ &\cdots\cdots\cdots\cdots & &\cdots\cdots\cdots \\ y_m' &= f_m(x, y_1, \cdots, y_m). & y_m(x_0) &= y_{m0}. \end{aligned}$$

$\mathbf{f}$ is assumed to be such that the problem has a unique solution $\mathbf{y}(x)$ on some open x-interval containing x_0. Our discussion will be independent of Chap. 4 on systems.

Before explaining solution methods it is important to note that (1) includes initial value problems for single mth-order ODEs,

$$y^{(m)} = f(x, y, y', y'', \cdots, y^{(m-1)}) \tag{2}$$

and initial conditions $y(x_0) = K_1$, $y'(x_0) = K_2, \cdots, y^{(m-1)}(x_0) = K_m$ as special cases.

Indeed, the connection is achieved by setting

$$(3) \qquad y_1 = y, \quad y_2 = y', \quad y_3 = y'', \quad \cdots, \quad y_m = y^{(m-1)}.$$

Then we obtain the system

$$(4) \qquad \begin{aligned} y_1' &= y_2 \\ y_2' &= y_3 \\ &\vdots \\ y_{m-1}' &= y_m \\ y_m' &= f(x, y_1, \cdots, y_m) \end{aligned}$$

and the initial conditions $y_1(x_0) = K_1, \quad y_2(x_0) = K_2, \quad \cdots, \quad y_m(x_0) = K_m.$

Euler Method for Systems

Methods for single first-order ODEs can be extended to systems (1) simply by writing vector functions $\mathbf{y}$ and $\mathbf{f}$ instead of scalar functions y and f, whereas x remains a scalar variable.

We begin with the Euler method. Just as for a single ODE this method will not be accurate enough for practical purposes, but it nicely illustrates the extension principle.

EXAMPLE 1 **Euler Method for a Second-Order ODE. Mass–Spring System**

Solve the initial value problem for a damped mass–spring system

$$y'' + 2y' + 0.75y = 0, \qquad y(0) = 3, \qquad y'(0) = -2.5$$

by the Euler method for systems with step $h = 0.2$ for x from 0 to 1 (where x is time).

Solution. The **Euler method** (3), Sec. 21.1, generalizes to systems in the form

$$(5) \qquad \mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{f}(x_n, \mathbf{y}_n),$$

in components

$$y_{1,n+1} = y_{1,n} + hf_1(x_n, y_{1,n}, y_{2,n})$$

$$y_{2,n+1} = y_{2,n} + hf_2(x_n, y_{1,n}, y_{2,n})$$

and similarly for systems of more than two equations. By (4) the given ODE converts to the system

$$y_1' = f_1(x, y_1, y_2) = y_2$$

$$y_2' = f_2(x, y_1, y_2) = -2y_2 - 0.75y_1.$$

Hence (5) becomes

$$y_{1,n+1} = y_{1,n} + 0.2y_{2,n}$$

$$y_{2,n+1} = y_{2,n} + 0.2(-2y_{2,n} - 0.75y_{1,n}).$$

The initial conditions are $y(0) = y_1(0) = 3$, $y'(0) = y_2(0) = -2.5$. The calculations are shown in Table 21.11 on the next page. As for single ODEs, the results would not be accurate enough for practical purposes. The example merely serves to illustrate the method because the problem can be readily solved exactly,

$$y = y_1 = 2e^{-0.5x} + e^{-1.5x}, \qquad \text{thus} \qquad y' = y_2 = -e^{-0.5x} - 1.5e^{-1.5x}.$$

Table 21.11 Euler Method for Systems in Example 1 (Mass–Spring System)

n	x_n	$y_{1,n}$	y_1 Exact (5D)	Error $\epsilon_1 = y_1 - y_{1,n}$	$y_{2,n}$	y_2 Exact (5D)	Error $\epsilon_2 = y_2 - y_{2,n}$
0	0.0	3.00000	3.00000	0.00000	−2.50000	−2.50000	0.00000
1	0.2	2.50000	2.55049	0.05049	−1.95000	−2.01606	−0.06606
2	0.4	2.11000	2.18627	0.76270	−1.54500	−1.64195	−0.09695
3	0.6	1.80100	1.88821	0.08721	−1.24350	−1.35067	−0.10717
4	0.8	1.55230	1.64183	0.08953	−1.01625	−1.12211	−0.10586
5	1.0	1.34905	1.43619	0.08714	−0.84260	−0.94123	−0.09863

Runge–Kutta Methods for Systems

As for Euler methods, we obtain RK methods for an initial value problem (1) simply by writing vector formulas for vectors with m components, which for $m = 1$ reduce to the previous scalar formulas.

Thus for the *classical* **RK method** *of fourth order* in Table 21.4 we obtain

(6a) $$\mathbf{y}(x_0) = \mathbf{y}_0 \qquad \text{(Initial values)}$$

and for each step $n = 0, 1, \cdots, N - 1$ we obtain the 4 auxiliary quantities

(6b)
$$\begin{aligned}
\mathbf{k}_1 &= h\mathbf{f}(x_n, \quad \mathbf{y}_n) \\
\mathbf{k}_2 &= h\mathbf{f}(x_n + \tfrac{1}{2}h, \quad \mathbf{y}_n + \tfrac{1}{2}\mathbf{k}_1) \\
\mathbf{k}_3 &= h\mathbf{f}(x_n + \tfrac{1}{2}h, \quad \mathbf{y}_n + \tfrac{1}{2}\mathbf{k}_2) \\
\mathbf{k}_4 &= h\mathbf{f}(x_n + h, \quad \mathbf{y}_n + \mathbf{k}_3)
\end{aligned}$$

and the new value [approximation of the solution $\mathbf{y}(x)$ at $x_{n+1} = x_0 + (n + 1)h$]

(6c) $$\mathbf{y}_{n+1} = \mathbf{y}_n + \tfrac{1}{6}(\mathbf{k}_1 + 2\mathbf{k}_2 + 2\mathbf{k}_3 + \mathbf{k}_4).$$

EXAMPLE 2 RK Method for Systems. Airy's Equation. Airy Function Ai(x)

Solve the initial value problem

$$y'' = xy, \qquad y(0) = 1/(3^{2/3} \cdot \Gamma(2/3)) = 0.35502\,805, \qquad y'(0) = -1/(3^{1/3} \cdot \Gamma(1/3)) = -0.25881\,940$$

by the Runge–Kutta method for systems with $h = 0.2$; do 5 steps. This is **Airy's equation,**[2] which arose in optics (see Ref. [A13], p. 188, listed in App. 1). Γ is the gamma function (see App. A3.1). The initial conditions are such that we obtain a standard solution, the **Airy function** Ai(x), a special function that has been thoroughly investigated; for numeric values, see Ref. [GR1], pp. 446, 475.

[2]Named after Sir GEORGE BIDELL AIRY (1801–1892), English mathematician, who is known for his work in elasticity and in PDEs.

Solution. For $y'' = xy$, setting $y_1 = y$, $y_2 = y_1' = y'$ we obtain the system (4)

$$y_1' = y_2$$
$$y_2' = xy_1.$$

Hence $\mathbf{f} = [f_1 \quad f_2]^\mathsf{T}$ in (1) has the components $f_1(x, y) = y_2$, $f_2(x, y) = xy_1$. We now write (6) in components. The initial conditions (6a) are $y_{1,0} = 0.35502\,805$, $y_{2,0} = -0.25881\,940$. In (6b) we have fewer subscripts by simply writing $\mathbf{k}_1 = \mathbf{a}$, $\mathbf{k}_2 = \mathbf{b}$, $\mathbf{k}_3 = \mathbf{c}$, $\mathbf{k}_4 = \mathbf{d}$, so that $\mathbf{a} = [a_1 \quad a_2]^\mathsf{T}$, etc. Then (6b) takes the form

(6b*)
$$\mathbf{a} = h\begin{bmatrix} y_{2,n} \\ x_n y_{1,n} \end{bmatrix}$$
$$\mathbf{b} = h\begin{bmatrix} y_{2,n} + \frac{1}{2}a_2 \\ (x_n + \frac{1}{2}h)(y_{1,n} + \frac{1}{2}a_1) \end{bmatrix}$$
$$\mathbf{c} = h\begin{bmatrix} y_{2,n} + \frac{1}{2}b_2 \\ (x_n + \frac{1}{2}h)(y_{1,n} + \frac{1}{2}b_1) \end{bmatrix}$$
$$\mathbf{d} = h\begin{bmatrix} y_{2,n} + c_2 \\ (x_n + h)(y_{1,n} + c_1) \end{bmatrix}.$$

For example, the second component of $\mathbf{b}$ is obtained as follows. $\mathbf{f}(x, \mathbf{y})$ has the second component $f_2(x, \mathbf{y}) = xy_1$. Now in $\mathbf{b}$ $(= \mathbf{k}_2)$ the first argument is

$$x = x_n + \tfrac{1}{2}h.$$

The second argument in $\mathbf{b}$ is

$$\mathbf{y} = \mathbf{y}_n + \tfrac{1}{2}\mathbf{a},$$

and the first component of this is

$$y_1 = y_{1,n} + \tfrac{1}{2}a_1.$$

Together,

$$xy_1 = (x_n + \tfrac{1}{2}h)(y_{1,n} + \tfrac{1}{2}a_1).$$

Similarly for the other components in (6b*). Finally,

(6c*)
$$\mathbf{y}_{n+1} = \mathbf{y}_n + \tfrac{1}{6}(\mathbf{a} + 2\mathbf{b} + 2\mathbf{c} + \mathbf{d}).$$

Table 21.12 shows the values $y(x) = y_1(x)$ of the Airy function $\mathrm{Ai}(x)$ and of its derivative $y'(x) = y_2(x)$ as well as of the (rather small!) error of $y(x)$. ∎

Table 21.12 RK Method for Systems: Values $y_{1,n}(x_n)$ of the Airy Function Ai(x) in Example 2

n	x_n	$y_{1,n}(x_n)$	$y_1(x_n)$ Exact (8D)	$10^8 \cdot$ Error of y_1	$y_{2,n}(x_n)$
0	0.0	0.35502 805	0.35502 805	0	−0.25881 940
1	0.2	0.30370 303	0.30370 315	12	−0.25240 464
2	0.4	0.25474 211	0.25474 235	24	−0.23583 073
3	0.6	0.20979 973	0.20980 006	33	−0.21279 185
4	0.8	0.16984 596	0.16984 632	36	−0.18641 171
5	1.0	0.13529 207	0.13529 242	35	−0.15914 687

Runge–Kutta–Nyström Methods (RKN Methods)

RKN methods are direct extensions of RK methods (Runge–Kutta methods) to second-order ODEs $y'' = f(x, y, y')$, as given by the Finnish mathematician E. J. Nyström [*Acta Soc. Sci. fenn.*, 1925, L, No. 13]. The best known of these uses the following formulas, where $n = 0, 1, \cdots, N - 1$ (N the number of steps):

(7a)
$$\begin{aligned} k_1 &= \tfrac{1}{2}hf(x_n, y_n, y_n') \\ k_2 &= \tfrac{1}{2}hf(x_n + \tfrac{1}{2}h, y_n + K, y_n' + k_1) \qquad \text{where } K = \tfrac{1}{2}h(y_n' + \tfrac{1}{2}k_1) \\ k_3 &= \tfrac{1}{2}hf(x_n + \tfrac{1}{2}h, y_n + K, y_n' + k_2) \\ k_4 &= \tfrac{1}{2}hf(x_n + h, y_n + L, y_n' + 2k_3) \qquad \text{where } L = h(y_n' + k_3). \end{aligned}$$

From this we compute the approximation y_{n+1} of $y(x_{n+1})$ at $x_{n+1} = x_0 + (n + 1)h$,

(7b)
$$y_{n+1} = y_n + h(y_n' + \tfrac{1}{3}(k_1 + k_2 + k_3)),$$

and the approximation y_{n+1}' of the derivative $y'(x_{n+1})$ needed in the next step,

(7c)
$$y_{n+1}' = y_n' + \tfrac{1}{3}(k_1 + 2k_2 + 2k_3 + k_4).$$

RKN for ODEs $y'' = f(x, y)$ Not Containing y'. Then $k_2 = k_3$ in (7), which makes the method particularly advantageous and reduces (7) to

(7*)
$$\begin{aligned} k_1 &= \tfrac{1}{2}hf(x_n, y_n) \\ k_2 &= \tfrac{1}{2}hf(x_n + \tfrac{1}{2}h, y_n + \tfrac{1}{2}h(y_n' + \tfrac{1}{2}k_1)) = k_3 \\ k_4 &= \tfrac{1}{2}hf(x_n + h, y_n + h(y_n' + k_2)) \\ y_{n+1} &= y_n + h(y_n' + \tfrac{1}{3}(k_1 + 2k_2)) \\ y_{n+1}' &= y_n' + \tfrac{1}{3}(k_1 + 4k_2 + k_4). \end{aligned}$$

EXAMPLE 3 **RKN Method. Airy's Equation. Airy Function Ai(x)**

For the problem in Example 2 and $h = 0.2$ as before we obtain from (7*) simply $k_1 = 0.1x_n y_n$ and

$$k_2 = k_3 = 0.1(x_n + 0.1)(y_n + 0.1y_n' + 0.05k_1), \qquad k_4 = 0.1(x_n + 0.2)(y_n + 0.2y_n' + 0.2k_2).$$

Table 21.13 shows the results. The accuracy is the same as in Example 2, but the work was much less. ■

Table 21.13 Runge–Kutta–Nyström Method Applied to Airy's Equation, Computation of the Airy Function y = Ai(x)

x_n	y_n	y_n'	$y(x)$ Exact (8D)	$10^8 \cdot$ Error of y_n
0.0	0.355 028 05	−0.258 819 40	0.355 028 05	0
0.2	0.303 703 04	−0.252 404 64	0.303 703 15	11
0.4	0.254 742 11	−0.235 830 70	0.254 742 35	24
0.6	0.209 799 74	−0.212 791 72	0.209 800 06	32
0.8	0.169 845 99	−0.186 411 34	0.169 846 32	33
1.0	0.135 292 18	−0.159 146 09	0.135 292 42	24

Our work in Examples 2 and 3 also illustrates that usefulness of methods for ODEs in the computation of values of **"higher transcendental functions."**

Backward Euler Method for Systems. Stiff Systems

The backward Euler formula (16) in Sec. 21.1 generalizes to systems in the form

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\,\mathbf{f}(x_{n+1}, \mathbf{y}_{n+1}) \qquad (n = 0, 1, \cdots). \tag{8}$$

This is again an implicit method, giving $\mathbf{y}_{n+1}$ implicitly for given $\mathbf{y}_n$. Hence (8) must be solved for $\mathbf{y}_{n+1}$. For a linear system this is shown in the next example. This example also illustrates that, similar to the case of a single ODE in Sec. 21.1, the method is very useful for **stiff systems.** These are systems of ODEs whose matrix has eigenvalues λ of very different magnitudes, having the effect that, just as in Sec. 21.1, the step in direct methods, RK for example, cannot be increased beyond a certain threshold without losing stability. ($\lambda = -1$ and -10 in Example 4, but larger differences do occur in applications.)

EXAMPLE 4 **Backward Euler Method for Systems of ODEs. Stiff Systems**

Compare the backward Euler method (8) with the Euler and the RK methods for numerically solving the initial value problem

$$y'' + 11y' + 10y = 10x + 11, \qquad y(0) = 2, \qquad y'(0) = -10$$

converted to a system of first-order ODEs.

Solution. The given problem can easily be solved, obtaining

$$y = e^{-x} + e^{-10x} + x$$

so that we can compute errors. Conversion to a system by setting $y = y_1$, $y' = y_2$ [see (4)] gives

$$\begin{aligned} y_1' &= y_2 & y_1(0) &= 2 \\ y_2' &= -10y_1 - 11y_2 + 10x + 11 & y_2(0) &= -10. \end{aligned}$$

The coefficient matrix

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -10 & -11 \end{bmatrix} \quad \text{has the characteristic determinant} \quad \begin{vmatrix} -\lambda & 1 \\ -10 & -\lambda - 11 \end{vmatrix}$$

whose value is $\lambda^2 + 11\lambda + 10 = (\lambda + 1)(\lambda + 10)$. Hence the eigenvalues are -1 and -10 as claimed above. The backward Euler formula is

$$\mathbf{y}_{n+1} = \begin{bmatrix} y_{1,n+1} \\ y_{2,n+1} \end{bmatrix} = \begin{bmatrix} y_{1,n} \\ y_{2,n} \end{bmatrix} + h \begin{bmatrix} y_{2,n+1} \\ -10y_{1,n+1} - 11y_{2,n+1} + 10x_{n+1} + 11 \end{bmatrix}.$$

Reordering terms gives the linear system in the unknowns $y_{1,n+1}$ and $y_{2,n+1}$

$$\begin{aligned} y_{1,n+1} - hy_{2,n+1} &= y_{1,n} \\ 10hy_{1,n+1} + (1 + 11h)y_{2,n+1} &= y_{2,n} + 10h(x_n + h) + 11h. \end{aligned}$$

The coefficient determinant is $D = 1 + 11h + 10h^2$, and Cramer's rule (in Sec. 7.6) gives the solution

$$\mathbf{y}_{n+1} = \frac{1}{D} \begin{bmatrix} (1 + 11h)y_{1,n} + hy_{2,n} + 10h^2x_n + 11h^2 + 10h^3 \\ -10hy_{1,n} + y_{2,n} + 10hx_n + 11h + 10h^2 \end{bmatrix}.$$

Table 21.14 Backward Euler Method (BEM) for Example 4. Comparison with Euler and RK

x	BEM $h = 0.2$	BEM $h = 0.4$	Euler $h = 0.1$	Euler $h = 0.2$	RK $h = 0.2$	RK $h = 0.3$	Exact
0.0	2.00000	2.00000	2.00000	2.00000	2.00000	2.00000	2.00000
0.2	1.36667		1.01000	0.00000	1.35207		1.15407
0.4	1.20556	1.31429	1.56100	2.04000	1.18144		1.08864
0.6	1.21574		1.13144	0.11200	1.18585	3.03947	1.15129
0.8	1.29460	1.35020	1.23047	2.20960	1.26168		1.24966
1.0	1.40599		1.34868	0.32768	1.37200		1.36792
1.2	1.53627	1.57243	1.48243	2.46214	1.50257	5.07569	1.50120
1.4	1.67954		1.62877	0.60972	1.64706		1.64660
1.6	1.83272	1.86191	1.78530	2.76777	1.80205		1.80190
1.8	1.99386		1.95009	0.93422	1.96535	8.72329	1.96530
2.0	2.16152	2.18625	2.12158	3.10737	2.13536		2.13534

Table 21.14 shows the following.

Stability of the backward Euler method for $h = 0.2$ and 0.4 (and in fact for any h; try $h = 5.0$) with decreasing accuracy for increasing h,

Stability of the Euler method for $h = 0.1$ but instability for $h = 0.2$,

Stability of RK for $h = 0.2$ but instability for $h = 0.3$.

Figure 451 shows the Euler method for $h = 0.18$, an interesting case with initial jumping (for about $x < 3$) but later monotone following the solution curve of $y = y_1$. See also CAS Experiment 21. ■

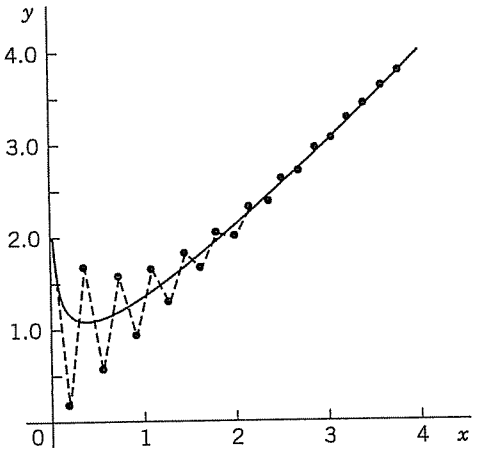


Fig. 451. Euler method with $h = 0.18$ in Example 4

PROBLEM SET 21.5

1. Verify the calculations in Example 1.

2–7 EULER FOR SYSTEMS AND SECOND-ORDER ODES

Solve by the Euler method:

2. $y_1' = -3y_1 + y_2$, $y_2' = y_1 - 3y_2$, $y_1(0) = 2$, $y_2(0) = 0$, $h = 0.1$, 5 steps

3. $y_1' = y_1$, $y_2' = y_2$, $y_1(0) = 1$, $y_2(0) = -1$, $h = 0.2$, 5 steps

4. $y_1' = y_1$, $y_2' = -y_2$, $y_1(0) = 2$, $y_2(0) = 2$, $h = 0.1$, 10 steps

5. $y'' + 4y = 0$, $y(0) = 1$, $y'(0) = 0$, $h = 0.2$, 5 steps

6. $y'' - y = x$, $y(0) = 1$, $y'(0) = -2$, $h = 0.1$, 5 steps

7. $y_1' = -y_1 + y_2$, $y_2' = -y_1 - y_2$, $y_1(0) = 0$, $y_2(0) = 4$, $h = 0.1$, 10 steps

8. Verify the formulas and calculations for the Airy equation in Example 2.

9–14 RK FOR SYSTEMS

Solve by the classical RK:

9. The system in Prob. 7. How much smaller is the error?

10. The ODE in Prob. 6. By what factor did the error decrease?

11. **Undamped Pendulum.** $y'' + \sin y = 0$, $y(\pi) = 0$, $y'(\pi) = 1$, $h = 0.2$, 5 steps. How does your result fit into Fig. 92 in Sec. 4.5?

12. **Bessel Function J_0.** $xy'' + y' + xy = 0$, $y(1) = 0.765198, y'(1) = -0.440051, h = 0.5$, 5 steps. (This gives the standard solution $J_0(x)$ in Fig. 107 in Sec. 5.5.)

13. $y_1' = -4y_1 + y_2$, $y_2' = y_1 - 4y_2$, $y_1(0) = 0$, $y_2(0) = 2$, $h = 0.1$, 5 steps

14. The system in Prob. 2. How much smaller is the error?

15. Verify the calculations for the Airy equation in Example 3.

16–19 RUNGE–KUTTA–NYSTRÖM METHOD

Do by RKN:

16. Prob. 12 (Bessel function J_0). Compare the results.

17. $y'' - xy' + 4y = 0$, $y(0) = 3$, $y'(0) = 0$, $h = 0.2$, 5 steps (Exact: $y = x^4 - 6x^2 + 3$.)

18. $(x^2 - x)y'' - xy' + y = 0$, $y(\frac{1}{2}) = 1 - \frac{1}{2}\ln 2$, $y'(\frac{1}{2}) = 1 - \ln 2$, $h = 0.1$, 4 steps

19. Prob. 11. Compare the results.

20. **CAS EXPERIMENT. Comparison of Methods.** (a) Write programs for RKN and RK for systems.

(b) Try them out for second-order ODEs of your choice to find out empirically which is better in specific cases.

(c) In using RKN, would it pay to first eliminate y' (see Prob. 29 in Problem Set 5.5)? Find out experimentally.

21. **CAS EXPERIMENT. Backward Euler and Stiffness.** Extend Example 4 as follows.

(a) Verify the values in Table 21.14 and show them graphically as in Fig. 451.

(b) Compute and graph Euler values for h near the "critical" $h = 0.18$ to determine more exactly when instability starts.

(c) Compute and graph RK values for values of h between 0.2 and 0.3 to find h for which the RK approximation begins to increase away from the exact solution.

(d) Compute and graph backward Euler values for large h; confirm stability and investigate the error increase for growing h.

21.4 Methods for Elliptic PDEs

The remaining sections of this chapter are devoted to numerics for PDEs (partial differential equations), particularly for the Laplace, Poisson, heat, and wave equations. These PDEs are basic in applications and, at the same time, are model cases of elliptic, parabolic, and hyperbolic PDEs, respectively. The definitions are as follows. (recall also Sec. 12.4).

A PDE is called **quasilinear** if it is linear in the highest derivatives. Hence a second-order quasilinear PDE in two independent variables x, y is of the form

$$au_{xx} + 2bu_{xy} + cu_{yy} = F(x, y, u, u_x, u_y). \tag{1}$$

u is an unknown function of x and y (a solution sought). F is a given function of the indicated variables.

Depending on the discriminant $ac - b^2$, the PDE (1) is said to be of

elliptic type if $ac - b^2 > 0$ (example: *Laplace equation*)

parabolic type if $ac - b^2 = 0$ (example: *heat equation*)

hyperbolic type if $ac - b^2 < 0$ (example: *wave equation*).

Here, in the heat and wave equations, y is time t. The *coefficients* a, b, c may be functions of x, y, so that the type of (1) may be different in different regions of the xy-plane. This classification is not merely a formal matter but is of great practical importance because the general behavior of solutions differs from type to type and so do the additional conditions (boundary and initial conditions) that must be taken into account.

Applications involving ***elliptic equations*** usually lead to boundary value problems in a region R, called a *first boundary value problem* or **Dirichlet problem** if u is prescribed on the boundary curve C of R, a *second boundary value problem* or **Neumann problem** if $u_n = \partial u/\partial n$ (normal derivative of u) is prescribed on C, and a *third* or **mixed problem** if u is prescribed on a part of C and u_n on the remaining part. C usually is a closed curve (or sometimes consists of two or more such curves).

Difference Equations for the Laplace and Poisson Equations

In this section we consider the **Laplace equation**

$$\nabla^2 u = u_{xx} + u_{yy} = 0 \tag{2}$$

and the **Poisson equation**

$$\nabla^2 u = u_{xx} + u_{yy} = f(x, y). \tag{3}$$

These are the most important elliptic PDEs in applications. To obtain methods of numeric solution, we replace the partial derivatives by corresponding **difference quotients,** as follows. By the Taylor formula,

$$\begin{aligned} &\text{(a)} \quad u(x + h, y) = u(x, y) + hu_x(x, y) + \tfrac{1}{2}h^2u_{xx}(x, y) + \tfrac{1}{6}h^3u_{xxx}(x, y) + \cdots \\ &\text{(b)} \quad u(x - h, y) = u(x, y) - hu_x(x, y) + \tfrac{1}{2}h^2u_{xx}(x, y) - \tfrac{1}{6}h^3u_{xxx}(x, y) + \cdots \end{aligned} \tag{4}$$

We subtract (4b) from (4a), neglect terms in h^3, h^4, $\cdots$, and solve for u_x. Then

$$u_x(x, y) \approx \frac{1}{2h}\,[u(x + h, y) - u(x - h, y)]. \tag{5a}$$

Similarly,

$$u(x, y + k) = u(x, y) + ku_y(x, y) + \tfrac{1}{2}k^2u_{yy}(x, y) + \cdots$$

and

$$u(x, y - k) = u(x, y) - ku_y(x, y) + \tfrac{1}{2}k^2u_{yy}(x, y) + \cdots.$$

By subtracting, neglecting terms in k^3, k^4, $\cdots$, and solving for u_y we obtain

$$u_y(x, y) \approx \frac{1}{2k}\,[u(x, y + k) - u(x, y - k)]. \tag{5b}$$

We now turn to second derivatives. Adding (4a) and (4b) and neglecting terms in $h^4, h^5, \cdots$, we obtain $u(x+h, y) + u(x-h, y) \approx 2u(x, y) + h^2 u_{xx}(x, y)$. Solving for u_{xx}, we have

(6a) $$u_{xx}(x, y) \approx \frac{1}{h^2}\,[u(x+h, y) - 2u(x, y) + u(x-h, y)].$$

Similarly,

(6b) $$u_{yy}(x, y) \approx \frac{1}{k^2}\,[u(x, y+k) - 2u(x, y) + u(x, y-k)].$$

We shall not need (see Prob. 1)

(6c) $$u_{xy}(x, y) \approx \frac{1}{4hk}\,[u(x+h, y+k) - u(x-h, y+k) - u(x+h, y-k) + u(x-h, y-k)].$$

Figure 452a shows the points $(x+h, y)$, $(x-h, y)$, $\cdots$ in (5) and (6).

We now substitute (6a) and (6b) into the ***Poisson equation*** (3), choosing $k = h$ to obtain a simple formula:

(7) $$u(x+h, y) + u(x, y+h) + u(x-h, y) + u(x, y-h) - 4u(x, y) = h^2 f(x, y).$$

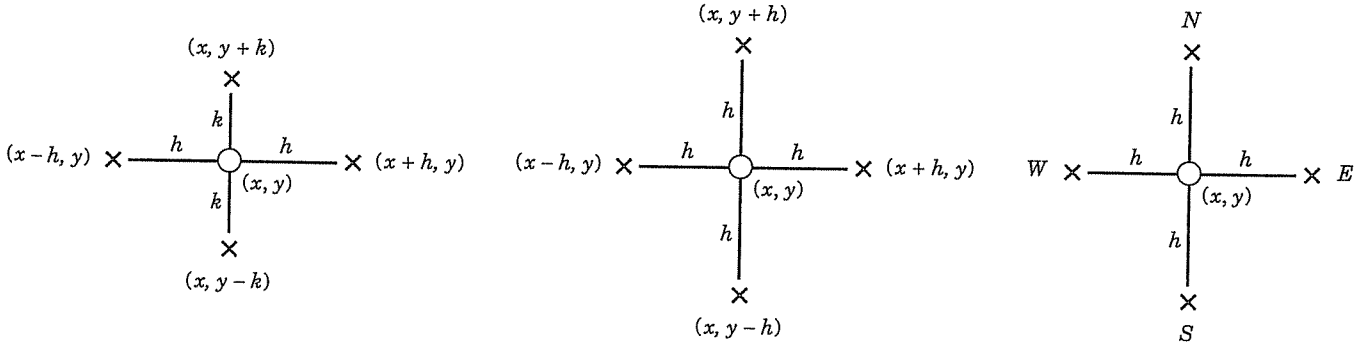
This is a **difference equation** corresponding to (3). Hence for the ***Laplace equation*** (2) the corresponding difference equation is

(8) $$u(x+h, y) + u(x, y+h) + u(x-h, y) + u(x, y-h) - 4u(x, y) = 0.$$

h is called the **mesh size.** Equation (8) relates u at (x, y) to u at the four neighboring points shown in Fig. 452b. It has a remarkable interpretation: u at (x, y) equals the mean of the values of u at the four neighboring points. This is an analog of the mean value property of harmonic functions (Sec. 18.6).

Those neighbors are often called E (East), N (North), W (West), S (South). Then Fig. 452b becomes Fig. 452c and (7) is

(7*) $$u(E) + u(N) + u(W) + u(S) - 4u(x, y) = h^2 f(x, y).$$



(a) Points in (5) and (6) (b) Points in (7) and (8) (c) Notation in (7*)

Fig. 452. Points and notation in (5)–(8) and (7*)

Our approximation of $h^2\nabla^2 u$ in (7) and (8) is a 5-point approximation with the coefficient scheme or **stencil** (also called *pattern, molecule,* or *star*)

$$\textbf{(9)}\quad \begin{Bmatrix} & 1 & \\ 1 & -4 & 1 \\ & 1 & \end{Bmatrix}. \quad \text{We may now write (7) as} \quad \begin{Bmatrix} & 1 & \\ 1 & -4 & 1 \\ & 1 & \end{Bmatrix} u = h^2 f(x, y).$$

Dirichlet Problem

In numerics for the Dirichlet problem in a region R we choose an h and introduce a square grid of horizontal and vertical straight lines of distance h. Their intersections are called **mesh points** (or *lattice points* or *nodes*). See Fig. 453.

Then we approximate the given PDE by a difference equation [(8) for the Laplace equation], which relates the unknown values of u at the mesh points in R to each other and to the given boundary values (details on p. 913). This gives a linear system of *algebraic* equations. By solving it we get approximations of the unknown values of u at the mesh points in R.

We shall see that the number of equations equals the number of unknowns. Now comes an important point. If the number of internal mesh points, call it p, is small, say, $p < 100$, then a direct solution method may be applied to that linear system of $p < 100$ equations in p unknowns. However, if p is large, a storage problem will arise. Now since each unknown u is related to only 4 of its neighbors, the coefficient matrix of the system is a **sparse matrix,** that is, a matrix with relatively few nonzero entries (for instance, 500 of 10000 when $p = 100$). Hence for large p we may avoid storage difficulties by using an iteration method, notably the Gauss–Seidel method (Sec. 20.3), which in PDEs is also called **Liebmann's method.** Remember that in this method we have the storage convenience that we can overwrite any solution component (value of u) as soon as a "new" value is available.

Both cases, large p and small p, are of interest to the engineer, large p if a fine grid is used to achieve high accuracy, and small p if the boundary values are known only rather inaccurately, so that a coarse grid will do it because in this case it would be meaningless to try for great accuracy in the interior of the region R.

We illustrate this approach with an example, keeping the number of equations small, for simplicity. As convenient ***notations for mesh points and corresponding values of the solution*** (and of approximate solutions) we use (see also Fig. 453)

$$\textbf{(10)}\qquad P_{ij} = (ih, jh), \qquad u_{ij} = u(ih, jh).$$

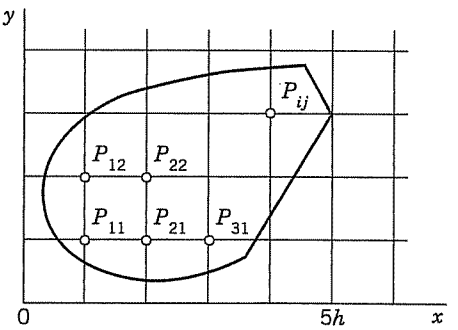


Fig. 453. Region in the xy-plane covered by a grid of mesh h, also showing mesh points $P_{11} = (h, h), \cdots, P_{ij} = (ih, jh), \cdots$

With this notation we can write (8) for any mesh point P_{ij} in the form

$$u_{i+1,j} + u_{i,j+1} + u_{i-1,j} + u_{i,j-1} - 4u_{ij} = 0. \tag{11}$$

EXAMPLE 1 Laplace Equation. Liebmann's Method

The four sides of a square plate of side 12 cm made of homogeneous material are kept at constant temperature 0°C and 100°C as shown in Fig. 454a. Using a (very wide) grid of mesh 4 cm and applying Liebmann's method (that is, Gauss–Seidel iteration), find the (steady-state) temperature at the mesh points.

Solution. In the case of independence of time, the heat equation (see Sec. 10.8)

$$u_t = c^2(u_{xx} + u_{yy})$$

reduces to the Laplace equation. Hence our problem is a Dirichlet problem for the latter. We choose the grid shown in Fig. 454b and consider the mesh points in the order P_{11}, P_{21}, P_{12}, P_{22}. We use (11) and, in each equation, take to the right all the terms resulting from the given boundary values. Then we obtain the system

$$\begin{aligned} -4u_{11} + u_{21} + u_{12} &= -200 \\ u_{11} - 4u_{21} + u_{22} &= -200 \\ u_{11} - 4u_{12} + u_{22} &= -100 \\ u_{21} + u_{12} - 4u_{22} &= -100. \end{aligned} \tag{12}$$

In practice, one would solve such a small system by the Gauss elimination, finding $u_{11} = u_{21} = 87.5$, $u_{12} = u_{22} = 62.5$.

More exact values (exact to 3S) of the solution of the actual problem [as opposed to its model (12)] are 88.1 and 61.9, respectively. (These were obtained by using Fourier series.) Hence the error is about 1%, which is surprisingly accurate for a grid of such a large mesh size h. If the system of equations were large, one would solve it by an indirect method, such as Liebmann's method. For (12) this is as follows. We write (12) in the form (divide by -4 and take terms to the right)

$$\begin{aligned} u_{11} &= 0.25u_{21} + 0.25u_{12} + 50 \\ u_{21} &= 0.25u_{11} + 0.25u_{22} + 50 \\ u_{12} &= 0.25u_{11} + 0.25u_{22} + 25 \\ u_{22} &= 0.25u_{21} + 0.25u_{12} + 25. \end{aligned}$$

These equations are now used for the Gauss–Seidel iteration. They are identical with (2) in Sec. 20.3, where $u_{11} = x_1$, $u_{21} = x_2$, $u_{12} = x_3$, $u_{22} = x_4$, and the iteration is explained there, with 100, 100, 100, 100 chosen as starting values. Some work can be saved by better starting values, usually by taking the average of the boundary values that enter into the linear system. The exact solution of the system is $u_{11} = u_{21} = 87.5$, $u_{12} = u_{22} = 62.5$, as you may verify.

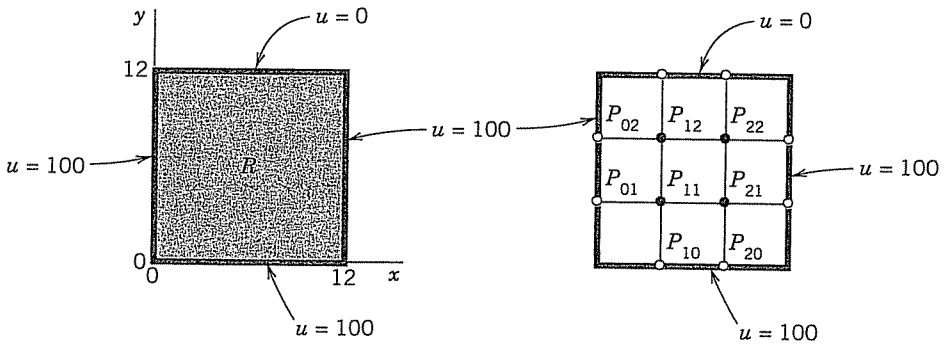


Fig. 454. Example 1

Remark. It is interesting to note that if we choose mesh $h = L/n$ (L = side of R) and consider the $(n-1)^2$ internal mesh points (i.e., mesh points not on the boundary) row by row in the order

$$P_{11}, P_{21}, \cdots, P_{n-1,1}, P_{12}, P_{22}, \cdots, P_{n-1,2}, \cdots,$$

then the system of equations has the $(n-1)^2 \times (n-1)^2$ coefficient matrix

(13) $$\mathbf{A} = \begin{bmatrix} \mathbf{B} & \mathbf{I} & & & & & \\ \mathbf{I} & \mathbf{B} & \mathbf{I} & & & & \\ & \cdot & & & & & \\ & & \cdot & & & & \\ & & & \mathbf{I} & \mathbf{B} & \mathbf{I} & \\ & & & & \mathbf{I} & \mathbf{B} \end{bmatrix}. \quad \text{Here} \quad \mathbf{B} = \begin{bmatrix} -4 & 1 & & & & \\ 1 & -4 & 1 & & & \\ & \cdot & & & & \\ & & \cdot & & & \\ & & & 1 & -4 & 1 \\ & & & & 1 & -4 \end{bmatrix}$$

is an $(n-1) \times (n-1)$ matrix. (In (12) we have $n = 3$, $(n-1)^2 = 4$ internal mesh points, two submatrices **B**, and two submatrices **I**.) The matrix **A** is nonsingular. This follows by noting that the off-diagonal entries in each row of **A** have the sum 3 (or 2), whereas each diagonal entry of **A** equals -4, so that nonsingularity is implied by Gerschgorin's theorem in Sec. 20.7 because no Gerschgorin disk can include 0. ∎

A matrix is called a **band matrix** if it has all its nonzero entries on the main diagonal and on sloping lines parallel to it (separated by sloping lines of zeros or not). For example, **A** in (13) is a band matrix. Although the Gauss elimination does not preserve zeros between bands, it does not introduce nonzero entries outside the limits defined by the original bands. Hence a band structure is advantageous. In (13) it has been achieved by carefully ordering the mesh points.

ADI Method

A matrix is called a **tridiagonal matrix** if it has all its nonzero entries on the main diagonal and on the two sloping parallels immediately above or below the diagonal. (See also Sec. 20.9.) In this case the Gauss elimination is particularly simple.

This raises the question of whether in the solution of the Dirichlet problem for the Laplace or Poisson equations one could obtain a system of equations whose coefficient matrix is tridiagonal. The answer is yes, and a popular method of that kind, called the **ADI method** (*alternating direction implicit method*) was developed by Peaceman and Rachford. The idea is as follows. The stencil in (9) shows that we could obtain a tridiagonal matrix if there were only the three points in a row (or only the three points in a column). This suggests that we write (11) in the form

(14a) $$u_{i-1,j} - 4u_{ij} + u_{i+1,j} = -u_{i,j-1} - u_{i,j+1}$$

so that the left side belongs to y-Row j only and the right side to x-Column i. Of course, we can also write (11) in the form

(14b) $$u_{i,j-1} - 4u_{ij} + u_{i,j+1} = -u_{i-1,j} - u_{i+1,j}$$

so that the left side belongs to Column i and the right side to Row j. In the ADI method we proceed by iteration. At every mesh point we choose an arbitrary starting value $u_{ij}^{(0)}$. In each step we compute new values at all mesh points. In one step we use an iteration

formula resulting from (14a) and in the next step an iteration formula resulting from (14b), and so on in alternating order.

In detail: suppose approximations $u_{ij}^{(m)}$ have been computed. Then, to obtain the next approximations $u_{ij}^{(m+1)}$, we substitute the $u_{ij}^{(m)}$ ***on the right*** side of (14a) and solve for the $u_{ij}^{(m+1)}$ on the left side; that is, we use

(15a)
$$u_{i-1,j}^{(m+1)} - 4u_{ij}^{(m+1)} + u_{i+1,j}^{(m+1)} = -u_{i,j-1}^{(m)} - u_{i,j+1}^{(m)}.$$

We use (15a) for a fixed j, that is, ***for a fixed row*** j, and for all internal mesh points in this row. This gives a linear system of N algebraic equations (N = number of internal mesh points per row) in N unknowns, the new approximations of u at these mesh points. Note that (15a) involves not only approximations computed in the previous step but also given boundary values. We solve the system (15a) (j fixed!) by Gauss elimination. Then we go to the next row, obtain another system of N equations and solve it by Gauss, and so on, until all rows are done. In the next step we ***alternate direction,*** that is, we compute the next approximations $u_{ij}^{(m+2)}$ column by column from the $u_{ij}^{(m+1)}$ and the given boundary values, using a formula obtained from (14b) by substituting the $u_{ij}^{(m+1)}$ ***on the right:***

(15b)
$$u_{i,j-1}^{(m+2)} - 4u_{ij}^{(m+2)} + u_{i,j+1}^{(m+2)} = -u_{i-1,j}^{(m+1)} - u_{i+1,j}^{(m+1)}.$$

For each fixed i, that is, ***for each column,*** this is a system of M equations (M = number of internal mesh points per column) in M unknowns, which we solve by Gauss elimination. Then we go to the next column, and so on, until all columns are done.

Let us consider an example that merely serves to explain the entire method.

EXAMPLE 2 Dirichlet Problem. ADI Method

Explain the procedure and formulas of the ADI method in terms of the problem in Example 1, using the same grid and starting values 100, 100, 100, 100.

Solution. While working, we keep an eye on Fig. 454b on p. 913 and the given boundary values. We obtain first approximations $u_{11}^{(1)}, u_{21}^{(1)}, u_{12}^{(1)}, u_{22}^{(1)}$ from (15a) with $m = 0$. We write boundary values contained in (15a) without an upper index, for better identification and to indicate that these given values remain the same during the iteration. From (15a) with $m = 0$ we have for $j = 1$ (first row) the system

$$(i = 1) \quad u_{01} - 4u_{11}^{(1)} + u_{21}^{(1)} = -u_{10} - u_{12}^{(0)}$$
$$(i = 2) \quad u_{11}^{(1)} - 4u_{21}^{(1)} + u_{31} = -u_{20} - u_{22}^{(0)}.$$

The solution is $u_{11}^{(1)} = u_{21}^{(1)} = 100$. For $j = 2$ (second row) we obtain from (15a) the system

$$(i = 1) \quad u_{02} - 4u_{12}^{(1)} + u_{22}^{(1)} = -u_{11}^{(0)} - u_{13}$$
$$(i = 2) \quad u_{12}^{(1)} - 4u_{22}^{(1)} + u_{32} = -u_{21}^{(0)} - u_{23}.$$

The solution is $u_{12}^{(1)} = u_{22}^{(1)} = 66.667$.

Second approximations $u_{11}^{(2)}, u_{21}^{(2)}, u_{12}^{(2)}, u_{22}^{(2)}$ are now obtained from (15b) with $m = 1$ by using the first approximations just computed and the boundary values. For $i = 1$ (first column) we obtain from (15b) the system

$$(j = 1) \quad u_{10} - 4u_{11}^{(2)} + u_{12}^{(2)} = -u_{01} - u_{21}^{(1)}$$
$$(j = 2) \quad u_{11}^{(2)} - 4u_{12}^{(2)} + u_{13} = -u_{02} - u_{22}^{(1)}.$$

The solution is $u_{11}^{(2)} = 91.11$, $u_{12}^{(2)} = 64.44$. For $i = 2$ (second column) we obtain from (15b) the system

$$(j = 1) \quad u_{20} - 4u_{21}^{(2)} + u_{22}^{(2)} = -u_{11}^{(1)} - u_{31}$$
$$(j = 2) \quad u_{21}^{(2)} - 4u_{22}^{(2)} + u_{23} = -u_{12}^{(1)} - u_{32}.$$

The solution is $u_{21}^{(2)} = 91.11$, $u_{22}^{(2)} = 64.44$.

In this example, which merely serves to explain the practical procedure in the ADI method, the accuracy of the second approximations is about the same as that of two Gauss–Seidel steps in Sec. 20.3 (where $u_{11} = x_1$, $u_{21} = x_2$, $u_{12} = x_3$, $u_{22} = x_4$), as the following table shows.

Method	u_{11}	u_{21}	u_{12}	u_{22}
ADI, 2nd approximations	91.11	91.11	64.44	64.44
Gauss–Seidel, 2nd approximations	93.75	90.62	65.62	64.06
Exact solution of (12)	87.50	87.50	62.50	62.50

Improving Convergence. Additional improvement of the convergence of the ADI method results from the following interesting idea. Introducing a parameter p, we can also write (11) in the form

$$\text{(16)} \quad \begin{aligned} &\text{(a)} \quad u_{i-1,j} - (2+p)u_{ij} + u_{i+1,j} = -u_{i,j-1} + (2-p)u_{ij} - u_{i,j+1} \\ &\text{(b)} \quad u_{i,j-1} - (2+p)u_{ij} + u_{i,j+1} = -u_{i-1,j} + (2-p)u_{ij} - u_{i+1,j}. \end{aligned}$$

This gives the more general ADI iteration formulas

$$\textbf{(17)} \quad \begin{aligned} &\textbf{(a)} \quad u_{i-1,j}^{(m+1)} - (2+p)u_{ij}^{(m+1)} + u_{i+1,j}^{(m+1)} = -u_{i,j-1}^{(m)} + (2-p)u_{ij}^{(m)} - u_{i,j+1}^{(m)} \\ &\textbf{(b)} \quad u_{i,j-1}^{(m+2)} - (2+p)u_{ij}^{(m+2)} + u_{i,j+1}^{(m+2)} = -u_{i-1,j}^{(m+1)} + (2-p)u_{ij}^{(m+1)} - u_{i+1,j}^{(m+1)}. \end{aligned}$$

For $p = 2$, this is (15). The parameter p may be used for improving convergence. Indeed, one can show that the ADI method converges for positive p, and that the optimum value for maximum rate of convergence is

$$\text{(18)} \qquad p_0 = 2 \sin \frac{\pi}{K}$$

where K is the larger of $M + 1$ and $N + 1$ (see above). Even better results can be achieved by letting p vary from step to step. More details of the ADI method and variants are discussed in Ref. [E25] listed in App. 1.

PROBLEM SET 21.4

1. Derive (5b), (6b), and (6c).

2–7 GAUSS ELIMINATION, GAUSS–SEIDEL ITERATION

For the grid in Fig. 455 compute the potential at the four internal points by Gauss and by 5 Gauss–Seidel steps with starting values 100, 100, 100, 100 (showing the details of your work) if the boundary values on the edges are:

2. $u = 0$ on the left, x^3 on the lower edge, $27 - 9y^2$ on the right, $x^3 - 27x$ on the upper edge.

3. $u(1, 0) = 60$, $u(2, 0) = 300$, $u = 100$ on the other three edges.

4. $u = x^4$ on the lower edge, $81 - 54y^2 + y^4$ on the right, $x^4 - 54x^2 + 81$ on the upper edge, y^4 on the left. Verify the exact solution $x^4 - 6x^2y^2 + y^4$ and determine the error.

5. $u = \sin \frac{1}{3}\pi x$ on the upper edge, 0 on the other edges. 10 steps.

6. $u = 220$ on the upper and lower edges, 110 on the left and right.

7. V_0 on the upper and lower edges, $-V_0$ on the left and right. Sketch the equipotential lines.

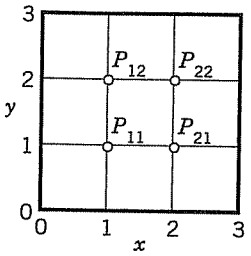


Fig. 455. Problems 2–7

8. Verify the calculations in Example 1. Find out experimentally how many steps are needed to obtain the solution of the linear system with an accuracy of 3S.

9. **(Use of symmetry)** Conclude from the boundary values in Example 1 that $u_{21} = u_{11}$ and $u_{22} = u_{12}$. Show that this leads to a system of two equations and solve it.

10. **(3 × 3 grid)** Solve Example 1, choosing $h = 3$ and starting values 100, 100, $\cdots$.

11. For the square $0 \leqq x \leqq 4$, $0 \leqq y \leqq 4$ let the boundary temperatures be 0°C on the horizontal and 50°C on the vertical edges. Find the temperatures at the interior points of a square grid with $h = 1$.

12. Using the answer to Prob. 11, try to sketch some isotherms.

13. Find the isotherms for the square and grid in Prob. 11 if $u = \sin \frac{1}{4}\pi x$ on the horizontal and $-\sin \frac{1}{4}\pi y$ on the vertical edges. Try to sketch some isotherms.

14. **(Influence of starting values)** Do Prob. 5 by Gauss–Seidel, starting from **0**. Compare and comment.

15. Find the potential in Fig. 456 using (a) the coarse grid, (b) the fine grid, and Gauss elimination. *Hint.* In (b), use symmetry; take $u = 0$ as the boundary value at the two points at which the potential has a jump.

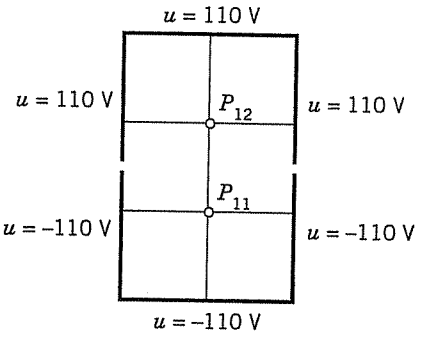


Fig. 456. Region and grids in Problem 15

16. **(ADI)** Apply the ADI method to the Dirichlet problem in Prob. 5, using the grid in Fig. 455, as before and starting values zero.

17. What p_0 in (18) should we choose for Prob. 16? Apply the ADI formulas (17) with $p_0 = 1.7$ to Prob. 16, performing 1 step. Illustrate the improved convergence by comparing with the corresponding values 0.077, 0.308 after the first step in Prob. 16. (Use the starting values zero.)

18. **CAS PROJECT. Laplace Equation. (a)** Write a program for Gauss–Seidel with 16 equations in 16 unknowns, composing the matrix (13) from the indicated 4 × 4 submatrices and including a transformation of the vector of the boundary values into the vector **b** of $\mathbf{Ax} = \mathbf{b}$.

(b) Apply the program to the square grid in $0 \leqq x \leqq 5$, $0 \leqq y \leqq 5$ with $h = 1$ and $u = 220$ on the upper and lower edges, $u = 110$ on the left edge and $u = -10$ on the right edge. Solve the linear system also by Gauss elimination. What accuracy is reached in the 20th Gauss–Seidel step?

21.5 Neumann and Mixed Problems. Irregular Boundary

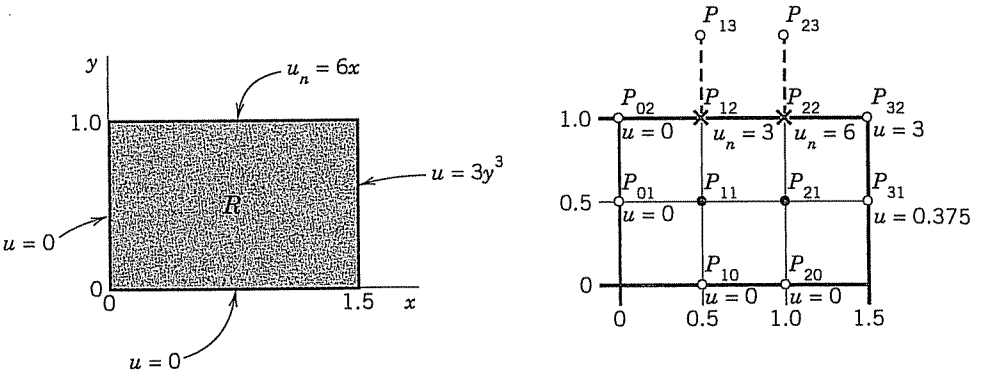
We continue our discussion of boundary value problems for elliptic PDEs in a region R in the xy-plane. The Dirichlet problem was studied in the last section. In solving **Neumann** and **mixed problems** (defined in the last section) we are confronted with a new situation, because there are boundary points at which the (outer) **normal derivative** $u_n = \partial u/\partial n$ of the solution is given, but u itself is unknown since it is not given. To handle such points we need a new idea. This idea is the same for Neumann and mixed problems. Hence we may explain it in connection with one of these two types of problems. We shall do so and consider a typical example as follows.

EXAMPLE 1 Mixed Boundary Value Problem for a Poisson Equation

Solve the mixed boundary value problem for the Poisson equation

$$\nabla^2 u = u_{xx} + u_{yy} = f(x, y) = 12xy$$

shown in Fig. 457a.



(a) Region R and boundary values (b) Grid ($h = 0.5$)

Fig. 457. Mixed boundary value problem in Example 1

Solution. We use the grid shown in Fig. 457b, where $h = 0.5$. We recall that (7) in Sec. 21.4 has the right side $h^2 f(x, y) = 0.5^2 \cdot 12xy = 3xy$. From the formulas $u = 3y^3$ and $u_n = 6x$ given on the boundary we compute the boundary data

$$(1) \qquad u_{31} = 0.375, \qquad u_{32} = 3, \qquad \frac{\partial u_{12}}{\partial n} = \frac{\partial u_{12}}{\partial y} = 6 \cdot 0.5 = 3, \qquad \frac{\partial u_{22}}{\partial n} = \frac{\partial u_{22}}{\partial y} = 6 \cdot 1 = 6.$$

P_{11} and P_{21} are internal mesh points and can be handled as in the last section. Indeed, from (7), Sec. 21.4, with $h^2 = 0.25$ and $h^2 f(x, y) = 3xy$ and from the given boundary values we obtain two equations corresponding to P_{11} and P_{21}, as follows (with -0 resulting from the left boundary).

$$(2a) \qquad \begin{aligned} -4u_{11} + u_{21} + u_{12} &= 12(0.5 \cdot 0.5) \cdot \tfrac{1}{4} - 0 = 0.75 \\ u_{11} - 4u_{21} + u_{22} &= 12(1 \cdot 0.5) \cdot \tfrac{1}{4} - 0.375 = 1.125 \end{aligned}$$

The only difficulty with these equations seems to be that they involve the unknown values u_{12} and u_{22} of u at P_{12} and P_{22} on the boundary, where the normal derivative $u_n = \partial u/\partial n = \partial u/\partial y$ is given, instead of u; but we shall overcome this difficulty as follows.

We consider P_{12} and P_{22}. The idea that will help us here is this. We imagine the region R to be extended above to the first row of external mesh points (corresponding to $y = 1.5$), and we assume that the Poisson equation also holds in the extended region. Then we can write down two more equations as before (Fig. 457b)

$$(2b) \qquad \begin{aligned} u_{11} - 4u_{12} + u_{22} + u_{13} &= 1.5 - 0 = 1.5 \\ u_{21} + u_{12} - 4u_{22} + u_{23} &= 3 - 3 = 0. \end{aligned}$$

On the right, 1.5 is $12xyh^2$ at (0.5, 1) and 3 is $12xyh^2$ at (1, 1) and 0 (at P_{02}) and 3 (at P_{32}) are given boundary values. We remember that we have not yet used the boundary condition on the upper part of the boundary of R, and we also notice that in (2b) we have introduced two more unknowns u_{13}, u_{23}. But we can now use that condition and get rid of u_{13}, u_{23} by applying the central difference formula for du/dy. From (1) we then obtain (see Fig. 457b)

$$3 = \frac{\partial u_{12}}{\partial y} \approx \frac{u_{13} - u_{11}}{2h} = u_{13} - u_{11}, \qquad \text{hence} \qquad u_{13} = u_{11} + 3$$

$$6 = \frac{\partial u_{22}}{\partial y} \approx \frac{u_{23} - u_{21}}{2h} = u_{23} - u_{21}, \qquad \text{hence} \qquad u_{23} = u_{21} + 6.$$

Substituting these results into (2b) and simplifying, we have

$$2u_{11} - 4u_{12} + u_{22} = 1.5 - 3 = -1.5$$
$$2u_{21} + u_{12} - 4u_{22} = 3 - 3 - 6 = -6.$$

Together with (2a) this yields, written in matrix form,

$$\text{(3)} \qquad \begin{bmatrix} -4 & 1 & 1 & 0 \\ 1 & -4 & 0 & 1 \\ 2 & 0 & -4 & 1 \\ 0 & 2 & 1 & -4 \end{bmatrix} \begin{bmatrix} u_{11} \\ u_{21} \\ u_{12} \\ u_{22} \end{bmatrix} = \begin{bmatrix} 0.75 \\ 1.125 \\ 1.5 - 3 \\ 0 - 6 \end{bmatrix} = \begin{bmatrix} 0.75 \\ 1.125 \\ -1.5 \\ -6 \end{bmatrix}.$$

(The entries 2 come from u_{13} and u_{23}, and so do -3 and -6 on the right). The solution of (3) (obtained by Gauss elimination) is as follows; the exact values of the problem are given in parentheses.

$$u_{12} = 0.866 \quad \text{(exact 1)} \qquad u_{22} = 1.812 \quad \text{(exact 2)}$$
$$u_{11} = 0.077 \quad \text{(exact 0.125)} \qquad u_{21} = 0.191 \quad \text{(exact 0.25).}$$

Irregular Boundary

We continue our discussion of boundary value problems for elliptic PDEs in a region R in the xy-plane. If R has a simple geometric shape, we can usually arrange for certain mesh points to lie on the boundary C of R, and then we can approximate partial derivatives as explained in the last section. However, if C intersects the grid at points that are not mesh points, then at points close to the boundary we must proceed differently, as follows.

The mesh point O in Fig. 458 is of that kind. For O and its neighbors A and P we obtain from Taylor's theorem

$$\text{(4)} \qquad \text{(a)} \quad u_A = u_O + ah\frac{\partial u_O}{\partial x} + \frac{1}{2}(ah)^2\frac{\partial^2 u_O}{\partial x^2} + \cdots$$
$$\text{(b)} \quad u_P = u_O - h\frac{\partial u_O}{\partial x} + \frac{1}{2}h^2\frac{\partial^2 u_O}{\partial x^2} + \cdots.$$

We disregard the terms marked by dots and eliminate $\partial u_O/\partial x$. Equation (4b) times a plus equation (4a) gives

$$u_A + au_P \approx (1 + a)u_O + \frac{1}{2}a(a + 1)h^2\frac{\partial^2 u_O}{\partial x^2}.$$

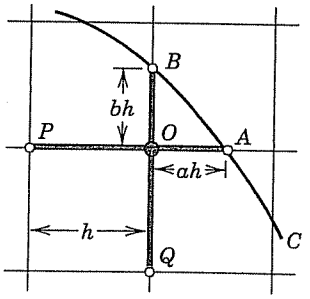


Fig. 458. Curved boundary C of a region R, a mesh point O near C, and neighbors A, B, P, Q

We solve this last equation algebraically for the derivative, obtaining

$$\frac{\partial^2 u_O}{\partial x^2} \approx \frac{2}{h^2}\left[\frac{1}{a(1+a)}\,u_A + \frac{1}{1+a}\,u_P - \frac{1}{a}\,u_O\right].$$

Similarly, by considering the points O, B, and Q,

$$\frac{\partial^2 u_O}{\partial y^2} \approx \frac{2}{h^2}\left[\frac{1}{b(1+b)}\,u_B + \frac{1}{1+b}\,u_Q - \frac{1}{b}\,u_O\right].$$

By addition,

$$\text{(5)}\quad \nabla^2 u_O \approx \frac{2}{h^2}\left[\frac{u_A}{a(1+a)} + \frac{u_B}{b(1+b)} + \frac{u_P}{1+a} + \frac{u_Q}{1+b} - \frac{(a+b)u_O}{ab}\right].$$

For example, if $a = \frac{1}{2}$, $b = \frac{1}{2}$, instead of the stencil (see Sec. 21.4)

$$\begin{Bmatrix} & 1 & \\ 1 & -4 & 1 \\ & 1 & \end{Bmatrix} \qquad \text{we now have} \qquad \begin{Bmatrix} & \frac{4}{3} & \\ \frac{2}{3} & -4 & \frac{4}{3} \\ & \frac{2}{3} & \end{Bmatrix}.$$

because $1/[a(1+a)] = \frac{4}{3}$, etc. The sum of all five terms still being zero (which is useful for checking).

Using the same ideas, you may show that in the case of Fig. 459.

$$\text{(6)}\quad \nabla^2 u_O \approx \frac{2}{h^2}\left[\frac{u_A}{a(a+p)} + \frac{u_B}{b(b+q)} + \frac{u_P}{p(p+a)} + \frac{u_Q}{q(q+b)} - \frac{ap+bq}{abpq}\,u_O\right],$$

a formula that takes care of all conceivable cases.

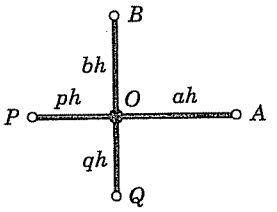


Fig. 459. Neighboring points A, B, P, Q of a mesh point O and notations in formula (6)

EXAMPLE 2 **Dirichlet Problem for the Laplace Equation. Curved Boundary**

Find the potential u in the region in Fig. 460 that has the boundary values given in that figure; here the curved portion of the boundary is an arc of the circle of radius 10 about (0, 0). Use the grid in the figure.

Solution. u is a solution of the Laplace equation. From the given formulas for the boundary values $u = x^3$, $u = 512 - 24y^2$, $\cdots$ we compute the values at the points where we need them; the result is shown in the figure. For P_{11} and P_{12} we have the usual regular stencil, and for P_{21} and P_{22} we use (6), obtaining

$$\text{(7)}\quad P_{11}, P_{12}: \begin{Bmatrix} & 1 & \\ 1 & -4 & 1 \\ & 1 & \end{Bmatrix}, \qquad P_{21}: \begin{Bmatrix} & 0.5 & \\ 0.6 & -2.5 & 0.9 \\ & 0.5 & \end{Bmatrix}, \qquad P_{22}: \begin{Bmatrix} & 0.9 & \\ 0.6 & -3 & 0.9 \\ & 0.6 & \end{Bmatrix}.$$

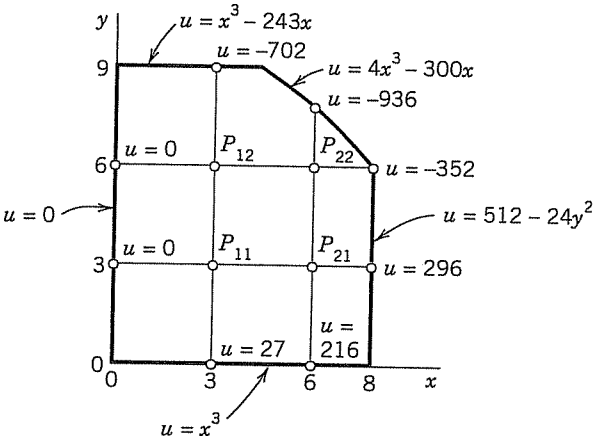


Fig. 460. Region, boundary values of the potential, and grid in Example 2

We use this and the boundary values and take the mesh points in the usual order P_{11}, P_{21}, P_{12}, P_{22}. Then we obtain the system

$$\begin{aligned}
-4u_{11} + u_{21} + u_{12} &= 0 - 27 &&= -27\\
0.6u_{11} - 2.5u_{21} + 0.5u_{22} &= -0.9 \cdot 296 - 0.5 \cdot 216 &&= -374.4\\
u_{11} - 4u_{12} + u_{22} &= 702 + 0 &&= 702\\
0.6u_{21} + 0.6u_{12} - 3u_{22} &= 0.9 \cdot 352 + 0.9 \cdot 936 &&= 1159.2.
\end{aligned}$$

In matrix form,

$$\text{(8)} \qquad \begin{bmatrix} -4 & 1 & 1 & 0 \\ 0.6 & -2.5 & 0 & 0.5 \\ 1 & 0 & -4 & 1 \\ 0 & 0.6 & 0.6 & -3 \end{bmatrix} \begin{bmatrix} u_{11} \\ u_{21} \\ u_{12} \\ u_{22} \end{bmatrix} = \begin{bmatrix} -27 \\ -374.4 \\ 702 \\ 1159.2 \end{bmatrix}.$$

Gauss elimination yields the (rounded) values

$$u_{11} = -55.6, \qquad u_{21} = 49.2, \qquad u_{12} = -298.5, \qquad u_{22} = -436.3.$$

Clearly, from a grid with so few mesh points we cannot expect great accuracy. The exact solution of the PDE (not of the difference equation) having the given boundary values is $u = x^3 - 3xy^2$ and yields the values

$$u_{11} = -54, \qquad u_{21} = 54, \qquad u_{12} = -297, \qquad u_{22} = -432.$$

In practice one would use a much finer grid and solve the resulting large system by an indirect method. ■

PROBLEM SET 21.5

1. Verify the calculation for the Poisson equation in Example 1. Check the values for (3) at the end.
2. Derive (5) in particular when $a = b = \frac{1}{2}$.
3. Derive the general stencil formula (6) in all detail.
4. Verify the calculation for the boundary value problem in Example 2.
5. Do Example 1 in the text for $\nabla^2 u = 0$ with grid and boundary data as before.

MIXED BOUNDARY VALUE PROBLEMS

6. Solve the mixed boundary value problem for the Laplace equation $\nabla^2 u = 0$ in the rectangle in Fig. 457a (using the grid in Fig. 457b) and the boundary conditions $u_x = 0$ on the left edge, $u_x = 3$ on the right edge, $u = x^2$ on the lower edge, and $u = x^2 - 1$ on the upper edge.
7. Solve Prob. 6 when $u_n = 1$ on the upper edge and $u = 1$ on the other edges.

8. Solve the mixed boundary value problem for the Poisson equation $\nabla^2 u = 2(x^2 + y^2)$ in the region and for the boundary conditions shown in Fig. 461, using the indicated grid.

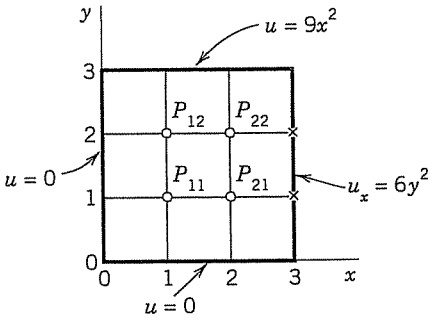


Fig. 461. Problems 8 and 10

9. **CAS EXPERIMENT. Mixed Problem.** Do Example 1 in the text with finer and finer grids of your choice and study the accuracy of the approximate values by comparing with the exact solution $u = 2xy^3$. Verify the latter.

10. Solve $\nabla^2 u = -\pi^2 y \sin \frac{1}{3}\pi x$ for the grid in Fig. 461 and $u_y(1, 3) = u_y(2, 3) = \frac{1}{2}\sqrt{243}$, $u = 0$ on the other three sides of the square.

IRREGULAR BOUNDARIES

11. Solve the Laplace equation in the region and for the boundary values shown in Fig. 462, using the indicated grid. (The sloping portion of the boundary is $y = 4.5 - x$.)

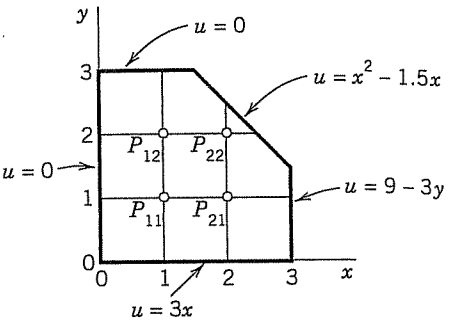


Fig. 462. Problem 11

12. If in Prob. 11 the axes are grounded ($u = 0$), what constant potential must the other portion of the boundary have in order to produce 100 volts at P_{11}?

13. What potential do we have in Prob. 11 if $u = 190$ volts on the axes and $u = 0$ on the other portion of the boundary?

14. Solve the Poisson equation $\nabla^2 u = 2$ in the region and for the boundary values shown in Fig. 463, using the grid also shown in the figure.

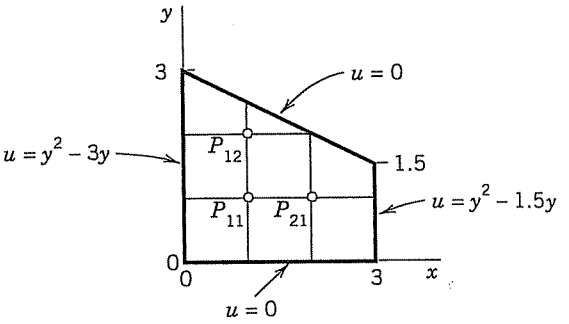


Fig. 463. Problem 14

21.6 Methods for Parabolic PDEs

The last two sections concerned elliptic PDEs, and we now turn to parabolic PDEs. Recall that the definitions of elliptic, parabolic, and hyperbolic PDEs were given in Sec. 21.4. There it was also mentioned that the general behavior of solutions differs from type to type, and so do the problems of practical interest. This reflects on numerics as follows.

For all three types, one replaces the PDE by a corresponding difference equation, but for *parabolic* and *hyperbolic* PDEs this does not automatically guarantee the **convergence** of the approximate solution to the exact solution as the mesh $h \to 0$; in fact, it does not even guarantee convergence at all. For these two types of PDEs one needs additional conditions (inequalities) to assure convergence and **stability,** the latter meaning that small perturbations in the initial data (or small errors at any time) cause only small changes at later times.

In this section we explain the numeric solution of the prototype of parabolic PDEs, the one-dimensional heat equation

$$u_t = c^2 u_{xx} \qquad (c \text{ constant}).$$

This PDE is usually considered for x in some fixed interval, say, $0 \leqq x \leqq L$, and time $t \geqq 0$, and one prescribes the initial temperature $u(x, 0) = f(x)$ (f given) and boundary conditions at $x = 0$ and $x = L$ for all $t \geqq 0$, for instance $u(0, t) = 0$, $u(L, t) = 0$. We may assume $c = 1$ and $L = 1$; this can always be accomplished by a linear transformation of x and t (Prob. 1). Then the **heat equation** and those conditions are

$$\text{(1)} \qquad u_t = u_{xx} \qquad 0 \leqq x \leqq 1,\ t \geqq 0$$

$$\text{(2)} \qquad u(x, 0) = f(x) \qquad \text{(Initial condition)}$$

$$\text{(3)} \qquad u(0, t) = u(1, t) = 0 \qquad \text{(Boundary conditions).}$$

A simple finite difference approximation of (1) is [see (6a) in Sec. 21.4; j is the number of the ***time step***]

$$\text{(4)} \qquad \frac{1}{k}(u_{i,j+1} - u_{ij}) = \frac{1}{h^2}(u_{i+1,j} - 2u_{ij} + u_{i-1,j}).$$

Figure 464 shows a corresponding grid and mesh points. The mesh size is h in the x-direction and k in the t-direction. Formula (4) involves the four points shown in Fig. 465. On the left in (4) we have used a *forward* difference quotient since we have no information for negative t at the start. From (4) we calculate $u_{i,j+1}$, which corresponds to time row $j + 1$, in terms of the three other u that correspond to time row j. Solving (4) for $u_{i,j+1}$, we have

$$\text{(5)} \qquad u_{i,j+1} = (1 - 2r)u_{ij} + r(u_{i+1,j} + u_{i-1,j}), \qquad r = \frac{k}{h^2}.$$

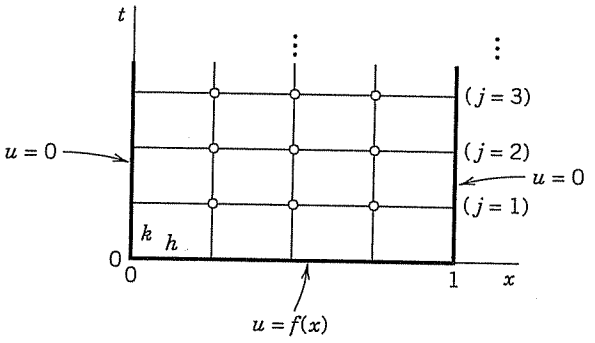


Fig. 464. Grid and mesh points corresponding to (4), (5)

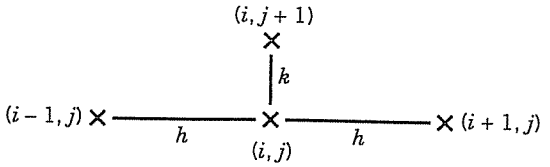


Fig. 465. The four points in (4) and (5)

Computations by this **explicit method** based on (5) are simple. However, it can be shown that crucial to the convergence of this method is the condition

$$r = \frac{k}{h^2} \leqq \frac{1}{2}. \tag{6}$$

That is, u_{ij} should have a positive coefficient in (5) or (for $r = \frac{1}{2}$) be absent from (5). Intuitively, (6) means that we should not move too fast in the t-direction. An example is given below.

Crank–Nicolson Method

Condition (6) is a handicap in practice. Indeed, to attain sufficient accuracy, we have to choose h small, which makes k very small by (6). For example, if $h = 0.1$, then $k \leqq 0.005$. Accordingly, we should look for a more satisfactory discretization of the heat equation.

A method that imposes no restriction on $r = k/h^2$ is the **Crank–Nicolson method,** which uses values of u at the six points in Fig. 466. The idea of the method is the replacement of the difference quotient on the right side of (4) by $\frac{1}{2}$ times the sum of two such difference quotients at two time rows (see Fig. 466). Instead of (4) we then have

$$\begin{aligned}\frac{1}{k}(u_{i,j+1} - u_{ij}) = {} & \frac{1}{2h^2}(u_{i+1,j} - 2u_{ij} + u_{i-1,j}) \\ & + \frac{1}{2h^2}(u_{i+1,j+1} - 2u_{i,j+1} + u_{i-1,j+1}).\end{aligned} \tag{7}$$

Multiplying by $2k$ and writing $r = k/h^2$ as before, we collect the terms corresponding to time row $j + 1$ on the left and the terms corresponding to time row j on the right:

$$(2 + 2r)u_{i,j+1} - r(u_{i+1,j+1} + u_{i-1,j+1}) = (2 - 2r)u_{ij} + r(u_{i+1,j} + u_{i-1,j}). \tag{8}$$

How do we use (8)? In general, the three values on the left are unknown, whereas the three values on the right are known. If we divide the x-interval $0 \leqq x \leqq 1$ in (1) into n equal intervals, we have $n - 1$ internal mesh points per time row (see Fig. 464, where $n = 4$). Then for $j = 0$ and $i = 1, \cdots, n - 1$, formula (8) gives a linear system of $n - 1$ equations for the $n - 1$ unknown values $u_{11}, u_{21}, \cdots, u_{n-1,1}$ in the first time row in terms of the initial values $u_{00}, u_{10}, \cdots, u_{n0}$ and the boundary values u_{01} $(= 0)$, u_{n1} $(= 0)$. Similarly for $j = 1$, $j = 2$, and so on; that is, for each time row we have to solve such a linear system of $n - 1$ equations resulting from (8).

Although $r = k/h^2$ is no longer restricted, smaller r will still give better results. In practice, one chooses a k by which one can save a considerable amount of work, without making r too large. For instance, often a good choice is $r = 1$ (which would be impossible in the previous method). Then (8) becomes simply

$$4u_{i,j+1} - u_{i+1,j+1} - u_{i-1,j+1} = u_{i+1,j} + u_{i-1,j}. \tag{9}$$

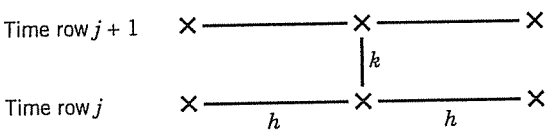


Fig. 466. The six points in the Crank–Nicolson formulas (7) and (8)

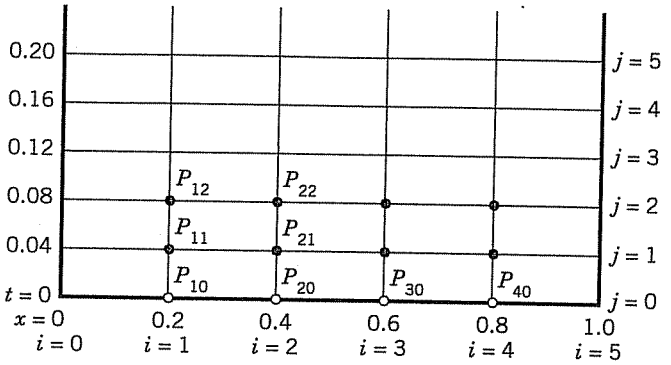


Fig. 467. Grid in Example 1

EXAMPLE 1 **Temperature in a Metal Bar. Crank–Nicolson Method, Explicit Method**

Consider a laterally insulated metal bar of length 1 and such that $c^2 = 1$ in the heat equation. Suppose that the ends of the bar are kept at temperature $u = 0°C$ and the temperature in the bar at some instant—call it $t = 0$—is $f(x) = \sin \pi x$. Applying the Crank–Nicolson method with $h = 0.2$ and $r = 1$, find the temperature $u(x, t)$ in the bar for $0 \leqq t \leqq 0.2$. Compare the results with the exact solution. Also apply (5) with an r satisfying (6), say, $r = 0.25$, and with values not satisfying (6), say, $r = 1$ and $r = 2.5$.

Solution by Crank–Nicolson. Since $r = 1$, formula (8) takes the form (9). Since $h = 0.2$ and $r = k/h^2 = 1$, we have $k = h^2 = 0.04$. Hence we have to do 5 steps. Figure 467 shows the grid. We shall need the initial values

$$u_{10} = \sin 0.2\pi = 0.587\,785, \qquad u_{20} = \sin 0.4\pi = 0.951\,057.$$

Also, $u_{30} = u_{20}$ and $u_{40} = u_{10}$. (Recall that u_{10} means u at P_{10} in Fig. 467, etc.) In each time row in Fig. 467 there are 4 internal mesh points. Hence in each time step we would have to solve 4 equations in 4 unknowns. But since the initial temperature distribution is symmetric with respect to $x = 0.5$, and $u = 0$ at both ends for all t, we have $u_{31} = u_{21}$, $u_{41} = u_{11}$ in the first time row and similarly for the other rows. This reduces each system to 2 equations in 2 unknowns. By (9), since $u_{31} = u_{21}$ and $u_{01} = 0$, for $j = 0$ these equations are

$$(i = 1) \qquad 4u_{11} - u_{21} = u_{00} + u_{20} = 0.951\,057$$

$$(i = 2) \qquad -u_{11} + 4u_{21} - u_{21} = u_{10} + u_{20} = 1.538\,842.$$

The solution is $u_{11} = 0.399\,274$, $u_{21} = 0.646\,039$. Similarly, for time row $j = 1$ we have the system

$$(i = 1) \qquad 4u_{12} - u_{22} = u_{01} + u_{21} = 0.646\,039$$

$$(i = 2) \qquad -u_{12} + 3u_{22} = u_{11} + u_{21} = 1.045\,313.$$

The solution is $u_{12} = 0.271\,221$, $u_{22} = 0.438\,844$, and so on. This gives the temperature distribution (Fig. 468):

t	$x = 0$	$x = 0.2$	$x = 0.4$	$x = 0.6$	$x = 0.8$	$x = 1$
0.00	0	0.588	0.951	0.951	0.588	0
0.04	0	0.399	0.646	0.646	0.399	0
0.08	0	0.271	0.439	0.439	0.271	0
0.12	0	0.184	0.298	0.298	0.184	0
0.16	0	0.125	0.202	0.202	0.125	0
0.20	0	0.085	0.138	0.138	0.085	0

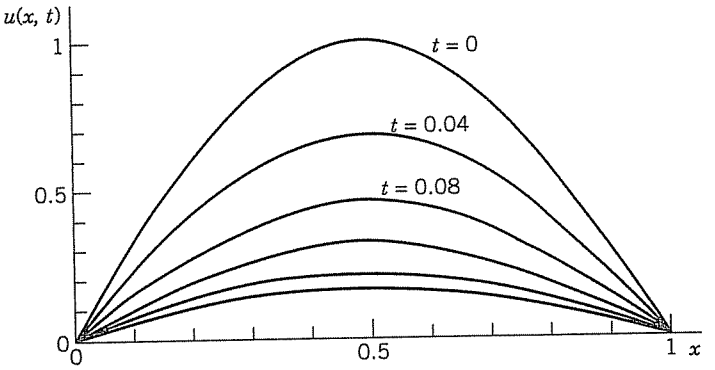


Fig. 468. Temperature distribution in the bar in Example 1

Comparison with the exact solution. The present problem can be solved exactly by separating variables (Sec. 12.5); the result is

$$u(x, t) = \sin \pi x \, e^{-\pi^2 t}. \tag{10}$$

Solution by the explicit method (5) with r = 0.25. For $h = 0.2$ and $r = k/h^2 = 0.25$ we have $k = rh^2 = 0.25 \cdot 0.04 = 0.01$. Hence we have to perform 4 times as many steps as with the Crank–Nicolson method! Formula (5) with $r = 0.25$ is

$$u_{i,j+1} = 0.25(u_{i-1,j} + 2u_{ij} + u_{i+1,j}). \tag{11}$$

We can again make use of the symmetry. For $j = 0$ we need $u_{00} = 0$, $u_{10} = 0.587\,785$ (see p. 925), $u_{20} = u_{30} = 0.951\,057$ and compute

$$u_{11} = 0.25(u_{00} + 2u_{10} + u_{20}) = 0.531\,657$$

$$u_{21} = 0.25(u_{10} + 2u_{20} + u_{30}) = 0.25(u_{10} + 3u_{20}) = 0.860\,239.$$

Of course we can omit the boundary terms $u_{01} = 0$, $u_{02} = 0, \cdots$ from the formulas. For $j = 1$ we compute

$$u_{12} = 0.25(2u_{11} + u_{21}) = 0.480\,888$$

$$u_{22} = 0.25(u_{11} + 3u_{21}) = 0.778\,094$$

and so on. We have to perform 20 steps instead of the 5 CN steps, but the numeric values show that the accuracy is only about the same as that of the Crank–Nicolson values CN. The exact 3D-values follow from (10).

t	$x = 0.2$			$x = 0.4$		
	CN	By (11)	Exact	CN	By (11)	Exact
0.04	0.399	0.393	0.396	0.646	0.637	0.641
0.08	0.271	0.263	0.267	0.439	0.426	0.432
0.12	0.184	0.176	0.180	0.298	0.285	0.291
0.16	0.125	0.118	0.121	0.202	0.191	0.196
0.20	0.085	0.079	0.082	0.138	0.128	0.132

Failure of (5) with r violating (6). Formula (5) with $h = 0.2$ and $r = 1$—which violates (6)—is

$$u_{i,j+1} = u_{i-1,j} - u_{ij} + u_{i+1,j}$$

and gives very poor values; some of these are

t	$x = 0.2$	Exact	$x = 0.4$	Exact
0.04	0.363	0.396	0.588	0.641
0.12	0.139	0.180	0.225	0.291
0.20	0.053	0.082	0.086	0.132

Formula (5) with an even larger $r = 2.5$ (and $h = 0.2$ as before) gives completely nonsensical results; some of these are

t	$x = 0.2$	Exact	$x = 0.4$	Exact
0.1	0.0265	0.2191	0.0429	0.3545
0.3	0.0001	0.0304	0.0001	0.0492.

PROBLEM SET 21.6

1. **(Nondimensional form)** Show that the heat equation $\tilde{u}_{\tilde{t}} = c^2 \tilde{u}_{\tilde{x}\tilde{x}}$, $0 \leqq \tilde{x} \leqq L$, can be transformed to the "nondimensional" standard form $u_t = u_{xx}$, $0 \leqq x \leqq 1$, by setting $x = \tilde{x}/L$, $t = c^2\tilde{t}/L^2$, $u = \tilde{u}/u_0$, where u_0 is any constant temperature.
2. Derive the difference approximation (4) of the heat equation.
3. Derive (5) from (4).
4. Using the explicit method [(5) with $h = 1$ and $k = 0.5$], find the temperature at $t = 2$ in a laterally insulated bar of length 10 with ends kept at temperature 0 and initial temperature $f(x) = x - 0.1x^2$.
5. Solve the heat problem (1)–(3) by Crank–Nicolson for $0 \leqq t \leqq 0.20$ with $h = 0.2$ and $k = 0.04$ when $f(x) = x$ if $0 \leqq x < \frac{1}{2}$, $f(x) = 1 - x$ if $\frac{1}{2} \leqq x \leqq 1$. Compare with the exact values for $t = 0.20$ obtained from the series (2 terms) in Sec. 12.5.
6. Solve Prob. 5 by the explicit method with $h = 0.2$ and $k = 0.01$. Do 8 steps. Compare the last values with the Crank–Nicolson 3S-values 0.107, 0.175 and the exact 3S-values 0.108, 0.175.
7. The accuracy of the explicit method depends on r ($\leqq \frac{1}{2}$). Illustrate this for Prob. 6, choosing $r = \frac{1}{2}$ (and $h = 0.2$ as before). Do 4 steps. Compare the values for $t = 0.04$ and 0.08 with the 3S-values in Prob. 6, which are 0.156, 0.254 ($t = 0.04$), 0.105, 0.170 ($t = 0.08$).
8. If the left end of a laterally insulated bar extending from $x = 0$ to $x = 1$ is insulated, the boundary condition at $x = 0$ is $u_n(0, t) = u_x(0, t) = 0$. Show that in the application of the explicit method given by (5), we can compute $u_{0,j+1}$ by the formula

$$u_{0,j+1} = (1 - 2r)u_{0j} + 2ru_{1j}.$$

Apply this with $h = 0.2$ and $r = 0.25$ to determine the temperature $u(x, t)$ in a laterally insulated bar extending from $x = 0$ to 1 if $u(x, 0) = 0$, the left end is insulated and the right end is kept at temperature $g(t) = \sin \frac{50}{3}\pi t$. *Hint.* Use $0 = \partial u_{0j}/\partial x = (u_{1j} - u_{-1,j})/2h$.

9. In a laterally insulated bar of length 1 let the initial temperature be $f(x) = x$ if $0 \leqq x \leqq 0.2$, $f(x) = 0.25(1 - x)$ if $0.2 \leqq x \leqq 1$. Let $u(0, t) = 0$, $u(1, t) = 0$ for all t. Apply the explicit method with $h = 0.2$, $k = 0.01$. Do 5 steps.
10. Solve Prob. 9 for $f(x) = x$ if $0 \leqq x \leqq 0.5$, $f(x) = 1 - x$ if $0.5 \leqq x \leqq 1$, all the other data being as before. Can you expect the solution to satisfy $u(x, t) = u(1 - x, t)$ for all t?
11. Solve Prob. 9 by (9) with $h = 0.2$, 2 steps. Compare with exact values obtained from the series in Sec. 12.5 (2 terms) with suitable coefficients.
12. **CAS EXPERIMENT. Comparison of Methods.**
    (a) Write programs for the explicit and the Crank–Nicolson methods.
    (b) Apply the programs to the heat problem of a laterally insulated bar of length 1 with $u(x, 0) = \sin \pi x$ and $u(0, t) = u(1, t) = 0$ for all t, using $h = 0.2$, $k = 0.01$ for the explicit method (20 steps), $h = 0.2$ and (9) for the Crank–Nicolson method (5 steps). Obtain exact 6D-values from a suitable series and compare.
    (c) Graph temperature curves in (b) in two figures similar to Fig. 296 in Sec. 12.6.
    (d) Experiment with smaller h (0.1, 0.05, etc.) for both methods to find out to what extent accuracy increases under systematic changes of h and k.

13–15 CRANK–NICOLSON

Solve (1)–(3) by Crank–Nicolson with $r = 1$ (5 steps), where:

13. $f(x) = x(1 - x)$, $h = 0.2$
14. $f(x) = x(1 - x)$, $h = 0.1$ (Compare with Prob. 13.)
15. $f(x) = 5x$ if $0 \leqq x < 0.2$, $f(x) = 1.25(1 - x)$ if $0.2 \leqq x \leqq 1$, $h = 0.2$

21.7 Method for Hyperbolic PDEs

In this section we consider the numeric solution of problems involving hyperbolic PDEs. We explain a standard method in terms of a typical setting for the prototype of a hyperbolic PDE, the **wave equation:**

(1) $$u_{tt} = u_{xx} \qquad 0 \leqq x \leqq 1,\ t \geqq 0$$

(2) $$u(x, 0) = f(x) \qquad \text{(Given initial displacement)}$$

(3) $$u_t(x, 0) = g(x) \qquad \text{(Given initial velocity)}$$

(4) $$u(0, t) = u(1, t) = 0 \qquad \text{(Boundary conditions).}$$

Note that an equation $u_{tt} = c^2 u_{xx}$ and another x-interval can be reduced to the form (1) by a linear transformation of x and t. This is similar to Sec. 21.6, Prob. 1.

For instance, (1)–(4) is the model of a vibrating elastic string with fixed ends at $x = 0$ and $x = 1$ (see Sec. 12.2). Although an analytic solution of the problem is given in (13), Sec. 12.4, we use the problem for explaining basic ideas of the numeric approach that are also relevant for more complicated hyperbolic PDEs.

Replacing the derivatives by difference quotients as before, we obtain from (1) [see (6) in Sec. 21.4 with $y = t$]

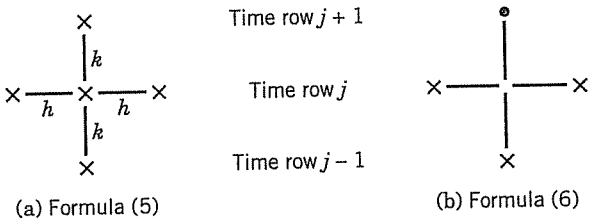
(5) $$\frac{1}{k^2}\,(u_{i,j+1} - 2u_{ij} + u_{i,j-1}) = \frac{1}{h^2}\,(u_{i+1,j} - 2u_{ij} + u_{i-1,j})$$

where h is the mesh size in x, and k is the mesh size in t. This difference equation relates 5 points as shown in Fig. 469a. It suggests a rectangular grid similar to the grids for parabolic equations in the preceding section. We choose $r^* = k^2/h^2 = 1$. Then u_{ij} drops out and we have

(6) $$u_{i,j+1} = u_{i-1,j} + u_{i+1,j} - u_{i,j-1} \qquad \text{(Fig. 469b).}$$

It can be shown that for $0 < r^* \leqq 1$ the present **explicit method** is stable, so that from (6) we may expect reasonable results for initial data that have no discontinuities. (For a hyperbolic PDE the latter would propagate into the solution domain—a phenomenon that would be difficult to deal with on our present grid. For unconditionally stable **implicit methods** see [E1] in App. 1.)

Equation (6) still involves 3 time steps $j - 1, j, j + 1$, whereas the formulas in the parabolic case involved only 2 time steps. Furthermore, we now have 2 initial conditions.



(a) Formula (5) (b) Formula (6)

Fig. 469. Mesh points used in (5) and (6)

So we ask how we get started and how we can use the initial condition (3). This can be done as follows.

From $u_t(x, 0) = g(x)$ we derive the difference formula

$$(7) \qquad \frac{1}{2k}(u_{i1} - u_{i,-1}) = g_i, \qquad \text{hence} \qquad u_{i,-1} = u_{i1} - 2kg_i$$

where $g_i = g(ih)$. For $t = 0$, that is, $j = 0$, equation (6) is

$$u_{i1} = u_{i-1,0} + u_{i+1,0} - u_{i,-1}.$$

Into this we substitute $u_{i,-1}$ as given in (7). We obtain $u_{i1} = u_{i-1,0} + u_{i+1,0} - u_{i1} + 2kg_i$ and by simplification

$$(8) \qquad u_{i1} = \tfrac{1}{2}(u_{i-1,0} + u_{i+1,0}) + kg_i.$$

This expresses u_{i1} in terms of the initial data. It is for the beginning only. Then use (6).

EXAMPLE 1 **Vibrating String, Wave Equation**

Apply the present method with $h = k = 0.2$ to the problem (1)–(4), where

$$f(x) = \sin \pi x, \qquad g(x) = 0.$$

Solution. The grid is the same as in Fig. 467, Sec. 21.6, except for the values of t, which now are 0.2, 0.4, $\cdots$ (instead of 0.04, 0.08, $\cdots$). The initial values $u_{00}, u_{10}, \cdots$ are the same as in Example 1, Sec. 21.6. From (8) and $g(x) = 0$ we have

$$u_{i1} = \tfrac{1}{2}(u_{i-1,0} + u_{i+1,0}).$$

From this we compute, using $u_{10} = u_{40} = \sin 0.2\pi = 0.587\,785$, $u_{20} = u_{30} = 0.951\,057$,

$$(i = 1) \quad u_{11} = \tfrac{1}{2}(u_{00} + u_{20}) = \tfrac{1}{2} \cdot 0.951\,057 = 0.475\,528$$

$$(i = 2) \quad u_{21} = \tfrac{1}{2}(u_{10} + u_{30}) = \tfrac{1}{2} \cdot 1.538\,842 = 0.769\,421$$

and $u_{31} = u_{21}$, $u_{41} = u_{11}$ by symmetry as in Sec. 21.6, Example 1. From (6) with $j = 1$ we now compute, using $u_{01} = u_{02} = \cdots = 0$,

$$(i = 1) \quad u_{12} = u_{01} + u_{21} - u_{10} = 0.769\,421 - 0.587\,785 = 0.181\,636$$

$$(i = 2) \quad u_{22} = u_{11} + u_{31} - u_{20} = 0.475\,528 + 0.769\,421 - 0.951\,057 = 0.293\,892,$$

and $u_{32} = u_{22}$, $u_{42} = u_{12}$ by symmetry; and so on. We thus obtain the following values of the displacement $u(x, t)$ of the string over the first half-cycle:

t	$x = 0$	$x = 0.2$	$x = 0.4$	$x = 0.6$	$x = 0.8$	$x = 1$
0.0	0	0.588	0.951	0.951	0.588	0
0.2	0	0.476	0.769	0.769	0.476	0
0.4	0	0.182	0.294	0.294	0.182	0
0.6	0	−0.182	−0.294	−0.294	−0.182	0
0.8	0	−0.476	−0.769	−0.769	−0.476	0
1.0	0	−0.588	−0.951	−0.951	−0.588	0

These values are exact to 3D (3 decimals), the exact solution of the problem being (see Sec. 12.3)

$$u(x, t) = \sin \pi x \cos \pi t.$$

The reason for the exactness follows from d'Alembert's solution (4), Sec. 12.4. (See Prob. 4, below.) ■

This is the end of Chap. 21 on numerics for ODEs and PDEs, a rapidly developing field of basic applications and interesting research, in which large-scale and complicated practical problems can now be attacked and solved by the computer.

PROBLEM SET 21.7

1–3 VIBRATING STRING

Solve (1)–(4) by the present method with $h = k = 0.2$ for the given initial deflection $f(x)$ and initial velocity 0 on the given t-interval.

1. $f(x) = 0.01x(1 - x)$, $0 \leqq t \leqq 2$
2. $f(x) = x^2(1 - x)$, $0 \leqq t \leqq 1$
3. $f(x) = x$ if $0 \leqq x \leqq 0.2$, $f(x) = 0.25(1 - x)$ if $0.2 < x \leqq 1$
4. Show that from d'Alembert's solution (13) in Sec. 12.4 with $c = 1$ it follows that (6) in the present section gives the exact value $u_{i,j+1} = u(ih, (j + 1)h)$.
5. If the string governed by the wave equation (1) starts from its equilibrium position with initial velocity $g(x) = \sin \pi x$, what is its displacement at time $t = 0.4$ and $x = 0.2, 0.4, 0.6, 0.8$? (Use the present method with $h = 0.2$, $k = 0.2$. Use (8). Compare with the exact values obtained from (12) in Sec. 12.4.)
6. Compute approximate values in Prob. 5, using a finer grid ($h = 0.1$, $k = 0.1$), and notice the increase in accuracy.
7. Illustrate the starting procedure when both f and g are not identically zero, say, $f(x) = 1 - \cos 2\pi x$, $g(x) = x - x^2$. Choose $h = k = 0.1$ and do 2 time steps.
8. Show that (12) in Sec. 12.4 gives as another starting formula

$$u_{i1} = \frac{1}{2}(u_{i+1,0} + u_{i-1,0}) + \frac{1}{2}\int_{x_i-k}^{x_i+k} g(s)\, ds$$

(where one can evaluate the integral numerically if necessary). In what case is this identical with (8)?

9. Compute u in Prob. 7 for $t = 0.1$ and $x = 0.1, 0.2, \cdots, 0.9$, using the formula in Prob. 8, and compare the values.
10. Solve (1)–(3) ($h = k = 0.2$, 5 time steps) subject to $f(x) = x^2$, $g(x) = 2x$, $u_x(0, t) = 2t$, $u(1, t) = (1 + t)^2$.

CHAPTER 21 REVIEW QUESTIONS AND PROBLEMS

1. Explain the Euler and Improved Euler methods in geometrical terms.
2. What are the local and global orders of a method? Give examples.
3. What do you know about error estimates? Why are they important?
4. How did we obtain numeric methods by using the Taylor series?
5. In each Runge–Kutta step we computed auxiliary values. How many? Why?
6. What are one-step and multistep methods? Give examples.
7. What is the idea of a predictor–corrector method? Mention some of these methods.
8. What is the idea of the Runge–Kutta–Fehlberg method?
9. How can Runge–Kutta be generalized to systems of ODEs?
10. What is automatic step size control? How is it done in practice?
11. Why and how did we use finite differences in this chapter?
12. Make a list of types of PDEs, corresponding problems, and methods for their numeric solution.
13. How did we approximate the Laplace equation? The Poisson equation?
14. Will a difference equation give exact solutions of a PDE?
15. How did we handle (a) irregularly shaped domains, (b) given normal derivatives at the boundary?
16. Solve $y' = 2xy$, $y(0) = 1$, by the Euler method with $h = 0.1$, 10 steps. Compute the error.
17. Solve $y' = 1 + y^2$, $y(0) = 0$, by the improved Euler method with $h = 0.1$, 5 steps. Compute the error.
18. Solve $y' = (x + y - 4)^2$, $y(0) = 4$, by RK with $h = 0.2$, 7 steps.

19. Solve Prob. 17 by RK with $h = 0.1$, 5 steps. Compute the error. Compare with Prob. 17.

20. **(Fair comparison)** Solve $y' = 2x^{-1}\sqrt{y - \ln x} + x^{-1}$, $y(1) = 0$ for $1 \leqq x \leqq 1.8$ (a) by the Euler method with $h = 0.1$, (b) by the improved Euler method with $h = 0.2$, (c) by RK with $h = 0.4$. Verify that the exact solution is $y = (\ln x)^2 + \ln x$. Compute and compare the errors. Why is the comparison fair?

21. Compute e^x for $x = 0, 0.1, \cdots, 1.0$ by applying RK to $y' = y$, $y(0) = 1$, $h = 0.1$. Show that the result is 5D-exact.

22. Solve $y' = (x + y)^2$, $y(0) = 0$ by RK with $h = 0.2$, 5 steps.

23. Show that by applying the method in Sec. 21.2 to a polynomial of first degree we obtain the multistep predictor and corrector formulas

$$y^*_{n+1} = y_n + \frac{h}{2}(3f_n - f_{n-1})$$

$$y_{n+1} = y_n + \frac{h}{2}(f^*_{n+1} + f_n)$$

where $f^*_{n+1} = f(x_{n+1}, y^*_{n+1})$.

24. Apply the multistep method in Prob. 23 to the initial value problem $y' = x + y$, $y(0) = 0$, choosing $h = 0.2$ and doing 5 steps. Compare with the exact values.

25. Solve $y' = (y - x - 1)^2 + 2$, $y(0) = 1$ for $0 \leqq x \leqq 1$ by Adams–Moulton with $h = 0.1$ and starting values 1, 1.200334589, 1.402709878, 1.609336039.

26. Solve $y'' + y = 0$, $y(0) = 0$, $y'(0) = 1$ by RKN with $h = 0.2$, 5 steps. Find the error.

27. Solve $y_1' = -4y_1 + 3y_2$, $y_2' = 5y_1 - 6y_2$, $y_1(0) = 3$, $y_2(0) = -5$, by RK for systems, $h = 0.1$, 5 steps.

28. Solve $y_1' = -5y_1 + 3y_2$, $y_2' = -3y_1 - 5y_2$, $y_1(0) = 2$, $y_2(0) = 2$ by RK for systems, $h = 0.1$, 5 steps.

29. Find rough approximate values of the electrostatic potential at P_{11}, P_{12}, P_{13} in Fig. 470 that lie in a field between conducting plates (in Fig. 470 appearing as sides of a rectangle) kept at potentials 0 and 110 volts as shown. (Use the indicated grid.)

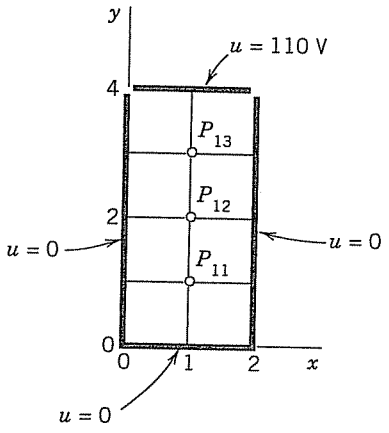


Fig. 470. Problem 29

30–32 POTENTIAL

Find the potential in Fig. 471, using the given grid and the boundary values:

30. $u = 70$ on the upper and left sides, $u = 0$ on the lower and right sides

31. $u(P_{10}) = u(P_{30}) = 960$, $u(P_{20}) = -480$, $u = 0$ elsewhere on the boundary

32. $u(P_{01}) = u(P_{03}) = u(P_{41}) = u(P_{43}) = 200$, $u(P_{10}) = u(P_{30}) = -400$, $u(P_{20}) = 1600$, $u(P_{02}) = u(P_{42}) = u(P_{14}) = u(P_{24}) = u(P_{34}) = 0$

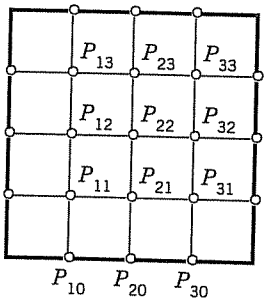


Fig. 471. Problems 30–32

33. Verify (13) in Sec. 21.4 for the system (12) and show that **A** in (12) is nonsingular.

34. Derive the difference approximation of the heat equation.

35. Solve the heat equation (1), Sec. 21.6, for the initial condition $f(x) = x$ if $0 \leqq x \leqq 0.2$, $f(x) = 0.25(1 - x)$ if $0.2 < x \leqq 1$ and boundary condition (3), Sec. 21.6, by the explicit method [formula (5) in Sec. 21.6] with $h = 0.2$ and $k = 0.01$ so that you get values of the temperature at time $t = 0.05$ as the answer.

36. A laterally insulated homogeneous bar with ends at $x = 0$ and $x = 1$ has initial temperature 0. Its left end is kept at 0, whereas the temperature at the right end varies sinusoidally according to

$$u(t, 1) = g(t) = \sin \tfrac{25}{3}\pi t.$$

Find the temperature $u(x, t)$ in the bar [solution of (1) in Sec. 21.6] by the explicit method with $h = 0.2$ and $r = 0.5$ (one period, that is, $0 \leqq t \leqq 0.24$).

37. Find $u(x, 0.12)$ and $u(x, 0.24)$ in Prob. 36 if the left end of the bar is kept at $-g(t)$ (instead of 0), all the other data being as before.

38. Find out how the results of Prob. 36 can be used for obtaining the results in Prob. 37. Use the values 0.054, 0.172, 0.325, 0.406 ($t = 0.12$, $x = 0.2, 0.4, 0.6, 0.8$) and -0.009, -0.086, -0.252, -0.353 ($t = 0.24$) from the answer to Prob. 36 to check your answer to Prob. 37.

39. Solve $u_t = u_{xx}$ ($0 \leqq x \leqq 1$, $t \geqq 0$), $u(x, 0) = x^2(1 - x)$, $u(0, t) = u(1, t) = 0$ by Crank–Nicolson with $h = 0.2$, $k = 0.04$, 5 time steps.

40. Find the solution of the vibrating string problem $u_{tt} = u_{xx}$, $u(x, 0) = x(1 - x)$, $u_t = 0$, $u(0, t) = u(1, t) = 0$ by the method in Sec. 21.7 with $h = 0.1$ and $k = 0.1$ for $t = 0.3$.

SUMMARY OF CHAPTER 21
Numerics for ODEs and PDEs

In this chapter we discussed numerics for ODEs (Secs. 21.1–21.3) and PDEs (Secs. 21.4–21.7). Methods for initial value problems

$$y' = f(x, y), \qquad y(x_0) = y_0 \tag{1}$$

involving a first-order ODE are obtained by truncating the Taylor series

$$y(x + h) = y(x) + hy'(x) + \frac{h^2}{2} y''(x) + \cdots$$

where, by (1), $y' = f$, $y'' = f' = \partial f/\partial x + (\partial f/\partial y)y'$, etc. Truncating after the term hy', we get the *Euler method,* in which we compute step by step

$$y_{n+1} = y_n + hf(x_n, y_n) \qquad (n = 0, 1, \cdots). \tag{2}$$

Taking one more term into account, we obtain the *improved Euler method.* Both methods show the basic idea but are too inaccurate in most cases.

Truncating after the term in h^4, we get the important classical **Runge–Kutta (RK) method** of fourth order. The crucial idea in this method is the replacement of the cumbersome evaluation of derivatives by the evaluation of $f(x, y)$ at suitable points (x, y); thus in each step we first compute four auxiliary quantities (Sec. 21.1)

$$\begin{aligned} k_1 &= hf(x_n, y_n) \\ k_2 &= hf(x_n + \tfrac{1}{2}h, y_n + \tfrac{1}{2}k_1) \\ k_3 &= hf(x_n + \tfrac{1}{2}h, y_n + \tfrac{1}{2}k_2) \\ k_4 &= hf(x_n + h, y_n + k_3) \end{aligned} \tag{3a}$$

and then the new value

$$y_{n+1} = y_n + \tfrac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4). \tag{3b}$$

Error and step size control are possible by step halving or by **RKF** (Runge–Kutta–Fehlberg).

The methods in Sec. 21.1 are **one-step methods** since they get y_{n+1} from the result y_n of a single step. A **multistep method** (Sec. 21.2) uses the values of $y_n, y_{n-1}, \cdots$ of several steps for computing y_{n+1}. Integrating cubic interpolation polynomials gives the **Adams–Bashforth predictor** (Sec. 21.2)

$$y^*_{n+1} = y_n + \frac{1}{24} h(55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3}) \tag{4a}$$

where $f_j = f(x_j, y_j)$, and an **Adams–Moulton corrector** (the actual new value)

$$(4b) \qquad y_{n+1} = y_n + \frac{1}{24} h(9f^*_{n+1} + 19f_n - 5f_{n-1} + f_{n-2}),$$

where $f^*_{n+1} = f(x_{n+1}, y^*_{n+1})$. Here, to get started, y_1, y_2, y_3 must be computed by the Runge–Kutta method or by some other accurate method.

Section 19.3 concerned the extension of Euler and RK methods to systems

$$\mathbf{y}' = \mathbf{f}(x, \mathbf{y}), \qquad \text{thus} \qquad y_j' = f_j(x, y_1, \cdots, y_m), \qquad j = 1, \cdots, m.$$

This includes single mth order ODEs, which are reduced to systems. Second-order equations can also be solved by **RKN** (Runge–Kutta–Nyström) **methods.** These are particularly advantageous for $y'' = f(x, y)$ with f not containing y'.

Numeric methods for PDEs are obtained by replacing partial derivatives by difference quotients. This leads to approximating difference equations, for the **Laplace equation** to

$$(5) \qquad u_{i+1,j} + u_{i,j+1} + u_{i-1,j} + u_{i,j-1} - 4u_{ij} = 0 \qquad \text{(Sec. 21.4)}$$

for the **heat equation** to

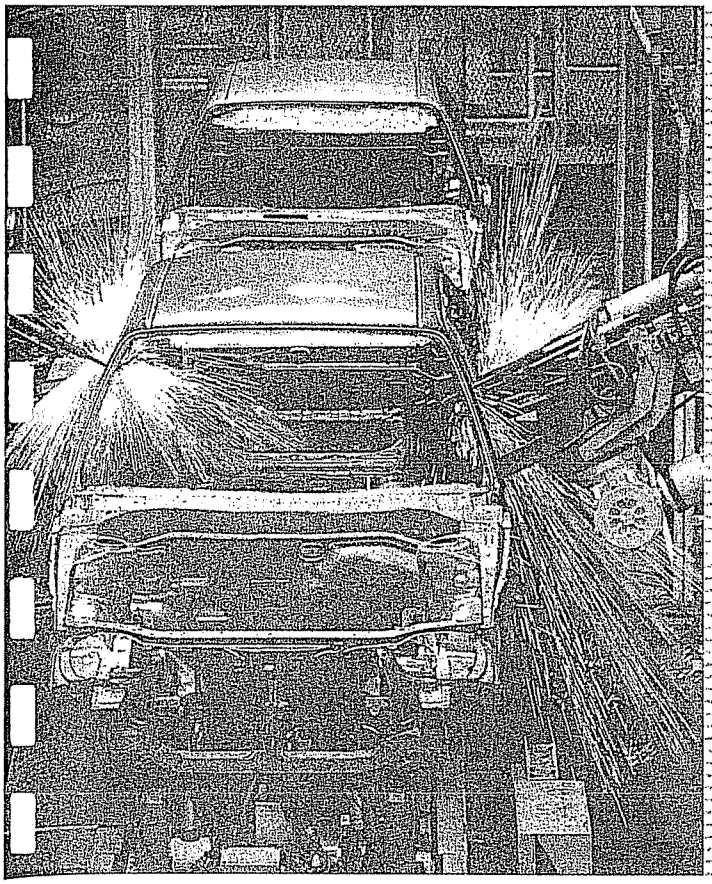
$$(6) \qquad \frac{1}{k}(u_{i,j+1} - u_{ij}) = \frac{1}{h^2}(u_{i+1,j} - 2u_{ij} + u_{i-1,j}) \qquad \text{(Sec. 21.6)}$$

and for the **wave equation** to

$$(7) \qquad \frac{1}{k^2}(u_{i,j+1} - 2u_{ij} + u_{i,j-1}) = \frac{1}{h^2}(u_{i+1,j} - 2u_{ij} + u_{i-1,j}) \qquad \text{(Sec. 21.7);}$$

here h and k are the mesh sizes of a grid in the x- and y-directions, respectively, where in (6) and (7) the variable y is time t.

These PDEs are *elliptic, parabolic,* and *hyperbolic,* respectively. Corresponding numeric methods differ, for the following reason. For elliptic PDEs we have boundary value problems, and we discussed for them the **Gauss–Seidel method** (also known as **Liebmann's method**) and the **ADI method** (Secs. 21.4, 21.5). For parabolic PDEs we are given one initial condition and boundary conditions, and we discussed an *explicit method* and the **Crank–Nicolson method** (Sec. 21.6). For hyperbolic PDEs, the problems are similar but we are given a second initial condition (Sec. 21.7).



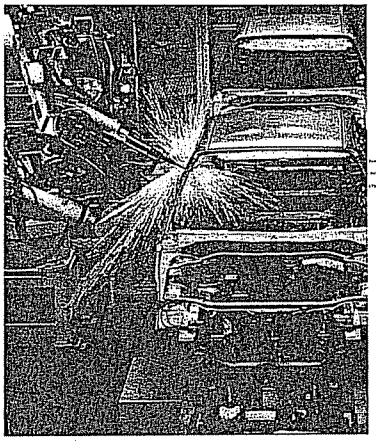
PART F

Optimization, Graphs

CHAPTER 22 **Unconstrained Optimization. Linear Programming**

CHAPTER 23 **Graphs. Combinatorial Optimization**

Ideas of optimization and application of graphs play an increasing role in engineering, computer science, systems theory, economics, and other areas. In the first chapter of this part we explain some basic concepts, methods, and results in unconstrained and constrained optimization. The second chapter is devoted to graphs and the corresponding so-called combinatorial optimization, a relatively new interesting area of ongoing applied and theoretical research.



CHAPTER 22

Unconstrained Optimization. Linear Programming

Optimization principles are of basic importance in modern engineering design and systems operation in various areas. The recent development has been influenced by computers capable of solving large-scale problems and by the creation of corresponding new optimization techniques, so that the entire field has become a large area of its own.

In the present chapter we give an introduction to the more important concepts, methods, and results on unconstrained optimization (the so-called **gradient method**) and constrained optimization (**linear programming**).

Prerequisite: a modest working knowledge of linear systems of equations
References and Answers to Problems: App. 1 Part F, App. 2.

22.1 Basic Concepts. Unconstrained Optimization

In an **optimization problem** the objective is to *optimize* (*maximize* or *minimize*) some function f. This function f is called the **objective function.**

For example, an objective function f to be *maximized* may be the revenue in a production of TV sets, the yield per minute in a chemical process, the mileage per gallon of a certain type of car, the hourly number of customers served in a bank, the hardness of steel, or the tensile strength of a rope.

Similarly, we may want to *minimize* f if f is the cost per unit of producing certain cameras, the operating cost of some power plant, the daily loss of heat in a heating system, the idling time of some lathe, or the time needed to produce a fender.

In most optimization problems the objective function f depends on several variables

$$x_1 \cdots, x_n.$$

These are called **control variables** because we can "control" them, that is, choose their values.

For example, the yield of a chemical process may depend on pressure x_1 and temperature x_2. The efficiency of a certain air-conditioning system may depend on temperature x_1, air pressure x_2, moisture content x_3, cross-sectional area of outlet x_4, and so on.

Optimization theory develops methods for optimal choices of $x_1, \cdots, x_n$, which maximize (or minimize) the objective function f, that is, methods for finding optimal values of $x_1, \cdots, x_n$.

In many problems the choice of values of $x_1, \cdots, x_n$ is not entirely free but is subject to some **constraints,** that is, additional restrictions arising from the nature of the problem and the variables.

For example, if x_1 is production cost, then $x_1 \geqq 0$, and there are many other variables (time, weight, distance traveled by a salesman, etc.) that can take nonnegative values only. Constraints can also have the form of equations (instead of inequalities).

We first consider **unconstrained optimization** in the case of a function $f(x_1, \cdots, x_n)$. We also write $\mathbf{x} = (x_1, \cdots, x_n)$ and $f(\mathbf{x})$, for convenience.

By definition, f has a **minimum** at a point $\mathbf{x} = \mathbf{X}_0$ in a region R (where f is defined) if

$$f(\mathbf{x}) \geqq f(\mathbf{X}_0)$$

for all $\mathbf{x}$ in R. Similarly, f has a **maximum** at $\mathbf{X}_0$ in R if

$$f(\mathbf{x}) \leqq f(\mathbf{X}_0)$$

for all $\mathbf{x}$ in R. Minima and maxima together are called **extrema.**

Furthermore, f is said to have a **local minimum** at $\mathbf{X}_0$ if

$$f(\mathbf{x}) \geqq f(\mathbf{X}_0)$$

for all $\mathbf{x}$ in a neighborhood of $\mathbf{X}_0$, say, for all $\mathbf{x}$ satisfying

$$|\mathbf{x} - \mathbf{X}_0| = [(x_1 - X_1)^2 + \cdots + (x_n - X_n)^2]^{1/2} < r,$$

where $\mathbf{X}_0 = (X_1, \cdots, X_n)$ and $r > 0$ is sufficiently small.

Similarly, f has a **local maximum** at $\mathbf{X}_0$ if $f(\mathbf{x}) \leqq f(\mathbf{X}_0)$ for all $\mathbf{x}$ satisfying $|\mathbf{x} - \mathbf{X}_0| < r$.

If f is differentiable and has an extremum at a point $\mathbf{X}_0$ in the *interior of a region* R (that is, not on the boundary), then the partial derivatives $\partial f/\partial x_1, \cdots, \partial f/\partial x_n$ must be zero at $\mathbf{X}_0$. These are the components of a vector that is called the **gradient** of f and denoted by grad f or ∇f. (For $n = 3$ this agrees with Sec. 9.7.) Thus

$$\textbf{(1)} \qquad \nabla f(\mathbf{X}_0) = \mathbf{0}.$$

A point $\mathbf{X}_0$ at which (1) holds is called a **stationary point** of f.

Condition (1) is necessary for an extremum of f at $\mathbf{X}_0$ in the interior of R, but is not sufficient. Indeed, if $n = 1$, then for $y = f(x)$, condition (1) is $y' = f'(X_0) = 0$; and, for instance, $y = x^3$ satisfies $y' = 3x^2 = 0$ at $x = X_0 = 0$ where f has no extremum but a point of inflection. Similarly, for $f(\mathbf{x}) = x_1x_2$ we have $\nabla f(\mathbf{0}) = \mathbf{0}$, and f does not have an extremum but has a saddle point at $\mathbf{0}$. Hence after solving (1), one must still find out whether one has obtained an extremum. In the case $n = 1$ the conditions $y'(X_0) = 0$, $y''(X_0) > 0$ guarantee a local minimum at X_0 and the conditions $y'(X_0) = 0$, $y''(X_0) < 0$ a local maximum, as is known from calculus. For $n > 1$ there exist similar criteria. However, in practice even solving (1) will often be difficult. For this reason, one generally prefers solution by iteration, that is, by a search process that starts at some point and moves stepwise to points at which f is smaller (if a minimum of f is wanted) or larger (in the case of a maximum).

The **method of steepest descent** or **gradient method** is of this type. We present it here in its standard form. (For refinements see Ref. [E25] listed in App. 1.)

The idea of this method is to find a minimum of $f(\mathbf{x})$ by repeatedly computing minima of a function $g(t)$ of a single variable t, as follows. Suppose that f has a minimum at $\mathbf{X}_0$ and we start at a point $\mathbf{x}$. Then we look for a minimum of f closest to $\mathbf{x}$ along the straight line in the direction of $-\nabla f(\mathbf{x})$, which is the direction of steepest descent (= direction of maximum decrease) of f at $\mathbf{x}$. That is, we determine the value of t and the corresponding point

$$\mathbf{z}(t) = \mathbf{x} - t\nabla f(\mathbf{x}) \tag{2}$$

at which the function

$$g(t) = f(\mathbf{z}(t)) \tag{3}$$

has a minimum. We take this $\mathbf{z}(t)$ as our next approximation to $\mathbf{X}_0$.

EXAMPLE 1 Method of Steepest Descent

Determine a minimum of

$$f(\mathbf{x}) = x_1^2 + 3x_2^2, \tag{4}$$

starting from $\mathbf{x}_0 = (6, 3) = 6\mathbf{i} + 3\mathbf{j}$ and applying the method of steepest descent.

Solution. Clearly, inspection shows that $f(\mathbf{x})$ has a minimum at $\mathbf{0}$. Knowing the solution gives us a better feel of how the method works. We obtain $\nabla f(\mathbf{x}) = 2x_1\mathbf{i} + 6x_2\mathbf{j}$ and from this

$$\mathbf{z}(t) = \mathbf{x} - t\nabla f(\mathbf{x}) = (1 - 2t)x_1\mathbf{i} + (1 - 6t)x_2\mathbf{j}$$

$$g(t) = f(\mathbf{z}(t)) = (1 - 2t)^2x_1^2 + 3(1 - 6t)^2x_2^2.$$

We now calculate the derivative

$$g'(t) = 2(1 - 2t)x_1^2(-2) + 6(1 - 6t)x_2^2(-6),$$

set $g'(t) = 0$, and solve for t, finding

$$t = \frac{x_1^2 + 9x_2^2}{2x_1^2 + 54x_2^2}.$$

Starting from $\mathbf{x}_0 = 6\mathbf{i} + 3\mathbf{j}$, we compute the values in Table 22.1, which are shown in Fig. 472.

Figure 472 suggests that in the case of slimmer ellipses ("a long narrow valley"), convergence would be poor. You may confirm this by replacing the coefficient 3 in (4) with a large coefficient. For more sophisticated descent and other methods, some of them also applicable to vector functions of vector variables, we refer to the references listed in Part F of App. 1; see also [E25]. ▩

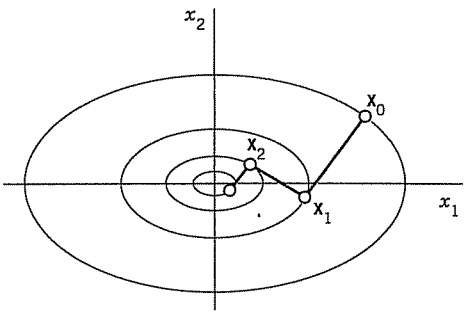


Fig. 472. Method of steepest descent in Example 1

Table 22.1 Method of Steepest Descent, Computations in Example 1

n	$\mathbf{x}$		t	$1 - 2t$	$1 - 6t$
0	6.000	3.000	0.210	0.581	−0.258
1	3.484	−0.774	0.310	0.381	−0.857
2	1.327	0.664	0.210	0.581	−0.258
3	0.771	−0.171	0.310	0.381	−0.857
4	0.294	0.147	0.210	0.581	−0.258
5	0.170	−0.038	0.310	0.381	−0.857
6	0.065	0.032			

PROBLEM SET 22.2

1. What happens if you apply the method of steepest descent to $f(x) = x_1^2 + x_2^2$?
2. Verify that in Example 1, successive gradients are orthogonal. What is the reason?

3–11 STEEPEST DESCENT

Do 3 steepest descent steps when:

3. $f(x) = 3x_1^2 + 2x_2^2 - 12x_1 + 16x_2$, $\mathbf{x}_0 = [1 \quad 1]^T$
4. $f(x) = x_1^2 + 2x_2^2 - x_1 - 6x_2$, $\mathbf{x}_0 = [0 \quad 0]^T$
5. $f(\mathbf{x}) = 0.5x_1^2 + 0.7x_2^2 - x_1 + 4.2x_2 + 1$, $x_0 = [-1 \quad 1]^T$
6. $f(\mathbf{x}) = x_1^2 + 0.1x_2^2 + 8x_1 + x_2 + 22.5$, $x_0 = [2 \quad -1]^T$
7. $f(\mathbf{x}) = 0.2x_1^2 + x_2^2 - 0.08x_1$, $x_0 = [4 \quad 4]^T$
8. $f(\mathbf{x}) = x_1^2 - x_2^2$, $\mathbf{x}_0 = [2 \quad 1]^T$, 5 steps. First guess. Then compute. Sketch your path.
9. $f(x) = x_1^2 + cx_2^2$, $\mathbf{x}_0 = [c \quad 1]^T$. Show that 2 steps give $[c \quad 1]^T$ times a factor, $-4c^2/(c^2 - 1)^2$. What can you conclude from this about the speed of convergence?
10. $f(\mathbf{x}) = x_1^2 - x_2$, $\mathbf{x}_0 = [1 \quad 1]^T$. Sketch your path. Predict the outcome of further steps.
11. $f(\mathbf{x}) = ax_1 + bx_2$, any $\mathbf{x}_0$. First guess, then compute.
12. **CAS EXPERIMENT. Steepest Descent. (a)** Write a program for the method.

    (b) Apply your program to $f(\mathbf{x}) = x_1^2 + 4x_2^2$, experimenting with respect to speed of convergence depending on the choice of $\mathbf{x}_0$.

    (c) Apply your program to $f(\mathbf{x}) = x_1^2 + x_2^4$ and to $f(\mathbf{x}) = x_1^4 + x_2^4$, $\mathbf{x}_0 = [2 \quad 1]^T$. Graph level curves and your path of descent. (Try to include graphing directly in your program.)

22.2 Linear Programming

Linear programming or **linear optimization** consists of methods for solving optimization problems ***with constraints,*** that is, methods for finding a maximum (or a minimum) $\mathbf{x} = [x_1, \cdots, x_n]$ of a ***linear*** objective function

$$z = f(x) = a_1x_1 + a_2x_2 + \cdots + a_nx_n$$

satisfying the constraints. The latter are **linear inequalities,** such as $3x_1 + 4x_2 \leqq 36$, or $x_1 \geqq 0$, etc. (examples below). Problems of this kind arise frequently, almost daily, for instance, in production, inventory management, bond trading, operation of power plants, routing delivery vehicles, airplane scheduling, and so on. Progress in computer technology has made it possible to solve programming problems involving hundreds or thousands or more variables. Let us explain the setting of a linear programming problem and the idea of a "geometric" solution, so that we shall see what is going on.

EXAMPLE 1 Production Plan

Energy Savers, Inc., produces heaters of types S and L. The wholesale price is \$40 per heater for S and \$88 for L. Two time constraints result from the use of two machines M_1 and M_2. On M_1 one needs 2 min for an S heater and 8 min for an L heater. On M_2 one needs 5 min for an S heater and 2 min for an L heater. Determine production figures x_1 and x_2 for S and L, respectively (number of heaters produced per hour) so that the hourly revenue

$$z = f(x) = 40x_1 + 88x_2$$

is maximum.

Solution. Production figures x_1 and x_2 must be nonnegative. Hence the objective function (to be maximized) and the four constraints are

$$\text{(0)} \qquad z = 40x_1 + 88x_2$$

$$\text{(1)} \qquad 2x_1 + 8x_2 \leqq 60 \text{ min time on machine } M_1$$

$$\text{(2)} \qquad 5x_1 + 2x_2 \leqq 60 \text{ min time on machine } M_2$$

$$\text{(3)} \qquad x_1 \geqq 0$$

$$\text{(4)} \qquad x_2 \geqq 0.$$

Figure 473 shows (0)–(4) as follows. Constancy lines

$$z = const$$

are marked (0). These are **lines of constant revenue.** Their slope is $-40/88 = -5/11$. To increase z we must move the line upward (parallel to itself), as the arrow shows. Equation (1) with the equality sign is marked (1). It intersects the coordinate axes at $x_1 = 60/2 = 30$ (set $x_2 = 0$) and $x_2 = 60/8 = 7.5$ (set $x_1 = 0$). The arrow marks the side on which the points (x_1, x_2) lie that satisfy the inequality in (1). Similarly for Eqs. (2)–(4). The blue quadrangle thus obtained is called the **feasibility region.** It is the set of all **feasible solutions,** meaning solutions that satisfy all four constraints. The figure also lists the revenue at O, A, B, C. The optimal solution is obtained by moving the line of constant revenue up as much as possible without leaving the feasibility region completely. Obviously, this optimum is reached when that line passes through B, the intersection (10, 5) of (1) and (2). We see that the optimal revenue

$$z_{\max} = 40 \cdot 10 + 88 \cdot 5 = \$840$$

is obtained by producing twice as many S heaters as L heaters. ■

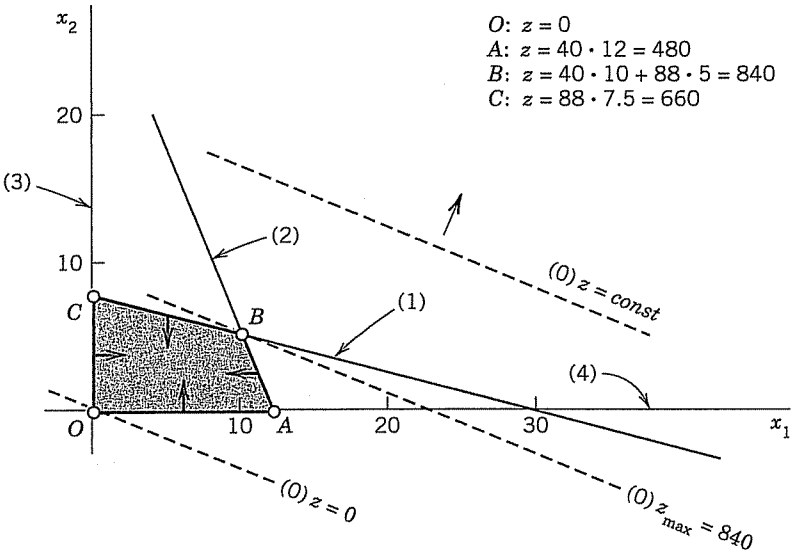


Fig. 473. Linear programming in Example 1

Note well that the problem in Example 1 or similar optimization problems *cannot* be solved by setting certain partial derivatives equal to zero, because crucial to such problems is the region in which the control variables are allowed to vary.

Furthermore, our "geometric" or graphic method illustrated in Example 1 is confined to two variables x_1, x_2. However, most practical problems involve much more than two variables, so that we need other methods of solution.

Normal Form of a Linear Programming Problem

To prepare for general solution methods, we show that constraints can be written more uniformly. Let us explain the idea in terms of (1),

$$2x_1 + 8x_2 \leqq 60.$$

This inequality implies $60 - 2x_1 - 8x_2 \geqq 0$ (and conversely), that is, the quantity

$$x_3 = 60 - 2x_1 - 8x_2$$

is nonnegative. Hence, our original inequality can now be written as an equation

$$2x_1 + 8x_2 + x_3 = 60,$$

where

$$x_3 \geqq 0.$$

x_3 is a nonnegative auxiliary variable introduced for converting inequalities to equations. Such a variable is called a **slack variable,** because it "takes up the slack" or difference between the two sides of the inequality.

EXAMPLE 2 Conversion of Inequalities by the Use of Slack Variables

With the help of two slack variables x_3, x_4 we can write the linear programming problem in Example 1 in the following form. ***Maximize***

$$f = 40x_1 + 88x_2$$

subject to the constraints

$$\begin{aligned} 2x_1 + 8x_2 + x_3 &= 60 \\ 5x_1 + 2x_2 + x_4 &= 60 \\ x_i &\geqq 0 \qquad (i = 1, \cdots, 4). \end{aligned}$$

We now have $n = 4$ variables and $m = 2$ (linearly independent) equations, so that two of the four variables, for example, x_1, x_2, determine the others. Also note that each of the four sides of the quadrangle in Fig. 473 now has an equation of the form $x_i = 0$:

$$\begin{aligned} &OA\text{:}\ x_2 = 0, \\ &AB\text{:}\ x_4 = 0, \\ &BC\text{:}\ x_3 = 0, \\ &CO\text{:}\ x_1 = 0. \end{aligned}$$

A vertex of the quadrangle is the intersection of two sides. Hence at a vertex, $n - m = 4 - 2 = 2$ of the variables are zero and the others are nonnegative. Thus at A we have $x_2 = 0$, $x_4 = 0$, and so on. ■

Our example suggests that a general linear optimization problem can be brought to the following **normal form.** *Maximize*

$$f = c_1x_1 + c_2x_2 + \cdots + c_nx_n \tag{5}$$

subject to the constraints

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + \cdots + a_{2n}x_n &= b_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\ a_{m1}x_1 + \cdots + a_{mn}x_n &= b_m \\ x_i \geqq 0 \qquad (i = 1, \cdots, n) & \end{aligned} \tag{6}$$

with all b_j nonnegative. (If a $b_j < 0$, multiply the equation by -1.) Here $x_1, \cdots, x_n$ include the slack variables (for which the c_j's in f are zero). We assume that the equations in (6) are linearly independent. Then, if we choose values for $n - m$ of the variables, the system uniquely determines the others. Of course, since we must have

$$x_1 \geqq 0, \cdots, x_n \geqq 0,$$

this choice is not entirely free.

Our problem also includes the **minimization** of an objective function f since this corresponds to maximizing $-f$ and thus needs no separate consideration.

An n-tuple $(x_1, \cdots, x_n)$ that satisfies all the constraints in (6) is called a *feasible point* or **feasible solution.** A feasible solution is called an **optimal solution** if for it the objective function f becomes maximum, compared with the values of f at all feasible solutions.

Finally, by a **basic feasible solution** we mean a feasible solution for which at least $n - m$ of the variables $x_1, \cdots, x_n$ are zero. For instance, in Example 2 we have $n = 4$, $m = 2$, and the basic feasible solutions are the four vertices O, A, B, C in Fig. 473. Here B is an optimal solution (the only one in this example).

The following theorem is fundamental.

THEOREM 1

Optimal Solution

Some optimal solution of a linear programming problem (5), (6) *is also a basic feasible solution of* (5), (6).

For a proof, see Ref. [F5], Chap. 3 (listed in App. 1). A problem can have many optimal solutions and not all of them may be *basic* feasible solutions; but the theorem guarantees that we can find an optimal solution by searching through the basic feasible solutions only. This is a great simplification; but since there are $\binom{n}{n-m} = \binom{n}{m}$ different ways of equating $n - m$ of the n variables to zero, considering all these possibilities, dropping those which are not feasible and then searching through the rest would still involve very much work, even when n and m are relatively small. Hence a systematic search is needed. We shall explain an important method of this type in the next section.

PROBLEM SET 22.2

1. What is the meaning of the slack variables x_3, x_4 in Example 2 in terms of the problem in Example 1?
2. Can we always expect a unique solution (as is the case in Example 1)?
3. Could we find a profit $f(x_1, x_2) = a_1x_1 + a_2x_2$ whose maximum is at an interior point of the quadrangle in Fig. 473? (Give a reason for your answer.)
4. Why are slack variables always nonnegative? How many of them do we need?

5–10 REGIONS AND CONSTRAINTS

Describe and graph the region in the first quadrant of the x_1x_2-plane determined by the inequalities:

5. $x_1 + 2x_2 \leqq 10$, $x_1 - x_2 \leqq 0$, $x_2 \geqq 2$

6. $-x_1 + x_2 \geqq 0$, $x_1 + x_2 \leqq 5$, $-2x_1 + x_2 \leqq 16$

7. $2.0x_1 + 6.0x_2 \leqq 18.0$, $5.0x_1 + 2.5x_2 \leqq 20.0$

8. $2x_1 - x_2 \geqq 6$, $4x_1 + 5x_2 \leqq 40$, $x_1 - 2x_2 \geqq -3$

9. $x_1 + x_2 \geqq 3$, $x_1 + x_2 \leqq 9$, $-x_1 + x_2 \geqq -3$, $-x_1 + x_2 \leqq 3$

10. $x_1 + x_2 \geqq 2$, $3x_1 + 5x_2 \geqq 15$, $2x_1 - x_2 \geqq -2$, $-x_1 + 2x_2 \leqq 10$

11–15 MAXIMIZATION AND MINIMIZATION

Maximize the given objective function f subject to the given constraints.

11. $f = -10x_1 + 2x_2$, $x_1 \geqq 0$, $x_2 \geqq 0$, $-x_1 + x_2 \geqq -1$, $x_1 + x_2 \leqq 6$, $x_2 \leqq 5$

12. $f = 3x_1 - 6x_2$, $4x_1 + x_2 \geqq 4$, $-x_1 + 2x_2 \geqq 6$, $x_1 + 2x_2 \leqq 14$

13. $f = 2x_1 + 3x_2$, $4x_1 + 3x_2 \geqq 12$, $x_1 - x_2 \geqq -3$, $x_2 \leqq 6$, $2x_1 - 3x_2 \leqq 0$

14. Minimize f in Prob. 13.

15. Minimize f in Prob. 11.

16. **(Maximum output)** Giant Ladders, Inc., wants to maximize its daily total output of large step ladders by producing x_1 of them by a process P_1 and x_2 by a process P_2, where P_1 requires 2 hours of labor and 4 machine hours per ladder, and P_2 requires 3 hours of labor and 2 machine hours. For this kind of work, 1200 hours of labor and 1600 hours on the machines are at most available per day. Find the optimal x_1 and x_2.

17. **(Maximum profit)** Universal Electric, Inc., manufactures and sells two models of lamps, L_1 and L_2, the profit being \$150 and \$100, respectively. The process involves two workers W_1 and W_2 who are available for this kind of work 100 and 80 hours per month, respectively. W_1 assembles L_1 in 20 min and L_2 in 30 min. W_2 paints L_1 in 20 min and L_2 in 10 min. Assuming that all lamps made can be sold without difficulty, determine production figures that maximize the profit.

18. **(Minimum cost)** Hardbrick, Inc., has two kilns. Kiln I can produce 3000 grey bricks, 2000 red bricks, and 300 glazed bricks daily. For Kiln II the corresponding figures are 2000, 5000, and 1500. Daily operating costs of Kilns I and II are \$400 and \$600, respectively. Find the number of days of operation of each kiln so that the operation cost in filling an order of 18000 grey, 34000 red, and 9000 glazed bricks is minimized.

19. **(Maximum profit)** United Metal, Inc., produces alloys B_1 (special brass) and B_2 (yellow tombac). B_1 contains 50% copper and 50% zinc. (Ordinary brass contains about 65% copper and 35% zinc.) B_2 contains 75% copper and 25% zinc. Net profits are \$120 per ton of B_1 and \$100 per ton of B_2. The daily copper supply is 45 tons. The daily zinc supply is 30 tons. Maximize the net profit of the daily production.

20. **(Nutrition)** Foods A and B have 600 and 500 calories, contain 15 g and 30 g of protein, and cost \$1.80 and \$2.10 per unit, respectively. Find the minimum cost diet of at least 3900 calories containing at least 150 g of protein.

22.3 Simplex Method

From the last section we recall the following. A linear optimization problem (linear programming problem) can be written in normal form; that is:

Maximize

(1) $$z = f(\mathbf{x}) = c_1x_1 + \cdots + c_nx_n$$

subject to the constraints

(2) $$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + \cdots + a_{2n}x_n &= b_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots &\cdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n &= b_m \\ x_i \geqq 0 \qquad (i = 1, \cdots, n)&. \end{aligned}$$

For finding an optimal solution of this problem, we need to consider only the **basic feasible solutions** (defined in Sec. 22.2), but there are still so many that we have to follow a systematic search procedure. In 1948 G. B. Dantzig published an iterative method, called the **simplex method,** for that purpose. In this method, one proceeds stepwise from one basic feasible solution to another in such a way that the objective function f always increases its value. Let us explain this method in terms of the example in the last section.

In its original form the problem concerned the maximization of the objective function

$$z = 40x_1 + 88x_2$$

subject to

$$\begin{aligned} 2x_1 + 8x_2 &\leqq 60 \\ 5x_1 + 2x_2 &\leqq 60 \\ x_1 \qquad &\geqq 0 \\ x_2 &\geqq 0. \end{aligned}$$

Converting the first two inequalities to equations by introducing two slack variables x_3, x_4, we obtained the **normal form** of the problem in Example 2. Together with the objective function (written as an equation $z - 40x_1 - 88x_2 = 0$) this normal form is

(3) $$\begin{aligned} z - 40x_1 - 88x_2 \qquad\qquad &= 0 \\ 2x_1 + 8x_2 + x_3 \qquad &= 60 \\ 5x_1 + 2x_2 \qquad + x_4 &= 60 \end{aligned}$$

where $x_1 \geqq 0, \cdots, x_4 \geqq 0$. This is a linear system of equations. To find an optimal solution of it, we may consider its **augmented matrix** (see Sec. 7.3)

(4) $$\begin{array}{c} \\ \mathbf{T}_0 = \end{array} \begin{array}{c} \begin{array}{cccccc} z & x_1 & x_2 & x_3 & x_4 & b \end{array} \\ \left[\begin{array}{c|cc|cc|c} 1 & -40 & -88 & 0 & 0 & 0 \\ \hline 0 & 2 & 8 & 1 & 0 & 60 \\ 0 & 5 & 2 & 0 & 1 & 60 \end{array}\right] \end{array}$$

This matrix is called a **simplex tableau** or **simplex table** (the ***initial** simplex table*). These are standard names. The dashed lines and the letters

$$z, \quad x_1, \quad \cdots, \quad b$$

are for ease in further manipulation.

Every simplex table contains two kinds of variables x_j. By **basic variables** we mean those whose columns have only one nonzero entry. Thus x_3, x_4 in (4) are basic variables and x_1, x_2 are **nonbasic variables.**

Every simplex table gives a basic feasible solution. It is obtained by setting the nonbasic variables to zero. Thus (4) gives the basic feasible solution

$$x_1 = 0, \qquad x_2 = 0, \qquad x_3 = 60/1 = 60, \qquad x_4 = 60/1 = 60, \qquad z = 0$$

with x_3 obtained from the second row and x_4 from the third.

The optimal solution (its location and value) is now obtained stepwise by pivoting, designed to take us to basic feasible solutions with higher and higher values of z until the maximum of z is reached. Here, the choice of the **pivot equation** and **pivot** are quite different from that in the Gauss elimination. The reason is that x_1, x_2, x_3, x_4 are restricted to nonnegative values.

***Step 1. Operation O_1*: Selection of the *Column* of the Pivot**
Select as the column of the pivot the first column with a negative entry in Row 1. In (4) this is Column 2 (because of the -40).

Operation O_2*: Selection of the *Row* of the Pivot.** Divide the right sides [60 and 60 in (4)] by the corresponding entries of the column just selected ($60/2 = 30$, $60/5 = 12$). Take as the pivot equation the equation that gives the ***smallest quotient. Thus the pivot is 5 because 60/5 is smallest.

***Operation O_3*: Elimination by Row Operations.** This gives zeros above and below the pivot (as in Gauss–Jordan, Sec. 7.8).

With the notation for row operations as introduced in Sec. 7.3, the calculations in Step 1 give from the simplex table $\mathbf{T}_0$ in (4) the following simplex table (augmented matrix), with the blue letters referring to the ***previous table.***

(5)
$$\mathbf{T}_1 = \begin{array}{c} \begin{array}{cccccc} z & x_1 & x_2 & x_3 & x_4 & b \end{array} \\ \left[\begin{array}{cccccc} 1 & 0 & -72 & 0 & 8 & 480 \\ 0 & 0 & 7.2 & 1 & -0.4 & 36 \\ 0 & 5 & 2 & 0 & 1 & 60 \end{array}\right] \end{array} \begin{array}{l} \text{Row 1 + 8 Row 3} \\ \text{Row 2 − 0.4 Row 3} \\ \end{array}$$

We see that basic variables are now x_1, x_3 and nonbasic variables are x_2, x_4. Setting the latter to zero, we obtain the basic feasible solution given by $\mathbf{T}_1$,

$$x_1 = 60/5 = 12, \qquad x_2 = 0, \qquad x_3 = 36/1 = 36, \qquad x_4 = 0, \qquad z = 480.$$

This is A in Fig. 473 (Sec. 22.2). We thus have moved from O: (0, 0) with $z = 0$ to A: (12, 0) with the greater $z = 480$. The reason for this increase is our elimination of a term $(-40x_1)$ with a negative coefficient. Hence ***elimination is applied only to negative entries*** in Row 1 but to no others. This motivates the selection of the *column* of the pivot.

We now motivate the selection of the *row* of the pivot. Had we taken the second row of $\mathbf{T}_0$ instead (thus 2 as the pivot), we would have obtained $z = 1200$ (verify!), but this line of constant revenue $z = 1200$ lies entirely outside the feasibility region in Fig. 473. This motivates our cautious choice of the entry 5 as our pivot because it gave the smallest quotient (60/5 = 12).

Step 2. The basic feasible solution given by (5) is not yet optimal because of the negative entry -72 in Row 1. Accordingly, we perform the operations O_1 to O_3 again, choosing a pivot in the column of -72.

Operation O_1. Select Column 3 of $\mathbf{T}_1$ in (5) as the column of the pivot (because $-72 < 0$).

Operation O_2. We have $36/7.2 = 5$ and $60/2 = 30$. Select 7.2 as the pivot (because $5 < 30$).

Operation O_3. Elimination by row operations gives

(6)
$$\mathbf{T}_2 = \begin{array}{c} \begin{array}{ccc|cc|c} z & x_1 & x_2 & x_3 & x_4 & b \end{array} \\ \left[\begin{array}{ccc|cc|c} 1 & 0 & 0 & 10 & 4 & 840 \\ 0 & 0 & 7.2 & 1 & -0.4 & 36 \\ 0 & 5 & 0 & -\dfrac{1}{3.6} & \dfrac{1}{0.9} & 50 \end{array}\right] \end{array} \begin{array}{l} \text{Row 1} + 10\ \text{Row 2} \\ \\ \text{Row 3} - \dfrac{2}{7.2}\ \text{Row 2} \end{array}$$

We see that now x_1, x_2 are basic and x_3, x_4 nonbasic. Setting the latter to zero, we obtain from $\mathbf{T}_2$ the basic feasible solution

$$x_1 = 50/5 = 10, \qquad x_2 = 36/7.2 = 5, \qquad x_3 = 0, \qquad x_4 = 0, \qquad z = 840.$$

This is B in Fig. 473 (Sec. 22.2). In this step, z has increased from 480 to 840, due to the elimination of -72 in $\mathbf{T}_1$. Since $\mathbf{T}_2$ contains no more negative entries in Row 1, we conclude that $z = f(10, 5) = 40 \cdot 10 + 88 \cdot 5 = 840$ is the maximum possible revenue. It is obtained if we produce twice as many S heaters as L heaters. This is the solution of our problem by the simplex method of linear programming. ▩

Minimization. If we want to ***minimize*** $z = f(\mathbf{x})$ (instead of maximize), we take as the columns of the pivots those whose entry in Row 1 is ***positive*** (instead of negative). In such a Column k we consider only positive entries t_{jk} and take as pivot a t_{jk} for which b_j/t_{jk} is smallest (as before). For examples, see the problem set.

PROBLEM SET 22.3

1–9 SIMPLEX METHOD

Write in normal form and solve by the simplex method, assuming all x_j to be nonnegative.

1. Maximize $f = 3x_1 + 2x_2$ subject to $3x_1 + 4x_2 \leqq 60$, $4x_1 + 3x_2 \leqq 60$, $10x_1 + 2x_2 \leqq 120$.
2. Prob. 16 in Problem Set 22.2.
3. Maximize the profit in the daily production of x_1 metal frames F_1 (\$90 profit/frame) and x_2 frames F_2 (\$50 profit/frame) under the restrictions $x_1 + 3x_2 \leqq 1800$ (material), $x_1 + x_2 \leqq 1000$ (machine hours), $3x_1 + x_2 \leqq 2400$ (labor).
4. Maximize $f = 2x_1 + 3x_2 + x_3$ subject to $x_1 + x_2 + x_3 \leqq 4.8$, $10x_1 + x_3 \leqq 9.9$, $x_2 - x_3 \leqq 0.2$.
5. The problem in the text with the order of the constraints interchanged.
6. Minimize $f = 4x_1 - 10x_2 - 20x_3$ subject to $3x_1 + 4x_2 + 5x_3 \leqq 60$, $2x_1 + x_2 \leqq 20$, $2x_1 + 3x_3 \leqq 30$.
7. Minimize $f = 5x_1 - 20x_2$ subject to $-2x_1 + 10x_2 \leqq 5$, $2x_1 + 5x_2 \leqq 10$.
8. Prob. 20 in Problem Set 22.2.

9. Maximize $f = 34x_1 + 29x_2 + 32x_3$ subject to $8x_1 + 2x_2 + x_3 \leqq 54$, $3x_1 + 8x_2 + 2x_3 \leqq 59$, $x_1 + x_2 + 5x_3 \leqq 39$.

10. **CAS PROJECT. Simplex Method.** (a) Write a program for graphing a region R in the first quadrant of the x_1x_2-plane determined by linear constraints.

(b) Write a program for maximizing $z = a_1x_1 + a_2x_2$ in R.

(c) Write a program for maximizing $z = a_1x_1 + \cdots + a_nx_n$ subject to linear constraints.

(d) Apply your programs to problems in this problem set and the previous one.

22.4 Simplex Method: Difficulties

We recall from the last section that in the simplex method we proceed stepwise from one basic feasible solution to another, thereby increasing the value of the objective function f until we reach an optimal solution. Occasionally (but rather infrequently in practice), two kinds of difficulties may occur.

The first of these is degeneracy. A **degenerate feasible solution** is a feasible solution at which more than the usual number $n - m$ of variables are zero. Here n is the number of variables (slack and others) and m the number of constraints (not counting the $x_j \geqq 0$ conditions). In the last section, $n = 4$ and $m = 2$, and the occurring basic feasible solutions were nondegenerate; $n - m = 2$ variables were zero in each such solution.

In the case of a degenerate feasible solution we do an extra elimination step in which a basic variable that is zero for that solution becomes nonbasic (and a nonbasic variable becomes basic instead). We explain this in a typical case. For more complicated cases and techniques (rarely needed in practice) see Ref. [F5] in App. 1.

EXAMPLE 1 **Simplex Method, Degenerate Feasible Solution**

AB Steel, Inc., produces two kinds of iron I_1, I_2 by using three kinds of raw material R_1, R_2, R_3 (scrap iron and two kinds of ore) as shown. Maximize the daily profit.

Raw Material	Raw Material Needed per Ton		Raw Material Available per Day (tons)
	Iron I_1	Iron I_2	
R_1	2	1	16
R_2	1	1	8
R_3	0	1	3.5
Net profit per ton	\$150	\$300	

Solution. Let x_1 and x_2 denote the amount (in tons) of iron I_1 and I_2, respectively, produced per day. Then our problem is as follows. Maximize

$$z = f(\mathbf{x}) = 150x_1 + 300x_2 \tag{1}$$

subject to the constraints $x_1 \geqq 0$, $x_2 \geqq 0$ and

$$\begin{aligned} 2x_1 + x_2 &\leqq 16 && \text{(raw material } R_1\text{)} \\ x_1 + x_2 &\leqq 8 && \text{(raw material } R_2\text{)} \\ x_2 &\leqq 3.5 && \text{(raw material } R_3\text{).} \end{aligned}$$

By introducing slack variables x_3, x_4, x_5 we obtain the normal form of the constraints

$$
\begin{aligned}
2x_1 + x_2 + x_3 &= 16 \\
x_1 + x_2 \quad + x_4 &= 8 \\
x_2 \quad\quad + x_5 &= 3.5 \\
x_i \geqq 0 \quad (i = 1, \cdots, 5).&
\end{aligned} \tag{2}
$$

As in the last section we obtain from (1) and (2) the initial simplex table

$$
\mathbf{T}_0 = \begin{array}{c} \begin{array}{ccccccc} z & x_1 & x_2 & x_3 & x_4 & x_5 & b \end{array} \\ \left[\begin{array}{c|cc|c|c|c|c} 1 & -150 & -300 & 0 & 0 & 0 & 0 \\ \hline 0 & 2 & 1 & 1 & 0 & 0 & 16 \\ 0 & 1 & 1 & 0 & 1 & 0 & 8 \\ 0 & 0 & 1 & 0 & 0 & 1 & 3.5 \end{array}\right] \end{array} . \tag{3}
$$

We see that x_1, x_2 are nonbasic variables and x_3, x_4, x_5 are basic. With $x_1 = x_2 = 0$ we have from (3) the basic feasible solution

$$
x_1 = 0, \quad x_2 = 0, \quad x_3 = 16/1 = 16, \quad x_4 = 8/1 = 8, \quad x_5 = 3.5/1 = 3.5, \quad z = 0.
$$

This is *O:* (0, 0) in Fig. 474. We have $n = 5$ variables x_j, $m = 3$ constraints, and $n - m = 2$ variables equal to zero in our solution, which thus is nondegenerate.

Step 1 of Pivoting

Operation O_1: Column Selection of Pivot. Column 2 (since $-150 < 0$).

Operation O_2: Row Selection of Pivot. $16/2 = 8$, $8/1 = 8$; 3.5/0 is not possible. Hence we could choose Row 2 or Row 3. We choose Row 2. The pivot is 2.

Operation O_3: Elimination by Row Operations. This gives the simplex table

$$
\mathbf{T}_1 = \begin{array}{c} \begin{array}{ccccccc} z & x_1 & x_2 & x_3 & x_4 & x_5 & b \end{array} \\ \left[\begin{array}{c|cc|c|c|c|c} 1 & 0 & -225 & 75 & 0 & 0 & 1200 \\ \hline 0 & 2 & 1 & 1 & 0 & 0 & 16 \\ 0 & 0 & \frac{1}{2} & -\frac{1}{2} & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 3.5 \end{array}\right] \end{array} \begin{array}{l} \\ \\ \text{Row 1} + 75\ \text{Row 2} \\ \\ \text{Row 3} - \frac{1}{2}\ \text{Row 2} \\ \text{Row 4} \end{array} \tag{4}
$$

We see that the basic variables are x_1, x_4, x_5 and the nonbasic are x_2, x_3. Setting the nonbasic variables to zero, we obtain from $\mathbf{T}_1$ the basic feasible solution

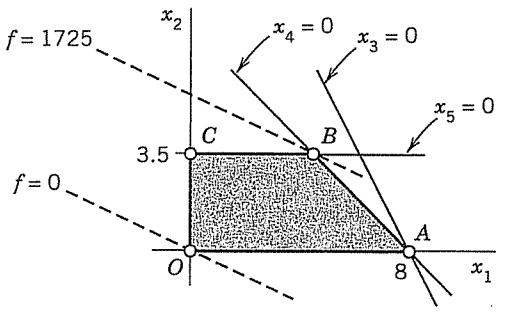


Fig. 474. Example 1, where *A* is degenerate

$$x_1 = 16/2 = 8, \quad x_2 = 0, \quad x_3 = 0, \quad x_4 = 0/1 = 0, \quad x_5 = 3.5/1 = 3.5, \quad z = 1200$$

This is *A:* (8, 0) in Fig. 474. This solution in degenerate because $x_4 = 0$ (in addition to $x_2 = 0$, $x_3 = 0$); geometrically: the straight line $x_4 = 0$ also passes through A. This requires the next step, in which x_4 will become nonbasic.

Step 2 of Pivoting

Operation O_1: Column Selection of Pivot. Column 3 (since $-225 < 0$).

Operation O_2: Row Selection of Pivot. $16/1 = 16$, $0/\frac{1}{2} = 0$. Hence $\frac{1}{2}$ must serve as the pivot.

Operation O_3: Elimination by Row Operations. This gives the following simplex table.

(5)
$$\mathbf{T}_2 = \begin{array}{c} \begin{array}{ccccccc} z & x_1 & x_2 & x_3 & x_4 & x_5 & b \end{array} \\ \left[\begin{array}{ccccccc} 1 & 0 & 0 & -150 & 450 & 0 & 1200 \\ 0 & 2 & 0 & 2 & -2 & 0 & 16 \\ 0 & 0 & \frac{1}{2} & -\frac{1}{2} & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -2 & 1 & 3.5 \end{array}\right] \end{array} \begin{array}{l} \\ \text{Row 1} + 450 \text{ Row 3} \\ \text{Row 2} - 2 \text{ Row 3} \\ \\ \text{Row 4} - 2 \text{ Row 3} \end{array}$$

We see that the basic variables are x_1, x_2, x_5 and the nonbasic are x_3, x_4. Hence x_4 has become nonbasic, as intended. By equating the nonbasic variables to zero we obtain from $\mathbf{T}_2$ the basic feasible solution

$$x_1 = 16/2 = 8, \quad x_2 = 0/\tfrac{1}{2} = 0, \quad x_3 = 0, \quad x_4 = 0, \quad x_5 = 3.5/1 = 3.5, \quad z = 1200.$$

This is still *A:* (8, 0) in Fig. 474 and z has not increased. But this opens the way to the maximum, which we reach in the next step.

Step 3 of Pivoting

Operation O_1: Column Selection of Pivot. Column 4 (since $-150 < 0$).

Operation O_2: Row Selection of Pivot. $16/2 = 8$, $0/(-\frac{1}{2}) = 0$, $3.5/1 = 3.5$. We can take 1 as the pivot. (With $-\frac{1}{2}$ as the pivot we would not leave A. Try it.)

Operation O_3: Elimination by Row Operations. This gives the simplex table

(6)
$$\mathbf{T}_3 = \begin{array}{c} \begin{array}{ccccccc} z & x_1 & x_2 & x_3 & x_4 & x_5 & b \end{array} \\ \left[\begin{array}{ccccccc} 1 & 0 & 0 & 0 & 150 & 150 & 1725 \\ 0 & 2 & 0 & 0 & 2 & -2 & 9 \\ 0 & 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & 1.75 \\ 0 & 0 & 0 & 1 & -2 & 1 & 3.5 \end{array}\right] \end{array} \begin{array}{l} \\ \text{Row 1} + 150 \text{ Row 4} \\ \text{Row 2} - 2 \text{ Row 4} \\ \text{Row 3} + \frac{1}{2} \text{ Row 2} \\ \end{array}$$

We see that basic variables are x_1, x_2, x_3 and nonbasic x_4, x_5. Equating the latter to zero we obtain from $\mathbf{T}_3$ the basic feasible solution

$$x_1 = 9/2 = 4.5, \quad x_2 = 1.75/\tfrac{1}{2} = 3.5, \quad x_3 = 3.5/1 = 3.5, \quad x_4 = 0, \quad x_5 = 0, \quad z = 1725.$$

This is *B:* (4.5, 3.5) in Fig. 474. Since Row 1 of $\mathbf{T}_3$ has no negative entries, we have reached the maximum daily profit $z_{\max} = f(4.5, 3.5) = 150 \cdot 4.5 + 300 \cdot 3.5 = \$1725$. This is obtained by using 4.5 tons of iron I_1 and 3.5 tons of iron I_2. ■

Difficulties in Starting

As a second kind of difficulty, it may sometimes be hard to find a basic feasible solution to start from. In such a case the idea of an **artificial variable** (or several such variables) is helpful. We explain this method in terms of a typical example.

EXAMPLE 2 Simplex Method: Difficult Start, Artificial Variable

Maximize

$$z = f(\mathbf{x}) = 2x_1 + x_2 \tag{7}$$

subject to the constraints $x_1 \geqq 0$, $x_2 \geqq 0$ and (Fig. 475)

$$x_1 - \tfrac{1}{2}x_2 \geqq 1$$
$$x_1 - x_2 \leqq 2$$
$$x_1 + x_2 \leqq 4.$$

Solution. By means of slack variables we achieve the normal form of the constraints

$$\begin{aligned} z - 2x_1 - x_2 &= 0 \\ x_1 - \tfrac{1}{2}x_2 - x_3 &= 1 \\ x_1 - x_2 + x_4 &= 2 \\ x_1 + x_2 + x_5 &= 4 \\ x_i \geqq 0 \quad (i = 1, \cdots, 5). & \end{aligned} \tag{8}$$

Note that the first slack variable is negative (or zero), which makes x_3 nonnegative within the feasibility region (and negative outside). From (7) and (8) we obtain the simplex table

z	x_1	x_2	x_3	x_4	x_5	b
1	−2	−1	0	0	0	0
0	1	$-\frac{1}{2}$	−1	0	0	1
0	1	−1	0	1	0	2
0	1	1	0	0	1	4

x_1, x_2 are nonbasic, and we would like to take x_3, x_4, x_5 as basic variables. By our usual process of equating the nonbasic variables to zero we obtain from this table

$$x_1 = 0, \quad x_2 = 0, \quad x_3 = 1/(-1) = -1, \quad x_4 = 2/1 = 2, \quad x_5 = 4/1 = 4, \quad z = 0.$$

$x_3 < 0$ indicates that (0, 0) lies outside the feasibility region. Since $x_3 < 0$, we cannot proceed immediately. Now, instead of searching for other basic variables, we use the following idea. Solving the second equation in (8) for x_3, we have

$$x_3 = -1 + x_1 - \tfrac{1}{2}x_2.$$

To this we now add a variable x_6 on the right,

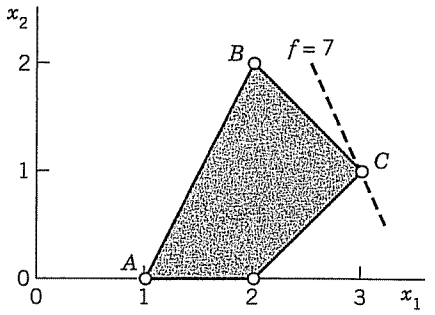


Fig. 475. Feasibility region in Example 2

(9) $$x_3 = -1 + x_1 - \tfrac{1}{2}x_2 + x_6.$$

x_6 is called an **artificial variable** and is subject to the constraint $x_6 \geqq 0$.

We must take care that x_6 (which is not part of the given problem!) will disappear eventually. We shall see that we can accomplish this by adding a term $-Mx_6$ with very large M to the objective function. Because of (7) and (9) (solved for x_6) this gives the modified objective function for this **"extended problem"**

(10) $$\hat{z} = z - Mx_6 = 2x_1 + x_2 - Mx_6 = (2 + M)x_1 + (1 - \tfrac{1}{2}M)x_2 - Mx_3 - M.$$

We see that the simplex table corresponding to (10) and (8) is

$$\mathbf{T}_0 = \begin{array}{c|cccccc|c} \hat{z} & x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & b \\ \hline 1 & -2-M & -1+\frac{1}{2}M & M & 0 & 0 & 0 & -M \\ \hline 0 & 1 & -\frac{1}{2} & -1 & 0 & 0 & 0 & 1 \\ 0 & 1 & -1 & 0 & 1 & 0 & 0 & 2 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 4 \\ 0 & 1 & -\frac{1}{2} & -1 & 0 & 0 & 1 & 1 \end{array}.$$

The last row of this table results from (9) written as $x_1 - \frac{1}{2}x_2 - x_3 + x_6 = 1$. We see that we can now start, taking x_4, x_5, x_6 as the basic variables and x_1, x_2, x_3 as the nonbasic variables. Column 2 has a negative first entry. We can take the second entry (1 in Row 2) as the pivot. This gives

$$\mathbf{T}_1 = \begin{array}{c|cccccc|c} \hat{z} & x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & b \\ \hline 1 & 0 & -2 & -2 & 0 & 0 & 0 & 2 \\ \hline 0 & 1 & -\frac{1}{2} & -1 & 0 & 0 & 0 & 1 \\ 0 & 0 & -\frac{1}{2} & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & \frac{3}{2} & 1 & 0 & 1 & 0 & 3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{array}.$$

This corresponds to $x_1 = 1$, $x_2 = 0$ (point A in Fig. 475), $x_3 = 0$, $x_4 = 1$, $x_5 = 3$, $x_6 = 0$. We can now drop Row 5 and Column 7. In this way we get rid of x_6, as wanted, and obtain

$$\mathbf{T}_2 = \begin{array}{c|cccccc|c} z & x_1 & x_2 & x_3 & x_4 & x_5 & b \\ \hline 1 & 0 & -2 & -2 & 0 & 0 & 2 \\ \hline 0 & 1 & -\frac{1}{2} & -1 & 0 & 0 & 1 \\ 0 & 0 & -\frac{1}{2} & 1 & 1 & 0 & 1 \\ 0 & 0 & \frac{3}{2} & 1 & 0 & 1 & 3 \end{array}.$$

In Column 3 we choose 3/2 as the next pivot. We obtain

$$\mathbf{T}_3 = \begin{array}{c|cccccc|c} z & x_1 & x_2 & x_3 & x_4 & x_5 & b \\ \hline 1 & 0 & 0 & -\frac{2}{3} & 0 & \frac{4}{3} & 6 \\ \hline 0 & 1 & 0 & -\frac{2}{3} & 0 & \frac{1}{3} & 2 \\ 0 & 0 & 0 & \frac{4}{3} & 1 & \frac{1}{3} & 2 \\ 0 & 0 & \frac{3}{2} & 1 & 0 & 1 & 3 \end{array}.$$

This corresponds to $x_1 = 2$, $x_2 = 2$ (this is B in Fig. 475), $x_3 = 0$, $x_4 = 2$, $x_5 = 0$. In Column 4 we choose 4/3 as the pivot, by the usual principle. This gives

$$\mathbf{T}_4 = \begin{array}{c} \begin{matrix} z & x_1 & x_2 & x_3 & x_4 & x_5 & b \end{matrix} \\ \left[\begin{array}{c|cc|cc|c} 1 & 0 & 0 & 0 & \frac{1}{2} & \frac{3}{2} & 7 \\ \hline 0 & 1 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 3 \\ 0 & 0 & 0 & \frac{4}{3} & 1 & \frac{1}{3} & 2 \\ 0 & 0 & \frac{3}{2} & 0 & -\frac{3}{4} & \frac{3}{4} & \frac{3}{2} \end{array}\right] \end{array}.$$

This corresponds to $x_1 = 3$, $x_2 = 1$ (point C in Fig. 475), $x_3 = \frac{3}{2}$, $x_4 = 0$, $x_5 = 0$. This is the maximum $f_{\max} = f(3, 1) = 7$. ■

PROBLEM SET 22.4

If in a step you have a choice between pivots, take the one that comes first in the column considered.

1. Maximize $z = f_1(\mathbf{x}) = 6x_1 + 12x_2$ subject to $0 \leqq x_1 \leqq 4$, $0 \leqq x_2 \leqq 4$, $6x_1 + 12x_2 \leqq 72$.
2. Do Prob. 1 with the last two constraints interchanged.
3. Maximize the daily output in producing x_1 glass plates by a process P_1 and x_2 glass plates by a process P_2 subject to the constraints (labor hours, machine hours, raw material supply)

$$2x_1 + 3x_2 \leqq 130, \qquad 3x_1 + 8x_2 \leqq 300,$$
$$4x_1 + 2x_2 \leqq 140.$$

4. Maximize $z = 300x_1 + 500x_2$ subject to $2x_1 + 8x_2 \leqq 60$, $2x_1 + x_2 \leqq 30$, $4x_1 + 4x_2 \leqq 60$.
5. Do Prob. 4 with the last two constraints interchanged. Comment on the resulting simplification.
6. Maximize the total output $f = x_1 + x_2 + x_3$ (production figures of three different production processes) subject to input constraints (limitation of machine time)

$$4x_1 + 5x_2 + 8x_3 \leqq 12,$$
$$8x_1 + 5x_2 + 4x_3 \leqq 12.$$

7. Maximize $f = 6x_1 + 6x_2 + 9x_3$ subject to $x_j \geqq 0$ $(j = 1, \cdots, 5)$, and $x_1 + x_3 + x_4 = 1$, $x_2 + x_3 + x_5 = 1$.
8. Using an artificial variable, minimize $f = 2x_1 - x_2$ subject to $x_1 \geqq 0$, $x_2 \geqq 0$, $x_1 + x_2 \geqq 5$, $-x_1 + x_2 \leqq 1$, $5x_1 + 4x_2 \leqq 40$.
9. Maximize $f = 4x_1 + x_2 + 2x_3$ subject to $x_1 \geqq 0$, $x_2 \geqq 0$, $x_3 \leqq 0$, $x_1 + x_2 + x_3 \leqq 1$, $x_1 + x_2 - x_3 \leqq 0$.
10. If one uses the method of artificial variables in a problem without solution, this nonexistence will become apparent by the fact that one cannot get rid of the artificial variable. Illustrate this by trying to maximize $f = 2x_1 + x_2$ subject to $x_1 \geqq 0$, $x_2 \geqq 0$, $2x_1 + x_2 \leqq 2$, $x_1 + 2x_2 \geqq 6$, $x_1 + x_2 \leqq 4$.

CHAPTER 22 REVIEW QUESTIONS AND PROBLEMS

1. What is the difference between constrained and unconstrained optimization?
2. State the idea and the basic formulas of the method of steepest descent.
3. Write down an algorithm for the method of steepest descent.
4. Design a "method of steepest ascent" for determining maxima.
5. What is linear programming? Its basic idea? An objective function?
6. Why can we not use methods of calculus for extrema in linear programming?
7. What are slack variables? Artificial variables? Why did we use them?
8. Apply the method of steepest descent to $f(\mathbf{x}) = x_1^2 + 1.5x_2^2$, starting from (6, 3). Do 3 steps. Why is the convergence faster than in Example 1, Sec. 22.1?
9. What does the method of steepest descent amount to in the case of a single variable?
10. In Prob. 8 start from $\mathbf{x}_0 = [1.5 \quad 1]^T$. Show that the next even-numbered approximations are $\mathbf{x}_2 = k\mathbf{x}_0$, $\mathbf{x}_4 = k^2\mathbf{x}_0$, etc., where $k = 0.04$.
11. What happens in Example 1 of Sec. 22.1 if you replace the function $f(\mathbf{x}) = x_1^2 + 3x_2^2$ by $f(\mathbf{x}) = x_1^2 + 5x_2^2$? Do 5 steps, starting from $\mathbf{x}_0 = [6 \quad 3]^T$. Is the convergence faster or slower?
12. Apply the method of steepest descent to $f(\mathbf{x}) = 9x_1^2 + x_2^2 + 18x_1 - 4x_2$, 5 steps, starting from $\mathbf{x}_0 = [2 \quad 4]^T$.
13. In Prob. 12, could you start from $[0 \quad 0]^T$ and do 5 steps?

14. Show that the gradients in Prob. 13 are orthogonal. Give a reason.

15–20 Graph or sketch the region in the first quadrant of the x_1x_2-plane determined by the following inequalities.

15. $x_1 + 3x_2 \leqq 6$, $2x_1 + x_2 \leqq 4$

16. $x_1 - 2x_2 \leqq -2$, $0.8x_1 + x_2 \leqq 6$

17. $-x_1 + x_2 \geqq 0$, $x_1 + x_2 \leqq 4$

18. $x_1 - 2x_2 \geqq -4$, $2x_1 + x_2 \leqq 12$, $x_1 + x_2 \leqq 8$

19. $x_1 + x_2 \leqq 5$, $x_2 \leqq 3$, $-x_1 + x_2 \leqq 2$

20. $x_1 + x_2 \geqq 2$, $2x_1 - 3x_2 \geqq -12$, $x_1 \leqq 15$

21–25 Maximize or minimize as indicated.

21. Maximize $f = 10x_1 + 20x_2$ subject to $x_1 \leqq 5$, $x_1 + x_2 \leqq 6$, $x_2 \leqq 4$.

22. Maximize $f = x_1 + x_2$ subject to $x_1 + 2x_2 \leqq 10$, $2x_1 + x_2 \leqq 10$, $x_2 \leqq 4$.

23. Minimize $f = 2x_1 - 10x_2$ subject to $x_1 - x_2 \leqq 4$, $2x_1 + x_2 \leqq 14$, $x_1 + x_2 \leqq 9$, $-x_1 + 3x_2 \leqq 15$.

24. A factory produces two kinds of gaskets, G_1, G_2, with net profit of \$60 and \$30, respectively. Maximize the total daily profit subject to the constraints (x_j = number of gaskets G_j produced per day)

$$40x_1 + 40x_2 \leqq 1800 \quad \text{(Machine hours)},$$

$$200x_1 + 20x_2 \leqq 6300 \quad \text{(Labor)}.$$

25. Maximize the daily output in producing x_1 chairs by a process P_1 and x_2 chairs by a process P_2 subject to $3x_1 + 4x_2 \leqq 550$ (machine hours), $5x_1 + 4x_2 \leqq 650$ (labor).

SUMMARY OF CHAPTER 22
Unconstrained Optimization. Linear Programming

In optimization problems we maximize or minimize an **objective function** $z = f(\mathbf{x})$ depending on control variables $x_1, \cdots, x_m$ whose domain is either unrestricted (**"unconstrained optimization,"** Sec. 22.1) or restricted by constraints in the form of inequalities or equations or both (**"constrained optimization,"** Sec. 22.2).

If the objective function is *linear* and the constraints are *linear inequalities* in $x_1, \cdots, x_m$, then by introducing **slack variables** $x_{m+1}, \cdots, x_n$ we can write the optimization problem in **normal form** with the objective function given by

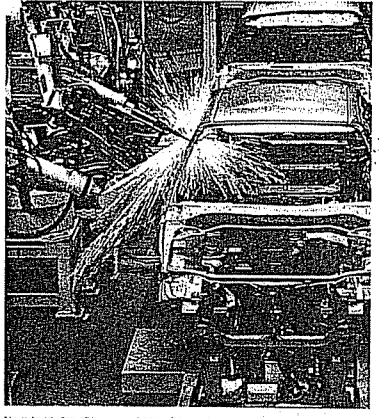
$$(1) \qquad f_1 = c_1x_1 + \cdots + c_nx_n$$

(where $c_{m+1} = \cdots = c_n = 0$) and the constraints given by

$$(2) \qquad \begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ &\cdots\cdots\cdots \\ &\cdots\cdots\cdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m \end{aligned}$$

$$x_1 \geqq 0, \cdots, x_n \geqq 0.$$

In this case we can then apply the widely used **simplex method** (Sec. 22.3), a systematic stepwise search through a very much reduced subset of all feasible solutions. Section 22.4 shows how to overcome difficulties with this method.



CHAPTER 23

Graphs. Combinatorial Optimization

Graphs and **digraphs** (= directed graphs) have developed into powerful tools in areas, such as electrical and civil engineering, communication networks, operations research, computer science, economics, industrial management, and marketing. An essential factor of this growth is the use of computers in large-scale optimization problems that can be modeled by graphs and solved by algorithms provided by graph theory. This approach yields models of general applicability and economic importance. It lies in the center of **combinatorial optimization,** a term denoting optimization problems that are of pronounced discrete or combinatorial structure.

This chapter gives an introduction to this wide area, which constitutes a shift of emphasis away from differential equations, eigenvalues, and so on, and is full of new ideas as well as open problems—in connection, for instance, with efficient computer algorithms. The classes of problems we shall consider include transportation of minimum cost or time, best assignment of workers to jobs, most efficient use of communication networks, and many others. Problems for these classes often form the core of larger and more involved practical problems.

Prerequisite: none.
References and Answers to Problems: App. 1 Part F, App. 2.

23.1 Graphs and Digraphs

Roughly, a *graph* consists of points, called *vertices,* and lines connecting them, called *edges.* For example, these may be four cities and five highways connecting them, as in Fig. 476. Or the points may represent some people, and we connect by an edge those who do business with each other. Or the vertices may represent computers in a network and the edges connections between them. Let us now give a formal definition.

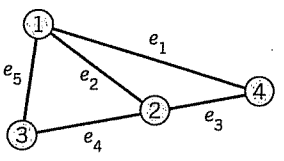


Fig. 476. Graph consisting of 4 vertices and 5 edges

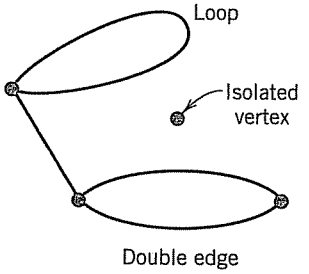


Fig. 477. Isolated vertex, loop, double edge. (Excluded by definition.)

DEFINITION

Graph

A **graph** G consists of two finite sets (sets having finitely many elements), a set V of points, called **vertices,** and a set E of connecting lines, called **edges,** such that each edge connects two vertices, called the *endpoints* of the edge. We write

$$G = (V, E).$$

Excluded are *isolated vertices* (vertices that are not endpoints of any edge), *loops* (edges whose endpoints coincide), and *multiple edges* (edges that have both endpoints in common. See Fig. 477.

CAUTION! Our three exclusions are practical and widely accepted, but not uniformly. For instance, some authors permit multiple edges and call graphs without them *simple graphs.* ■

We denote vertices by letters, $u, v, \cdots$ or $v_1, v_2, \cdots$ or simply by numbers $1, 2, \cdots$ (as in Fig. 476). We denote edges by $e_1, e_2, \cdots$ or by their two endpoints; for instance, $e_1 = (1, 4)$, $e_2 = (1, 2)$ in Fig. 476.

An edge (v_i, v_j) is called **incident** with the vertex v_i (and conversely); similarly, (v_i, v_j) is *incident* with v_j. The number of edges incident with a vertex v is called the **degree** of v. Two vertices are called **adjacent** in G if they are connected by an edge in G (that is, if they are the two endpoints of some edge in G).

We meet graphs in different fields under different names: as "networks" in electrical engineering, "structures" in civil engineering, "molecular structures" in chemistry, "organizational structures" in economics, "sociograms," "road maps," "telecommunication networks," and so on.

Digraphs (Directed Graphs)

Nets of one-way streets, pipeline networks, sequences of jobs in construction work, flows of computation in a computer, producer–consumer relations, and many other applications suggest the idea of a "digraph" (= directed graph), in which each edge has a direction (indicated by an arrow, as in Fig. 478).

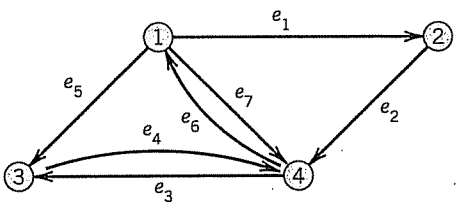


Fig. 478. Digraph

DEFINITION

Digraph (Directed Graph)

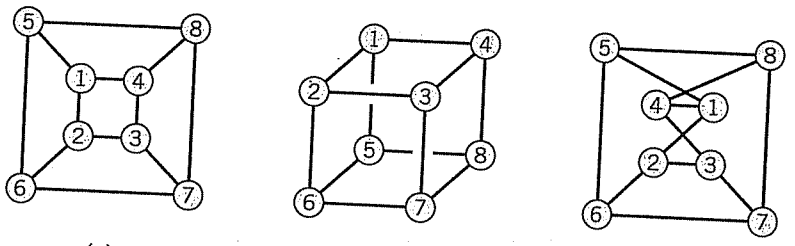
A **digraph** $G = (V, E)$ is a graph in which each edge $e = (i, j)$ has a direction from its *"initial point" i* to its *"terminal point" j.*

Two edges connecting the same two points i, j are now permitted, provided they have opposite directions, that is, they are (i, j) and (j, i). *Example.* (1, 4) and (4, 1) in Fig. 478.

A **subgraph** or subdigraph of a given graph or digraph $G = (V, E)$, respectively, is a graph or digraph obtained by deleting some of the edges and vertices of G, retaining the other edges of G (together with their pairs of endpoints). For instance, e_1, e_3 (together with the vertices 1, 2, 4) form a subgraph in Fig. 476, and e_3, e_4, e_5 (together with the vertices 1, 3, 4) form a subdigraph in Fig. 478.

Computer Representation of Graphs and Digraphs

Drawings of graphs are useful to people in explaining or illustrating specific situations. Here one should be aware that a graph may be sketched in various ways; see Fig. 479. For handling graphs and digraphs in computers, one uses matrices or lists as appropriate data structures, as follows.



(a) (b) (c)

Fig. 479. Different sketches of the same graph

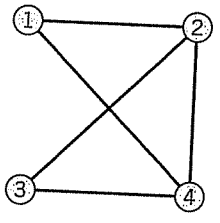
Adjacency Matrix of a Graph G: Matrix $\mathbf{A} = [a_{ij}]$ with entries

$$a_{ij} = \begin{cases} 1 & \text{if } G \text{ has an edge } (i, j), \\ 0 & \text{else.} \end{cases}$$

Thus $a_{ij} = 1$ if and only if two vertices i and j are adjacent in G. Here, by definition, no vertex is considered to be adjacent to itself; thus, $a_{ii} = 0$. $\mathbf{A}$ is symmetric, $a_{ij} = a_{ji}$. (Why?)

The adjacency matrix of a graph is generally much smaller than the so-called *incidence matrix* (see Probs. 21, 22) and is preferred over the latter if one decides to store a graph in a computer in matrix form.

EXAMPLE 1 **Adjacency Matrix of a Graph**



Vertex	1	2	3	4
Vertex 1	0	1	0	1
2	1	0	1	1
3	0	1	0	1
4	1	1	1	0

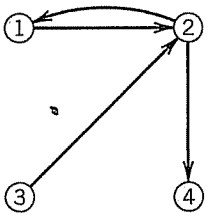
■

Adjacency Matrix of a Digraph G: Matrix $\mathbf{A} = [a_{ij}]$ with entries

$$a_{ij} = \begin{cases} 1 & \text{if } G \text{ has a directed edge } (i, j), \\ 0 & \text{else.} \end{cases}$$

This matrix $\mathbf{A}$ is not symmetric. (Why?)

EXAMPLE 2 Adjacency Matrix of a Digraph



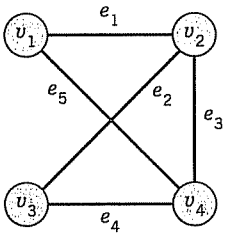
To vertex	1	2	3	4
From vertex 1	0	1	0	0
2	1	0	0	1
3	0	1	0	0
4	0	0	0	0

■

Lists. The **vertex incidence list** of a graph shows for each vertex the incident edges. The **edge incidence list** shows for each edge its two endpoints. Similarly for a *digraph;* in the vertex list, outgoing edges then get a minus sign, and in the edge list we now have *ordered* pairs of vertices.

EXAMPLE 3 Vertex Incidence List and Edge Incidence List of a Graph

This graph is the same as in Example 1, except for notation.



Vertex	Incident Edges
v_1	e_1, e_5
v_2	e_1, e_2, e_3
v_3	e_2, e_4
v_4	e_3, e_4, e_5

Edge	Endpoints
e_1	v_1, v_2
e_2	v_2, v_3
e_3	v_2, v_4
e_4	v_3, v_4
e_5	v_1, v_4

■

"Sparse graphs" are graphs with few edges (far fewer than the maximum possible number $n(n-1)/2$, where n is the number of vertices). For these graphs, matrices are not efficient. *Lists* then have the advantage of requiring much less storage and being easier to handle; they can be ordered, sorted, or manipulated in various other ways directly within the computer. For instance, in tracing a "walk" (a connected sequence of edges with pairwise common endpoints), one can easily go back and forth between the two lists just discussed, instead of scanning a large column of a matrix for a single 1.

Computer science has developed more refined lists, which, in addition to the actual content, contain "pointers" indicating the preceding item or the next item to be scanned or both items (in the case of a "walk": the preceding edge or the subsequent one). For details, see Refs. [E16] and [F7].

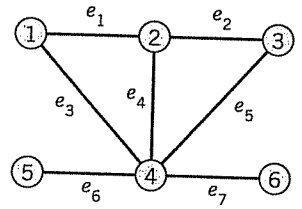
This section was devoted to basic concepts and notations needed throughout this chapter, in which we shall discuss some of the most important classes of combinatorial optimization problems. This will at the same time help us to become more and more familiar with graphs and digraphs.

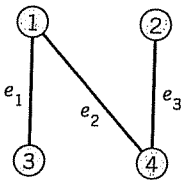
PROBLEM SET 23.1

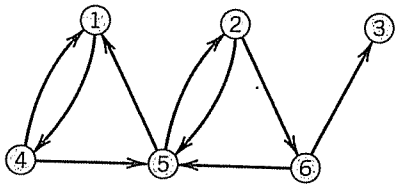
1. Sketch the graph consisting of the vertices and edges of a square. Of a tetrahedron.
2. Worker W_1 can do jobs J_1 and J_3, worker W_2 job J_4, worker W_3 jobs J_2 and J_3. Represent this by a graph.
3. Explain how the following may be regarded as graphs or digraphs: flight connections between given cities; memberships of some persons in some committees; relations between chapters of a book; a tennis tournament; a family tree.
4. How would you represent a net of one-way and two-way streets by a digraph?
5. Give further examples of situations that could be represented by a graph or digraph.
6. Find the adjacency matrix of the graph in Fig. 476.
7. When will the adjacency matrix of a graph be symmetric? Of a digraph?

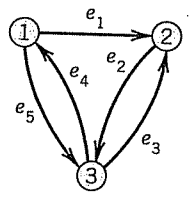
8–13 ADJACENCY MATRIX

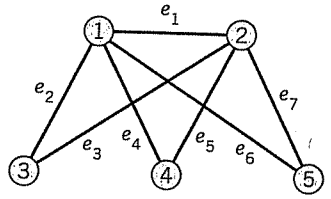
Find the adjacency matrix of the graph or digraph.

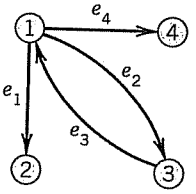
8.


9.


10.


11.


12.


13.


Sketch the graph whose adjacency matrix is:

14.
$$\begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

15.
$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

16.
$$\begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

Sketch the digraph whose adjacency matrix is:

17. The matrix in Prob. 14.
18. The matrix in Prob. 16.
19. **(Complete graph)** Show that a graph G with n vertices can have at most $n(n-1)/2$ edges, and G has exactly $n(n-1)/2$ edges if G is *complete*, that is, if every pair of vertices of G is joined by an edge. (Recall that loops and multiple edges are excluded.)
20. In what case are all the off-diagonal entries of the adjacency matrix of a graph G equal to 1?

Incidence Matrix of a Graph: Matrix $\mathbf{B} = [b_{jk}]$ with entries

$$b_{jk} = \begin{cases} 1 & \text{if vertex } j \text{ is an endpoint of edge } e_k \\ 0 & \text{otherwise.} \end{cases}$$

Find the incidence matrix of:

21. The graph in Prob. 9.
22. The graph in Prob. 8.

Incidence Matrix of a Digraph: Matrix $\widetilde{\mathbf{B}} = [\widetilde{b}_{jk}]$ with entries

$$\widetilde{b}_{jk} = \begin{cases} -1 & \text{if edge } e_k \text{ leaves vertex } j \\ 1 & \text{if edge } e_k \text{ enters vertex } j \\ 0 & \text{otherwise.} \end{cases}$$

Find the incidence matrix of:

23. The digraph in Prob. 11.
24. The digraph in Prob. 13.
25. Make a vertex incidence list of the digraph in Prob. 13.

23.2 Shortest Path Problems. Complexity

Beginning in this section, we shall discuss some of the most important classes of optimization problems that concern graphs and digraphs as they arise in applications. Basic ideas and algorithms will be explained and illustrated by small graphs, but you should keep in mind that real-life problems may often involve many ***thousands or even millions of vertices and edges*** (think of telephone networks, worldwide air travel, companies that have offices and stores in all larger cities). Then reliable and efficient systematic methods are an absolute necessity—solution by inspection or by trial and error would no longer work, even if "nearly optimal" solutions are acceptable.

We begin with **shortest path problems,** as they arise, for instance, in designing shortest (or least expensive, or fastest) routes for a traveling salesman, for a cargo ship, etc. Let us first explain what we mean by a path.

In a graph $G = (V, E)$ we can walk from a vertex v_1 along some edges to some other vertex v_k. Here we can

(A) make no restrictions, or

(B) require that each *edge* of G be traversed at most once, or

(C) require that each *vertex* be visited at most once.

In case (A) we call this a **walk.** Thus a walk from v_1 to v_k is of the form

$$(1) \qquad (v_1, v_2), (v_2, v_3), \cdots, (v_{k-1}, v_k),$$

where some of these edges or vertices may be the same. In case (B), where each *edge* may occur at most once, we call the walk a **trail.** Finally, in case (C), where each *vertex* may occur at most once (and thus each edge automatically occurs at most once), we call the trail a **path.**

We admit that a walk, trail, or path may end at the vertex it started from, in which case we call it **closed;** then $v_k = v_1$ in (1).

A closed path is called a **cycle.** *A cycle has at least three edges* (because we do not have double edges; see Sec. 23.1). Figure 480 illustrates all these concepts.

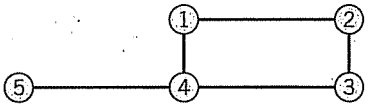


Fig. 480. Walk, trail, path, cycle

1 − 2 − 3 − 2 is a walk (not a trail).
4 − 1 − 2 − 3 − 4 − 5 is a trail (not a path).
1 − 2 − 3 − 4 − 5 is a path (not a cycle).
1 − 2 − 3 − 4 − 1 is a cycle.

Shortest Path

To define the concept of a shortest path, we assume that $G = (V, E)$ is a **weighted graph,** that is, each edge (v_i, v_j) in G has a given *weight* or *length* $l_{ij} > 0$. Then a **shortest path** $v_1 \to v_k$ (with fixed v_1 and v_k) is a path (1) such that the sum of the lengths of its edges

$$l_{12} + l_{23} + l_{34} + \cdots + l_{k-1,k}$$

(l_{12} = length of (v_1, v_2), etc.) is minimum (as small as possible among all paths from v_1 to v_k). Similarly, a **longest path** $v_1 \to v_k$ is one for which that sum is maximum.

Shortest (and longest) path problems are among the most important optimization problems. Here, "length" l_{ij} (often also called "cost" or "weight") can be an actual length measured in miles or travel time or gasoline expenses, but it may also be something entirely different.

For instance, the *"traveling salesman problem"* requires the determination of a shortest **Hamiltonian**[1] **cycle** in a graph, that is, a cycle that contains all the vertices of the graph.

As another example, by choosing the "most profitable" route $v_1 \to v_k$, a salesman may want to maximize Σl_{ij}, where l_{ij} is his expected commission minus his travel expenses for going from town i to town j.

In an investment problem, i may be the day an investment is made, j the day it matures, and l_{ij} the resulting profit, and one gets a graph by considering the various possibilities of investing and reinvesting over a given period of time.

Shortest Path if All Edges Have Length $l = 1$

Obviously, if all edges have length 1, then a shortest path $v_1 \to v_k$ is one that has the smallest number of edges among all paths $v_1 \to v_k$ in a given graph G. For this problem we discuss a BFS algorithm. BFS stands for **Breadth First Search.** This means that in each step the algorithm visits *all neighboring* (all adjacent) vertices of a vertex reached, as opposed to a DFS algorithm (**Depth First Search** algorithm), which makes a long trail (as in a maze). This widely used BFS algorithm is shown in Table 23.1.

We want to find a shortest path in G from a vertex s (***start***) to a vertex t (***terminal***). To guarantee that there is a path from s to t, we make sure that G does not consist of separate portions. Thus we assume that G is **connected,** that is, for any two vertices v and w there is a path $v \to w$ in G. (Recall that a vertex v is called **adjacent** to a vertex u if there is an edge (u, v) in G.)

Table 23.1 Moore's BFS for Shortest Path (All Lengths One)

Proceedings of the International Symposium for Switching Theory, Part II, pp. 285–292. Cambridge: Harvard University Press, 1959.

ALGORITHM MOORE $[G = (V, E), s, t]$

This algorithm determines a shortest path in a connected graph $G = (V, E)$ from a vertex s to a vertex t.

INPUT: Connected graph $G = (V, E)$, in which one vertex is denoted by s and one by t, and each edge (i, j) has length $l_{ij} = 1$. Initially all vertices are unlabeled.

OUTPUT: A shortest path $s \to t$ in $G = (V, E)$

1. Label s with 0.
2. Set $i = 0$.
3. Find all *unlabeled* vertices adjacent to a vertex labeled i.
4. Label the vertices just found with $i + 1$.
5. If vertex t is labeled, then "backtracking" gives the shortest path

$$k\ (=\text{label of } t),\ k-1,\ k-2,\ \cdots,\ 0$$

OUTPUT $k, k-1, k-2, \cdots, 0$. Stop

Else increase i by 1. Go to Step 3.

End MOORE

[1] WILLIAM ROWAN HAMILTON (1805–1865), Irish mathematician, known for his work in dynamics.

EXAMPLE 1 Application of Moore's BFS Algorithm

Find a shortest path $s \to t$ in the graph G shown in Fig. 481.

Solution. Figure 481 shows the labels. The blue edges form a shortest path (length 4). There is another shortest path $s \to t$. (Can you find it?) Hence in the program we must introduce a rule that makes backtracking unique because otherwise the computer would not know what to do next if at some step there is a choice (for instance, in Fig. 481 when it got back to the vertex labeled 2). The following rule seems to be natural.

Backtracking rule. Using the numbering of the vertices from 1 to n (not the labeling!), at each step, if a vertex labeled i is reached, take as the next vertex that with the smallest number (not label!) among all the vertices labeled $i - 1$. ■

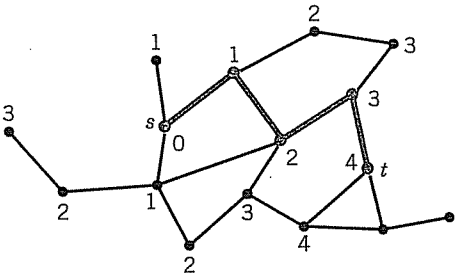


Fig. 481. Example 1, given graph and result of labeling

Complexity of an Algorithm

Complexity *of Moore's algorithm.* To find the vertices to be labeled 1, we have to scan all edges incident with s. Next, when $i = 1$, we have to scan all edges incident with vertices labeled 1, etc. Hence each edge is scanned twice. These are $2m$ operations (m = number of edges of G). This is a function $c(m)$. Whether it is $2m$ or $5m + 3$ or $12m$ is not so essential; it *is* essential that $c(m)$ is proportional to m (not m^2, for example); it is of the "order" m. We write for any function $am + b$ simply $O(m)$, for any function $am^2 + bm + d$ simply $O(m^2)$, and so on; here, O suggests **order**. The underlying idea and practical aspect are as follows.

In judging an algorithm, we are mostly interested in its behavior for very large problems (large m in the present case), since these are going to determine the limits of the applicability of the algorithm. Thus, the essential item is the fastest growing term (am^2 in $am^2 + bm + d$, etc.) since it will overwhelm the others when m is large enough. Also, a constant factor in this term is not very essential; for instance, the difference between two algorithms of orders, say, $5m^2$ and $8m^2$ is generally not very essential and can be made irrelevant by a modest increase in the speed of computers. However, it does make a great practical difference whether an algorithm is of order m or m^2 or of a still higher power m^p. And the biggest difference occurs between these "polynomial orders" and "exponential orders," such as 2^m.

For instance, on a computer that does 10^9 operations per second, a problem of size $m = 50$ will take 0.3 second with an algorithm that requires m^5 operations, but 13 days with an algorithm that requires 2^m operations. But this is not our only reason for regarding polynomial orders as good and exponential orders as bad. Another reason is the ***gain in using a faster computer.*** For example let two algorithms be $O(m)$ and $O(m^2)$. Then, since $1000 = 31.6^2$, an increase in speed by a factor 1000 has the effect that per hour we can do problems 1000 and 31.6 times as big, respectively. But since $1000 = 2^{9.97}$, with an algorithm that is $O(2^m)$, all we gain is a relatively modest increase of 10 in problem size because $2^{9.97} \cdot 2^m = 2^{m+9.97}$.

The **symbol O** is quite practical and commonly used whenever the order of growth is essential, but not the specific form of a function. Thus if a function $g(m)$ is of the form

$$g(m) = kh(m) + \text{more slowly growing terms} \qquad (k \neq 0, \text{ constant}),$$

we say that $g(m)$ is of the *order* $h(m)$ and write

$$g(m) = O(h(m)).$$

For instance,

$$am + b = O(m), \qquad am^2 + bm + d = O(m^2), \qquad 5 \cdot 2^m + 3m^2 = O(2^m).$$

We want an algorithm $\mathcal{A}$ to be "efficient," that is, "good" with respect to

(i) *Time* (number $c_{\mathcal{A}}(m)$ of computer operations), or

(ii) *Space* (storage needed in the internal memory)

or both. Here $c_{\mathcal{A}}$ suggests **"complexity"** of $\mathcal{A}$. Two popular choices for $c_{\mathcal{A}}$ are

(*Worst case*) $c_{\mathcal{A}}(m)$ = longest time $\mathcal{A}$ takes for a problem of size m,

(*Average case*) $c_{\mathcal{A}}(m)$ = average time $\mathcal{A}$ takes for a problem of size m.

In problems on graphs, the "size" will often be m (number of edges) or n (number of vertices). For our present simple algorithm, $c_{\mathcal{A}}(m) = 2m$ in both cases.

For a "good" algorithm $\mathcal{A}$, we want that $c_{\mathcal{A}}(m)$ does not grow too fast. Accordingly, we call $\mathcal{A}$ **efficient** if $c_{\mathcal{A}}(m) = O(m^k)$ for some integer $k \geqq 0$; that is, $c_{\mathcal{A}}$ may contain only powers of m (or functions that grow even more slowly, such as $\ln m$), but no exponential functions. Furthermore, we call $\mathcal{A}$ **polynomially bounded** if $\mathcal{A}$ is efficient when we choose the "worst case" $c_{\mathcal{A}}(m)$. These conventional concepts have intuitive appeal, as our discussion shows.

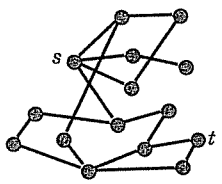
Complexity should be investigated for every algorithm, so that one can also compare different algorithms for the same task. This may often exceed the level in this chapter; accordingly, we shall confine ourselves to a few occasional comments in this direction.

PROBLEM SET 23.2

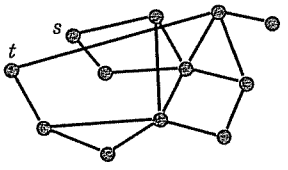
1–6 SHORTEST PATH

Find a shortest path P: $s \to t$ and its length by Moore's BFS algorithm; sketch the graph with the labels and indicate P by heavier lines (as in Fig. 481).

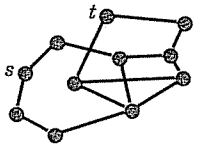
1.



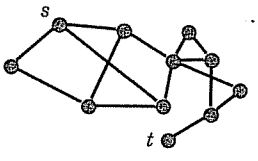
2.



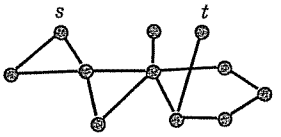
3.



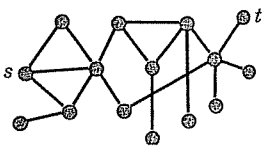
4.



5.



6.



7. **(Nonuniqueness)** A shortest path $s \to t$ for given s and t need not be unique. Illustrate this by finding another shortest path $s \to t$ in Example 1 in the text.

8. **(Maximum length)** If P is a shortest path between any two vertices in a graph with n vertices, how many edges can P at most have? In a complete graph (with all edges of length 1)? Give a reason.

9. **(Moore's algorithm)** Show that if a vertex v has label $\lambda(v) = k$, then there is a path $s \to v$ of length k.

10. Call the length of a shortest path $s \to v$ the **distance** of v from s. Show that if v has distance l, it has label $\lambda(v) = l$.

11. **(Hamiltonian cycle)** Find and sketch a Hamiltonian cycle in the graph of Prob. 3.

12. Find and sketch a Hamiltonian cycle in the graph of a dodecahedron, which has 12 pentagonal faces and 20 vertices (Fig. 482). This is a problem Hamilton himself considered.

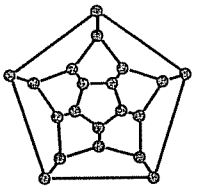


Fig. 482. Problem 12

13. Find and sketch a Hamiltonian cycle in Fig. 479, Sec. 23.1.

14. **(Euler graph)** An *Euler graph* G is a graph that has a closed Euler trail. An **Euler trail** is a trail that contains every edge of G exactly once. Which subgraph with four edges of the graph in Example 1, Sec. 23.1, is an Euler graph?

15. Is the graph in Fig. 483 an Euler graph? (Give a reason.)

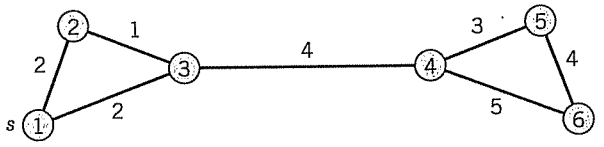


Fig. 483. Problems 15, 17

16. Find 4 different closed Euler trails in Fig. 484.

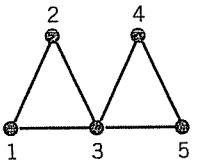


Fig. 484. Problem 16

17. The **postman problem** is the problem of finding a closed walk W: $s \to s$ (s the post office) in a graph G with edges (i, j) of length $l_{ij} > 0$ such that every edge of G is traversed at least once and the length of W is minimum. Find a solution for the graph in Fig. 483 by inspection. (The problem is also called the *Chinese postman problem* since it was published in the journal *Chinese Mathematics* 1 (1962), 273–277.)

18. Show that the length of a shortest postman trail is the same for every starting vertex.

19. **(Order)** Show that $O(m^3) + O(m^3) = O(m^3)$ and $kO(m^p) = O(m^p)$.

20. Show that $\sqrt{1 + m^2} = O(m)$, $0.02e^m + 100m^2 = O(e^m)$.

21. If we switch from one computer to another that is 100 times as fast, what is our gain in problem size per hour in the use of an algorithm that is $O(m)$, $O(m^2)$, $O(m^5)$, $O(e^m)$?

22. **CAS PROBLEM. Moore's Algorithm.** Write a computer program for the algorithm in Table 23.1. Test the program with the graph in Example 1. Apply it to Probs. 1–3 and to some graphs of your own choice.

23.3 Bellman's Principle. Dijkstra's Algorithm

We continue our discussion of the shortest path problem in a graph G. The last section concerned the special case that all edges had length 1. But in most applications the edges (i, j) will have any lengths $l_{ij} > 0$, and we now turn to this general case, which is of greater practical importance. We write $l_{ij} = \infty$ for any edge (i, j) that does not exist in G (setting $\infty + a = \infty$ for any number a, as usual).

We consider the problem of finding shortest paths from a given vertex, denoted by 1 and called the **origin,** to ***all*** other vertices 2, 3, $\cdots$, n of G. We let L_j denote the length of a shortest path P_j: $1 \to j$ in G.

THEOREM 1

Bellman's Minimality Principle or Optimality Principle[2]

If P_j: $1 \to j$ is a shortest path from 1 to j in G and (i, j) is the last edge of P_j (Fig. 485), *then P_i: $1 \to i$ [obtained by dropping (i, j) from P_j] is a shortest path $1 \to i$.*

[2]RICHARD BELLMAN (1920–1984), American mathematician, known for his work in dynamic programming.

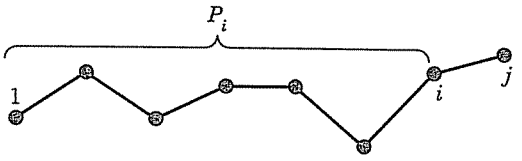


Fig. 485. Paths P and P_i in Bellman's minimality principle

PROOF Suppose that the conclusion is false. Then there is a path P_i^*: $1 \to i$ that is shorter than P_i. Hence if we now add (i, j) to P_i^*, we get a path $1 \to j$ that is shorter than P_j. This contradicts our assumption that P_j is shortest. ∎

From Bellman's principle we can derive basic equations as follows. For fixed j we may obtain various paths $1 \to j$ by taking shortest paths P_i for various i for which there is in G an edge (i, j), and add (i, j) to the corresponding P_i. These paths obviously have lengths $L_i + l_{ij}$ (L_i = length of P_i). We can now take the minimum over i, that is, pick an i for which $L_i + l_{ij}$ is smallest. By the Bellman principle, this gives a shortest path $1 \to j$. It has the length

$$\text{(1)} \qquad \begin{aligned} L_1 &= 0 \\ L_j &= \min_{i \neq j} (L_i + l_{ij}), \end{aligned} \qquad j = 2, \cdots, n.$$

These are the **Bellman equations.** Since $l_{ii} = 0$ by definition, instead of $\min_{i \neq j}$ we can simply write $\min_i$. These equations suggest the idea of one of the best-known algorithms for the shortest path problem, as follows.

Dijkstra's Algorithm for Shortest Paths

Dijkstra's[3] algorithm is shown in Table 23.2, where a **connected graph** G is a graph in which for any two vertices v and w in G there is a path $v \to w$. The algorithm is a labeling procedure. At each stage of the computation, each vertex v gets a label, either

$$\text{(PL) a } \textit{permanent label} = \text{length } L_v \text{ of a shortest path } 1 \to v$$

or

$$\text{(TL) a } \textit{temporary label} = \text{upper bound } \tilde{L}_v \text{ for the length of a shortest path } 1 \to v.$$

We denote by $\mathcal{PL}$ and $\mathcal{TL}$ the sets of vertices with a permanent label and with a temporary label, respectively. The algorithm has an initial step in which vertex 1 gets the permanent label $L_1 = 0$ and the other vertices get temporary labels, and then the algorithm alternates between Steps 2 and 3. In Step 2 the idea is to pick k "minimally." In Step 3 the idea is that the upper bounds will in general improve (decrease) and must be updated accordingly. Namely, the new temporary label $\tilde{L}_j$ of vertex j will be the old one if there is no improvement or it will be $L_k + l_{kj}$ if there is.

[3]EDSGER WYBE DIJKSTRA (1930–2002), Dutch computer scientist, 1972 recipient of the ACM Turing Award. His algorithm appeared in *Numerische Mathematik* **1** (1959), 269–271.

Table 23.2 Dijkstra's Algorithm for Shortest Paths

ALGORITHM DIJKSTRA [$G = (V, E)$, $V = \{1, \cdots, n\}$, l_{ij} for all (i, j) in E]

Given a connected graph $G = (V, E)$ with vertices $1, \cdots, n$ and edges (i, j) having lengths $l_{ij} > 0$, this algorithm determines the lengths of shortest paths from vertex 1 to the vertices $2, \cdots, n$.

INPUT: Number of vertices n, edges (i, j), and lengths l_{ij}

OUTPUT: Lengths L_j of shortest paths $1 \to j$, $j = 2, \cdots, n$

1. *Initial step*

Vertex 1 gets PL: $L_1 = 0$.
Vertex j $(= 2, \cdots, n)$ gets TL: $\tilde{L}_j = l_{1j}$ $(= \infty$ if there is no edge $(1, j)$ in $G)$.
Set $\mathcal{PL} = \{1\}$, $\mathcal{TL} = \{2, 3, \cdots, n\}$.

2. *Fixing a permanent label*

Find a k in $\mathcal{TL}$ for which $\tilde{L}_k$ is miminum, set $L_k = \tilde{L}_k$. Take the smallest k if there are several. Delete k from $\mathcal{TL}$ and include it in $\mathcal{PL}$.
If $\mathcal{TL} = \varnothing$ (that is, $\mathcal{TL}$ is empty) then

OUTPUT $L_2, \cdots, L_n$. Stop

Else continue (that is, go to Step 3).

3. *Updating temporary labels*

For all j in $\mathcal{TL}$, set $\tilde{L}_j = \min_k \{\tilde{L}_j, L_k + l_{kj}\}$ (that is, take the smaller of $\tilde{L}_j$ and $L_k + l_{kj}$ as your new $\tilde{L}_j$).

Go to Step 2.

End DIJKSTRA

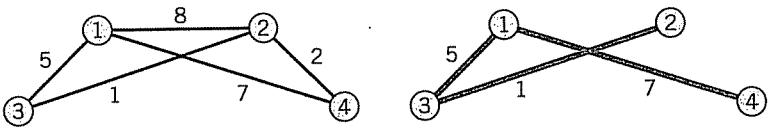
EXAMPLE 1 **Application of Dijkstra's Algorithm**

Applying Dijkstra's algorithm to the graph in Fig. 486a, find shortest paths from vertex 1 to vertices 2, 3, 4.

Solution. We list the steps and computations.

1.	$L_1 = 0, \tilde{L}_2 = 8, \tilde{L}_3 = 5, \tilde{L}_4 = 7,$	$\mathcal{PL} = \{1\},$	$\mathcal{TL} = \{2, 3, 4\}$
2.	$L_3 = \min\{\tilde{L}_2, \tilde{L}_3, \tilde{L}_4\} = 5, k = 3,$	$\mathcal{PL} = \{1, 3\},$	$\mathcal{TL} = \{2, 4\}$
3.	$\tilde{L}_2 = \min\{8, L_3 + l_{32}\} = \min\{8, 5 + 1\} = 6$		
	$\tilde{L}_4 = \min\{7, L_3 + l_{34}\} = \min\{7, \infty\} = 7$		
2.	$L_2 = \min\{\tilde{L}_2, \tilde{L}_4\} = \min\{6, 7\} = 6, k = 2,$	$\mathcal{PL} = \{1, 2, 3\},$	$\mathcal{TL} = \{4\}$
3.	$\tilde{L}_4 = \min\{7, L_2 + l_{24}\} = \min\{7, 6 + 2\} = 7$		
2.	$L_4 = 7, k = 4$	$\mathcal{PL} = \{1, 2, 3, 4\},$	$\mathcal{TL} = \varnothing.$

Figure 486b shows the resulting shortest paths, of lengths $L_2 = 6$, $L_3 = 5$, $L_4 = 7$. ■



(a) Given graph G (b) Shortest paths in G

Fig. 486. Example 1

Complexity. *Dijkstra's algorithm is* $O(n^2)$.

PROOF Step 2 requires comparison of elements, first $n-2$, the next time $n-3$, etc., a total of $(n-2)(n-1)/2$. Step 3 requires the same number of comparisons, a total of $(n-2)(n-1)/2$, as well as additions, first $n-2$, the next time $n-3$, etc., again a total of $(n-2)(n-1)/2$. Hence the total number of operations is $3(n-2)(n-1)/2 = O(n^2)$. ■

PROBLEM SET 23.3

1. The net of roads in Fig. 487 connecting four villages is to be reduced to minimum length, but so that one can still reach every village from every other village. Which of the roads should be retained? Find the solution (a) by inspection, (b) by Dijkstra's algorithm.

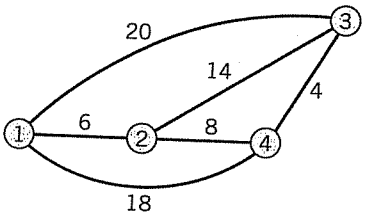
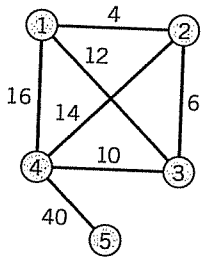


Fig. 487. Problem 1

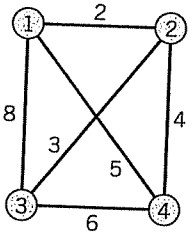
2–7 DIJKSTRA'S ALGORITHM

Find shortest paths for the following graphs.

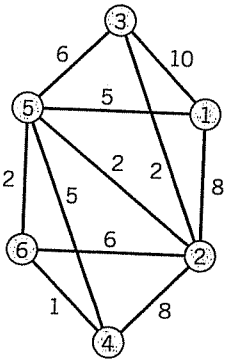
2.



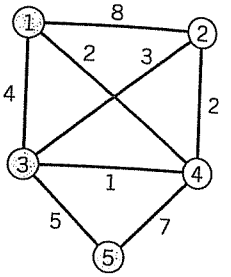
3.



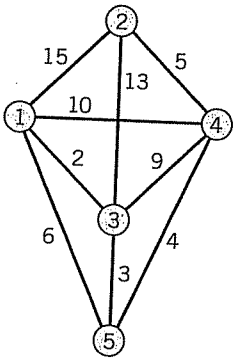
4.



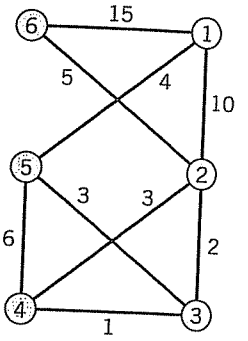
5.



6.



7.



8. Show that in Dijkstra's algorithm, for L_k there is a path P: $1 \to k$ of length L_k.

9. Show that in Dijkstra's algorithm, at each instant the demand on storage is light (data for less than n edges).

10. CAS PROBLEM. Dijkstra's Algorithm. Write a program and apply it to Probs. 2–4.

23.4 Shortest Spanning Trees: Greedy Algorithm

So far we have discussed shortest path problems. We now turn to a particularly important kind of graph, called a tree, along with related optimization problems that arise quite often in practice.

By definition, a **tree** T is a graph that is connected and has no cycles. **"Connected"** was defined in Sec. 23.3; it means that there is a path from any vertex in T to any other vertex in T. A **cycle** is a path $s \to t$ of at least three edges that is closed ($t = s$); see also Sec. 23.2. Figure 488a shows an example.

CAUTION! The terminology varies; *cycles* are sometimes also called *circuits*.

A **spanning tree** T in a given connected graph $G = (V, E)$ is a tree containing *all* the n vertices of G. See Fig. 488b. Such a tree has $n - 1$ edges. (Proof?)

A **shortest spanning tree** T in a connected graph G (whose edges (i, j) have lengths $l_{ij} > 0$) is a spanning tree for which Σl_{ij} (sum over all edges of T) is minimum compared to Σl_{ij} for any other spanning tree in G.

Trees are among the most important types of graphs, and they occur in various applications. Familiar examples are family trees and organization charts. Trees can be used to exhibit, organize, or analyze electrical networks, producer–consumer and other business relations, information in database systems, syntactic structure of computer programs, etc. We mention a few specific applications that need no lengthy additional explanations.

The set of shortest paths from vertex 1 to the vertices $2, \cdots, n$ in the last section forms a spanning tree.

Railway lines connecting a number of cities (the vertices) can be set up in the form of a spanning tree, the "length" of a line (edge) being the construction cost, and one wants to minimize the total construction cost. Similarly for bus lines, where "length" may be the average annual operating cost. Or for steamship lines (freight lines), where "length" may be profit and the goal is the maximization of total profit. Or in a network of telephone lines between some cities, a shortest spanning tree may simply represent a selection of lines that connect all the cities at minimal cost. In addition to these examples we could mention others from distribution networks, and so on.

We shall now discuss a simple algorithm for the problem of finding a shortest spanning tree. This algorithm (Table 23.3) is particularly suitable for sparse graphs (graphs with very few edges; see Sec. 23.1).

Table 23.3 Kruskal's Greedy Algorithm for Shortest Spanning Trees

Proceedings of the American Mathematical Society 7 (1956), 48–50.

ALGORITHM KRUSKAL [$G = (V, E)$, l_{ij} for all (i, j) in E]

Given a connected graph $G = (V, E)$ with edges (i, j) having length $l_{ij} > 0$, the algorithm determines a shortest spanning tree T in G.

INPUT: Edges (i, j) of G and their lengths l_{ij}

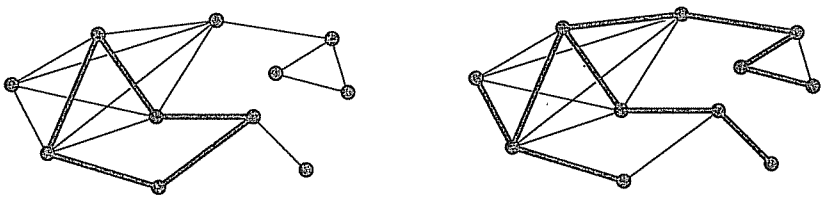
OUTPUT: Shortest spanning tree T in G

1. Order the edges of G in ascending order of length.
2. Choose them in this order as edges of T, rejecting an edge only if it forms a cycle with edges already chosen.

If $n - 1$ edges have been chosen, then

OUTPUT T (= the set of edges chosen). Stop

End KRUSKAL



(a) A cycle (b) A spanning tree

Fig. 488. Example of (a) a cycle, (b) a spanning tree in a graph

EXAMPLE 1 Application of Kruskal's Algorithm

Using Kruskal's algorithm, we shall determine a shortest spanning tree in the graph in Fig. 489.

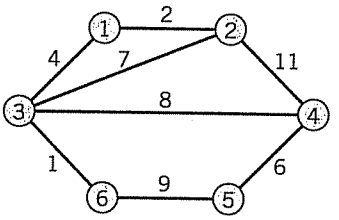


Fig. 489. Graph in Example 1

Solution. See Table 23.4. In some of the intermediate stages the edges chosen form a *disconnected* graph (see Fig. 490); this is typical. We stop after $n - 1 = 5$ choices since a spanning tree has $n - 1$ edges. In our problem the edges chosen are in the upper part of the list. This is typical of problems of any size; in general, edges farther down in the list have a smaller chance of being chosen. ▩

Table 23.4 Solution in Example 1

Edge	Length	Choice
(3, 6)	1	1st
(1, 2)	2	2nd
(1, 3)	4	3rd
(4, 5)	6	4th
(2, 3)	7	Reject
(3, 4)	8	5th
(5, 6)	9	
(2, 4)	11	

The efficiency of Kruskal's method is greatly increased by

Double Labeling of Vertices. *Each vertex i carries a double label (r_i, p_i), where*

$$r_i = \textit{Root of the subtree to which } i \textit{ belongs,}$$
$$p_i = \textit{Predecessor of } i \textit{ in its subtree,}$$
$$p_i = 0 \textit{ for roots.}$$

This simplifies

Rejecting. If (i, j) *is next in the list to be considered, reject* (i, j) *if* $r_i = r_j$ (that is, i and j are in the same subtree, so that they are already joined by edges and (i, j) would thus create a cycle). *If* $r_i \neq r_j$, *include* (i, j) *in* T.

If there are several choices for r_i, choose the smallest. If subtrees merge (become a single tree), retain the smallest root as the root of the new subtree.

For Example 1 the double-label list is shown in Table 23.5. In storing it, at each instant one may retain only the latest double label. We show all double labels in order to exhibit the process in all its stages. Labels that remain unchanged are not listed again. Underscored are the two 1's that are the common root of vertices 2 and 3, the reason for rejecting the edge (2, 3). By reading for each vertex the latest label we can read from this list that 1 is the vertex we have chosen as a root and the tree is as shown in the last part of Fig. 490.

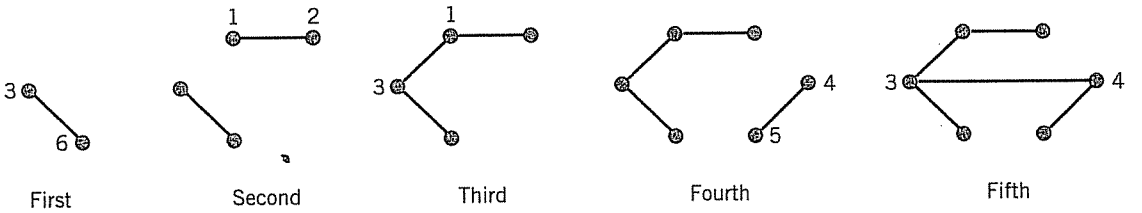


Fig. 490. Choice process in Example 1

This is made possible by the predecessor label that each vertex carries. Also, for accepting or rejecting an edge we have to make only one comparison (the roots of the two endpoints of the edge).

Ordering is the more expensive part of the algorithm. It is a standard process in data processing for which various methods have been suggested (see **Sorting** in Ref. [E25] listed in App. 1). For a complete list of m edges, an algorithm would be $O(m \log_2 m)$, but since the $n - 1$ edges of the tree are most likely to be found earlier, by inspecting the q ($< m$) topmost edges, for such a list of q edges one would have $O(q \log_2 m)$.

Table 23.5 List of Double Labels in Example 1

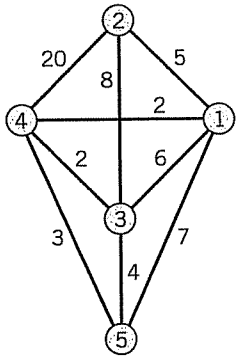
Vertex	Choice 1 (3, 6)	Choice 2 (1, 2)	Choice 3 (1, 3)	Choice 4 (4, 5)	Choice 5 (3, 4)
1		(1, 0)			
2		(1, 1)			
3	(3, 0)		(1, 1)		
4				(4, 0)	(1, 3)
5				(4, 4)	(1, 4)
6	(3, 3)		(1, 3)		

PROBLEM SET 23.4

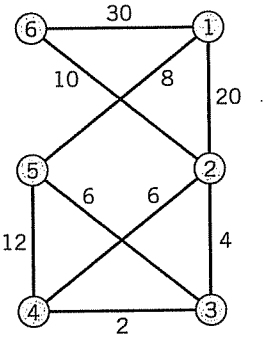
1–6 KRUSKAL'S ALGORITHM

Find a shortest spanning tree by Kruskal's algorithm.

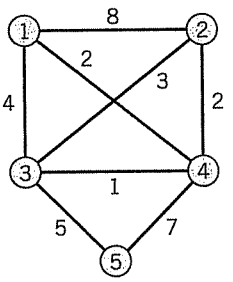
1.



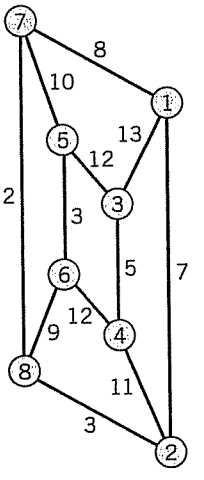
2.



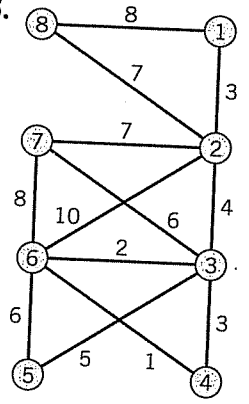
3.



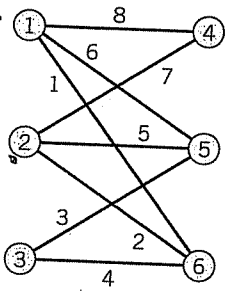
4.



5.



6.



7. CAS PROBLEM. Kruskal's Algorithm. Write a corresponding program. (Sorting is discussed in Ref. [E25] listed in App. 1.)

8. Design an algorithm for obtaining longest spanning trees.

9. Apply the algorithm in Prob. 8 to the graph in Example 1. Compare with the result in Example 1.

10. To get a minimum spanning tree, instead of adding shortest edges, one could think of deleting longest edges. For what graphs would this be feasible? Describe an algorithm for this.

11. Apply the method suggested in Prob. 10 to the graph in Example 1. Do you get the same tree?

12. Find a shortest spanning tree in the complete graph of all possible 15 connections between the six cities given (distances by airplane, in miles, rounded). Can you think of a practical application of the result?

	Dallas	Denver	Los Angeles	New York	Washington, DC
Chicago	800	900	1800	700	650
Dallas		650	1300	1350	1200
Denver			850	1650	1500
Los Angeles				2500	2350
New York					200

13. (Forest) A (not necessarily connected) graph without cycles is called a *forest.* Give typical examples of applications in which graphs occur that are forests or trees.

14–20 GENERAL PROPERTIES OF TREES

Prove:

14. (Uniqueness) The path connecting any two vertices u and v in a tree is unique.

15. If in a graph any two vertices are connected by a unique path, the graph is a tree.

16. If a graph has no cycles, it must have at least 2 vertices of degree 1 (definition in Sec. 23.1).

17. A tree with exactly two vertices of degree 1 must be a path.

18. A tree with n vertices has $n - 1$ edges. (Proof by induction.)

19. If two vertices in a tree are joined by a new edge, a cycle is formed.

20. A graph with n vertices is a tree if and only if it has $n - 1$ edges and has no cycles.

23.5 Shortest Spanning Trees: Prim's Algorithm

Prim's algorithm shown in Table 23.6 is another popular algorithm for the shortest spanning tree problem (see Sec. 23.4). This algorithm avoids ordering edges and gives a tree T at each stage, a property that Kruskal's algorithm in the last section did not have (look back at Fig. 490 if you did not notice it).

In Prim's algorithm, starting from any single vertex, which we call 1, we "grow" the tree T by adding edges to it, one at a time, according to some rule (in Table 23.6) until T finally becomes a *spanning* tree, which is shortest.

We denote by U the set of vertices of the growing tree T and by S the set of its edges. Thus, initially $U = \{1\}$ and $S = \emptyset$; at the end, $U = V$, the vertex set of the given graph $G = (V, E)$, whose edges (i, j) have length $l_{ij} > 0$, as before.

Thus at the beginning (Step 1) the labels

$$\lambda_2, \cdots, \lambda_n \qquad \text{of the vertices} \qquad 2, \cdots, n$$

are the lengths of the edges connecting them to vertex 1 (or ∞ if there is no such edge in G). And we pick (Step 2) the shortest of these as the first edge of the growing tree T and include its other end j in U (choosing the smallest j if there are several, to make the process unique). Updating labels in Step 3 (at this stage and at any later stage) concerns each vertex k not yet in U. Vertex k has label $\lambda_k = l_{i(k),k}$ from before. If $l_{jk} < \lambda_k$, this means that k is closer to the new member j just included in U than k is to its old "closest neighbor" $i(k)$ in U. Then we update the label of k, replacing $\lambda_k = l_{i(k),k}$ by $\lambda_k = l_{jk}$ and setting $i(k) = j$. If, however, $l_{jk} \geqq \lambda_k$ (the *old* label of k), we don't touch the old label. Thus the label λ_k always identifies the closest neighbor of k in U, and this is updated in Step 3 as U and the tree T grow. From the final labels we can backtrack the final tree, and from their numeric values we compute the total length (sum of the lengths of the edges) of this tree.

Table 23.6 Prim's Algorithm for Shortest Spanning Trees

Bell System Technical Journal **36** (1957), 1389–1401.

For an improved version of the algorithm, see Cheriton and Tarjan, *SIAM Journal on Computation* **5** (1976), 724–742.

ALGORITHM PRIM [$G = (V, E)$, $V = \{1, \cdots, n\}$, l_{ij} for all (i, j) in E]

Given a connected graph $G = (V, E)$ with vertices $1, 2, \cdots, n$ and edges (i, j) having length $l_{ij} > 0$, this algorithm determines a shortest spanning tree T in G and its length $L(T)$.

INPUT: n, edges (i, j) of G and their lengths l_{ij}
OUTPUT: Edge set S of a shortest spanning tree T in G; $L(T)$
[*Initially, all vertices are unlabeled.*]

1. *Initial step*
Set $i(k) = 1$, $U = \{1\}$, $S = \varnothing$.
Label vertex k ($= 2, \cdots, n$) with $\lambda_k = l_{ik}$ [$= \infty$ if G has no edge $(1, k)$].

2. *Addition of an edge to the tree T*
Let λ_j be the smallest λ_k for vertex k not in U. Include vertex j in U and edge $(i(j), j)$ in S.
If $U = V$ then compute
$L(T) = \Sigma l_{ij}$ (sum over all edges in S)
OUTPUT S, $L(T)$. Stop
[*S is the edge set of a shortest spanning tree T in G.*]
Else continue (that is, go to Step 3).

3. *Label updating*
For every k not in U, if $l_{jk} < \lambda_k$, then set $\lambda_k = l_{jk}$ and $i(k) = j$.
Go to Step 2.

End PRIM

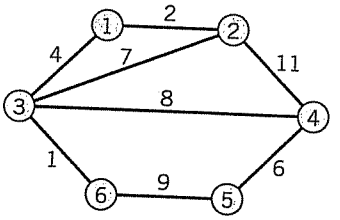


Fig. 491. Graph in Example 1

EXAMPLE 1 **Application of Prim's Algorithm**

Find a shortest spanning tree in the graph in Fig. 491 (which is the same as in Example 1, Sec. 23.4, so that we can compare).

Solution. The steps are as follows.

1. $i(k) = 1$, $U = \{1\}$, $S = \emptyset$, initial labels see Table 23.7.
2. $\lambda_2 = l_{12} = 2$ is smallest, $U = \{1, 2\}$, $S = \{(1, 2)\}$
3. Update labels as shown in Table 23.7, column (I).
2. $\lambda_3 = l_{13} = 4$ is smallest, $U = \{1, 2, 3\}$, $S = \{(1, 2), (1, 3)\}$
3. Update labels as shown in Table 23.7, column (II).
2. $\lambda_6 = l_{36} = 1$ is smallest, $U = \{1, 2, 3, 6\}$, $S = \{(1, 2), (1, 3), (3, 6)\}$
3. Update labels as shown in Table 23.7, column (III).
2. $\lambda_4 = l_{34} = 8$ is smallest, $U = \{1, 2, 3, 4, 6\}$, $S = \{(1, 2), (1, 3), (3, 4), (3, 6)\}$
3. Update labels as shown in Table 23.7, column (IV).
2. $\lambda_5 = l_{45} = 6$ is smallest, $U = V$, $S = (1, 2), (1, 3), (3, 4), (3, 6), (4, 5)$. Stop.

The tree is the same as in Example 1, Sec. 23.4. Its length is 21. You will find it interesting to compare the growth process of the present tree with that in Sec. 23.4. ■

Table 23.7 Labeling of Vertices in Example 1

Vertex	Initial Label	Relabeling			
		(I)	(II)	(III)	(IV)
2	$l_{12} = 2$	—	—	—	—
3	$l_{13} = 4$	$l_{13} = 4$	—	—	—
4	∞	$l_{24} = 11$	$l_{34} = 8$	$l_{34} = 8$	—
5	∞	∞	∞	$l_{65} = 9$	$l_{45} = 6$
6	∞	∞	$l_{36} = 1$	—	—

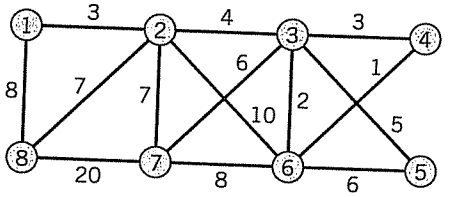
PROBLEM SET 23.5

1–7 PRIM'S ALGORITHM

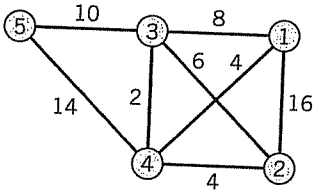
Find a shortest spanning tree by Prim's algorithm. Sketch it.

1. For the graph in Prob. 1, Sec. 23.4
2. For the graph in Prob. 2, Sec. 23.4
3. For the graph in Prob. 4, Sec. 23.4

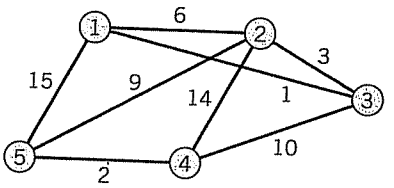
4.

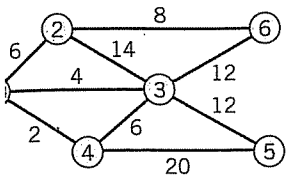


5.



6.





omplexity) Show that Prim's algorithm has nplexity $O(n^2)$.

w does Prim's algorithm prevent the formation of cles as one grows T?

Sec. 23.4

r a complete graph (or one that is almost complete), our data is an $n \times n$ distance table (as in Prob. 12, c. 23.4), show that the present algorithm [which is n^2)] cannot easily be replaced by an algorithm of ler less than $O(n^2)$.

what case will Prim's algorithm give $S = E$ as the al result?

AM PROJECT. Center of a Graph and Related ncepts. (a) Distance, eccentricity. Call the length a shortest path $u \to v$ in a graph $G = (V, E)$ the *tance* $d(u, v)$ from u to v. For fixed u, call the atest $d(u, v)$ as v ranges over V the *eccentricity* $\epsilon(u)$ u. Find the eccentricity of vertices 1, 2, 3 in the ph in Prob. 7.

sting to c

(b) Diameter, radius, center. The *diameter* $d(G)$ of a graph $G = (V, E)$ is the maximum of $d(u, v)$ as u and v vary over V, and the *radius* $r(G)$ is the smallest eccentricity $\epsilon(v)$ of the vertices v. A vertex v with $\epsilon(v) = r(G)$ is called a *central vertex*. The set of all central vertices is called the *center* of G. Find $d(G)$, $r(G)$ and the center of the graph in Prob. 7.

(c) What are the diameter, radius, and center of the spanning tree in Example 1?

(d) Explain how the idea of a center can be used in setting up an emergency service facility on a transportation network. In setting up a fire station, a shopping center. How would you generalize the concepts in the case of two or more such facilities?

(e) Show that a tree T whose edges all have length 1 has center consisting of either one vertex or two adjacent vertices.

(f) Set up an algorithm of complexity $O(n)$ for finding the center of a tree T.

13. What would the result be if you applied Prim's algorithm to a graph that is not connected?

14. CAS PROBLEM. Prim's Algorithm. Write a program and apply it to Probs. 4–6.

6 Flows in Networks

After shortest path problems and problems for trees, as a third large area in combinatorial optimization we discuss **flow problems in networks** (electrical, water, communication, traffic, business connections, etc.), turning from graphs to digraphs (directed graphs; see Sec. 23.1).

By definition, a **network** is a digraph $G = (V, E)$ in which each edge (i, j) has assigned to it a **capacity** $c_{ij} > 0$ [= maximum possible flow along (i, j)], and at one vertex, s, called the **source,** a flow is produced that flows along the edges of the digraph G to another vertex, t, called the **target** or **sink,** where the flow disappears.

In applications, this may be the flow of electricity in wires, of water in pipes, of cars on roads, of people in a public transportation system, of goods from a producer to consumers, of e-mail from senders to recipients over the Internet, and so on.

We denote the flow along a (directed!) edge (i, j) by f_{ij} and impose two conditions:

1. For each edge (i, j) in G the flow does not exceed the capacity c_{ij},

$$\textbf{(1)} \qquad 0 \leqq f_{ij} \leqq c_{ij} \qquad \textbf{("Edge condition").}$$

2. For each vertex i, not s or t,

$$\text{Inflow} = \text{Outflow} \qquad \textbf{("Vertex condition," "Kirchhoff's law");}$$

in a formula,

$$(2) \qquad \underbrace{\sum_k f_{ki}}_{\text{Inflow}} - \underbrace{\sum_j f_{ij}}_{\text{Outflow}} = \begin{cases} 0 \text{ if vertex } i \neq s,\ i \neq t, \\ -f \text{ at the source } s, \\ f \text{ at the target (sink) } t, \end{cases}$$

where f is the total flow (and at s the inflow is zero, whereas at t the outflow is zero). Figure 492 illustrates the notation (for some hypothetical figures).

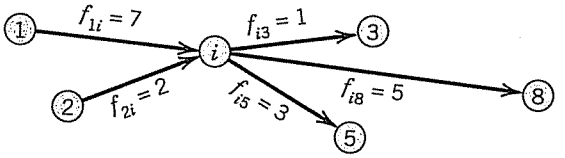


Fig. 492. Notation in (2): inflow and outflow for a vertex i (not s or t)

Paths

By a **path** $v_1 \to v_k$ from a vertex v_1 to a vertex v_k in a digraph G we mean a sequence of edges

$$(v_1, v_2), (v_2, v_3), \cdots, (v_{k-1}, v_k),$$

regardless of their directions in G, that forms a path as in a graph (see Sec. 23.2). Hence when we travel along this path from v_1 to v_k we may traverse some edge ***in*** its given direction—then we call it a **forward edge** of our path—or ***opposite to*** its given direction—then we call it a **backward edge** of our path. In other words, our path consists of one-way streets, and forward edges (backward edges) are those that we travel *in the right direction* (*in the wrong direction*). Figure 493 shows a forward edge (u, v) and a backward edge (w, v) of a path $v_1 \to v_k$.

CAUTION! Each edge in a network has a given direction, *which we cannot change.* Accordingly, if (u, v) is a forward edge in a path $v_1 \to v_k$, then (u, v) can become a backward edge only in another path $x_1 \to x_j$ in which it is an edge and is traversed in the opposite direction as one goes from x_1 to x_j; see Fig. 494. Keep this in mind, to avoid misunderstandings.

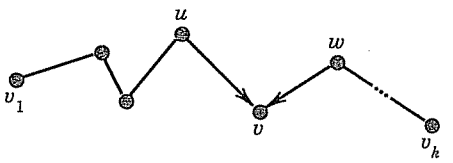


Fig. 493. Forward edge (u, v) and backward edge (w, v) of a path $v_1 \to v_k$

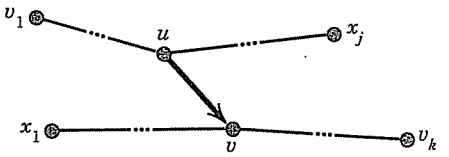


Fig. 494. Edge (u, v) as forward edge in the path $v_1 \to v_k$ and as backward edge in the path $x_1 \to x_j$

Flow Augmenting Paths

Our goal will be to ***maximize the flow*** from the source s to the target t of a given network. We shall do this by developing methods for increasing an existing flow (including the special case in which the latter is zero). The idea then is to find a path P: $s \to t$ all of whose edges are not fully used, so that we can push additional flow through P. This suggests the following concept.

DEFINITION

Flow Augmenting Path

A *flow augmenting path* in a network with a given flow f_{ij} on each edge (i, j) is a path P: $s \to t$ such that

(i) no forward edge is used to capacity; thus $f_{ij} < c_{ij}$ for these;

(ii) no backward edge has flow 0; thus $f_{ij} > 0$ for these.

EXAMPLE 1 **Flow Augmenting Paths**

Find flow augmenting paths in the network in Fig. 495, where the first number is the capacity and the second number a given flow.

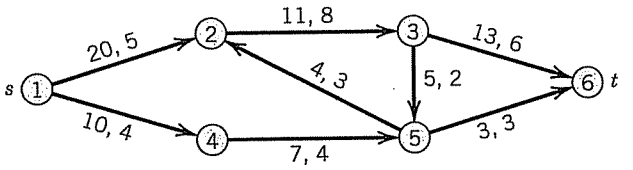


Fig. 495. Network in Example 1
First number = Capacity, Second number = Given flow

Solution. In practical problems, networks are large and one needs a ***systematic method for augmenting flows, which we discuss in the next section.*** In our small network, which should help to illustrate and clarify the concepts and ideas, we can find flow augmenting paths by inspection and augment the existing flow $f = 9$ in Fig. 495. (The outflow from s is $5 + 4 = 9$, which equals the inflow $6 + 3$ into t.)

We use the notation

$$\Delta_{ij} = c_{ij} - f_{ij} \qquad \text{for forward edges}$$

$$\Delta_{ij} = f_{ij} \qquad \text{for backward edges}$$

$$\Delta = \min \Delta_{ij} \qquad \text{taken over all edges of a path.}$$

From Fig. 495 we see that a flow augmenting path P_1: $s \to t$ is P_1: $1 - 2 - 3 - 6$ (Fig. 496), with $\Delta_{12} = 20 - 5 = 15$, etc., and $\Delta = 3$. Hence we can use P_1 to increase the given flow 9 to $f = 9 + 3 = 12$. All three edges of P_1 are forward edges. We augment the flow by 3. Then the flow in each of the edges of P_1 is increased by 3, so that we now have $f_{12} = 8$ (instead of 5), $f_{23} = 11$ (instead of 8), and $f_{36} = 9$ (instead of 6). Edge (2, 3) is now used to capacity. The flow in the other edges remains as before.

We shall now try to increase the flow in this network in Fig. 495 beyond $f = 12$.

There is another flow augmenting path P_2: $s \to t$, namely, P_2: $1 - 4 - 5 - 3 - 6$ (Fig. 496). It shows how a backward edge comes in and how it is handled. Edge (3, 5) is a backward edge. It has flow 2, so that $\Delta_{35} = 2$. We compute $\Delta_{14} = 10 - 4 = 6$, etc. (Fig. 496) and $\Delta = 2$. Hence we can use P_2 for another augmentation to get $f = 12 + 2 = 14$. The new flow is shown in Fig. 497. No further augmentation is possible. We shall confirm later that $f = 14$ is maximum. ■

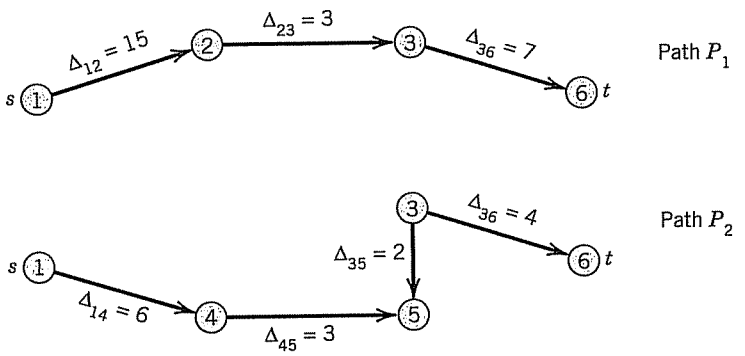


Fig. 496. Flow augmenting paths in Example 1

Cut Sets

A "cut set" is a set of edges in a network. The underlying idea is simple and natural. If we want to find out what is flowing from s to t in a network, we may cut the network somewhere between s and t (Fig. 497 shows an example) and see what is flowing in the edges hit by the cut, because any flow from s to t must sometimes pass through some of these edges. These form what is called a **cut set.** [In Fig. 497, the cut set consists of the edges (2, 3), (5, 2), (4, 5).] We denote this cut set by (S, T). Here S is the set of vertices on that side of the cut on which s lies ($S = \{s, 2, 4\}$ for the cut in Fig. 497) and T is the set of the other vertices ($T = \{3, 5, t\}$ in Fig. 497). We say that a cut ***"partitions"*** the vertex set V into two parts S and T. Obviously, the corresponding cut set (S, T) consists of all the edges in the network with one end in S and the other end in T.

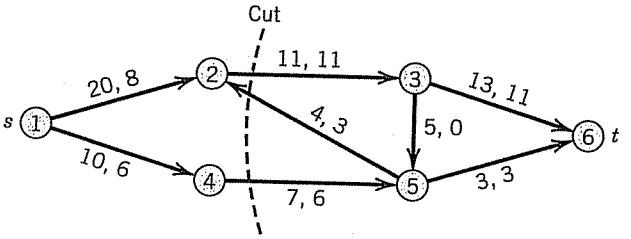


Fig. 497. Maximum flow in Example 1

By definition, the **capacity** cap (S, T) of a cut set (S, T) is the sum of the capacities of all **forward edges** in (S, T) (forward edges only!), that is, the edges that are directed *from S to T,*

(3) $$\text{cap}\,(S, T) = \Sigma c_{ij} \qquad \text{[sum over the forward edges of } (S, T)\text{]}.$$

Thus, cap $(S, T) = 11 + 7 = 18$ in Fig. 497.

The other edges (directed *from T to S*) are called **backward edges** of the cut set (S, T), and by the **net flow** through a cut set we mean the sum of the flows in the forward edges minus the sum of the flows in the backward edges of the cut set.

CAUTION! Distinguish well between forward and backward edges in a cut set and in a path: (5, 2) in Fig. 497 is a backward edge for the cut shown but a forward edge in the path $1 - 4 - 5 - 2 - 3 - 6$.

For the cut in Fig. 497 the net flow is $11 + 6 - 3 = 14$. For the same cut in Fig. 495 (not indicated there), the net flow is $8 + 4 - 3 = 9$. In both cases it equals the flow f. We claim that this is not just by chance, but cuts do serve the purpose for which we have introduced them:

THEOREM 1

Net Flow in Cut Sets

Any given flow in a network G is the net flow through any cut set (S, T) of G.

PROOF By Kirchhoff's law (2), multiplied by -1, at a vertex i we have

(4) $$\underbrace{\sum_j f_{ij}}_{\text{Outflow}} - \underbrace{\sum_l f_{li}}_{\text{Inflow}} = \begin{cases} 0 & \text{if } i \neq s, t, \\ f & \text{if } i = s. \end{cases}$$

Here we can sum over j and l from 1 to n (= number of vertices) by putting $f_{ij} = 0$ for $j = i$ and also for edges without flow or nonexisting edges; hence we can write the two sums as one,

$$\sum_j (f_{ij} - f_{ji}) = \begin{cases} 0 & \text{if } i \neq s, t, \\ f & \text{if } i = s. \end{cases}$$

We now sum over all i in S. Since s is in S, this sum equals f:

$$\sum_{i \in S} \sum_{j \in V} (f_{ij} - f_{ji}) = f. \tag{5}$$

We claim that in this sum, only the edges belonging to the cut set contribute. Indeed, edges with both ends in T cannot contribute, since we sum only over i in S; but edges (i, j) with both ends in S contribute $+f_{ij}$ at one end and $-f_{ij}$ at the other, a total contribution of 0. Hence the left side of (5) equals the net flow through the cut set. By (5), this is equal to the flow f and proves the theorem. ∎

This theorem has the following consequence, which we shall also need later in this section.

THEOREM 2

Upper Bound for Flows

A flow f in a network G cannot exceed the capacity of any cut set (S, T) in G.

PROOF By Theorem 1 the flow f equals the net flow through the cut set, $f = f_1 - f_2$, where f_1 is the sum of the flows through the forward edges and f_2 ($\geqq 0$) is the sum of the flows through the backward edges of the cut set. Thus $f \leqq f_1$. Now f_1 cannot exceed the sum of the capacities of the forward edges; but this sum equals the capacity of the cut set, by definition. Together, $f \leqq \text{cap } (S, T)$, as asserted. ∎

Cut sets will now bring out the full importance of augmenting paths:

THEOREM 3

Main Theorem. Augmenting Path Theorem for Flows

A flow from s to t in a network G is maximum if and only if there does not exist a flow augmenting path $s \to t$ in G.

PROOF **(a)** If there is a flow augmenting path P: $s \to t$, we can use it to push through it an additional flow. Hence the given flow cannot be maximum.

(b) On the other hand, suppose that there is no flow augmenting path $s \to t$ in G. Let S_0 be the set of all vertices i (including s) such that there is a flow augmenting path $s \to i$, and let T_0 be the set of the other vertices in G. Consider any edge (i, j) with i in S_0 and j in T_0. Then we have a flow augmenting path $s \to i$ since i is in S_0, but $s \to i \to j$ is not flow augmenting because j is not in S_0. Hence we must have

$$f_{ij} = \begin{cases} c_{ij} \\ 0 \end{cases} \quad \text{if } (i, j) \text{ is a } \begin{cases} \text{forward} \\ \text{backward} \end{cases} \text{ edge of the path } s \to i \to j. \tag{6}$$

Otherwise we could use (i, j) to get a flow augmenting path $s \to i \to j$. Now (S_0, T_0) defines a cut set (since t is in T_0; why?). Since by (6), forward edges are used to capacity and backward edges carry no flow, the net flow through the cut set (S_0, T_0) equals the sum of the capacities of the forward edges, which is cap (S_0, T_0) by definition. This net flow equals the given flow f by Theorem 1. Thus $f = \text{cap}\ (S_0, T_0)$. We also have $f \leqq \text{cap}\ (S_0, T_0)$ by Theorem 2. Hence f must be maximum since we have reached equality. ∎

The end of this proof yields another basic result (by Ford and Fulkerson, *Canadian Journal of Mathematics* **8** (1956), 399–404), namely, the so-called

THEOREM 4

Max-Flow Min-Cut Theorem

The maximum flow in any network G equals the capacity of a **"minimum cut set"** (= a cut set of minimum capacity) *in G.*

PROOF We have just seen that $f = \text{cap}\ (S_0, T_0)$ for a maximum flow f and a suitable cut set (S_0, T_0). Now by Theorem 2 we also have $f \leqq \text{cap}\ (S, T)$ for this f and any cut set (S, T) in G. Together, cap $(S_0, T_0) \leqq$ cap (S, T). Hence (S_0, T_0) is a minimum cut set.

The existence of a maximum flow in this theorem follows for rational capacities from the algorithm in the next section and for arbitrary capacities from the Edmonds–Karp BFS also in that section. ∎

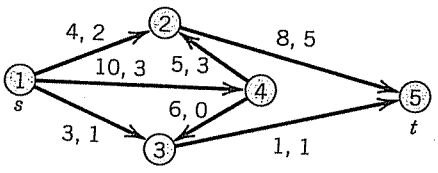
The two basic tools in connection with networks are flow augmenting paths and cut sets. In the next section we show how flow augmenting paths can be used in an algorithm for maximum flows.

PROBLEM SET 23.6

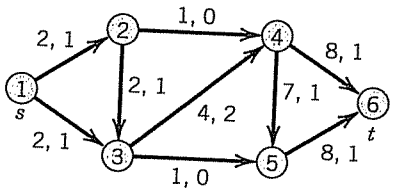
1–4 FLOW AUGMENTING PATHS

Find flow augmenting paths:

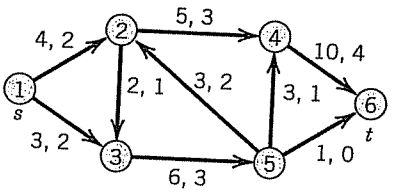
1.



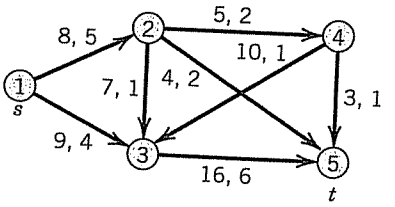
2.



3.



4.



5–8 MAXIMUM FLOW

Find the maximum flow by inspection:

5. In Prob. 1.
6. In Prob. 2.
7. In Prob. 3.
8. In Prob. 4.

9–11 CAPACITY

In Fig. 495 find T and cap (S, T) if S equals

9. $\{1, 2, 3\}$
10. $\{1, 2, 4, 5\}$
11. $\{1, 3, 5\}$

12. Find a minimum cut set in Fig. 495 and verify that its capacity equals the maximum flow $f = 14$.
13. Find examples of flow augmenting paths and the maximum flow in the network in Fig. 498.

14–16 CAPACITY

In Fig. 498 find T and cap (S, T) if S equals

14. $\{1, 2, 4\}$
15. $\{1, 2, 4, 6\}$
16. $\{1, 2, 3, 4, 5\}$

17. In Fig. 498 find a minimum cut set and its capacity.

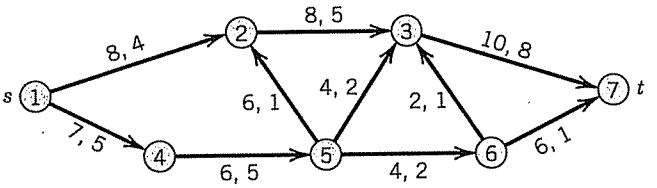


Fig. 498. Problems 13–17

18. Why are backward edges not considered in the definition of the capacity of a cut set?
19. In which case can an edge (i, j) be used as a forward as well as a backward edge of a path in a network with a given flow?
20. **(Incremental network)** Sketch the network in Fig. 498, and on each edge (i, j) write $c_{ij} - f_{ij}$ and f_{ij}. Do you recognize that from this "incremental network" one can more easily see flow augmenting paths?

23.7 Maximum Flow: Ford–Fulkerson Algorithm

Flow augmenting paths, as discussed in the last section, are used as the basic tool in the Ford–Fulkerson[4] algorithm in Table 23.8 on the next page in which a given flow (for instance, zero flow in all edges) is increased until it is maximum. The algorithm accomplishes the increase by a stepwise construction of flow augmenting paths, one at a time, until no further such paths can be constructed, which happens precisely when the flow is maximum.

In Step 1, an initial flow may be given. In Step 3, a vertex j can be labeled if there is an edge (i, j) with i labeled and

$$c_{ij} > f_{ij} \qquad \textbf{("forward edge")}$$

or if there is an edge (j, i) with i labeled and

$$f_{ji} > 0 \qquad \textbf{("backward edge").}$$

To **scan** a labeled vertex i means to label every unlabeled vertex j adjacent to i that can be labeled. Before scanning a labeled vertex i, scan all the vertices that got labeled before i. This **BFS (Breadth First Search)** strategy was suggested by Edmonds and Karp in 1972 (*Journal of the Association for Computing Machinery* **19,** 248–64). It has the effect that one gets shortest possible augmenting paths.

[4]LESTER RANDOLPH FORD (born 1927) and DELBERT RAY FULKERSON (1924–1976), American mathematicians known for their pioneering work on flow algorithms.

Table 23.8 Ford–Fulkerson Algorithm for Maximum Flow
Canadian Journal of Mathematics **9** (1957), 210–218

ALGORITHM FORD–FULKERSON

[$G = (V, E)$, vertices 1 ($= s$), $\cdots$, n ($= t$), edges (i, j), c_{ij}]
This algorithm computes the maximum flow in a network G with source s, sink t, and capacities $c_{ij} > 0$ of the edges (i, j).

INPUT: n, $s = 1$, $t = n$, edges (i, j) of G, c_{ij}
OUTPUT: Maximum flow f in G

1. Assign an initial flow f_{ij} (for instance, $f_{ij} = 0$ for all edges), compute f.
2. Label s by $\varnothing$. Mark the other vertices *"unlabeled."*
3. Find a labeled vertex i that has not yet been scanned. Scan i as follows. For every unlabeled adjacent vertex j, if $c_{ij} > f_{ij}$, compute

$$\Delta_{ij} = c_{ij} - f_{ij} \quad \text{and} \quad \Delta_j = \begin{cases} \Delta_{1j} & \text{if } i = 1 \\ \min(\Delta_i, \Delta_{ij}) & \text{if } i > 1 \end{cases}$$

and label j with a *"forward label"* (i^+, Δ_j); or if $f_{ji} > 0$, compute

$$\Delta_j = \min(\Delta_i, f_{ji})$$

and label j by a "backward label" (i^-, Δ_j).

If no such j exists then OUTPUT f. Stop

[*f is the maximum flow.*]

Else continue (that is, go to Step 4).

4. Repeat Step 3 until t is reached.

[*This gives a flow augmenting path P: $s \to t$.*]

If it is impossible to reach t then OUTPUT f. Stop

[*f is the maximum flow.*]

Else continue (that is, go to Step 5).

5. Backtrack the path P, using the labels.
6. Using P, augment the existing flow by Δ_t. Set $f = f + \Delta_t$.
7. Remove all labels from vertices 2, $\cdots$, n. Go to Step 3.

End FORD–FULKERSON

EXAMPLE 1 Ford–Fulkerson Algorithm

Applying the Ford–Fulkerson algorithm, determine the maximum flow for the network in Fig. 499 (which is the same as that in Example 1, Sec. 23.6, so that we can compare).

Solution. The algorithm proceeds as follows.

1. An initial flow $f = 9$ is given.
2. Label s ($= 1$) by $\varnothing$. Mark 2, 3, 4, 5, 6 "unlabeled."

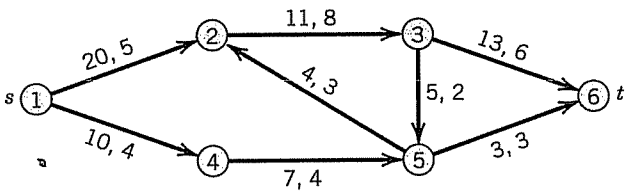


Fig. 499. Network in Example 1 with capacities (first numbers) and given flow

3. Scan 1.

Compute $\Delta_{12} = 20 - 5 = 15 = \Delta_2$. Label 2 by $(1^+, 15)$.

Compute $\Delta_{14} = 10 - 4 = 6 = \Delta_4$. Label 4 by $(1^+, 6)$.

4. Scan 2.

Compute $\Delta_{23} = 11 - 8 = 3$, $\Delta_3 = \min(\Delta_2, 3) = 3$. Label 3 by $(2^+, 3)$.

Compute $\Delta_5 = \min(\Delta_2, 3) = 3$. Label 5 by $(2^-, 3)$.

Scan 3.

Compute $\Delta_{36} = 13 - 6 = 7$, $\Delta_6 = \Delta_t = \min(\Delta_3, 7) = 3$. Label 6 by $(3^+, 3)$.

5. P: $1 - 2 - 3 - 6$ $(= t)$ is a flow augmenting path.

6. $\Delta_t = 3$. Augmentation gives $f_{12} = 8$, $f_{23} = 11$, $f_{36} = 9$, other f_{ij} unchanged. Augmented flow $f = 9 + 3 = 12$.

7. Remove labels on vertices $2, \cdots, 6$. Go to Step 3.

3. Scan 1.

Compute $\Delta_{12} = 20 - 8 = 12 = \Delta_2$. Label 2 by $(1^+, 12)$.

Compute $\Delta_{14} = 10 - 4 = 6 = \Delta_4$. Label 4 by $(1^+, 6)$.

4. Scan 2.

Compute $\Delta_5 = \min(\Delta_2, 3) = 3$. Label 5 by $(2^-, 3)$.

Scan 4. [*No vertex left for labeling.*]

Scan 5.

Compute $\Delta_3 = \min(\Delta_5, 2) = 2$. Label 3 by $(5^-, 2)$.

Scan 3.

Compute $\Delta_{36} = 13 - 9 = 4$, $\Delta_6 = \min(\Delta_3, 4) = 2$. Label 6 by $(3^+, 2)$.

5. P: $1 - 2 - 5 - 3 - 6$ $(= t)$ is a flow augmenting path.

6. $\Delta_t = 2$. Augmentation gives $f_{12} = 10$, $f_{52} = 1$, $f_{35} = 0$, $f_{36} = 11$, other f_{ij} unchanged. Augmented flow $f = 12 + 2 = 14$.

7. Remove labels on vertices $2, \cdots, 6$. Go to Step 3.

One can now scan 1 and then scan 2, as before, but in scanning 4 and then 5 one finds that no vertex is left for labeling. Thus one can no longer reach t. Hence the flow obtained (Fig. 500) is maximum, in agreement with our result in the last section. ▩

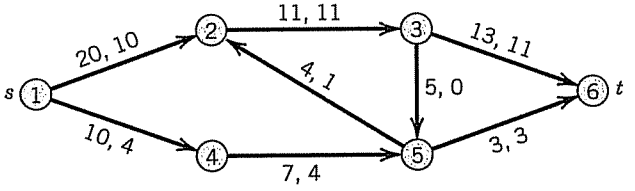


Fig. 500. Maximum flow in Example 1

PROBLEM SET 23.7

1. Do the computations indicated near the end of Example 1 in detail.
2. Solve Example 1 by Ford–Fulkerson with initial flow 0. Is it more work than in Example 1?
3. Which are the "bottleneck" edges by which the flow in Example 1 is actually limited? Hence which capacities could be decreased without decreasing the maximum flow?

4–7 MAXIMUM FLOW

Find the maximum flow by Ford–Fulkerson:

4. In Prob. 2, Sec. 23.6.
5. In Prob. 1, Sec. 23.6.
6. In Prob. 4, Sec. 23.6.
7. In Prob. 3, Sec. 23.6.
8. What is the (simple) reason that Kirchhoff's law is preserved in augmenting a flow by the use of a flow augmenting path?
9. How does Ford–Fulkerson prevent the formation of cycles?
10. How can you see that Ford–Fulkerson follows a BFS technique?
11. Are the consecutive flow augmenting paths produced by Ford–Fulkerson unique?
12. **(Integer flow theorem)** Prove that if the capacities in a network G are integers, then a maximum flow exists and is an integer.
13. **CAS PROBLEM. Ford–Fulkerson.** Write a program and apply it to Probs. 4–7.
14. If the Ford–Fulkerson algorithm stops without reaching t, show that the edges with one end labeled and the other end unlabeled form a cut set (S, T) whose capacity equals the maximum flow.
15. **(Several sources and sinks)** If a network has several sources $s_1, \cdots, s_k$, show that it can be reduced to the case of a single-source network by introducing a new vertex s and connecting s to $s_1, \cdots, s_k$ by k edges of capacity ∞. Similarly if there are several sinks. Illustrate this idea by a network with two sources and two sinks.
16. Find the maximum flow in the network in Fig. 501 with two sources (factories) and two sinks (consumers).
17. Find a minimum cut set in Fig. 499 and its capacity.
18. Show that in a network G with all $c_{ij} = 1$, the maximum flow equals the number of edge-disjoint paths $s \to t$.
19. In Prob. 17, the cut set contains precisely all forward edges used to capacity by the maximum flow (Fig. 500). Is this just by chance?
20. Show that in a network G with capacities all equal to 1, the capacity of a minimum cut set (S, T) equals the minimum number q of edges whose deletion destroys all directed paths $s \to t$. (A **directed path** $v \to w$ is a path in which each edge has the direction in which it is traversed in going from v to w.)

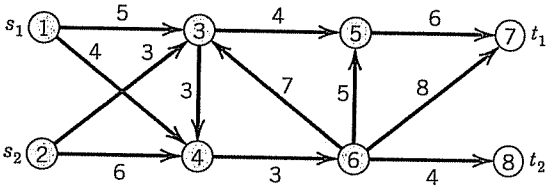


Fig. 501. Problem 16

23.8 Bipartite Graphs. Assignment Problems

From digraphs we return to graphs and discuss another important class of combinatorial optimization problems that arises in **assignment problems** of workers to jobs, jobs to machines, goods to storage, ships to piers, classes to classrooms, exams to time periods, and so on. To explain the problem, we need the following concepts.

A **bipartite graph** $G = (V, E)$ is a graph in which the vertex set V is partitioned into two sets S and T (without common elements, by the definition of a partition) such that every edge of G has one end in S and the other in T. Hence there are no edges in G that have both ends in S or both ends in T. Such a graph $G = (V, E)$ is also written $G = (S, T; E)$.

Figure 502 shows an illustration. V consists of seven elements, three workers a, b, c, making up the set S, and four jobs 1, 2, 3, 4, making up the set T. The edges indicate that worker a can do the jobs 1 and 2, worker b the jobs 1, 2, 3, and worker c the job 4. The problem is to assign one job to each worker so that every worker gets one job to do. This suggests the next concept, as follows.

DEFINITION

Maximum Cardinality Matching

A **matching** in $G = (S, T; E)$ is a set M of edges of G such that no two of them have a vertex in common. If M consists of the greatest possible number of edges, we call it a **maximum cardinality matching** in G.

For instance, a matching in Fig. 502 is $M_1 = \{(a, 2), (b, 1)\}$. Another is $M_2 = \{(a, 1), (b, 3), (c, 4)\}$; obviously, this is of maximum cardinality.

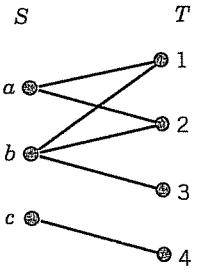


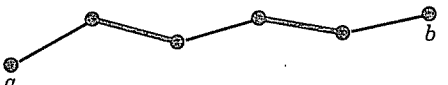
Fig. 502. Bipartite graph in the assignment of a set $S = \{a, b, c\}$ of workers to a set $T = \{1, 2, 3, 4\}$ of jobs

A vertex v is **exposed** (or *not covered*) by a matching M if v is not an endpoint of an edge of M. This concept, which always refers to some matching, will be of interest when we begin to augment given matchings (below). If a matching leaves no vertex exposed, we call it a **complete matching.** Obviously, a complete matching can exist only if S and T consist of the same number of vertices.

We now want to show how one can stepwise increase the cardinality of a matching M until it becomes maximum. Central in this task is the concept of an augmenting path.

An **alternating path** is a path that consists alternately of edges in M and not in M (Fig. 503A). An **augmenting path** is an alternating path both of whose endpoints (a and b in Fig. 503B) are exposed. By dropping from the matching M the edges that are on an augmenting path P (two edges in Fig. 503B) and adding to M the other edges of P (three in the figure), we get a new matching, with one more edge than M. This is how we use an augmenting path in ***augmenting a given matching*** by one edge. We assert that this will always lead, after a number of steps, to a maximum cardinality matching. Indeed, the basic role of augmenting paths is expressed in the following theorem.

(A) Alternating path



(B) Augmenting path P

Fig. 503. Alternating and augmenting paths. Heavy edges are those belonging to a matching M.

THEOREM 1

Augmenting Path Theorem for Bipartite Matching

A matching M in a bipartite graph $G = (S, T; E)$ is of maximum cardinality if and only if there does not exist an augmenting path P with respect to M.

PROOF **(a)** We show that if such a path P exists, then M is not of maximum cardinality. Let P have q edges belonging to M. Then P has $q + 1$ edges not belonging to M. (In Fig. 503B we have $q = 2$.) The endpoints a and b of P are exposed, and all the other vertices on P are endpoints of edges in M, by the definition of an alternating path. Hence if an edge of M is not an edge of P, it cannot have an endpoint on P since then M would not be a matching. Consequently, the edges of M not on P, together with the $q + 1$ edges of P not belonging to M form a matching of cardinality one more than the cardinality of M because we omitted q edges from M and added $q + 1$ instead. Hence M cannot be of maximum cardinality.

(b) We now show that if there is no augmenting path for M, then M is of maximum cardinality. Let M^* be a maximum cardinality matching and consider the graph H consisting of all edges that belong either to M or to M^*, but not to both. Then it is possible that two edges of H have a vertex in common, but three edges cannot have a vertex in common since then two of the three would have to belong to M (or to M^*), violating that M and M^* are matchings. So every v in V can be in common with two edges of H or with one or none. Hence we can characterize each "component" (= maximal *connected* subset) of H as follows.

(A) A component of H can be a closed path with an *even* number of edges (in the case of an *odd* number, two edges from M or two from M^* would meet, violating the matching property). See (A) in Fig. 504.

(B) A component of H can be an open path P with the same number of edges from M and edges from M^*, for the following reason. P must be alternating, that is, an edge of M is followed by an edge of M^*, etc. (since M and M^* are matchings). Now if P had an edge more from M^*, then P would be augmenting for M [see (B2) in Fig. 504], contradicting our assumption that there is no augmenting path for M. If P had an edge more from M, it would be augmenting for M^* [see (B3) in Fig. 504], violating the maximum cardinality of M^*, by part (a) of this proof. Hence in each component of H, the two matchings have the same number of edges. Adding to this the number of edges that belong to both M and M^* (which we left aside when we made up H), we conclude that M and M^* must have the same number of edges. Since M^* is of maximum cardinality, this shows that the same holds for M, as we wanted to prove. ∎

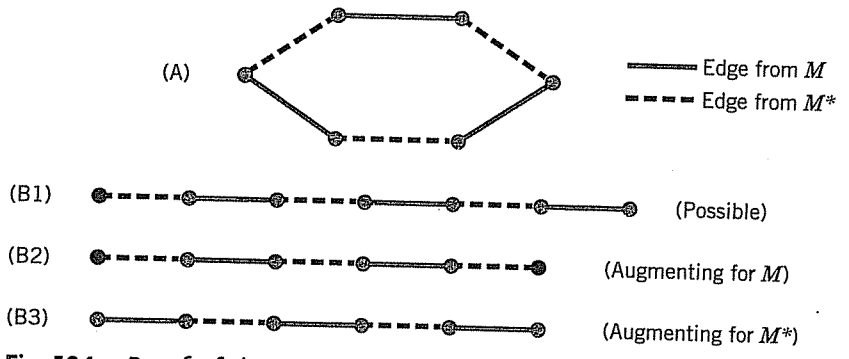


Fig. 504. Proof of the augmenting path theorem for bipartite matching

This theorem suggests the algorithm in Table 23.9 for obtaining augmenting paths, in which vertices are labeled for the purpose of backtracking paths. Such a label is *in addition* to the number of the vertex, which is also retained. Clearly, to get an augmenting path, one must start from an *exposed* vertex, and then trace an alternating path until one arrives at another *exposed* vertex. After Step 3 all vertices in S are labeled. In Step 4, the set T contains at least one exposed vertex, since otherwise we would have stopped at Step 1.

Table 23.9 Bipartite Maximum Cardinality Matching

ALGORITHM MATCHING $[G = (S, T; E), M, n]$

This algorithm determines a maximum cardinality matching M in a bipartite graph G by augmenting a given matching in G.

INPUT: Bipartite graph $G = (S, T; E)$ with vertices $1, \cdots, n$, matching M in G (for instance, $M = \emptyset$)

OUTPUT: Maximum cardinality matching M in G

1. If there is no exposed vertex in S then

OUTPUT M. Stop

[*M is of maximum cardinality in G.*]

Else label all *exposed* vertices *in* S with $\emptyset$.

2. For each i in S and edge (i, j) *not* in M, label j with i, unless already labeled.

3. For each *nonexposed* j in T, label i with j, where i is the other end of the unique edge (i, j) in M.

4. Backtrack the alternating paths P ending on an exposed vertex in T by using the labels on the vertices.

5. If no P in Step 4 is augmenting then

OUTPUT M. Stop

[*M is of maximum cardinality in G.*]

Else augment M by using an augmenting path P.

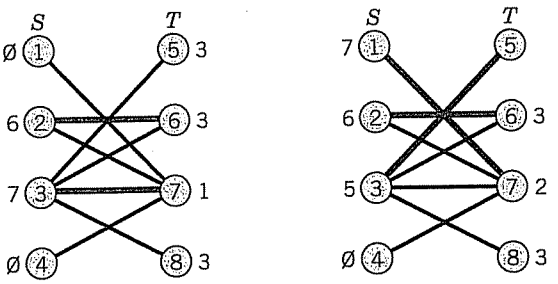
Remove all labels.

Go to Step 1.

End MATCHING

EXAMPLE 1 Maximum Cardinality Matching

Is the matching M_1 in Fig. 505a of maximum cardinality? If not, augment it until maximum cardinality is reached.



(a) Given graph and matching M_1

(b) Matching M_2 and new labels

Fig. 505. Example 1

Solution. We apply the algorithm.

1. Label 1 and 4 with ∅.
2. Label 7 with 1. Label 5, 6, 8 with 3.
3. Label 2 with 6, and 3 with 7.

   [*All vertices are now labeled as shown in Fig.* 474a.]
4. P_1: 1 — 7 — 3 — 5. [*By backtracking,* P_1 *is augmenting.*]

   P_2: 1 — 7 — 3 — 8. [P_2 *is augmenting.*]
5. Augment M_1 by using P_1, dropping (3, 7) from M_1 and including (1, 7) and (3, 5). Remove all labels. Go to Step 1.

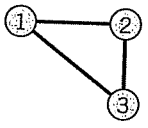
   Figure 474b *shows the resulting matching* $M_2 = \{(1, 7), (2, 6), (3, 5)\}$.

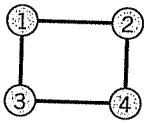
1. Label 4 with ∅.
2. Label 7 with 2. Label 6 and 8 with 3.
3. Label 1 with 7, and 2 with 6, and 3 with 5.
4. P_3: 5 — 3 — 8. [P_3 *is alternating but not augmenting.*]
5. Stop. M_2 *is of maximum cardinality* (*namely,* 3).

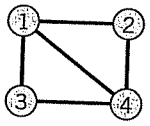
PROBLEM SET 23.8

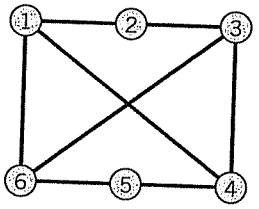
1–6 BIPARTITE OR NOT?

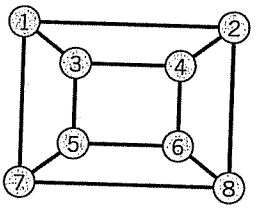
Are the following graphs bipartite? If your answer is yes, find S and T.

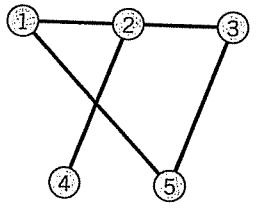
1. 

2. 

3. 

4. 

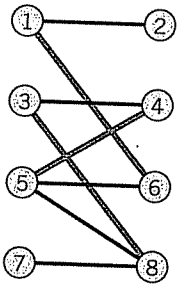
5. 

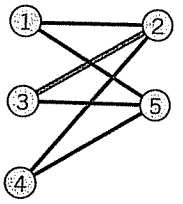
6. 

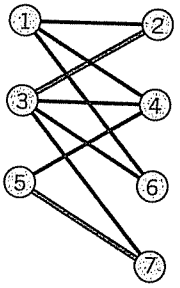
7. Can you obtain the answer to Prob. 3 from that to Prob. 1?

8–10 AUGMENTING PATHS

Find an augmenting path:

8. 

9. 

10. 

11–13 MAXIMUM CARDINALITY MATCHING

Augmenting the given matching, find a maximum cardinality matching:

11. In Prob. 9.
12. In Prob. 8.
13. In Prob. 10.

14. (Scheduling and matching) Three teachers x_1, x_2, x_3 teach four classes y_1, y_2, y_3, y_4 for these numbers of periods:

	y_1	y_2	y_3	y_4
x_1	1	0	1	1
x_2	1	1	1	1
x_3	0	1	1	1

Show that this arrangement can be represented by a bipartite graph G and that a teaching schedule for one period corresponds to a matching in G. Set up a teaching schedule with the smallest possible number of periods.

15. (Vertex coloring and exam scheduling) What is the smallest number of exam periods for six subjects a, b, c, d, e, f if some of the students simultaneously take a, b, f, some c, d, e, some a, c, e, and some c, e? Solve this as follows. Sketch a graph with six vertices $a, \cdots, f$ and join vertices if they represent subjects simultaneously taken by some students. Color the vertices so that adjacent vertices receive different colors. (Use numbers 1, 2, $\cdots$ instead of actual colors if you want.) What is the minimum number of colors you need? For any graph G, this minimum number is called the (vertex) **chromatic number** $\chi_v(G)$. Why is this the answer to the problem? Write down a possible schedule.

16. How many colors do you need in vertex coloring the graph in Prob. 5?

17. Show that all trees can be vertex colored with two colors.

18. (Harbor management) How many piers does a harbor master need for accommodating six cruise ships $S_1, \cdots, S_6$ with expected dates of arrival A and departure D in July, $(A, D) = (10, 13), (13, 15), (14, 17), (12, 15), (16, 18), (14, 17)$, respectively, if each pier can accommodate only one ship, arrival being at 6 a:m and departures at 11 p:m? *Hint.* Join S_i and S_j by an edge if their intervals overlap. Then color vertices.

19. What would be the answer to Prob. 18 if only the five ships $S_1, \cdots, S_5$ had to be accommodated?

20. (Complete bipartite graphs) A bipartite graph $G = (S, T; E)$ is called *complete* if every vertex in S is joined to every vertex in T by an edge, and is denoted by K_{n_1,n_2}, where n_1 and n_2 are the numbers of vertices in S and T, respectively. How many edges does this graph have?

21. (Planar graph) A *planar graph* is a graph that can be drawn on a sheet of paper so that no two edges cross. Show that the complete graph K_4 with four vertices is planar. The complete graph K_5 with five vertices is not planar. Make this plausible by attempting to draw K_5 so that no edges cross. Interpret the result in terms of a net of roads between five cities.

22. (Bipartite graph $K_{3,3}$ not planar) Three factories 1, 2, 3 are each supplied underground by water, gas, and electricity, from points A, B, C, respectively. Show that this can be represented by $K_{3,3}$ (the complete bipartite graph $G = (S, T; E)$ with S and T consisting of three vertices each) and that eight of the nine supply lines (edges) can be laid out without crossing. Make it plausible that $K_{3,3}$ is not planar by attempting to draw the ninth line without crossing the others.

23. (Four- (vertex) color theorem) The famous *four-color theorem* states that one can color the vertices of any *planar* graph (so that adjacent vertices get different colors) with at most four colors. It had been conjectured for a long time and was eventually proved in 1976 by Appel and Haken [Illinois J. Math **21** (1977), 429–567]. Can you color the complete graph K_5 with four colors? Does the result contradict the four-color theorem? (For more details, see Ref. [F8] in App. 1.)

24. (Edge coloring) The *edge chromatic number* $\chi_e(G)$ of a graph G is the minimum number of colors needed for coloring the edges of G so that incident edges get different colors. Clearly, $\chi_e(G) \geqq \max d(u)$, where $d(u)$ is the degree of vertex u. If $G = (S, T; E)$ is bipartite, the equality sign holds. Prove this for $K_{n,n}$.

25. Vizing's theorem states that for any graph G (without multiple edges!), $\max d(u) \leqq \chi_e(G) \leqq \max d(u) + 1$. Give an example of a graph for which $\chi_e(G)$ does exceed $\max d(u)$.

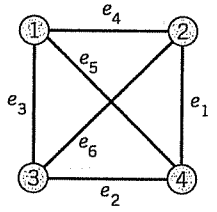
CHAPTER 23 REVIEW QUESTIONS AND PROBLEMS

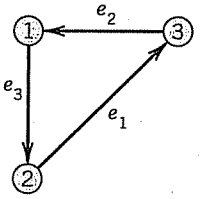
1. What is a graph? A digraph? A tree? A cycle? A path?
2. State from memory how you can handle graphs and digraphs on computers.
3. Describe situations and problems that can be modeled using graphs or digraphs.
4. What is a shortest path problem? Give applications.
5. What is BFS? DFS? In what connection did these concepts occur?
6. Give some applications in which spanning trees play a role.
7. What are bipartite graphs? What applications motivate this concept?

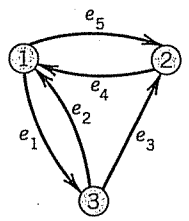
8. What is the traveling salesman problem?
9. What is a network? What optimization problems are connected with it?
10. Can a forward edge in one path be a backward edge in another path? In a cut set? Explain.
11. There is a famous theorem on cut sets. Can you remember and explain it?

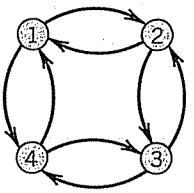
12–17 MATRICES FOR GRAPHS OR DIGRAPHS

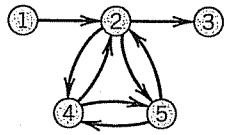
Find the adjacency matrix of:

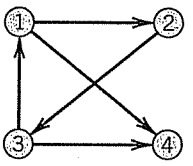
12. 

13. 

14. 

15. 

16. 

17. 

18–20 GIVEN ADJACENCY MATRIX

Sketch the graph whose adjacency matrix is:

18. $\begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$

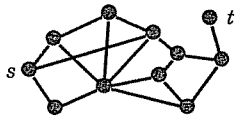
19. $\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$

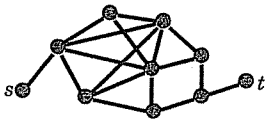
20. $\begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}$

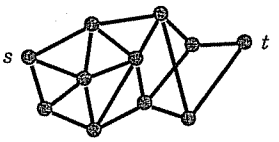
21. Make a vertex incidence list of the digraph in Prob. 13.
22. Make a vertex incidence list of the digraph in Prob. 14.

23–28 SHORTEST PATHS

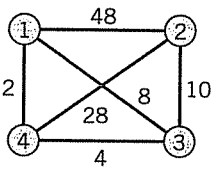
Find a shortest path and its length by Moore's BFS algorithm, assuming that all the edges have length 1:

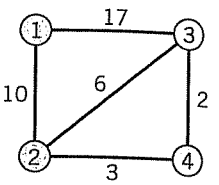
23. 

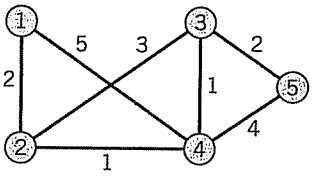
24. 

25. 

Find shortest paths by Dijkstra's algorithm:

26. 

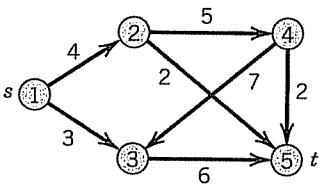
27. 

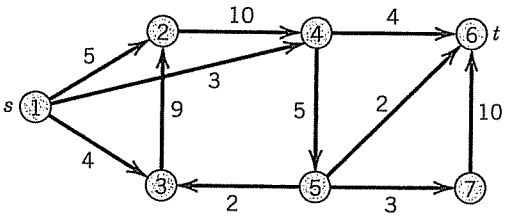
28. 

29. **(Shortest spanning tree)** Find a shortest spanning tree for the graph in Prob. 26.
30. Find a shortest spanning tree in Prob. 27.
31. **Cayley's theorem** states that the number of spanning trees in a complete graph with n vertices is n^{n-2}. Verify this for $n = 2, 3, 4$.
32. Show that $O(m^3) + O(m^2) = O(m^3)$.

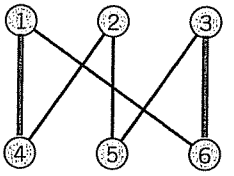
33–34 MAXIMUM FLOW.

Find the maximum flow, where the given numbers are capacities:

33. 

34. 

35. Company A has offices in Chicago, Los Angeles, and New York, Company B in Boston and New York, Company C in Chicago, Dallas, and Los Angeles. Represent this by a bipartite graph.
36. **(Maximum cardinality matching).** Augmenting the given matching, find a maximum cardinality matching:



SUMMARY OF CHAPTER 23
Graphs and Combinatorial Optimization

Combinatorial optimization concerns optimization problems of a discrete or combinatorial structure. It uses graphs and digraphs (Sec. 23.1) as basic tools.

A **graph** $G = (V, E)$ consists of a set V of **vertices** $v_1, v_2, \cdots, v_n$, (often simply denoted by $1, 2, \cdots, n$) and a set E of **edges** $e_1, e_2, \cdots, e_m$, each of which connects two vertices. We also write (i, j) for an edge with vertices i and j as endpoints. A **digraph** (= directed graph) is a graph in which each edge has a direction (indicated by an arrow). For handling graphs and digraphs in computers, one can use *matrices* or *lists* (Sec. 23.1).

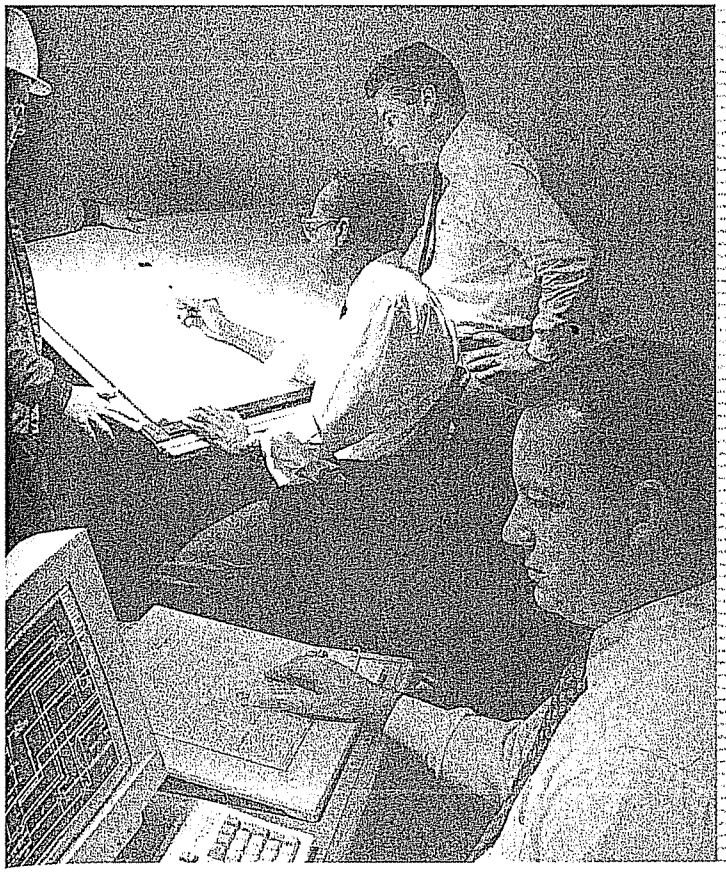
This chapter is devoted to important classes of optimization problems for graphs and digraphs that all arise from practical applications, and corresponding algorithms, as follows.

In a **shortest path problem** (Sec. 23.2) we determine a path of minimum length (consisting of edges) from a vertex s to a vertex t in a graph whose edges (i, j) have a "length" $l_{ij} > 0$, which may be an actual length or a travel time or cost or an electrical resistance [if (i, j) is a wire in a net], and so on. **Dijkstra's algorithm** (Sec. 23.3) or, when all $l_{ij} = 1$, **Moore's algorithm** (Sec. 23.2) are suitable for these problems.

A **tree** is a graph that is connected and has no **cycles** (no closed paths). Trees are very important in practice. A *spanning tree* in a graph G is a tree containing *all* the vertices of G. If the edges of G have lengths, we can determine a **shortest spanning tree,** for which the sum of the lengths of all its edges is minimum, by **Kruskal's algorithm** or **Prim's algorithm** (Secs. 23.4, 23.5).

A **network** (Sec. 23.6) is a digraph in which each edge (i, j) has a *capacity* $c_{ij} > 0$ [= maximum possible flow along (i, j)] and at one vertex, the *source* s, a flow is produced that flows along the edges to a vertex t, the *sink* or *target,* where the flow disappears. The problem is to maximize the flow, for instance, by applying the **Ford–Fulkerson algorithm** (Sec. 23.7), which uses **flow augmenting paths** (Sec. 23.6). Another related concept is that of a **cut set,** as defined in Sec. 23.6.

A **bipartite graph** $G = (V, E)$ (Sec. 23.8) is a graph whose vertex set V consists of two parts S and T such that every edge of G has one end in S and the other in T, so that there are no edges connecting vertices in S or vertices in T. A **matching** in G is a set of edges, no two of which have an endpoint in common. The problem then is to find a **maximum cardinality matching** in G, that is, a matching M that has a maximum number of edges. For an algorithm, see Sec. 23.8.



PART G

Probability, Statistics

CHAPTER 24 **Data Analysis. Probability Theory**

CHAPTER 25 **Mathematical Statistics**

Probability theory (Chap. 24) provides models of probability distributions (theoretical models of the observable reality involving chance effects) to be tested by statistical methods, and it will also supply the mathematical foundation of these methods in Chap. 25.

Modern **mathematical statistics** (Chap. 25) has various engineering applications, for instance, in testing materials, control of production processes, quality control of production outputs, performance tests of systems, robotics, and automatization in general, production planning, marketing analysis, and so on.

To this we could add a long list of fields of applications, for instance, in agriculture, biology, computer science, demography, economics, geography, management of natural resources, medicine, meteorology, politics, psychology, sociology, traffic control, urban planning, etc. Although these applications are very heterogeneous, we shall see that most statistical methods are universal in the sense that each of them can be applied in various fields.

Additional Software for Probability and Statistics

See also the list of software at the beginning of Part E on Numerical Analysis.

DATA DESK. Data Description, Inc., Ithaca, NY. Phone 1-800-573-5121 or (607) 257-1000, website at www.datadescription.com.

MINITAB. Minitab, Inc., College Park, PA. Phone 1-800-448-3555 or (814) 238-3280, website at www.minitab.com.

SAS. SAS Institute, Inc., Cary, NC. Phone 1-800-727-0025 or (919) 677-8000, website at www.sas.com.

S-PLUS. Insightful Corporation, Inc., Seattle, WA. Phone 1-800-569-0123 or (206) 283-8802, website at www.insightful.com.

SPSS. SPSS, Inc., Chicago, IL. Phone 1-800-543-2185 or (312) 651-3000, website at www.spss.com.

STATISTICA. StatSoft, Inc., Tulsa, OK. Phone (918) 749-1119, website at www.statsoft.com.

CHAPTER 24

Data Analysis. Probability Theory

We first show how to handle data numerically or in terms of graphs, and how to extract information (average size, spread of data, etc.) from them. If these data are influenced by "chance," by factors whose effect we cannot predict exactly (e.g., weather data, stock prices, lifespans of tires, etc.), we have to rely on **probability theory.** This theory originated in games of chance, such as flipping coins, rolling dice, or playing cards. Nowadays it gives mathematical models of chance processes called *random experiments* or, briefly, **experiments**. In such an experiment we observe a **random variable** X, that is, a function whose values in a **trial** (a performance of an experiment) occur "by chance" (Sec. 24.3) according to a **probability distribution** that gives the individual probabilities with which possible values of X may occur in the long run. (Example: Each of the six faces of a die should occur with the same probability, 1/6.) Or we may simultaneously observe more than one random variable, for instance, height *and* weight of persons or hardness *and* tensile strength of steel. This is discussed in Sec. 24.9, which will also give the basis for the mathematical justification of the statistical methods in Chap. 25.

Prerequisite: Calculus.
References and Answers to Problems: App. 1 Part G, App. 2.

24.1 Data Representation. Average. Spread

Data can be represented numerically or graphically in various ways. For instance, your daily newspaper may contain tables of stock prices and money exchange rates, curves or bar charts illustrating economical or political developments, or pie charts showing how your tax dollar is spent. And there are numerous other representations of data for special purposes.

In this section we discuss the use of standard representations of data in statistics. (For these, software packages, such as DATA DESK and MINITAB, are available, and Maple or Mathematica may also be helpful; see pp. 778 and 991) We explain corresponding concepts and methods in terms of typical examples, beginning with

(1) 89 84 87 81 89 86 91 90 78 89 87 99 83 89.

These are $n = 14$ measurements of the tensile strength of sheet steel in kg/mm^2, recorded in the order obtained and rounded to integer values. To see what is going on, we **sort** these data, that is, we order them by size,

(2) 78 81 83 84 86 87 87 89 89 89 89 90 91 99.

Sorting is a standard process on the computer; see Ref. [E25], listed in App. 1.

Graphic Representation of Data

We shall now discuss standard graphic representations used in statistics for obtaining information on properties of data.

Stem-and-Leaf Plot

This is one of the simplest but most useful representations of data. For (1) it is shown in Fig. 506. The numbers in (1) range from 78 to 99; see (2). We divide these numbers into 5 groups, 75–79, 80–84, 85–89, 90–94, 95–99. The integers in the tens position of the groups are 7, 8, 8, 9, 9. These form the *stem* in Fig. 506. The first *leaf* is 8 (representing 78). The second leaf is 134 (representing 81, 83, 84), and so on.

The number of times a value occurs is called its **absolute frequency.** Thus 78 has absolute frequency 1, the value 89 has absolute frequency 4, etc. The column to the extreme left in Fig. 506 shows the **cumulative absolute frequencies,** that is, the sum of the absolute frequencies of the values up to the line of the leaf. Thus, the number 4 in the second line on the left shows that (1) has 4 values up to and including 84. The number 11 in the next line shows that there are 11 values not exceeding 89, etc. Dividing the cumulative absolute frequencies by n (= 14 in Fig. 506) gives the **cumulative relative frequencies.**

Histogram

For large sets of data, histograms are better in displaying the distribution of data than stem-and-leaf plots. The principle is explained in Fig. 507. (An application to a larger data set is shown in Sec. 25.7). The bases of the rectangles in Fig. 507 are the x-intervals (known as **class intervals**) 74.5–79.5, 79.5–84.5, 84.5–89.5, 89.5–94.5, 94.5–99.5, whose midpoints (known as **class marks**) are $x = 77, 82, 87, 92, 97$, respectively. The height of a rectangle with class mark x is the **relative class frequency** $f_{\text{rel}}(x)$, defined as the number of data values in that class interval, divided by n (= 14 in our case). Hence the areas of the rectangles are proportional to these relative frequencies, so that histograms give a good impression of the distribution of data.

Center and Spread of Data: Median, Quartiles

As a center of the location of data values we can simply take the **median,** the data value that falls in the middle when the values are ordered. In (2) we have 14 values. The seventh of them is 87, the eighth is 89, and we split the difference, obtaining the median 88. (In general, we would get a fraction.)

The spread (variability) of the data values can be measured by the **range** $R = x_{\max} - x_{\min}$, the largest minus the smallest data values, $R = 99 - 78 = 21$ in (2).

Leaf unit = 1.0

1	7	8
4	8	134
11	8	6779999
13	9	01
14	9	9

Fig. 506. Stem-and-leaf plot of the data in (1) and (2)

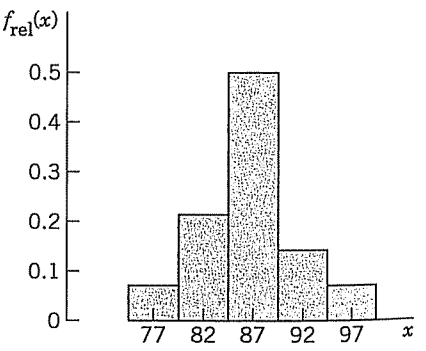


Fig. 507. Histogram of the data in (1) and (2) (grouped as in Fig. 506)

Better information gives the **interquartile range** IQR $= q_U - q_L$. Here the **upper quartile** q_U is the middle value among the data values *above* the median. The **lower quartile** q_L is the middle value among the data values *below* the median. Thus in (2) we have $q_U = 89$ (the fourth value from the end), $q_L = 84$ (the fourth value from the beginning), and IQR $= 89 - 84 = 5$. The median is also called the **middle quartile** and is denoted by q_M. The rule of "splitting the difference" (just applied to the middle quartile) is equally well used for the other quartiles if necessary.

Boxplot

The **boxplot** of (1) in Fig. 508 is obtained from the five numbers $x_{\min}$, q_L, q_M, q_U, $x_{\max}$ just determined. The box extends from q_L to q_U. Hence it has the height IQR. The position of the median in the box shows that the data distribution is not symmetric. The two lines extend from the box to $x_{\min}$ below and to $x_{\max}$ above. Hence they mark the range R.

Boxplots are particularly suitable for making comparisons. For example, Fig. 508 shows boxplots of the data sets (1) and

(3) 91 89 93 91 87 94 92 85 91 90 96 93 89

(consisting of $n = 13$ values). Ordering gives

(4) 85 87 89 89 90 91 91 91 92 93 93 94 96

(tensile strength, as before). From the plot we immediately see that the box of (3) is shorter than the box of (1) (indicating the higher quality of the steel sheets!) and that q_M is located in the middle of the box (showing the more symmetric form of the distribution). Finally, $x_{\max}$ is closer to q_U for (3) than it is for (1), a fact that we shall discuss later.

For plotting the box of (3) we took from (4) the values $x_{\min} = 85$, $q_L = 89$, $q_M = 91$, $q_U = 93$, $x_{\max} = 96$.

Outliers

An **outlier** is a value that appears to be uniquely different from the rest of the data set. It might indicate that something went wrong with the data collection process. In connection with quartiles an outlier is conventionally defined as a value more than a distance of 1.5 IQR from either end of the box.

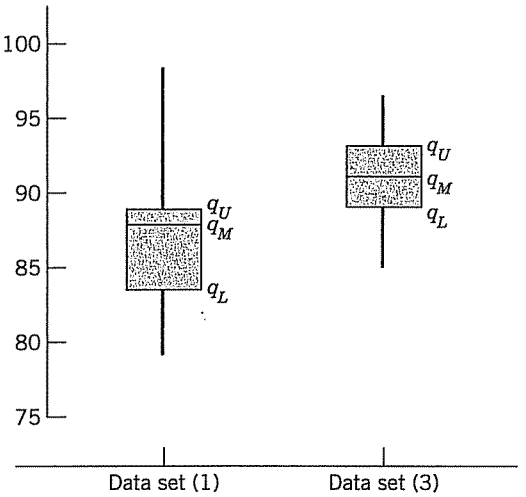


Fig. 508. Boxplots of data sets (1) and (3)

For the data in (1) we have IQR = 5, $q_L = 84$, $q_U = 89$. Hence outliers are smaller than $84 - 7.5$ or larger than $89 + 7.5$, so that 99 is an outlier [see (2)]. The data (3) have no outliers, as you can readily verify.

Mean. Standard Deviation. Variance

Medians and quartiles are easily obtained by ordering and counting, practically without calculation. But they do not give full information on data: you can change data values to some extent without changing the median. Similarly for the quartiles.

The average size of the data values can be measured in a more refined way by the **mean**

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j = \frac{1}{n}(x_1 + x_2 + \cdots + x_n). \tag{5}$$

This is the arithmetic mean of the data values, obtained by taking their sum and dividing by the data *size* n. Thus in (1),

$$\bar{x} = \tfrac{1}{14}(89 + 84 + \cdots + 89) = \tfrac{611}{7} \approx 87.3.$$

Every data value contributes, and changing one of them will change the mean.

Similarly, the spread (variability) of the data values can be measured in a more refined way by the **standard deviation** s or by its square, the **variance**

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{n} (x_j - \bar{x})^2 = \frac{1}{n-1}\left[(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2\right]. \tag{6}$$

Thus, to obtain the variance of the data, take the difference $x_j - \bar{x}$ of each data value from the mean, square it, take the sum of these n squares, and divide it by $n - 1$ (not n, as we motivate in Sec. 25.2). To get the standard deviation s, take the square root of s^2.

For example, using $\bar{x} = 611/7$, we get for the data (1) the variance

$$s^2 = \tfrac{1}{13}\left[(89 - \tfrac{611}{7})^2 + (84 - \tfrac{611}{7})^2 + \cdots + (89 - \tfrac{611}{7})^2\right] = \tfrac{176}{7} \approx 25.14.$$

Hence the standard deviation is $s = \sqrt{176/7} \approx 5.014$. Note that the standard deviation has the same dimension as the data values (kg/mm^2, see at the beginning), which is an advantage. On the other hand, the variance is preferable to the standard deviation in developing statistical methods, as we shall see in Chap. 25.

CAUTION! Your CAS (Maple, for instance) may use $1/n$ instead of $1/(n-1)$ in (6), but the latter is better when n is small (see Sec. 25.2).

PROBLEM SET 24.1

1–10 DATA REPRESENTATIONS

Represent the data by a stem-and-leaf plot, a histogram, and a boxplot:

1. 20 21 20 19 20 19 21 19

2. 7 6 4 0 7 1 2 4 6 6

3. 56 58 54 33 41 30 44 37 51 46 56 38 38 49 39

4. 12.1 10 12.4 10.5 9.2 17.2 11.4 11.8 14.7 9.9

5. 70.6 70.9 69.1 71.3 70.5 69.7 71.5 69.8 71.1 68.9 70.3 69.2 71.2 70.4 72.8

6. −0.52 0.11 −0.48 0.94 0.24 −0.19 −0.55

7. Reaction time [sec] of an automatic switch
2.3 2.2 2.4 2.5 2.3 2.3 2.4 2.1 2.5 2.4
2.6 2.3 2.5 2.1 2.4 2.2 2.3 2.5 2.4 2.4

8. Carbon content [%] of coal
89 90 89 84 80 88 90 89 88 90 85
87 86 82 85 76 89 87 86 86

9. Weight of filled bottles [g] in an automatic filling process
403 399 398 401 400 401 401

10. Gasoline consumption [gallons per mile] of six cars of the same model
14.0 14.5 13.5 14.0 14.5 14.0

11–16 AVERAGE AND SPREAD

Find the mean and compare it with the median. Find the standard deviation and compare it with the interquartile range.

11. The data in Prob. 1.

12. The data in Prob. 2.

13. The data in Prob. 5.

14. The data in Prob. 6.

15. The data in Prob. 9.

16. 5 22 7 23 6. Why is $|\bar{x} - q_M|$ so large?

17. Construct the simplest possible data with $\bar{x} = 100$ but $q_M = 0$.

18. (Mean) Prove that $\bar{x}$ must always lie between the smallest and the largest data values.

19. (Outlier, reduced data) Calculate s for the data 4 1 3 10 2. Then reduce the data by deleting the outlier and calculate s. Comment.

20. WRITING PROJECT. Average and Spread. Compare Q_M, IQR and $\bar{x}$, s, illustrating the advantages and disadvantages with examples of your own.

24.2 Experiments, Outcomes, Events

We now turn to **probability theory.** This theory has the purpose of providing mathematical models of situations affected or even governed by "chance effects," for instance, in weather forecasting, life insurance, quality of technical products (computers, batteries, steel sheets, etc.), traffic problems, and, of course, games of chance with cards or dice. And the accuracy of these models can be tested by suitable observations or experiments—this is a main purpose of **statistics** to be explained in Chap. 25.

We begin by defining some standard terms. An **experiment** is a process of measurement or observation, in a laboratory, in a factory, on the street, in nature, or wherever; so "experiment" is used in a rather general sense. Our interest is in experiments that involve **randomness,** chance effects, so that we cannot predict a result exactly. A **trial** is a single performance of an experiment. Its result is called an **outcome** or a **sample point.** n trials then give a **sample** of **size** n consisting of n sample points. The **sample space** S of an experiment is the set of all possible outcomes.

EXAMPLES 1–6 Random Experiments. Sample Spaces

(1) Inspecting a lightbulb. S = {Defective, Nondefective}.

(2) Rolling a die. $S = \{1, 2, 3, 4, 5, 6\}$.

(3) Measuring tensile strength of wire. S the numbers in some interval.

(4) Measuring copper content of brass. S: 50% to 90%, say.

(5) Counting daily traffic accidents in New York. S the integers in some interval.

(6) Asking for opinion about a new car model. S = {Like, Dislike, Undecided}. ■

The subsets of S are called **events** and the outcomes **simple events.**

EXAMPLE 7 Events

In (2), events are $A = \{1, 3, 5\}$ (*"Odd number"*), $B = \{2, 4, 6\}$ (*"Even number"*), $C = \{5, 6\}$, etc. Simple events are $\{1\}, \{2\}, \cdots, \{6\}$. ■

If in a trial an outcome a happens and $a \in A$ (a *is an element of* A), we say that A happens. For instance, if a die turns up a 3, the event A: *Odd number* happens. Similarly, if C in Example 1 happens (meaning 5 or 6 turns up), then $D = \{4, 5, 6\}$ happens. Also note that S happens in each trial, meaning that *some* event of S always happens. All this is quite natural.

Unions, Intersections, Complements of Events

In connection with basic probability laws we shall need the following concepts and facts about events (subsets) $A, B, C, \cdots$ of a given sample space S.

The **union** $A \cup B$ of A and B consists of all points in A or B or both.

The **intersection** $A \cap B$ of A and B consists of all points that are in both A and B.

If A and B have no points in common, we write

$$A \cap B = \varnothing$$

where $\varnothing$ is the *empty set* (set with no elements) and we call A and B **mutually exclusive** (or **disjoint**) because in a trial the occurrence of A *excludes* that of B (and conversely)—if your die turns up an odd number, it cannot turn up an even number in the same trial. Similarly, a coin cannot turn up *Head* and *Tail* at the same time.

Complement A^c of A. This is the set of all the points of S ***not*** in A. Thus,

$$A \cap A^c = \varnothing, \qquad A \cup A^c = S.$$

In Example 7 we have $A^c = B$, hence $A \cup A^c = \{1, 2, 3, 4, 5, 6\} = S$.

Another notation for the complement of A is $\overline{A}$ (instead of A^c), but we shall not use this because in set theory $\overline{A}$ is used to denote the *closure* of A (not needed in our work).

Unions and intersections of more events are defined similarly. The **union**

$$\bigcup_{j=1}^{m} A_j = A_1 \cup A_2 \cup \cdots \cup A_m$$

of events $A_1, \cdots, A_m$ consists of all points that are in at least one A_j. Similarly for the union $A_1 \cup A_2 \cup \cdots$ of infinitely many subsets $A_1, A_2, \cdots$ of an *infinite* sample space S (that is, S consists of infinitely many points). The **intersection**

$$\bigcap_{j=1}^{m} A_j = A_1 \cap A_2 \cap \cdots \cap A_m$$

of $A_1, \cdots, A_m$ consists of the points of S that are in each of these events. Similarly for the intersection $A_1 \cap A_2 \cap \cdots$ of infinitely many subsets of S.

Working with events can be illustrated and facilitated by **Venn diagrams**[1] for showing unions, intersections, and complements, as in Figs. 509 and 510, which are typical examples that give the idea.

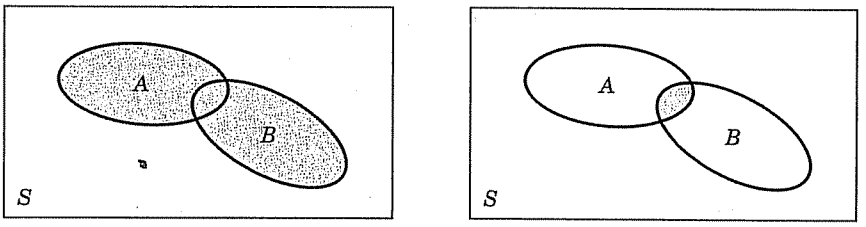
EXAMPLE 8 **Unions and Intersections of 3 Events**

In rolling a die, consider the events

A: *Number greater than* 3, B: *Number less than* 6, C: *Even number.*

Then $A \cap B = \{4, 5\}$, $B \cap C = \{2, 4\}$, $C \cap A = \{4, 6\}$, $A \cap B \cap C = \{4\}$. Can you sketch a Venn diagram of this? Furthermore, $A \cup B = S$, hence $A \cup B \cup C = S$ (why?). ■

[1]JOHN VENN (1834–1923), English mathematician.



Union $A \cup B$ Intersection $A \cap B$

Fig. 509. Venn diagrams showing two events A and B in a sample space S and their union $A \cup B$ (colored) and intersection $A \cap B$ (colored)

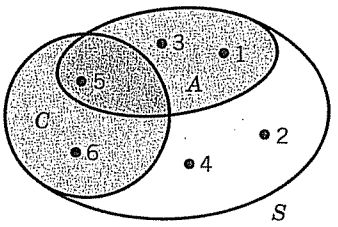


Fig. 510. Venn diagram for the experiment of rolling a die, showing S, $A = \{1, 3, 5\}$, $C = \{5, 6\}$, $A \cup C = \{1, 3, 5, 6\}$, $A \cap C = \{5\}$

PROBLEM SET 24.2

1–9 SAMPLE SPACES, EVENTS

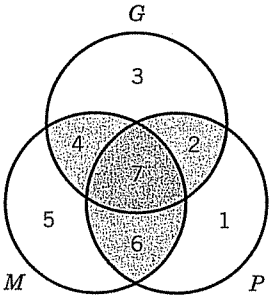
Graph a sample space for the experiment:

1. Tossing 2 coins
2. Drawing 4 screws from a lot of right-handed and left-handed screws
3. Rolling 2 dice
4. Tossing a coin until the first *Head* appears
5. Rolling a die until the first *"Six"* appears
6. Drawing bolts from a lot of 20, containing one defective D, until D is drawn, one at a time and assuming **sampling without replacement,** that is, bolts drawn are not returned to the lot
7. Recording the lifetime of each of 3 lightbulbs
8. Choosing a committee of 3 from a group of 5 people
9. Recording the daily maximum temperature X and the maximum air pressure Y at some point in a city

10. In Prob. 3, circle and mark the events *A: Equal faces*, *B: Sum exceeds* 9, *C: Sum equals* 7.
11. In rolling 2 dice, are the events *A: Sum divisible by* 3 and *B: Sum divisible by* 5 mutually exclusive?
12. Answer the question in Prob. 11 for rolling 3 dice.
13. In Prob. 5 list the outcomes that make up the event *E:* First *"Six"* in rolling at most 3 times. Describe E^c.
14. List all 8 subsets of the sample space $S = \{a, b, c\}$.

15–20 VENN DIAGRAMS

15. In connection with a trip to Europe by some students, consider the events P that they see Paris, G that they have a good time, and M that they run out of money, and describe in words the events 1, · · ·, 7 in the diagram.



Problem 15

16. Using Venn diagrams, graph and check the rules

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C).$$

17. **(De Morgan's laws)** Using Venn diagrams, graph and check *De Morgan's laws*

$$(A \cup B)^c = A^c \cap B^c$$

$$(A \cap B)^c = A^c \cup B^c.$$

18. Using a Venn diagram, show that $A \subseteq B$ if and only if $A \cap B = A$.

19. Show that, by the definition of complement, for any subset A of a sample space S,

$$(A^c)^c = A, \qquad S^c = \varnothing, \qquad \varnothing^c = S,$$
$$A \cup A^c = S, \qquad A \cap A^c = \varnothing.$$

20. Using a Venn diagram, show that $A \subseteq B$ if and only if $A \cup B = B$.

24.3 Probability

The "probability" of an event A in an experiment is supposed to measure how frequently A is *about* to occur if we make many trials. If we flip a coin, then heads H and tails T will appear *about* equally often—we say that H and T are **"equally likely."** Similarly, for a regularly shaped die of homogeneous material (**"fair die"**) each of the six outcomes $1, \cdots, 6$ will be equally likely. These are examples of experiments in which the sample space S consists of finitely many outcomes (points) that for reasons of some symmetry can be regarded as equally likely. This suggests the following definition.

DEFINITION 1

First Definition of Probability

If the sample space S of an experiment consists of finitely many outcomes (points) that are equally likely, then the probability $P(A)$ of an event A is

(1)
$$P(A) = \frac{\text{Number of points in } A}{\text{Number of points in } S}.$$

From this definition it follows immediately that, in particular,

(2)
$$P(S) = 1.$$

EXAMPLE 1 **Fair Die**

In rolling a fair die once, what is the probability $P(A)$ of A of obtaining a 5 or a 6? The probability of B: *"Even number"?*

Solution. The six outcomes are equally likely, so that each has probability 1/6. Thus $P(A) = 2/6 = 1/3$ because $A = \{5, 6\}$ has 2 points, and $P(B) = 3/6 = 1/2$. ■

Definition 1 takes care of many games as well as some practical applications, as we shall see, but certainly not of all experiments, simply because in many problems we do not have finitely many equally likely outcomes. To arrive at a more general definition of probability, we regard ***probability as the counterpart of relative frequency.*** Recall from Sec. 24.1 that the **absolute frequency** $f(A)$ of an event A in n trials is the number of times A occurs, and the **relative frequency** of A in these trials is $f(A)/n$; thus

(3)
$$f_{\text{rel}}(A) = \frac{f(A)}{n} = \frac{\text{Number of times } A \text{ occurs}}{\text{Number of trials}}.$$

Now if A did not occur, then $f(A) = 0$. If A always occurred, then $f(A) = n$. These are the extreme cases. Division by n gives

$$0 \leqq f_{\text{rel}}(A) \leqq 1. \tag{4*}$$

In particular, for $A = S$ we have $f(S) = n$ because S always occurs (meaning that some event always occurs; if necessary, see Sec. 24.2, after Example 7). Division by n gives

$$f_{\text{rel}}(S) = 1. \tag{5*}$$

Finally, if A and B are mutually exclusive, they cannot occur together. Hence the absolute frequency of their union $A \cup B$ must equal the sum of the absolute frequencies of A and B. Division by n gives the same relation for the relative frequencies,

$$f_{\text{rel}}(A \cup B) = f_{\text{rel}}(A) + f_{\text{rel}}(B) \qquad (A \cap B = \varnothing). \tag{6*}$$

We are now ready to extend the definition of probability to experiments in which equally likely outcomes are not available. Of course, the extended definition should include Definition 1. Since probabilities are supposed to be the theoretical counterpart of relative frequencies, we choose the properties in (4*), (5*), (6*) as axioms. (Historically, such a choice is the result of a long process of gaining experience on what might be best and most practical.)

DEFINITION 2

General Definition of Probability

Given a sample space S, with each event A of S (subset of S) there is associated a number $P(A)$, called the **probability** of A, such that the following **axioms of probability** are satisfied.

1. For every A in S,

$$0 \leqq P(A) \leqq 1. \tag{4}$$

2. The entire sample space S has the probability

$$P(S) = 1. \tag{5}$$

3. For mutually exclusive events A and B ($A \cap B = \varnothing$; see Sec. 24.2),

$$P(A \cup B) = P(A) + P(B) \qquad (A \cap B = \varnothing). \tag{6}$$

If S is infinite (has infinitely many points), Axiom 3 has to be replaced by

3′. For mutually exclusive events $A_1, A_2, \cdots$,

$$P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots. \tag{6′}$$

In the infinite case the subsets of S on which $P(A)$ is defined are restricted to form a so-called *σ-algebra,* as explained in Ref. [GR6] (not [G6]!) in App. 1. This is of no practical consequence to us.

Basic Theorems of Probability

We shall see that the axioms of probability will enable us to build up probability theory and its application to statistics. We begin with three basic theorems. The first of them is useful if we can get the probability of the complement A^c more easily than $P(A)$ itself.

THEOREM 1

Complementation Rule

For an event A and its complement A^c in a sample space S,

$$P(A^c) = 1 - P(A). \tag{7}$$

PROOF By the definition of complement (Sec. 24.2), we have $S = A \cup A^c$ and $A \cap A^c = \emptyset$. Hence by Axioms 2 and 3,

$$1 = P(S) = P(A) + P(A^c), \qquad \text{thus} \qquad P(A^c) = 1 - P(A). \qquad \blacksquare$$

EXAMPLE 2 **Coin Tossing**

Five coins are tossed simultaneously. Find the probability of the event *A: At least one head turns up.* Assume that the coins are fair.

Solution. Since each coin can turn up heads or tails, the sample space consists of $2^5 = 32$ outcomes. Since the coins are fair, we may assign the same probability (1/32) to each outcome. Then the event A^c (*No heads turn up*) consists of only 1 outcome. Hence $P(A^c) = 1/32$, and the answer is $P(A) = 1 - P(A^c) = 31/32$. $\blacksquare$

The next theorem is a simple extension of Axiom 3, which you can readily prove by induction.

THEOREM 2

Addition Rule for Mutually Exclusive Events

For mutually exclusive events $A_1, \cdots, A_m$ in a sample space S,

$$P(A_1 \cup A_2 \cup \cdots A_m) = P(A_1) + P(A_2) + \cdots + P(A_m). \tag{8}$$

EXAMPLE 3 **Mutually Exclusive Events**

If the probability that on any workday a garage will get 10–20, 21–30, 31–40, over 40 cars to service is 0.20, 0.35, 0.25, 0.12, respectively, what is the probability that on a given workday the garage gets at least 21 cars to service?

Solution. Since these are mutually exclusive events, Theorem 2 gives the answer $0.35 + 0.25 + 0.12 = 0.72$. Check this by the complementation rule. $\blacksquare$

In many cases, events will not be mutually exclusive. Then we have

THEOREM 3

Addition Rule for Arbitrary Events

For events A and B in a sample space,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{9}$$

PROOF C, D, E in Fig. 511 make up $A \cup B$ and are mutually exclusive (disjoint). Hence by Theorem 2,

$$P(A \cup B) = P(C) + P(D) + P(E).$$

This gives (9) because on the right $P(C) + P(D) = P(A)$ by Axiom 3 and disjointness; and $P(E) = P(B) - P(D) = P(B) - P(A \cap B)$, also by Axiom 3 and disjointness. ▩

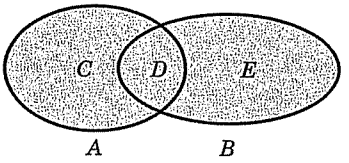


Fig. 511. Proof of Theorem 3

Note that for mutually exclusive events A and B we have $A \cap B = \varnothing$ by definition and, by comparing (9) and (6),

(10) $$P(\varnothing) = 0.$$

(Can you also prove this by (5) and (7)?)

EXAMPLE 4 **Union of Arbitrary Events**

In tossing a fair die, what is the probability of getting an odd number or a number less than 4?

Solution. Let A be the event *"Odd number"* and B the event *"Number less than* 4." Then Theorem 3 gives the answer

$$P(A \cup B) = \tfrac{3}{6} + \tfrac{3}{6} - \tfrac{2}{6} = \tfrac{2}{3}$$

because $A \cap B =$ *"Odd number less than* 4" $= \{1, 3\}$. ▩

Conditional Probability. Independent Events

Often it is required to find the probability of an event B under the condition that an event A occurs. This probability is called the **conditional probability** *of B given A* and is denoted by $P(B|A)$. In this case A serves as a new (reduced) sample space, and that probability is the fraction of $P(A)$ which corresponds to $A \cap B$. Thus

(11) $$P(B|A) = \frac{P(A \cap B)}{P(A)} \qquad [P(A) \neq 0].$$

Similarly, the *conditional probability of A given B* is

(12) $$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad [P(B) \neq 0].$$

Solving (11) and (12) for $P(A \cap B)$, we obtain

THEOREM 4 **Multiplication Rule**

If A and B are events in a sample space S and $P(A) \neq 0$, $P(B) \neq 0$, then

(13) $$P(A \cap B) = P(A)P(B|A) = P(B)P(A|B).$$

EXAMPLE 5 Multiplication Rule

In producing screws, let A mean "screw too slim" and B "screw too short." Let $P(A) = 0.1$ and let the conditional probability that a slim screw is also too short be $P(B|A) = 0.2$. What is the probability that a screw that we pick randomly from the lot produced will be both too slim and too short?

Solution. $P(A \cap B) = P(A)P(B|A) = 0.1 \cdot 0.2 = 0.02 = 2\%$, by Theorem 4. ∎

Independent Events. If events A and B are such that

$$P(A \cap B) = P(A)P(B), \tag{14}$$

they are called **independent events.** Assuming $P(A) \neq 0$, $P(B) \neq 0$, we see from (11)–(13) that in this case

$$P(A|B) = P(A), \qquad P(B|A) = P(B).$$

This means that the probability of A does not depend on the occurrence or nonoccurrence of B, and conversely. This justifies the term "independent."

Independence of m Events. Similarly, m events $A_1, \cdots, A_m$ are called **independent** if

$$P(A_1 \cap \cdots \cap A_m) = P(A_1) \cdots P(A_m) \tag{15a}$$

as well as for every k different events $A_{j_1}, A_{j_2}, \cdots, A_{j_k}$,

$$P(A_{j_1} \cap A_{j_2} \cap \cdots \cap A_{j_k}) = P(A_{j_1})P(A_{j_2}) \cdots P(A_{j_k}) \tag{15b}$$

where $k = 2, 3, \cdots, m - 1$.

Accordingly, three events A, B, C are independent if and only if

$$\begin{aligned} P(A \cap B) &= P(A)P(B), \\ P(B \cap C) &= P(B)P(C), \\ P(C \cap A) &= P(C)P(A), \\ P(A \cap B \cap C) &= P(A)P(B)P(C). \end{aligned} \tag{16}$$

Sampling. Our next example has to do with randomly drawing objects, *one at a time,* from a given set of objects. This is called **sampling from a population,** and there are two ways of sampling, as follows.

1. In **sampling with replacement,** the object that was drawn at random is placed back to the given set and the set is mixed thoroughly. Then we draw the next object at random.
2. In **sampling without replacement** the object that was drawn is put aside.

EXAMPLE 6 Sampling With and Without Replacement

A box contains 10 screws, three of which are defective. Two screws are drawn at random. Find the probability that none of the two screws is defective.

Solution. We consider the events

A: First drawn screw nondefective.

B: Second drawn screw nondefective.

Clearly, $P(A) = \frac{7}{10}$ because 7 of the 10 screws are nondefective and we sample at random, so that each screw has the same probability ($\frac{1}{10}$) of being picked. If we sample with replacement, the situation before the second drawing is the same as at the beginning, and $P(B) = \frac{7}{10}$. The events are independent, and the answer is

$$P(A \cap B) = P(A)P(B) = 0.7 \cdot 0.7 = 0.49 = 49\%.$$

If we sample without replacement, then $P(A) = \frac{7}{10}$, as before. If A has occurred, then there are 9 screws left in the box, 3 of which are defective. Thus $P(B|A) = \frac{6}{9} = \frac{2}{3}$, and Theorem 4 yields the answer

$$P(A \cap B) = \frac{7}{10} \cdot \frac{2}{3} \approx 47\%.$$

Is it intuitively clear that this value must be smaller than the preceding one? ▩

PROBLEM SET 24.3

1. Three screws are drawn at random from a lot of 100 screws, 10 of which are defective. Find the probability that the screws drawn will be nondefective in drawing (a) with replacement, (b) without replacement.
2. In Prob. 1 find the probability of *E: At least* 1 *defective* (i) directly, (ii) by using complements; in both cases (a) and (b).
3. If we inspect paper by drawing 5 sheets without replacement from every batch of 500, what is the probability of getting 5 clean sheets although 2% of the sheets contain spots? First guess.
4. Under what conditions will it make *practically* no difference whether we sample with or without replacement? Give numeric examples.
5. If you need a right-handed screw from a box containing 20 right-handed and 5 left-handed screws, what is the probability that you get at least one right-handed screw in drawing 2 screws with replacement?
6. If in Prob. 5 you draw without replacement, does the probability decrease or increase? First think, then calculate.
7. What gives the greater probability of hitting some target at least once: (a) hitting in a shot with probability 1/2 and firing 1 shot, or (b) hitting in a shot with probability 1/4 and firing 2 shots? First guess. Then calculate.
8. Suppose that we draw cards repeatedly and with replacement from a file of 100 cards, 50 of which refer to male and 50 to female persons. What is the probability of obtaining the second "female" card before the third "male" card?
9. What is the complementary event of the event considered in Prob. 8? Calculate its probability and use it to check your result in Prob. 8.
10. In rolling two fair dice, what is the probability of obtaining a sum greater than 4 but not exceeding 7?
11. In rolling two fair dice, what is the probability of obtaining equal numbers or numbers with an even product?
12. Solve Prob. 11 by considering complements.
13. A motor drives an electric generator. During a 30-day period, the motor needs repair with probability 8% and the generator needs repair with probability 4%. What is the probability that during a given period, the entire apparatus (consisting of a motor and a generator) will need repair?
14. If a circuit contains 3 automatic switches and we want that, with a probability of 95%, during a given time interval they are all working, what probability of failure per time interval can we admit for a single switch?
15. If a certain kind of tire has a life exceeding 25 000 miles with probability 0.95, what is the probability that a set of 4 of these tires on a car will last longer than 25 000 miles?
16. In Prob. 15, what is the probability that at least one of the tires will not last for 25 000 miles?
17. A pressure control apparatus contains 4 valves. The apparatus will not work unless all valves are operative. If the probability of failure of each valve during some interval of time is 0.03, what is the corresponding probability of failure of the apparatus?
18. Show that if B is a subset of A, then $P(B) \leqq P(A)$.
19. Extending Theorem 4, show that $P(A \cap B \cap C) = P(A)P(B|A)P(C|A \cap B)$.
20. You may wonder whether in (16) the last relation follows from the others, but the answer is no. To see this, imagine that a chip is drawn from a box containing 4 chips numbered 000, 011, 101, 110, and let A, B, C be the events that the first, second, and third digit, respectively, on the drawn chip is 1. Show that then the first three formulas in (16) hold but the last one does not hold.

24.4 Permutations and Combinations

Permutations and combinations help in finding probabilities $P(A) = a/k$ by ***systematically counting*** the number a of points of which an event A consists; here, k is the number of points of the sample space S. The practical difficulty is that a may often be surprisingly large, so that actual counting becomes hopeless. For example, if in assembling some instrument you need 10 different screws in a certain order and you want to draw them randomly from a box (which contains nothing else) the probability of obtaining them in the required order is only 1/3 628 800 because there are

$$10! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 \cdot 7 \cdot 8 \cdot 9 \cdot 10 = 3\,628\,800$$

orders in which they can be drawn. Similarly, in many other situations the numbers of orders, arrangements, etc. are often incredibly large. (If you are unimpressed, take 20 screws—how much bigger will the number be?)

Permutations

A **permutation** of given things (*elements* or *objects*) is an arrangement of these things in a row in some order. For example, for three letters a, b, c there are $3! = 1 \cdot 2 \cdot 3 = 6$ permutations: *abc*, *acb*, *bac*, *bca*, *cab*, *cba*. This illustrates (a) in the following theorem.

THEOREM 1

Permutations

(a) ***Different things.*** *The number of permutations of n different things taken all at a time is*

(1) $$n! = 1 \cdot 2 \cdot 3 \cdots n$$ (read "*n factorial*").

(b) ***Classes of equal things.*** *If n given things can be divided into c classes of alike things differing from class to class, then the number of permutations of these things taken all at a time is*

(2) $$\frac{n!}{n_1!n_2!\cdots n_c!} \qquad (n_1 + n_2 + \cdots + n_c = n)$$

where n_j is the number of things in the jth class.

PROOF **(a)** There are n choices for filling the first place in the row. Then $n - 1$ things are still available for filling the second place, etc.

(b) n_1 alike things in class 1 make $n_1!$ permutations collapse into a single permutation (those in which class 1 things occupy the same n_1 positions), etc., so that (2) follows from (1). ∎

EXAMPLE 1 **Illustration of Theorem 1(b)**

If a box contains 6 red and 4 blue balls, the probability of drawing first the red and then the blue balls is

$$P = 6!4!/10! = 1/210 \approx 0.5\%.$$ ■

A **permutation of n things taken k at a time** is a permutation containing only k of the n given things. Two such permutations consisting of the same k elements, in a different order, are different, by definition. For example, there are 6 different permutations of the three letters a, b, c, taken two letters at a time, ab, ac, bc, ba, ca, cb.

A **permutation of n things taken k at a time with repetitions** is an arrangement obtained by putting any given thing in the first position, any given thing, including a repetition of the one just used, in the second, and continuing until k positions are filled. For example, there are $3^2 = 9$ different such permutations of a, b, c taken 2 letters at a time, namely, the preceding 6 permutations and aa, bb, cc. You may prove (see Team Project 18):

THEOREM 2

Permutations

The number of different permutations of n different things taken k at a time without repetitions is

(3a)
$$n(n-1)(n-2)\cdots(n-k+1) = \frac{n!}{(n-k)!}$$

and with repetitions is

(3b)
$$n^k.$$

EXAMPLE 2 **Illustration of Theorem 2**

In a coded telegram the letters are arranged in groups of five letters, called *words*. From (3b) we see that the number of different such words is

$$26^5 = 11\,881\,376.$$

From (3a) it follows that the number of different such words containing each letter no more than once is

$$26!/(26-5)! = 26 \cdot 25 \cdot 24 \cdot 23 \cdot 22 = 7\,893\,600.$$ ■

Combinations

In a permutation, the order of the selected things is essential. In contrast, a **combination** of given things means any selection of one or more things *without regard to order*. There are two kinds of combinations, as follows.

The number of **combinations of n different things, taken k at a time, without repetitions** is the number of sets that can be made up from the n given things, each set containing k different things and no two sets containing exactly the same k things.

The number of **combinations of n different things, taken k at a time, with repetitions** is the number of sets that can be made up of k things chosen from the given n things, each being used as often as desired.

For example, there are three combinations of the three letters a, b, c, taken two letters at a time, without repetitions, namely, ab, ac, bc, and six such combinations with repetitions, namely, ab, ac, bc, aa, bb, cc.

THEOREM 3

Combinations

The number of different combinations of n different things taken, k at a time, without repetitions, is

$$\text{(4a)} \qquad \binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)\cdots(n-k+1)}{1\cdot 2\cdots k},$$

and the number of those combinations with repetitions is

$$\text{(4b)} \qquad \binom{n+k-1}{k}.$$

PROOF The statement involving (4a) follows from the first part of Theorem 2 by noting that there are $k!$ *permutations* of k things from the given n things that differ by the order of the elements (see Theorem 1), but there is only a single *combination* of those k things of the type characterized in the first statement of Theorem 3. The last statement of Theorem 3 can be proved by induction (see Team Project 18). ▩

EXAMPLE 3 **Illustration of Theorem 3**

The number of samples of five lightbulbs that can be selected from a lot of 500 bulbs is [see (4a)]

$$\binom{500}{5} = \frac{500!}{5!495!} = \frac{500\cdot 499\cdot 498\cdot 497\cdot 496}{1\cdot 2\cdot 3\cdot 4\cdot 5} = 255\,244\,687\,600. \qquad ▩$$

Factorial Function

In (1)–(4) the **factorial function** is basic. By definition,

$$\text{(5)} \qquad 0! = 1.$$

Values may be computed recursively from given values by

$$\text{(6)} \qquad (n+1)! = (n+1)n!.$$

For large n the function is very large (see Table A3 in App. 5). A convenient approximation for large n is the **Stirling formula**[2]

$$\text{(7)} \qquad n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \qquad (e = 2.718\cdots)$$

[2]JAMES STIRLING (1692–1770), Scots mathematician.

where $\sim$ is read **"asymptotically equal"** and means that the ratio of the two sides of (7) approaches 1 as n approaches infinity.

EXAMPLE 4 **Stirling Formula**

$n!$	By (7)	Exact Value	Relative Error
4!	23.5	24	2.1%
10!	3 598 696	3 628 800	0.8%
20!	$2.422\,79 \cdot 10^{18}$	2 432 902 008 176 640 000	0.4%

Binomial Coefficients

The **binomial coefficients** are defined by the formula

$$\binom{a}{k} = \frac{a(a-1)(a-2)\cdots(a-k+1)}{k!} \qquad (k \geqq 0, \text{ integer}). \tag{8}$$

The numerator has k factors. Furthermore, we define

$$\binom{a}{0} = 1, \qquad \text{in particular,} \qquad \binom{0}{0} = 1. \tag{9}$$

For integer $a = n$ we obtain from (8)

$$\binom{n}{k} = \binom{n}{n-k} \qquad (n \geqq 0,\ 0 \leqq k \leqq n). \tag{10}$$

Binomial coefficients may be computed recursively, because

$$\binom{a}{k} + \binom{a}{k+1} = \binom{a+1}{k+1} \qquad (k \geqq 0, \text{ integer}). \tag{11}$$

Formula (8) also yields

$$\binom{-m}{k} = (-1)^k \binom{m+k-1}{k} \qquad \begin{matrix}(k \geqq 0, \text{ integer}) \\ (m > 0).\end{matrix} \tag{12}$$

There are numerous further relations; we mention two important ones,

$$\sum_{s=0}^{n-1} \binom{k+s}{k} = \binom{n+k}{k+1} \qquad \begin{matrix}(k \geqq 0,\ n \geqq 1, \\ \text{both integer})\end{matrix} \tag{13}$$

and

$$\sum_{k=0}^{r} \binom{p}{k}\binom{q}{r-k} = \binom{p+q}{r} \qquad (r \geqq 0, \text{ integer}). \tag{14}$$

PROBLEM SET 24.4

1. List all permutations of four digits 1, 2, 3, 4, taken all at a time.
2. List (a) all permutations, (b) all combinations without repetitions, (c) all combinations with repetitions, of 5 letters a, e, i, o, u taken 2 at a time.
3. In how many ways can we assign 8 workers to 8 jobs (one worker to each job and conversely)?
4. How many samples of 4 objects can be drawn from a lot of 80 objects?
5. In how many different ways can we choose a committee of 3 from 20 persons? First guess.
6. In how many different ways can we select a committee consisting of 3 engineers, 2 biologists, and 2 chemists from 10 engineers, 5 biologists, and 6 chemists? First guess.
7. Of a lot of 10 items, 2 are defective. (a) Find the number of different samples of 4. Find the number of samples of 4 containing (b) no defectives, (c) 1 defective, (d) 2 defectives.
8. If a cage contains 100 mice, two of which are male, what is the probability that the two male mice will be included if 12 mice are randomly selected?
9. An urn contains 2 blue, 3 green, and 4 red balls. We draw 1 ball at random and put it aside. Then we draw the next ball, and so on. Find the probability of drawing at first the 2 blue balls, then the 3 green ones, and finally the red ones.
10. By what factor is the probability in Prob. 9 decreased if the number of balls is doubled (4 blue, etc.)?
11. Determine the number of different bridge hands. (A bridge hand consists of 13 cards selected from a full deck of 52 cards.)
12. In how many different ways can 5 people be seated at a round table?
13. If 3 suspects who committed a burglary and 6 innocent persons are lined up, what is the probability that a witness who is not sure and has to pick three persons will pick the three suspects by chance? That the witness picks 3 innocent persons by chance?
14. **(Birthday problem)** What is the probability that in a group of 20 people (that includes no twins) at least two have the same birthday, if we assume that the probability of having birthday on a given day is 1/365 for every day. First guess.
15. How many different license plates showing 5 symbols, namely, 2 letters followed by 3 digits, could be made?
16. How many automobile registrations may the police have to check in a hit-and-run accident if a witness reports KDP5 and cannot remember the last two digits on the license plate but is certain that all three digits were different?
17. **CAS PROJECT. Stirling formula. (a)** Using (7), compute approximate values of $n!$ for $n = 1, \cdots, 20$.

    (b) Determine the relative error in (a). Find an empirical formula for that relative error.

    (c) An upper bound for that relative error is $e^{1/12n} - 1$. Try to relate your empirical formula to this.

    (d) Search through the literature for further information on Stirling's formula. Write a short report about your findings, arranged in logical order and illustrated with numeric examples.
18. **TEAM PROJECT. Permutations, Combinations. (a)** Prove Theorem 2.

    (b) Prove the last statement of Theorem 3.

    (c) Derive (11) from (8).

    (d) By the **binomial theorem,**

$$(a + b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k},$$

so that $a^k b^{n-k}$ has the coefficient $\binom{n}{k}$. Can you conclude this from Theorem 3 or is this a mere coincidence?

(e) Prove (14) by using the binomial theorem.

(f) Collect further formulas for binomial coefficients from the literature and illustrate them numerically.

24.5 Random Variables. Probability Distributions

In Sec. 24.1 we considered frequency distributions of data. These distributions show the absolute or relative frequency of the data values. Similarly, a **probability distribution** or, briefly, a **distribution,** shows the probabilities of events in an experiment. The quantity that we observe in an experiment will be denoted by X and called a **random variable** (or

stochastic variable) because the value it will assume in the next trial depends on chance, on **randomness**—if you roll a dice, you get one of the numbers from 1 to 6, but you don't know which one will show up next. Thus X = *Number a die turns up* is a random variable. So is X = *Elasticity of rubber* (elongation at break). ("Stochastic" means related to chance.)

If we ***count*** (cars on a road, defective screws in a production, tosses until a die shows the first Six), we have a **discrete random variable and distribution.** If we ***measure*** (electric voltage, rainfall, hardness of steel), we have a **continuous random variable and distribution.** Precise definitions follow. In both cases the distribution of X is determined by the **distribution function**

$$F(x) = P(X \leqq x); \tag{1}$$

this is the probability that in a trial, X will assume any value not exceeding x.

CAUTION! The terminology is not uniform. $F(x)$ is sometimes also called the **cumulative distribution function.**

For (1) to make sense in both the discrete and the continuous case we formulate conditions as follows.

DEFINITION

Random Variable

A **random variable** X is a function defined on the sample space S of an experiment. Its values are real numbers. For every number a the probability

$$P(X = a)$$

with which X assumes a is defined. Similarly, for any interval I the probability

$$P(X \in I)$$

with which X assumes any value in I is defined.

Although this definition is very general, practically only a very small number of distributions will occur over and over again in applications.

From (1) we obtain the fundamental formula for the probability corresponding to an interval $a < x \leqq b$,

$$P(a < X \leqq b) = F(b) - F(a). \tag{2}$$

This follows because $X \leqq a$ (*"X assumes any value not exceeding a"*) and $a < X \leqq b$ (*"X assumes any value in the interval* $a < x \leqq b$*"*) are mutually exclusive events, so that by (1) and Axiom 3 of Definition 2 in Sec. 24.3

$$\begin{aligned} F(b) = P(X \leqq b) &= P(X \leqq a) + P(a < X \leqq b) \\ &= F(a) + P(a < X \leqq b) \end{aligned}$$

and subtraction of $F(a)$ on both sides gives (2).

Discrete Random Variables and Distributions

By definition, a random variable X and its distribution are **discrete** if X assumes only finitely many or at most countably many values $x_1, x_2, x_3, \cdots$, called the **possible values** of X, with positive probabilities $p_1 = P(X = x_1), p_2 = P(X = x_2), p_3 = P(X = x_3), \cdots$, whereas the probability $P(X \in I)$ is zero for any interval I containing no possible value.

Clearly, the discrete distribution of X is also determined by the **probability function** $f(x)$ of X, defined by

$$f(x) = \begin{cases} p_j & \text{if } x = x_j \\ 0 & \text{otherwise} \end{cases} \qquad (j = 1, 2, \cdots), \tag{3}$$

From this we get the values of the **distribution function** $F(x)$ by taking sums,

$$F(x) = \sum_{x_j \leqq x} f(x_j) = \sum_{x_j \leqq x} p_j \tag{4}$$

where for any given x we sum all the probabilities p_j for which x_j is smaller than or equal to that of x. This is a **step function** with upward jumps of size p_j at the possible values x_j of X and constant in between.

EXAMPLE 1 **Probability Function and Distribution Function**

Figure 512 shows the probability function $f(x)$ and the distribution function $F(x)$ of the discrete random variable

$$X = \textit{Number a fair die turns up.}$$

X has the possible values $x = 1, 2, 3, 4, 5, 6$ with probability 1/6 each. At these x the distribution function has upward jumps of magnitude 1/6. Hence from the graph of $f(x)$ we can construct the graph of $F(x)$, and conversely.

In Figure 512 (and the next one) at each jump the ***fat dot*** indicates the ***function value at the jump!*** ■

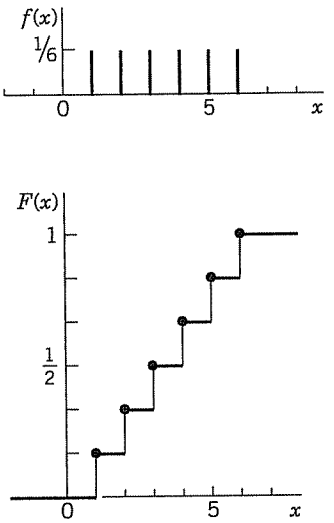


Fig. 512. Probability function $f(x)$ and distribution function $F(x)$ of the random variable X = *Number obtained in tossing a fair die once*

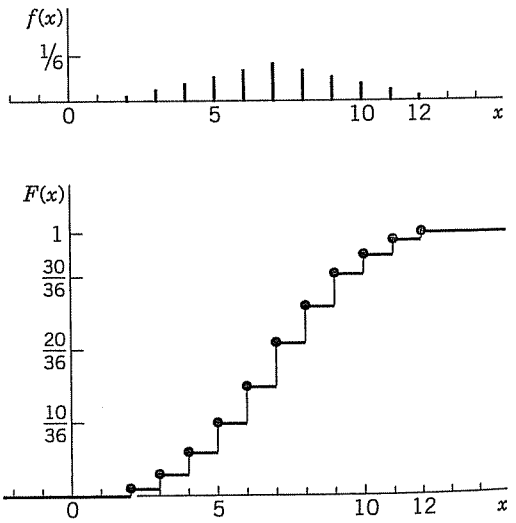


Fig. 513. Probability function $f(x)$ and distribution function $F(x)$ of the random variable X = *Sum of the two numbers obtained in tossing two fair dice once*

EXAMPLE 2 **Probability Function and Distribution Function**

The random variable X = *Sum of the two numbers two fair dice turn up* is discrete and has the possible values $2\ (= 1 + 1), 3, 4, \cdots, 12\ (= 6 + 6)$. There are $6 \cdot 6 = 36$ equally likely outcomes (1, 1) (1, 2), $\cdots$, (6, 6), where the first number is that shown on the first die and the second number that on the other die. Each such outcome has probability 1/36. Now $X = 2$ occurs in the case of the outcome (1, 1); $X = 3$ in the case of the two outcomes (1, 2) and (2, 1); $X = 4$ in the case of the three outcomes (1, 3), (2, 2), (3, 1); and so on. Hence $f(x) = P(X = x)$ and $F(x) = P(X \leqq x)$ have the values

x	2	3	4	5	6	7	8	9	10	11	12
$f(x)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36
$F(x)$	1/36	3/36	6/36	10/36	15/36	21/36	26/36	30/36	33/36	35/36	36/36

Figure 513 shows a bar chart of this function and the graph of the distribution function, which is again a step function, with jumps (of different height!) at the possible values of X. ▩

Two useful formulas for discrete distributions are readily obtained as follows. For the probability corresponding to intervals we have from (2) and (4)

(5) $$P(a < X \leqq b) = F(b) - F(a) = \sum_{a<x_j \leqq b} p_j \qquad (X \text{ discrete}).$$

This is the sum of all probabilities p_j for which x_j satisfies $a < x_j \leqq b$. (Be careful about $<$ and $\leqq$!) From this and $P(S) = 1$ (Sec. 24.3) we obtain the following formula.

(6) $$\sum_j p_j = 1 \qquad \text{(sum of all probabilities)}.$$

EXAMPLE 3 **Illustration of Formula (5)**

In Example 2, compute the probability of a sum of at least 4 and at most 8.

Solution. $P(3 < X \leqq 8) = F(8) - F(3) = \frac{26}{36} - \frac{3}{36} = \frac{23}{36}$. ▩

EXAMPLE 4 **Waiting Time Problem. Countably Infinite Sample Space**

In tossing a fair coin, let X = *Number of trials until the first head appears.* Then, by independence of events (Sec. 24.3),

$$P(X = 1) = P(H) = \tfrac{1}{2} \qquad (H = \text{Head})$$

$$P(X = 2) = P(TH) = \tfrac{1}{2} \cdot \tfrac{1}{2} = \tfrac{1}{4} \qquad (T = \text{Tail})$$

$$P(X = 3) = P(TTH) = \tfrac{1}{2} \cdot \tfrac{1}{2} \cdot \tfrac{1}{2} = \tfrac{1}{8}, \qquad \text{etc.}$$

and in general $P(X = n) = (\frac{1}{2})^n$, $n = 1, 2, \cdots$. Also, (6) can be confirmed by the sum formula for the geometric series,

$$\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots = -1 + \frac{1}{1 - \frac{1}{2}}$$

$$= -1 + 2 = 1. \qquad ▩$$

Continuous Random Variables and Distributions

Discrete random variables appear in experiments in which we ***count*** (defectives in a production, days of sunshine in Chicago, customers standing in a line, etc.). Continuous random variables appear in experiments in which we ***measure*** (lengths of screws, voltage in a power line, Brinell hardness of steel, etc.). By definition, a random variable X and its distribution are *of continuous type* or, briefly, **continuous,** if its distribution function $F(x)$ [defined in (1)] can be given by an integral

$$F(x) = \int_{-\infty}^{x} f(v)\, dv \tag{7}$$

(we write v because x is needed as the upper limit of the integral) whose integrand $f(x)$, called the **density** of the distribution, is nonnegative, and is continuous, perhaps except for finitely many x-values. Differentiation gives the relation of f to F as

$$f(x) = F'(x) \tag{8}$$

for every x at which $f(x)$ is continuous.

From (2) and (7) we obtain the very important formula for the probability corresponding to an interval:

$$P(a < X \leqq b) = F(b) - F(a) = \int_{a}^{b} f(v)\, dv. \tag{9}$$

This is the analog of (5).

From (7) and $P(S) = 1$ (Sec. 24.3) we also have the analog of (6):

$$\int_{-\infty}^{\infty} f(v)\, dv = 1. \tag{10}$$

Continuous random variables are ***simpler than discrete ones*** with respect to intervals. Indeed, in the continuous case the four probabilities corresponding to $a < X \leqq b$, $a < X < b$, $a \leqq X < b$, and $a \leqq X \leqq b$ with any fixed a and b ($> a$) are all the same. Can you see why? (*Answer.* This probability is the area under the density curve, as in Fig. 514, and does not change by adding or subtracting a single point in the interval of integration.) This is different from the discrete case! (Explain.)

The next example illustrates notations and typical applications of our present formulas.

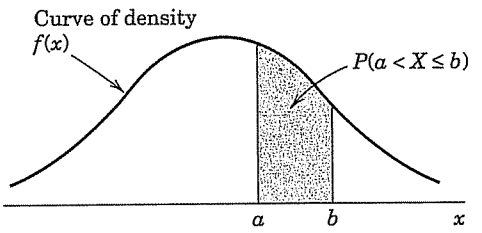


Fig. 514. Example illustrating formula (9)

EXAMPLE 5 Continuous Distribution

Let X have the density function $f(x) = 0.75(1 - x^2)$ if $-1 \leqq x \leqq 1$ and zero otherwise. Find the distribution function. Find the probabilities $P(-\frac{1}{2} \leqq X \leqq \frac{1}{2})$ and $P(\frac{1}{4} \leqq X \leqq 2)$. Find x such that $P(X \leqq x) = 0.95$.

Solution. From (7) we obtain $F(x) = 0$ if $x \leqq -1$,

$$F(x) = 0.75 \int_{-1}^{x} (1 - v^2)\, dv = 0.5 + 0.75x - 0.25x^3 \qquad \text{if } -1 < x \leqq 1,$$

and $F(x) = 1$ if $x > 1$. From this and (9) we get

$$P(-\tfrac{1}{2} \leqq X \leqq \tfrac{1}{2}) = F(\tfrac{1}{2}) - F(-\tfrac{1}{2}) = 0.75 \int_{-1/2}^{1/2} (1 - v^2)\, dv = 68.75\%$$

(because $P(-\frac{1}{2} \leqq X \leqq \frac{1}{2}) = P(-\frac{1}{2} < X \leqq \frac{1}{2})$ for a continuous distribution) and

$$P(\tfrac{1}{4} \leqq X \leqq 2) = F(2) - F(\tfrac{1}{4}) = 0.75 \int_{1/4}^{1} (1 - v^2)\, dv = 31.64\%.$$

(Note that the upper limit of integration is 1, not 2. Why?) Finally,

$$P(X \leqq x) = F(x) = 0.5 + 0.75x - 0.25x^3 = 0.95.$$

Algebraic simplification gives $3x - x^3 = 1.8$. A solution is $x = 0.73$, approximately.

Sketch $f(x)$ and mark $x = -\frac{1}{2}, \frac{1}{2}, \frac{1}{4}$, and 0.73, so that you can see the results (the probabilities) as areas under the curve. Sketch also $F(x)$. ■

Further examples of continuous distributions are included in the next problem set and in later sections.

PROBLEM SET 24.5

1. Graph the probability function $f(x) = kx^2$ ($x = 1, 2, 3, 4, 5$; k suitable) and the distribution function.
2. Graph the density function $f(x) = kx^2$ ($0 \leqq x \leqq 5$; k suitable) and the distribution function.
3. **(Uniform distribution)** Graph f and F when the density is $f(x) = k = const$ if $-4 \leqq x \leqq 4$ and 0 elsewhere.
4. In Prob. 3 find $P(0 \leqq x \leqq 4)$ and c such that $P(-c < X < c) = 95\%$.
5. Graph f and F when $f(-2) = f(2) = 1/8$, $f(-1) = f(1) = 3/8$. Can f have further positive values?
6. Graph the distribution function $F(x) = 1 - e^{-3x}$ if $x > 0$, $F(x) = 0$ if $x \leqq 0$, and the density $f(x)$. Find x such that $F(x) = 0.9$.
7. Let X be the number of years before a particular type of machine will need replacement. Assume that X has the probability function $f(1) = 0.1$, $f(2) = 0.2$, $f(3) = 0.2$, $f(4) = 0.2$, $f(5) = 0.3$. Graph f and F. Find the probability that the machine needs no replacement during the first 3 years.
8. If X has the probability function $f(x) = k/2^x$ ($x = 0, 1, 2, \cdots$), what are k and $P(X \geqq 4)$?
9. Find the probability that none of the three bulbs in a traffic signal must be replaced during the first 1200 hours of operation if the probability that a bulb must be replaced is a random variable X with density $f(x) = 6[0.25 - (x - 1.5)^2]$ when $1 \leqq x \leqq 2$ and $f(x) = 0$ otherwise, where x is time measured in multiples of 1000 hours.
10. Suppose that certain bolts have length $L = 200 + X$ mm, where X is a random variable with density $f(x) = \frac{3}{4}(1 - x^2)$ if $-1 \leqq x \leqq 1$ and 0 otherwise. Determine c so that with a probability of 95% a bolt will have any length between $200 - c$ and $200 + c$. *Hint:* See also Example 5.
11. Let X [millimeters] be the thickness of washers a machine turns out. Assume that X has the density $f(x) = kx$ if $1.9 < x < 2.1$ and 0 otherwise. Find k. What is the probability that a washer will have thickness between 1.95 mm and 2.05 mm?

12. Suppose that in an automatic process of filling oil into cans, the content of a can (in gallons) is $Y = 50 + X$, where X is a random variable with density $f(x) = 1 - |x|$ when $|x| \leqq 1$ and 0 when $|x| > 1$. Graph $f(x)$ and $F(x)$. In a lot of 100 cans, about how many will contain 50 gallons or more? What is the probability that a can will contain less than 49.5 gallons? Less than 49 gallons?

13. Let the random variable X with density $f(x) = ke^{-x}$ if $0 \leqq x \leqq 2$ and 0 otherwise (x = time measured in years) be the time after which certain ball bearings are worn out. Find k and the probability that a bearing will last at least 1 year.

14. Let X be the ratio of sales to profits of some firm. Assume that X has the distribution function $F(x) = 0$ if $x < 2$, $F(x) = (x^2 - 4)/5$ if $2 \leqq x < 3$, $F(x) = 1$ if $x \geqq 3$. Find and graph the density. What is the probability that X is between 2.5 (40% profit) and 5 (20% profit)?

15. Show that $b < c$ implies $P(X \leqq b) \leqq P(X \leqq c)$.

16. If the diameter X of axles has the density $f(x) = k$ if $119.9 \leqq x \leqq 120.1$ and 0 otherwise, how many defectives will a lot of 500 axles approximately contain if defectives are axles slimmer than 119.92 or thicker than 120.08?

17. Let X be a random variable that can assume every real value. What are the complements of the events $X \leqq b$, $X < b$, $X \geqq c$, $X > c$, $b \leqq X \leqq c$, $b < X \leqq c$?

18. A box contains 4 right-handed and 6 left-handed screws. Two screws are drawn at random without replacement. Let X be the number of left-handed screws drawn. Find the probabilities $P(X = 0)$, $P(X = 1)$, $P(X = 2)$, $P(1 < X < 2)$, $P(X \leqq 1)$, $P(X \geqq 1)$, $P(X > 1)$, and $P(0.5 < X < 10)$.

24.6 Mean and Variance of a Distribution

The mean μ and variance σ^2 of a random variable X and of its distribution are the theoretical counterparts of the mean $\bar{x}$ and variance s^2 of a frequency distribution in Sec. 24.1 and serve a similar purpose. Indeed, the mean characterizes the central location and the variance the spread (the variability) of the distribution. The **mean** μ (mu) is defined by

$$\text{(1)} \qquad \text{(a)} \quad \mu = \sum_j x_j f(x_j) \qquad \text{(Discrete distribution)}$$

$$\text{(b)} \quad \mu = \int_{-\infty}^{\infty} x f(x)\, dx \qquad \text{(Continuous distribution)}$$

and the **variance** σ^2 (sigma square) by

$$\text{(2)} \qquad \text{(a)} \quad \sigma^2 = \sum_j (x_j - \mu)^2 f(x_j) \qquad \text{(Discrete distribution)}$$

$$\text{(b)} \quad \sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, dx \qquad \text{(Continuous distribution).}$$

σ (the positive square root of σ^2) is called the **standard deviation** of X and its distribution. f is the probability function or the density, respectively, in (a) and (b).

The mean μ is also denoted by $E(X)$ and is called the **expectation** *of* X because it gives the average value of X to be expected in many trials. Quantities such as μ and σ^2 that measure certain properties of a distribution are called **parameters.** μ and σ^2 are the two most important ones. From (2) we see that

$$\text{(3)} \qquad \sigma^2 > 0$$

(except for a discrete "distribution" with only one possible value, so that $\sigma^2 = 0$). We assume that μ and σ^2 exist (are finite), as is the case for practically all distributions that are useful in applications.

EXAMPLE 1 **Mean and Variance**

The random variable X = *Number of heads in a single toss of a fair coin* has the possible values $X = 0$ and $X = 1$ with probabilities $P(X = 0) = \frac{1}{2}$ and $P(X = 1) = \frac{1}{2}$. From (1a) we thus obtain the mean $\mu = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = \frac{1}{2}$, and (2a) yields the variance

$$\sigma^2 = (0 - \tfrac{1}{2})^2 \cdot \tfrac{1}{2} + (1 - \tfrac{1}{2})^2 \cdot \tfrac{1}{2} = \tfrac{1}{4}.$$

EXAMPLE 2 **Uniform Distribution. Variance Measures Spread**

The distribution with the density

$$f(x) = \frac{1}{b - a} \qquad \text{if} \qquad a < x < b$$

and $f = 0$ otherwise is called the **uniform distribution** on the interval $a < x < b$. From (1b) (or from Theorem 1, below) we find that $\mu = (a + b)/2$, and (2b) yields the variance

$$\sigma^2 = \int_a^b \left(x - \frac{a + b}{2}\right)^2 \frac{1}{b - a}\, dx = \frac{(b - a)^2}{12}.$$

Figure 515 illustrates that the spread is large if and only if σ^2 is large.

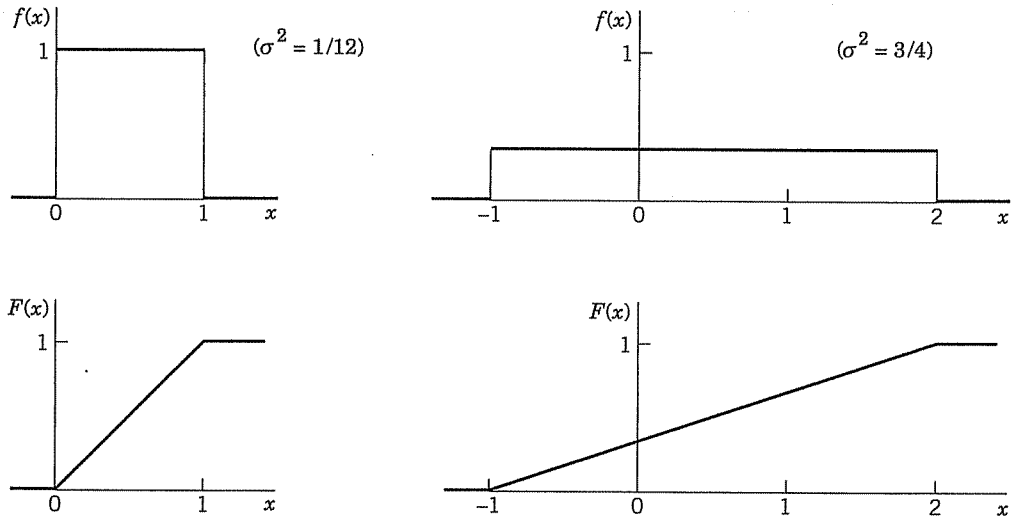


Fig. 515. Uniform distributions having the same mean (0.5) but different variances σ^2

Symmetry. We can obtain the mean μ without calculation if a distribution is symmetric. Indeed, you may prove

THEOREM 1

Mean of a Symmetric Distribution

If a distribution is **symmetric** *with respect to $x = c$, that is, $f(c - x) = f(c + x)$, then $\mu = c$.* (Examples 1 and 2 illustrate this.)

Transformation of Mean and Variance

Given a random variable X with mean μ and variance σ^2, we want to calculate the mean and variance of $X^* = a_1 + a_2X$, where a_1 and a_2 are given constants. This problem is important in statistics, where it appears often.

THEOREM 2

Transformation of Mean and Variance

(a) *If a random variable X has mean μ and variance σ^2, then the random variable*

$$X^* = a_1 + a_2X \qquad (a_2 > 0) \tag{4}$$

has the mean μ^* and variance σ^{*2}, where

$$\mu^* = a_1 + a_2\mu \quad \text{and} \quad \sigma^{*2} = a_2{}^2\sigma^2. \tag{5}$$

(b) *In particular, the* **standardized random variable** *Z corresponding to X, given by*

$$Z = \frac{X - \mu}{\sigma} \tag{6}$$

has the mean 0 and the variance 1.

PROOF We prove (5) for a continuous distribution. To a small interval I of length Δx on the x-axis there corresponds the probability $f(x)\Delta x$ [approximately; the area of a rectangle of base Δx and height $f(x)$]. Then the probability $f(x)\Delta x$ must equal that for the corresponding interval on the x^*-axis, that is, $f^*(x^*)\Delta x^*$, where f^* is the density of X^* and Δx^* is the length of the interval on the x^*-axis corresponding to I. Hence for differentials we have $f^*(x^*)\,dx^* = f(x)\,dx$. Also, $x^* = a_1 + a_2x$ by (4), so that (1b) applied to X^* gives

$$\begin{aligned}\mu^* &= \int_{-\infty}^{\infty} x^* f^*(x^*)\,dx^* \\ &= \int_{-\infty}^{\infty} (a_1 + a_2x) f(x)\,dx \\ &= a_1 \int_{-\infty}^{\infty} f(x)\,dx + a_2 \int_{-\infty}^{\infty} xf(x)\,dx.\end{aligned}$$

On the right the first integral equals 1, by (10) in Sec. 24.5. The second integral is μ. This proves (5) for μ^*. It implies

$$x^* - \mu^* = (a_1 + a_2x) - (a_1 + a_2\mu) = a_2(x - \mu).$$

From this and (2) applied to X^*, again using $f^*(x^*)\,dx^* = f(x)\,dx$, we obtain the second formula in (5),

$$\sigma^{*2} = \int_{-\infty}^{\infty} (x^* - \mu^*)^2 f^*(x^*)\,dx^* = a_2{}^2 \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\,dx = a_2{}^2\sigma^2.$$

For a discrete distribution the proof of (5) is similar.

Choosing $a_1 = -\mu/\sigma$ and $a_2 = 1/\sigma$ we obtain (6) from (4), writing $X^* = Z$. For these a_1, a_2 formula (5) gives $\mu^* = 0$ and $\sigma^{*2} = 1$, as claimed in (b). ∎

Expectation, Moments

Recall that (1) defines the expectation (the mean) of X, the value of X to be expected on the average, written $\mu = E(X)$. More generally, if $g(x)$ is nonconstant and continuous for all x, then $g(X)$ is a random variable. Hence its *mathematical expectation* or, briefly, its **expectation** $E(g(X))$ is the value of $g(X)$ to be expected on the average, defined [similarly to (1)] by

$$\text{(7)} \qquad E(g(X)) = \sum_j g(x_j)f(x_j) \qquad \text{or} \qquad E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)\,dx.$$

In the first formula, f is the probability function of the discrete random variable X. In the second formula, f is the density of the continuous random variable X. Important special cases are the ***k*th moment** of X (where $k = 1, 2, \cdots$)

$$\text{(8)} \qquad E(X^k) = \sum_j x_j^k f(x_j) \qquad \text{or} \qquad \int_{-\infty}^{\infty} x^k f(x)\,dx$$

and the ***k*th central moment** of X ($k = 1, 2, \cdots$)

$$\text{(9)} \qquad E([X - \mu]^k) = \sum_j (x_j - \mu)^k f(x_j) \qquad \text{or} \qquad \int_{-\infty}^{\infty} (x - \mu)^k f(x)\,dx.$$

This includes the first moment, the **mean** of X

$$\text{(10)} \qquad \mu = E(X) \qquad [(8) \text{ with } k = 1].$$

It also includes the second central moment, the **variance** of X

$$\text{(11)} \qquad \sigma^2 = E([X - \mu]^2) \qquad [(9) \text{ with } k = 2].$$

For later use you may prove

$$\text{(12)} \qquad E(1) = 1.$$

PROBLEM SET 24.6

1–6 MEAN, VARIANCE

Find the mean and the variance of the random variable X with probability function or density $f(x)$.

1. $f(x) = 2x \quad (0 \leqq x \leqq 1)$
2. $f(0) = 0.512, \quad f(1) = 0.384, \quad f(2) = 0.096, \quad f(3) = 0.008$
3. $X =$ *Number a fair die turns up*
4. $Y = -4X + 5$ with X as in Prob. 1
5. Uniform distribution on [0, 8]
6. $f(x) = 2e^{-2x} \quad (x \geqq 0)$
7. What is the expected daily profit if a store sells X air conditioners per day with probability $f(10) = 0.1$, $f(11) = 0.3$, $f(12) = 0.4$, $f(13) = 0.2$ and the profit per conditioner is \$55?
8. What is the mean life of a light bulb whose life X [hours] has the density $f(x) = 0.001e^{-0.001x}$ $(x \geqq 0)$?
9. If the mileage (in multiples of 1000 mi) after which a tire must be replaced is given by the random variable X with density $f(x) = \theta e^{-\theta x}$ $(x > 0)$, what mileage can you expect to get on one of these tires? Let $\theta = 0.04$ and find the probability that a tire will last at least 40000 mi.

10. What sum can you expect in rolling a fair die 10 times? Do it. Repeat this experiment 20 times and record how the sum varies.

11. A small filling station is supplied with gasoline every Saturday afternoon. Assume that its volume X of sales in ten thousands of gallons has the probability density $f(x) = 6x(1 - x)$ if $0 \leqq x \leqq 1$ and 0 otherwise. Determine the mean, the variance, and the standardized variable.

12. What capacity must the tank in Prob. 11 have in order that the probability that the tank will be emptied in a given week be 5%?

13. Let X [cm] be the diameter of bolts in a production. Assume that X has the density $f(x) = k(x - 0.9)(1.1 - x)$ if $0.9 < x < 1.1$ and 0 otherwise. Determine k, sketch $f(x)$, and find μ and σ^2.

14. Suppose that in Prob. 13, a bolt is regarded as being defective if its diameter deviates from 1.00 cm by more than 0.09 cm. What percentage of defective bolts should we then expect?

15. For what choice of the maximum possible deviation c from 1.00 cm shall we obtain 3% defectives in Probs. 13 and 14?

16. **TEAM PROJECT. Means, Variances, Expectations.**
(a) Show that $E(X - \mu) = 0$, $\sigma^2 = E(X^2) - \mu^2$.

(b) Prove (10)–(12).

(c) Find all the moments of the uniform distribution on an interval $a \leqq x \leqq b$.

(d) The **skewness** γ of a random variable X is defined by

$$\gamma = \frac{1}{\sigma^3} E([X - \mu]^3). \tag{13}$$

Show that for a symmetric distribution (whose third central moment exists) the skewness is zero.

(e) Find the skewness of the distribution with density $f(x) = xe^{-x}$ when $x > 0$ and $f(x) = 0$ otherwise. Sketch $f(x)$.

(f) Calculate the skewness of a few simple discrete distributions of your own choice.

(g) Find a nonsymmetric discrete distribution with 3 possible values, mean 0, and skewness 0.

24.7 Binomial, Poisson, and Hypergeometric Distributions

These are the three most important ***discrete*** distributions, with numerous applications.

Binomial Distribution

The **binomial distribution** occurs in games of chance (rolling a die, see below, etc.), quality inspection (e.g., counting of the number of defectives), opinion polls (counting number of employees favoring certain schedule changes, etc.), medicine (e.g., recording the number of patients recovered by a new medication), and so on. The conditions of its occurrence are as follows.

We are interested in the number of times an event A occurs in n independent trials. In each trial the event A has the same probability $P(A) = p$. Then in a trial, A will not occur with probability $q = 1 - p$. In n trials the random variable that interests us is

$$X = \textit{Number of times the event } A \textit{ occurs in } n \textit{ trials.}$$

X can assume the values $0, 1, \cdots, n$, and we want to determine the corresponding probabilities. Now $X = x$ means that A occurs in x trials and in $n - x$ trials it does not occur. This may look as follows.

(1)
$$\underbrace{A \quad A \cdots A}_{x \text{ times}} \qquad \underbrace{B \quad B \cdots B}_{n-x \text{ times}}.$$

Here $B = A^c$ is the complement of A, meaning that A does not occur (Sec. 24.2). We now use the assumption that the trials are independent, that is, they do not influence each other. Hence (1) has the probability (see Sec. 24.3 on independent events)

(1*)
$$\underbrace{pp \cdots p}_{x \text{ times}} \cdot \underbrace{qq \cdots q}_{n-x \text{ times}} = p^x q^{n-x}.$$

Now (1) is just one order of arranging x A's and $n - x$ B's. We now use Theorem 1(b) in Sec. 24.4, which gives the number of permutations of n things (the n outcomes of the n trials) consisting of 2 classes, class 1 containing the $n_1 = x$ A's and class 2 containing the $n - n_1 = n - x$ B's. This number is

$$\frac{n!}{x!(n-x)!} = \binom{n}{x}.$$

Accordingly, (1*) multiplied by this binomial coefficient gives the probability $P(X = x)$ of $X = x$, that is, of obtaining A precisely x times in n trials. Hence X has the probability function

(2)
$$f(x) = \binom{n}{x} p^x q^{n-x} \qquad (x = 0, 1, \cdots, n)$$

and $f(x) = 0$ otherwise. The distribution of X with probability function (2) is called the **binomial distribution** or *Bernoulli distribution.* The occurrence of A is called *success* (regardless of what it actually is; it may mean that you miss your plane or lose your watch) and the nonoccurrence of A is called *failure.* Figure 516 shows typical examples. Numeric values can be obtained from Table A5 in App. 5 or from your CAS.

The mean of the binomial distribution is (see Team Project 16)

(3)
$$\mu = np$$

and the variance is (see Team Project 16)

(4)
$$\sigma^2 = npq.$$

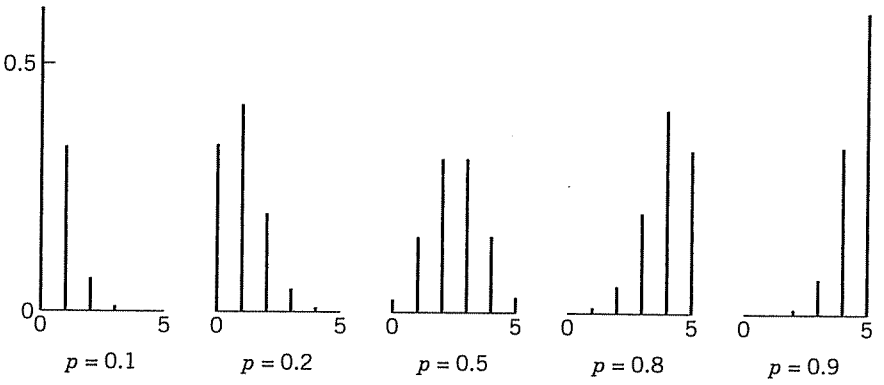


Fig. 516. Probability function (2) of the binomial distribution for $n = 5$ and various values of p

For the ***symmetric case*** of equal chance of success and failure ($p = q = 1/2$) this gives the mean $n/2$, the variance $n/4$, and the probability function

$$(2^*) \qquad f(x) = \binom{n}{x}\left(\frac{1}{2}\right)^n \qquad (x = 0, 1, \cdots, n).$$

EXAMPLE 1 **Binomial Distribution**

Compute the probability of obtaining at least two *"Six"* in rolling a fair die 4 times.

Solution. $p = P(A) = P(\text{"Six"}) = 1/6$, $q = 5/6$, $n = 4$. The event *"At least two 'Six'"* occurs if we obtain 2 or 3 or 4 "Six." Hence the answer is

$$P = f(2) + f(3) + f(4) = \binom{4}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^2 + \binom{4}{3}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right) + \binom{4}{4}\left(\frac{1}{6}\right)^4$$

$$= \frac{1}{6^4}(6 \cdot 25 + 4 \cdot 5 + 1) = \frac{171}{1296} = 13.2\%. \qquad \blacksquare$$

Poisson Distribution

The discrete distribution with infinitely many possible values and probability function

$$(5) \qquad f(x) = \frac{\mu^x}{x!}e^{-\mu} \qquad (x = 0, 1, \cdots)$$

is called the **Poisson distribution,** named after S. D. Poisson (Sec. 18.5). Figure 517 shows (5) for some values of μ. It can be proved that this distribution is obtained as a limiting case of the binomial distribution, if we let $p \to 0$ and $n \to \infty$ so that the mean $\mu = np$ approaches a finite value. (For instance, $\mu = np$ may be kept constant.) The Poisson distribution has the mean μ and the variance (see Team Project 16)

$$(6) \qquad \sigma^2 = \mu.$$

Figure 517 gives the impression that with increasing mean the spread of the distribution increases, thereby illustrating formula (6), and that the distribution becomes more and more (approximately) symmetric.

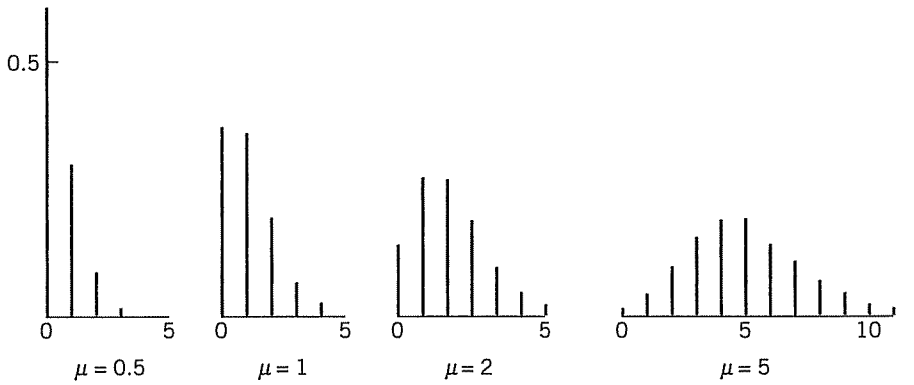


Fig. 517. Probability function (5) of the Poisson distribution for various values of μ

EXAMPLE 2 **Poisson Distribution**

If the probability of producing a defective screw is $p = 0.01$, what is the probability that a lot of 100 screws will contain more than 2 defectives?

Solution. The complementary event is A^c: *Not more than* 2 *defectives.* For its probability we get from the binomial distribution with mean $\mu = np = 1$ the value [see (2)]

$$P(A^c) = \binom{100}{0} 0.99^{100} + \binom{100}{1} 0.01 \cdot 0.99^{99} + \binom{100}{2} 0.01^2 \cdot 0.99^{98}.$$

Since p is very small, we can approximate this by the much more convenient Poisson distribution with mean $\mu = np = 100 \cdot 0.01 = 1$, obtaining [see (5)]

$$P(A^c) \approx e^{-1}\left(1 + 1 + \frac{1}{2}\right)$$

$$= 91.97\%.$$

Thus $P(A) = 8.03\%$. Show that the binomial distribution gives $P(A) = 7.94\%$, so that the Poisson approximation is quite good. ■

EXAMPLE 3 **Parking Problems. Poisson Distribution**

If on the average, 2 cars enter a certain parking lot per minute, what is the probability that during any given minute 4 or more cars will enter the lot?

Solution. To understand that the Poisson distribution is a model of the situation, we imagine the minute to be divided into very many short time intervals, let p be the (constant) probability that a car will enter the lot during any such short interval, and assume independence of the events that happen during those intervals. Then we are dealing with a binomial distribution with very large n and very small p, which we can approximate by the Poisson distribution with

$$\mu = np = 2,$$

because 2 cars enter on the average. The complementary event of the event "4 cars or more during a given minute" is "3 *cars or fewer enter the lot*" and has the probability

$$f(0) + f(1) + f(2) + f(3) = e^{-2}\left(\frac{2^0}{0!} + \frac{2^1}{1!} + \frac{2^2}{2!} + \frac{2^3}{3!}\right)$$

$$= 0.857.$$

Answer: 14.3%. (Why did we consider that complement?) ■

Sampling with Replacement

This means that we draw things from a given set one by one, and after each trial we replace the thing drawn (put it back to the given set and mix) before we draw the next thing. This guarantees independence of trials and leads to the **binomial distribution.** Indeed, if a box contains N things, for example, screws, M of which are defective, the probability of drawing a defective screw in a trial is $p = M/N$. Hence the probability of drawing a nondefective screw is $q = 1 - p = 1 - M/N$, and (2) gives the probability of drawing x defectives in n trials in the form

(7) $$f(x) = \binom{n}{x}\left(\frac{M}{N}\right)^x\left(1 - \frac{M}{N}\right)^{n-x} \qquad (x = 0, 1, \cdots, n).$$

Sampling without Replacement. Hypergeometric Distribution

Sampling without replacement means that we return no screw to the box. Then we no longer have independence of trials (why?), and instead of (7) the probability of drawing x defectives in n trials is

$$\textbf{(8)} \qquad f(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \qquad (x = 0, 1, \cdots, n).$$

The distribution with this probability function is called the **hypergeometric distribution** (because its moment generating function (see Team Project 16) can be expressed by the hypergeometric function defined in Sec. 5.4, a fact that we shall not use).

Derivation of (8). By (4a) in Sec. 24.4 there are

(a) $\binom{N}{n}$ different ways of picking n things from N,

(b) $\binom{M}{x}$ different ways of picking x defectives from M,

(c) $\binom{N-M}{n-x}$ different ways of picking $n - x$ nondefectives from $N - M$,

and each way in (b) combined with each way in (c) gives the total number of mutually exclusive ways of obtaining x defectives in n drawings without replacement. Since (a) is the total number of outcomes and we draw at random, each such way has the probability $1\Big/\binom{N}{n}$. From this, (8) follows. ▩

The hypergeometric distribution has the mean (Team Project 16)

$$\textbf{(9)} \qquad \mu = n\frac{M}{N}$$

and the variance

$$\textbf{(10)} \qquad \sigma^2 = \frac{nM(N-M)(N-n)}{N^2(N-1)}.$$

EXAMPLE 4 Sampling with and without Replacement

We want to draw random samples of two gaskets from a box containing 10 gaskets, three of which are defective. Find the probability function of the random variable X = *Number of defectives in the sample.*

Solution. We have $N = 10$, $M = 3$, $N - M = 7$, $n = 2$. For sampling with replacement, (7) yields

$$f(x) = \binom{2}{x}\left(\frac{3}{10}\right)^x\left(\frac{7}{10}\right)^{2-x}, \qquad f(0) = 0.49, \quad f(1) = 0.42, \quad f(2) = 0.09.$$

For sampling without replacement we have to use (8), finding

$$f(x) = \binom{3}{x}\binom{7}{2-x}\Big/\binom{10}{2}, \qquad f(0) = f(1) = \frac{21}{45} \approx 0.47, \quad f(2) = \frac{3}{45} \approx 0.07.$$ ▩

If N, M, and $N - M$ are large compared with n, then it does not matter too much whether we sample with or without replacement, and in this case the hypergeometric distribution may be approximated by the binomial distribution (with $p = M/N$), *which is somewhat simpler.*

Hence in sampling from an indefinitely large population (**"infinite population"**) *we may use the binomial distribution, regardless of whether we sample with or without replacement.*

PROBLEM SET 24.7

1. Four fair coins are tossed simultaneously. Find the probability function of the random variable X = *Number of heads* and compute the probabilities of obtaining no heads, precisely 1 head, at least 1 head, not more than 3 heads.
2. If the probability of hitting a target in a single shot is 10% and 10 shots are fired independently, what is the probability that the target will be hit at least once?
3. In Prob. 2, if the probability of hitting would be 5% and we fired 20 shots, would the probability of hitting at least once be less than, equal to, or greater than in Prob. 2? Guess first, then compute.
4. Suppose that 3% of bolts made by a machine are defective, the defectives occurring at random during production. If the bolts are packaged 50 per box, what is the Poisson approximation of the probability that a given box will contain $x = 0, 1, \cdots, 5$ defectives?
5. Let X be the number of cars per minute passing a certain point of some road between 8 A.M. and 10 A.M. on a Sunday. Assume that X has a Poisson distribution with mean 5. Find the probability of observing 3 or fewer cars during any given minute.
6. Suppose that a telephone switchboard of some company on the average handles 300 calls per hour, and that the board can make at most 10 connections per minute. Using the Poisson distribution, estimate the probability that the board will be overtaxed during a given minute. (Use Table A6 in App. 5 or your CAS.)
7. **(Rutherford–Geiger experiments)** In 1910, E. Rutherford and H. Geiger showed experimentally that the number of alpha particles emitted per second in a radioactive process is a random variable X having a Poisson distribution. If X has mean 0.5, what is the probability of observing two or more particles during any given second?
8. A process of manufacturing screws is checked every hour by inspecting n screws selected at random from that hour's production. If one or more screws are defective, the process is halted and carefully examined. How large should n be if the manufacturer wants the probability to be about 95% that the process will be halted when 10% of the screws being produced are defective? (Assume independence of the quality of any screw of that of the other screws.)
9. Suppose that in the production of 50-Ω resistors, nondefective items are those that have a resistance between 45 Ω and 55 Ω and the probability of a resistor's being defective is 0.2%. The resistors are sold in lots of 100, with the guarantee that all resistors are nondefective. What is the probability that a given lot will violate this guarantee? (Use the Poisson distribution.)
10. Let $p = 1\%$ be the probability that a certain type of lightbulb will fail in a 24-hr test. Find the probability that a sign consisting of 10 such bulbs will burn 24 hours with no bulb failures.
11. Guess how much less the probability in Prob. 10 would be if the sign consisted of 100 bulbs. Then calculate.
12. Suppose that a certain type of magnetic tape contains, on the average, 2 defects per 100 meters. What is the probability that a roll of tape 300 meters long will contain (a) x defects, (b) no defects?
13. Suppose that a test for extrasensory perception consists of naming (in any order) 3 cards randomly drawn from a deck of 13 cards. Find the probability that by chance alone, the person will correctly name (a) no cards, (b) 1 card, (c) 2 cards, (d) 3 cards.
14. A carton contains 20 fuses, 5 of which are defective. Find the probability that, if a sample of 3 fuses is chosen from the carton by random drawing without replacement, x fuses in the sample will be defective.
15. **(Multinomial distribution)** Suppose a trial can result in precisely one of k mutually exclusive events $A_1, \cdots, A_k$ with probabilities $p_1, \cdots, p_k$, respectively, where $p_1 + \cdots + p_k = 1$. Suppose that n independent trials are performed. Show that the probability of getting x_1 A_1's, $\cdots$, x_k A_k's is

$$f(x_1, \cdots, x_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$$

where $0 \leqq x_j \leqq n$, $j = 1, \cdots, k$, and $x_1 + \cdots + x_k = n$. The distribution having this

probability function is called the *multinomial distribution.*

16. TEAM PROJECT. Moment Generating Function. The moment generating function $G(t)$ is defined by

$$G(t) = E(e^{tX_j}) = \sum_j e^{tx_j} f(x_j)$$

or

$$G(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx$$

where X is a discrete or continuous random variable, respectively.

(a) Assuming that termwise differentiation and differentiation under the integral sign are permissible, show that $E(X^k) = G^{(k)}(0)$, where $G^{(k)} = d^k G/dt^k$, in particular, $\mu = G'(0)$.

(b) Show that the binomial distribution has the moment generating function

$$G(t) = \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x q^{n-x} = \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x q^{n-x} = (pe^t + q)^n.$$

(c) Using (b), prove (3).

(d) Prove (4).

(e) Show that the Poisson distribution has the moment generating function $G(t) = e^{-\mu} e^{\mu e^t}$ and prove (6).

(f) Prove $x \binom{M}{x} = M \binom{M-1}{x-1}$.

Using this, prove (9).

24.8 Normal Distribution

Turning from discrete to continuous distributions, in this section we discuss the normal distribution. This is the most important continuous distribution because in applications many random variables are **normal random variables** (that is, they have a normal distribution) or they are approximately normal or can be transformed into normal random variables in a relatively simple fashion. Furthermore, the normal distribution is a useful approximation of more complicated distributions, and it also occurs in the proofs of various statistical tests.

The **normal distribution** or *Gauss distribution* is defined as the distribution with the density

$$\textbf{(1)} \qquad f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \qquad (\sigma > 0)$$

where exp is the exponential function with base $e = 2.718\cdots$. This is simpler than it may at first look. $f(x)$ has these features (see also Fig. 518).

1. μ is the mean and σ the standard deviation.

2. $1/(\sigma\sqrt{2\pi})$ is a constant factor that makes the area under the curve of $f(x)$ from $-\infty$ to ∞ equal to 1, as it must be by (10), Sec. 24.5.

3. The curve of $f(x)$ is symmetric with respect to $x = \mu$ because the exponent is quadratic. Hence for $\mu = 0$ it is symmetric with respect to the y-axis $x = 0$ (Fig. 518, *"bell-shaped curves"*).

4. The exponential function in (1) goes to zero very fast—the faster the smaller the standard deviation σ is, as it should be (Fig. 518).

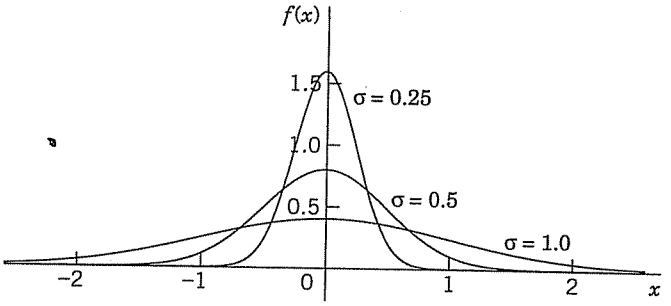


Fig. 518. Density (1) of the normal distribution with $\mu = 0$ for various values of σ

Distribution Function $F(x)$

From (7) in Sec. 24.5 and (1) we see that the normal distribution has the **distribution function**

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left[-\frac{1}{2}\left(\frac{v-\mu}{\sigma}\right)^2\right] dv. \tag{2}$$

Here we needed x as the upper limit of integration and wrote v (instead of x) in the integrand.

For the corresponding **standardized normal distribution** with mean 0 and standard deviation 1 we denote $F(x)$ by $\Phi(z)$. Then we simply have from (2)

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-u^2/2}\, du. \tag{3}$$

This integral cannot be integrated by one of the methods of calculus. But this is no serious handicap because its values can be obtained from Table A7 in App. 5 or from your CAS. These values are needed in working with the normal distribution. The curve of $\Phi(z)$ is S-shaped. It increases monotone (why?) from 0 to 1 and intersects the vertical axis at 1/2 (why?), as shown in Fig. 519.

Relation Between $F(x)$ and $\Phi(z)$. Although your CAS will give you values of $F(x)$ in (2) with any μ and σ directly, it is important to comprehend that and why any such an $F(x)$ can be expressed in terms of the tabulated standard $\Phi(z)$, as follows.

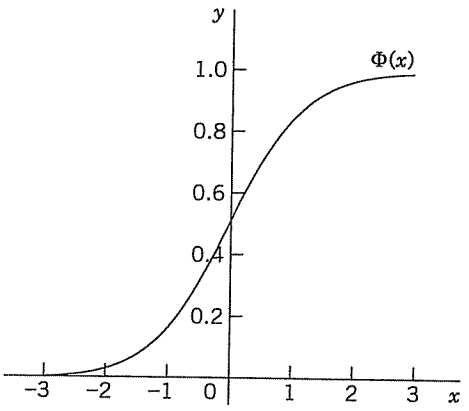


Fig. 519. Distribution function $\Phi(z)$ of the normal distribution with mean 0 and variance 1

THEOREM 1

Use of the Normal Table A7 in App. 5

The distribution function $F(x)$ of the normal distribution with any μ and σ [see (2)] *is related to the standardized distribution function $\Phi(z)$ in* (3) *by the formula*

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right). \tag{4}$$

PROOF Comparing (2) and (3) we see that we should set

$$u = \frac{v-\mu}{\sigma}. \qquad \text{Then } v = x \text{ gives} \qquad u = \frac{x-\mu}{\sigma}$$

as the new upper limit of integration. Also $v - \mu = \sigma u$, thus $dv = \sigma\, du$. Together, since σ drops out,

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{(x-\mu)/\sigma} e^{-u^2/2}\, \sigma\, du = \Phi\left(\frac{x-\mu}{\sigma}\right). \qquad \blacksquare$$

Probabilities corresponding to intervals will be needed quite frequently in statistics in Chap. 25. These are obtained as follows.

THEOREM 2

Normal Probabilities for Intervals

The probability that a normal random variable X with mean μ and standard deviation σ assume any value in an interval $a < x \leqq b$ is

$$P(a < X \leqq b) = F(b) - F(a) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right). \tag{5}$$

PROOF Formula (2) in Sec. 24.5 gives the first equality in (5), and (4) in this section gives the second equality. $\blacksquare$

Numeric Values

In practical work with the normal distribution it is good to remember that about 2/3 of all values of X to be observed will lie between $\mu \pm \sigma$, about 95% between $\mu \pm 2\sigma$, and practically all between the **three-sigma limits** $\mu \pm 3\sigma$. More precisely, by Table A7 in App. 5,

$$\begin{aligned} &\textbf{(a)} \quad P(\mu - \sigma < X \leqq \mu + \sigma) \approx 68\% \\ &\textbf{(b)} \quad P(\mu - 2\sigma < X \leqq \mu + 2\sigma) \approx 95.5\% \\ &\textbf{(c)} \quad P(\mu - 3\sigma < X \leqq \mu + 3\sigma) \approx 99.7\%. \end{aligned} \tag{6}$$

Formulas (6a) and (6b) are illustrated in Fig. 520.

The formulas in (6) show that a value deviating from μ by more than σ, 2σ, or 3σ will occur in one of about 3, 20, and 300 trials, respectively.

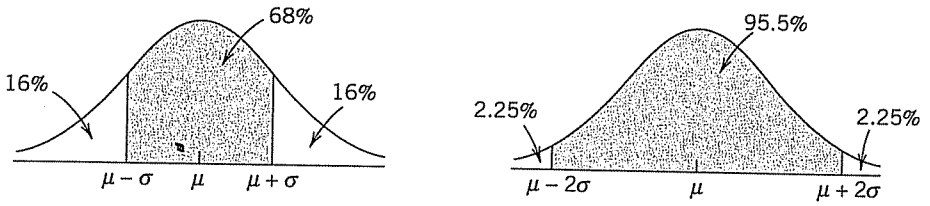


Fig. 520. Illustration of formula (6)

In tests (Chap. 25) we shall ask conversely for the intervals that correspond to certain given probabilities; practically most important are the probabilities of 95%, 99%, and 99.9%. For these, Table A8 in App. 5 gives the answers $\mu \pm 2\sigma$, $\mu \pm 2.5\sigma$, and $\mu \pm 3.3\sigma$, respectively. More precisely,

$$
\begin{aligned}
&\text{(a)} \quad P(\mu - 1.96\sigma < X \leqq \mu + 1.96\sigma) = 95\% \\
(7) \qquad &\text{(b)} \quad P(\mu - 2.58\sigma < X \leqq \mu + 2.58\sigma) = 99\% \\
&\text{(c)} \quad P(\mu - 3.29\sigma < X \leqq \mu + 3.29\sigma) = 99.9\%.
\end{aligned}
$$

Working With the Normal Tables A7 and A8 in App. 5

There are two normal tables in App. 5, Tables A7 and A8. If you want probabilities, use Table A7. If probabilities are given and corresponding intervals or x-values are wanted, use Table A8. The following examples are typical. Do them with care, verifying all values, and don't just regard them as dull exercises for your software. Make sketches of the density to see whether the results look reasonable.

EXAMPLE 1 **Reading Entries from Table A7**

If X is standardized normal (so that $\mu = 0$, $\sigma = 1$), then

$$P(X \leqq 2.44) = 0.9927 \approx 99\tfrac{1}{4}\%$$

$$P(X \leqq -1.16) = 1 - \Phi(1.16) = 1 - 0.8770 = 0.1230 = 12.3\%$$

$$P(X \geqq 1) = 1 - P(X \leqq 1) = 1 - 0.8413 = 0.1587 \text{ by (7), Sec. 24.3}$$

$$P(1.0 \leqq X \leqq 1.8) = \Phi(1.8) - \Phi(1.0) = 0.9641 - 0.8413 = 0.1228.$$ ■

EXAMPLE 2 **Probabilities for Given Intervals, Table A7**

Let X be normal with mean 0.8 and variance 4 (so that $\sigma = 2$). Then by (4) and (5)

$$P(X \leqq 2.44) = F(2.44) = \Phi\left(\frac{2.44 - 0.80}{2}\right) = \Phi(0.82) = 0.7939 \approx 80\%$$

or if you like it better (similarly in the other cases)

$$P(X \leqq 2.44) = P\left(\frac{X - 0.80}{2} \leqq \frac{2.44 - 0.80}{2}\right) = P(Z \leqq 0.82) = 0.7939$$

$$P(X \geqq 1) = 1 - P(X \leqq 1) = 1 - \Phi\left(\frac{1 - 0.8}{2}\right) = 1 - 0.5398 = 0.4602$$

$$P(1.0 \leqq X \leqq 1.8) = \Phi(0.5) - \Phi(0.1) = 0.6915 - 0.5398 = 0.1517.$$ ■

EXAMPLE 3 **Unknown Values c for Given Probabilities, Table A8**

Let X be normal with mean 5 and variance 0.04 (hence standard deviation 0.2). Find c or k corresponding to the given probability

$$P(X \leqq c) = 95\%, \qquad \Phi\left(\frac{c-5}{0.2}\right) = 95\%, \qquad \frac{c-5}{0.2} = 1.645, \qquad c = 5.329$$

$$P(5-k \leqq X \leqq 5+k) = 90\%, \qquad 5+k = 5.329 \qquad \text{(as before; why?)}$$

$$P(X \geqq c) = 1\%, \qquad \text{thus } P(X \leqq c) = 99\%, \qquad \frac{c-5}{0.2} = 2.326, \qquad c = 5.465. \quad \blacksquare$$

EXAMPLE 4 **Defectives**

In a production of iron rods let the diameter X be normally distributed with mean 2 in. and standard deviation 0.008 in.

(a) What percentage of defectives can we expect if we set the tolerance limits at 2 ± 0.02 in.?

(b) How should we set the tolerance limits to allow for 4% defectives?

Solution. **(a)** $1\frac{1}{4}\%$ because from (5) and Table A7 we obtain for the complementary event the probability

$$\begin{aligned} P(1.98 \leqq X \leqq 2.02) &= \Phi\left(\frac{2.02-2.00}{0.008}\right) - \Phi\left(\frac{1.98-2.00}{0.008}\right) \\ &= \Phi(2.5) - \Phi(-2.5) \\ &= 0.9938 - (1 - 0.9938) \\ &= 0.9876 \\ &\approx 98\tfrac{3}{4}\%. \end{aligned}$$

(b) 2 ± 0.0164 because for the complementary event we have

$$0.96 = P(2-c \leqq X \leqq 2+c)$$

or

$$0.98 = P(X \leqq 2+c)$$

so that Table A8 gives

$$0.98 = \Phi\left(\frac{2+c-2}{0.008}\right),$$

$$\frac{2+c-2}{0.008} = 2.054, \qquad c = 0.0164. \quad \blacksquare$$

Normal Approximation of the Binomial Distribution

The probability function of the binomial distribution is (Sec. 24.7)

$$(8) \qquad f(x) = \binom{n}{x} p^x q^{n-x} \qquad (x = 0, 1, \cdots, n).$$

If n is large, the binomial coefficients and powers become very inconvenient. It is of great practical (and theoretical) importance that in this case the normal distribution provides a good approximation of the binomial distribution, according to the following theorem, one of the most important theorems in all probability theory.

THEOREM 3

Limit Theorem of De Moivre and Laplace

For large n,

$$f(x) \sim f^*(x) \qquad (x = 0, 1, \cdots, n). \tag{9}$$

Here f is given by (8). *The function*

$$f^*(x) = \frac{1}{\sqrt{2\pi}\sqrt{npq}}\, e^{-z^2/2}, \qquad z = \frac{x - np}{\sqrt{npq}} \tag{10}$$

is the density of the normal distribution with mean $\mu = np$ *and variance* $\sigma^2 = npq$ *(the mean and variance of the binomial distribution). The symbol* $\sim$ (read **asymptotically equal**) *means that the ratio of both sides approaches* 1 *as n approaches* ∞. *Furthermore, for any nonnegative integers a and b* $(> a)$,

$$P(a \leqq X \leqq b) = \sum_{x=a}^{b} \binom{n}{x} p^x q^{n-x} \sim \Phi(\beta) - \Phi(\alpha), \tag{11}$$

$$\alpha = \frac{a - np - 0.5}{\sqrt{npq}}, \qquad \beta = \frac{b - np + 0.5}{\sqrt{npq}}.$$

A proof of this theorem can be found in [G3] listed in App. 1. The proof shows that the term 0.5 in α and β is a correction caused by the change from a discrete to a continuous distribution.

PROBLEM SET 24.8

1–13 NORMAL DISTRIBUTION

1. Let X be normal with mean 80 and variance 9. Find $P(X > 83)$, $P(X < 81)$, $P(X < 80)$, and $P(78 < X < 82)$.

2. Let X be normal with mean 120 and variance 16. Find $P(X \leqq 126)$, $P(X > 116)$, $P(125 < X < 130)$.

3. Let X be normal with mean 14 and variance 4. Determine c such that $P(X \leqq c) = 95\%$, $P(X \leqq c) = 5\%$, $P(X \leqq c) = 99.5\%$.

4. Let X be normal with mean 4.2 and variance 0.04. Find c such that $P(X \leqq c) = 50\%$, $P(X > c) = 10\%$, $P(-c < X - 4.2 \leqq c) = 99\%$.

5. If the lifetime X of a certain kind of automobile battery is normally distributed with a mean of 4 yr and a standard deviation of 1 yr, and the manufacturer wishes to guarantee the battery for 3 yr, what percentage of the batteries will he have to replace under the guarantee?

6. If the standard deviation in Prob. 5 were smaller, would that percentage be smaller or larger?

7. A manufacturer knows from experience that the resistance of resistors he produces is normal with mean $\mu = 150\ \Omega$ and standard deviation $\sigma = 5\ \Omega$. What percentage of the resistors will have resistance between 148 Ω and 152 Ω? Between 140 Ω and 160 Ω?

8. The breaking strength X [kg] of a certain type of plastic block is normally distributed with a mean of 1250 kg and a standard deviation of 55 kg. What is the maximum load such that we can expect no more than 5% of the blocks to break?

9. A manufacturer produces airmail envelopes whose weight is normal with mean $\mu = 1.950$ grams and standard deviation $\sigma = 0.025$ grams. The envelopes are sold in lots of 1000. How many envelopes in a lot will be heavier than 2 grams?

10. If the resistance X of certain wires in an electrical network is normal with mean 0.01 Ω and standard deviation 0.001 Ω, how many of 1000 wires will meet the specification that they have resistance between 0.009 and 0.011 Ω?

11. If the mathematics scores of the SAT college entrance exams are normal with mean 480 and standard deviation 100 (these are about the actual values over the past years) and if some college sets 500 as the minimum score for new students, what percent of students will not reach that score?

12. If the monthly machine repair and maintenance cost X in a certain factory is known to be normal with mean \$12000 and standard deviation \$2000, what is the probability that the repair cost for the next month will exceed the budgeted amount of \$15000?

13. If sick-leave time X used by employees of a company in one month is (very roughly) normal with mean 1000 hours and standard deviation 100 hours, how much time t should be budgeted for sick leave during the next month if t is to be exceeded with probability of only 20%?

14. TEAM PROJECT. Normal Distribution. **(a)** Derive the formulas in (6) and (7) from the appropriate normal table.

(b) Show that $\Phi(-z) = 1 - \Phi(z)$. Give an example.

(c) Find the points of inflection of the curve of (1).

(d) Considering $\Phi^2(\infty)$ and introducing polar coordinates in the double integral (a standard trick worth remembering), prove

$$(12) \qquad \Phi(\infty) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-u^2/2}\, du = 1.$$

(e) Show that σ in (1) is indeed the standard deviation of the normal distribution. [Use (12).]

(f) Bernoulli's law of large numbers. In an experiment let an event A have probability p $(0 < p < 1)$, and let X be the number of times A happens in n independent trials. Show that for any given $\epsilon > 0$,

$$P\left(\left|\frac{X}{n} - p\right| \leqq \epsilon\right) \to 1 \qquad \text{as } n \to \infty.$$

(g) Transformation. If X is normal with mean μ and variance σ^2, show that $X^* = c_1X + c_2$ $(c_1 > 0)$ is normal with mean $\mu^* = c_1\mu + c_2$ and variance $\sigma^{*2} = c_1{}^2\sigma^2$.

15. WRITING PROJECT. Use of Tables. Give a systematic discussion of the use of Tables A7 and A8 for obtaining $P(X < b)$, $P(X > a)$, $P(a < X < b)$, $P(X < c) = k$, $P(X > c) = k$, as well as $P(\mu - c < X < \mu + c) = k$; include simple examples. If you have a CAS, describe to what extent it makes the use of those tables superfluous; give examples.

24.9 Distributions of Several Random Variables

Distributions of two or more random variables are of interest for two reasons:

1. They occur in experiments in which we observe several random variables, for example, carbon content X and hardness Y of steel, amount of fertilizer X and yield of corn Y, height X_1, weight X_2, and blood pressure X_3 of persons, and so on.

2. They will be needed in the mathematical justification of the methods of statistics in Chap. 25.

In this section we consider two random variables X and Y or, as we also say, a **two-dimensional random variable** (X, Y). For (X, Y) the outcome of a trial is a pair of numbers $X = x$, $Y = y$, briefly $(X, Y) = (x, y)$, which we can plot as a point in the XY-plane.

The **two-dimensional probability distribution** of the random variable (X, Y) is given by the **distribution function**

$$(1) \qquad F(x, y) = P(X \leqq x, Y \leqq y).$$

This is the probability that in a trial, X will assume any value not greater than x and in the same trial, Y will assume any value not greater than y. This corresponds to the blue region in Fig. 521, which extends to $-\infty$ to the left and below. $F(x, y)$ determines the

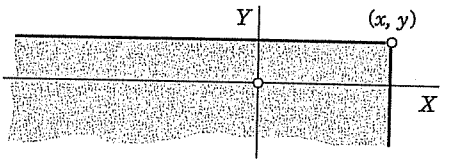


Fig. 521. Formula (1)

probability distribution uniquely, because in analogy to formula (2) in Sec. 24.5, that is, $P(a < X \leqq b) = F(b) - F(a)$, we now have for a rectangle (see Prob. 14)

$$(2) \quad P(a_1 < X \leqq b_1, \quad a_2 < Y \leqq b_2) = F(b_1, b_2) - F(a_1, b_2) - F(b_1, a_2) + F(a_1, a_2).$$

As before, in the two-dimensional case we shall also have discrete and continuous random variables and distributions.

Discrete Two-Dimensional Distributions

In analogy to the case of a single random variable (Sec. 24.5), we call (X, Y) and its distribution **discrete** if (X, Y) can assume only finitely many or at most countably infinitely many pairs of values (x_1, y_1), (x_2, y_2), $\cdots$ with positive probabilities, whereas the probability for any domain containing none of those values of (X, Y) is zero.

Let (x_i, y_j) be any of those pairs and let $P(X = x_i, Y = y_j) = p_{ij}$ (where we admit that p_{ij} may be 0 for certain pairs of subscripts i, j). Then we define the **probability function** $f(x, y)$ of (X, Y) by

$$(3) \qquad f(x, y) = p_{ij} \quad \text{if} \quad x = x_i, \, y = y_j \qquad \text{and} \qquad f(x, y) = 0 \quad \text{otherwise;}$$

here, $i = 1, 2, \cdots$ and $j = 1, 2, \cdots$ independently. In analogy to (4), Sec. 24.5, we now have for the distribution function the formula

$$(4) \qquad F(x, y) = \sum_{x_i \leqq x} \sum_{y_j \leqq y} f(x_i, y_j).$$

Instead of (6) in Sec. 24.5 we now have the condition

$$(5) \qquad \sum_i \sum_j f(x_i, y_j) = 1.$$

EXAMPLE 1 **Two-Dimensional Discrete Distribution**

If we simultaneously toss a dime and a nickel and consider

X = *Number of heads the dime turns up,*

Y = *Number of heads the nickel turns up,*

then X and Y can have the values 0 or 1, and the probability function is

$$f(0, 0) = f(1, 0) = f(0, 1) = f(1, 1) = \tfrac{1}{4}, \quad f(x, y) = 0 \text{ otherwise.}$$

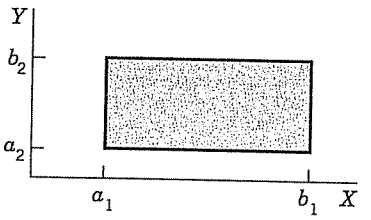


Fig. 522. Notion of a two-dimensional distribution

Continuous Two-Dimensional Distributions

In analogy to the case of a single random variable (Sec. 24.5) we call (X, Y) and its distribution **continuous** if the corresponding distribution function $F(x, y)$ can be given by a double integral

$$F(x, y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f(x^*, y^*)\, dx^*\, dy^* \tag{6}$$

whose integrand f, called the **density** of (X, Y), is nonnegative everywhere, and is continuous, possibly except on finitely many curves.

From (6) we obtain the probability that (X, Y) assume any value in a rectangle (Fig. 522) given by the formula

$$P(a_1 < X \leqq b_1, \quad a_2 < Y \leqq b_2) = \int_{a_2}^{b_2} \int_{a_1}^{b_1} f(x, y)\, dx\, dy. \tag{7}$$

EXAMPLE 2 **Two-Dimensional Uniform Distribution in a Rectangle**

Let R be the rectangle $\alpha_1 < x \leqq \beta_1$, $\alpha_2 < y \leqq \beta_2$. The density (see Fig. 523)

$$f(x, y) = 1/k \quad \text{if } (x, y) \text{ is in } R, \qquad f(x, y) = 0 \text{ otherwise} \tag{8}$$

defines the so-called **uniform distribution** *in the rectangle* R; here $k = (\beta_1 - \alpha_1)(\beta_2 - \alpha_2)$ is the area of R. The distribution function is shown in Fig. 524. ■

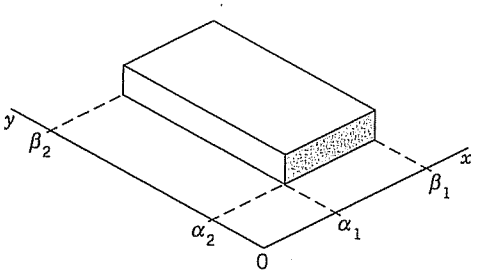


Fig. 523. Density function (8) of the uniform distribution

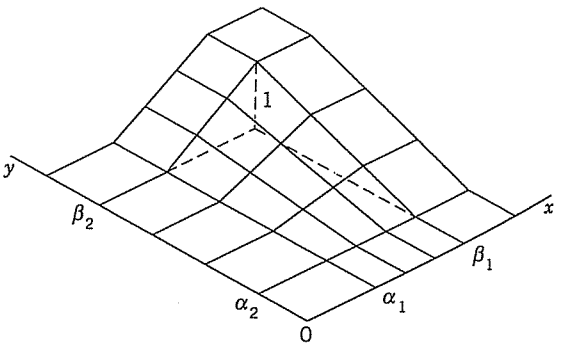


Fig. 524. Distribution function of the uniform distribution defined by (8)

Marginal Distributions of a Discrete Distribution

This is a rather natural idea, without counterpart for a single random variable. It amounts to being interested only in one of the two variables in (X, Y), say, X, and asking for its

distribution, called the **marginal distribution** of X in (X, Y). So we ask for the probability $P(X = x, Y \text{ arbitrary})$. Since (X, Y) is discrete, so is X. We get its probability function, call it $f_1(x)$, from the probability function $f(x, y)$ of (X, Y) by summing over y:

$$f_1(x) = P(X = x, Y \text{ arbitrary}) = \sum_y f(x, y) \tag{9}$$

where we sum all the values of $f(x, y)$ that are not 0 for that x.

From (9) we see that the distribution function of the marginal distribution of X is

$$F_1(x) = P(X \leqq x, Y \text{ arbitrary}) = \sum_{x^* \leqq x} f_1(x^*). \tag{10}$$

Similarly, the probability function

$$f_2(y) = P(X \text{ arbitrary}, Y = y) = \sum_x f(x, y) \tag{11}$$

determines the **marginal distribution** of Y in (X, Y). Here we sum all the values of $f(x, y)$ that are not zero for the corresponding y. The distribution function of this marginal distribution is

$$F_2(y) = P(X \text{ arbitrary}, Y \leqq y) = \sum_{y^* \leqq y} f_2(y^*). \tag{12}$$

EXAMPLE 3 **Marginal Distributions of a Discrete Two-Dimensional Random Variable**

In drawing 3 cards with replacement from a bridge deck let us consider

$$(X, Y), \qquad X = \textit{Number of queens}, \qquad Y = \textit{Number of kings or aces}.$$

The deck has 52 cards. These include 4 queens, 4 kings, and 4 aces. Hence in a single trial a queen has probability $4/52 = 1/13$ and a king or ace $8/52 = 2/13$. This gives the probability function of (X, Y),

$$f(x, y) = \frac{3!}{x!\, y!\, (3 - x - y)!} \left(\frac{1}{13}\right)^x \left(\frac{2}{13}\right)^y \left(\frac{10}{13}\right)^{3-x-y} \qquad (x + y \leqq 3)$$

and $f(x, y) = 0$ otherwise. Table 24.1 shows in the center the values of $f(x, y)$ and on the right and lower margins the values of the probability functions $f_1(x)$ and $f_2(y)$ of the marginal distributions of X and Y, respectively. ▩

Table 24.1 Values of the Probability Functions $f(x, y)$, $f_1(x)$, $f_2(y)$ in Drawing Three Cards with Replacement from a Bridge Deck, where X is the Number of Queens Drawn and Y is the Number of Kings or Aces Drawn

x \ y	0	1	2	3	$f_1(x)$
0	$\frac{1000}{2197}$	$\frac{600}{2197}$	$\frac{120}{2197}$	$\frac{8}{2197}$	$\frac{1728}{2197}$
1	$\frac{300}{2197}$	$\frac{120}{2197}$	$\frac{12}{2197}$	0	$\frac{432}{2197}$
2	$\frac{30}{2197}$	$\frac{6}{2197}$	0	0	$\frac{36}{2197}$
3	$\frac{1}{2197}$	0	0	0	$\frac{1}{2197}$
$f_2(y)$	$\frac{1331}{2197}$	$\frac{726}{2197}$	$\frac{132}{2197}$	$\frac{8}{2197}$	

Marginal Distributions of a Continuous Distribution

This is conceptually the same as for discrete distributions, with probability functions and sums replaced by densities and integrals. For a continuous random variable (X, Y) with density $f(x, y)$ we now have the **marginal distribution** of X in (X, Y), defined by the distribution function

$$F_1(x) = P(X \leqq x, -\infty < Y < \infty) = \int_{-\infty}^{x} f_1(x^*)\, dx^* \tag{13}$$

with the density f_1 of X obtained from $f(x, y)$ by integration over y,

$$f_1(x) = \int_{-\infty}^{\infty} f(x, y)\, dy. \tag{14}$$

Interchanging the roles of X and Y, we obtain the **marginal distribution** of Y in (X, Y) with the distribution function

$$F_2(y) = P(-\infty < X < \infty, Y \leqq y) = \int_{-\infty}^{y} f_2(y^*)\, dy^* \tag{15}$$

and density

$$f_2(y) = \int_{-\infty}^{\infty} f(x, y)\, dx. \tag{16}$$

Independence of Random Variables

X and Y in a (discrete or continuous) random variable (X, Y) are said to be **independent** if

$$F(x, y) = F_1(x)F_2(y) \tag{17}$$

holds for all (x, y). Otherwise these random variables are said to be **dependent.** These definitions are suggested by the corresponding definitions for events in Sec. 24.3.

Necessary and sufficient for independence is

$$f(x, y) = f_1(x)f_2(y) \tag{18}$$

for all x and y. Here the f's are the above probability functions if (X, Y) is discrete or those densities if (X, Y) is continuous. (See Prob. 20.)

EXAMPLE 4 **Independence and Dependence**

In tossing a dime and a nickel, X = *Number of heads on the dime,* Y = *Number of heads on the nickel* may assume the values 0 or 1 and are independent. The random variables in Table 24.1 are dependent. ■

Extension of Independence to n-Dimensional Random Variables. This will be needed throughout Chap. 25. The distribution of such a random variable $\mathbf{X} = (X_1, \cdots, X_n)$ is determined by a **distribution function** of the form

$$F(x_1, \cdots, x_n) = P(X_1 \leqq x_1, \cdots, X_n \leqq x_n).$$

The random variables $X_1, \cdots, X_n$ are said to be **independent** if

$$F(x_1, \cdots, x_n) = F_1(x_1)F_2(x_2) \cdots F_n(x_n) \tag{19}$$

for all $(x_1, \cdots, x_n)$. Here $F_j(x_j)$ is the distribution function of the marginal distribution of X_j in $\mathbf{X}$, that is,

$$F_j(x_j) = P(X_j \leqq x_j, X_k \text{ arbitrary}, k = 1, \cdots, n, k \neq j).$$

Otherwise these random variables are said to be **dependent.**

Functions of Random Variables

When $n = 2$, we write $X_1 = X, X_2 = Y, x_1 = x, x_2 = y$. Taking a nonconstant continuous function $g(x, y)$ defined for all x, y, we obtain a random variable $Z = g(X, Y)$. For example, if we roll two dice and X and Y are the numbers the dice turn up in a trial, then $Z = X + Y$ is the sum of those two numbers (see Fig. 513 in Sec. 24.5).

In the case of a ***discrete*** random variable (X, Y) we may obtain the probability function $f(z)$ of $Z = g(X, Y)$ by summing all $f(x, y)$ for which $g(x, y)$ equals the value of z considered; thus

$$f(z) = P(Z = z) = \sum\sum_{g(x,y)=z} f(x, y). \tag{20}$$

Hence the distribution function of Z is

$$F(z) = P(Z \leqq z) = \sum\sum_{g(x,y)\leqq z} f(x, y) \tag{21}$$

where we sum all values of $f(x, y)$ for which $g(x, y) \leqq z$.

In the case of a ***continuous*** random variable (X, Y) we similarly have

$$F(z) = P(Z \leqq z) = \iint_{g(x,y)\leqq z} f(x, y)\, dx\, dy \tag{22}$$

where for each z we integrate the density $f(x, y)$ of (X, Y) over the region $g(x, y) \leqq z$ in the xy-plane, the boundary curve of this region being $g(x, y) = z$.

Addition of Means

The number

$$E(g(X, Y)) = \begin{cases} \sum_x \sum_y g(x, y)f(x, y) & [(X, Y) \text{ discrete}] \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y)f(x, y)\, dx\, dy & [(X, Y) \text{ continuous}] \end{cases} \tag{23}$$

is called the *mathematical expectation* or, briefly, the **expectation of** $g(X, Y)$. Here it is assumed that the double series converges absolutely and the integral of $|g(x, y)|f(x, y)$ over the xy-plane exists (is finite). Since summation and integration are linear processes, we have from (23)

$$E(ag(X, Y) + bh(X, Y)) = aE(g(X, Y)) + bE(h(X, Y)). \tag{24}$$

An important special case is

$$E(X + Y) = E(X) + E(Y),$$

and by induction we have the following result.

THEOREM 1

Addition of Means

The mean (expectation) of a sum of random variables equals the sum of the means (expectations), that is,

$$E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n). \tag{25}$$

Furthermore, we readily obtain

THEOREM 2

Multiplication of Means

The mean (expectation) of the product of ***independent*** *random variables equals the product of the means (expectations), that is,*

$$E(X_1X_2 \cdots X_n) = E(X_1)E(X_2) \cdots E(X_n). \tag{26}$$

PROOF If X and Y are independent random variables (both discrete or both continuous), then $E(XY) = E(X)E(Y)$. In fact, in the discrete case we have

$$E(XY) = \sum_x \sum_y xyf(x, y) = \sum_x xf_1(x) \sum_y yf_2(y) = E(X)E(Y),$$

and in the continuous case the proof of the relation is similar. Extension to n independent random variables gives (26), and Theorem 2 is proved. ∎

Addition of Variances

This is another matter of practical importance that we shall need. As before, let $Z = X + Y$ and denote the mean and variance of Z by μ and σ^2. Then we first have (see Team Project 16(a) in Problem Set 24.6)

$$\sigma^2 = E([Z - \mu]^2) = E(Z^2) - [E(Z)]^2.$$

From (24) we see that the first term on the right equals

$$E(Z^2) = E(X^2 + 2XY + Y^2) = E(X^2) + 2E(XY) + E(Y^2).$$

For the second term on the right we obtain from Theorem 1

$$[E(Z)]^2 = [E(X) + E(Y)]^2 = [E(X)]^2 + 2E(X)E(Y) + [E(Y)]^2.$$

By substituting these expressions into the formula for σ^2 we have

$$\sigma^2 = E(X^2) - [E(X)]^2 + E(Y^2) - [E(Y)]^2 + 2[E(XY) - E(X)E(Y)].$$

From Team Project 16, Sec. 24.6, we see that the expression in the first line on the right is the sum of the variances of X and Y, which we denote by ${\sigma_1}^2$ and ${\sigma_2}^2$, respectively. The quantity in the second line (except for the factor 2) is

(27) $$\sigma_{XY} = E(XY) - E(X)E(Y)$$

and is called the **covariance** of X and Y. Consequently, our result is

(28) $$\sigma^2 = {\sigma_1}^2 + {\sigma_2}^2 + 2\,\sigma_{XY}.$$

If X and Y are independent, then

$$E(XY) = E(X)E(Y);$$

hence $\sigma_{XY} = 0$, and

(29) $$\sigma^2 = {\sigma_1}^2 + {\sigma_2}^2.$$

Extension to more than two variables gives the basic

THEOREM 3

Addition of Variances

The variance of the sum of ***independent*** *random variables equals the sum of the variances of these variables.*

CAUTION! In the numerous applications of Theorems 1 and 3 we must always remember that Theorem 3 holds only for ***independent*** variables.

This is the end of Chap. 24 on probability theory. Most of the concepts, methods, and special distributions discussed in this chapter will play a fundamental role in the next chapter, which deals with methods of **statistical inference,** that is, conclusions from samples to populations, whose unknown properties we want to know and try to discover by looking at suitable properties of samples that we have obtained.

PROBLEM SET 24.9

1. Let $f(x, y) = k$ when $8 \leqq x \leqq 12$ and $0 \leqq y \leqq 2$ and zero elsewhere. Find k. Find $P(X \leqq 11, 1 \leqq Y \leqq 1.5)$ and $P(9 \leqq X \leqq 13, Y \leqq 1)$.

2. Find $P(X > 2, Y > 2)$ and $P(X \leqq 1, Y \leqq 1)$ if (X, Y) has the density $f(x, y) = 1/8$ if $x \geqq 0, y \geqq 0, x + y \leqq 4$.

3. Let $f(x, y) = k$ if $x > 0$, $y > 0$, $x + y < 3$ and 0 otherwise. Find k. Sketch $f(x, y)$. Find $P(X + Y \leqq 1)$, $P(Y > X)$.

4. Find the density of the marginal distribution of X in Prob 2.

5. Find the density of the marginal distribution of Y in Fig. 523.

6. If certain sheets of wrapping paper have a mean weight of 10 g each, with a standard deviation of 0.05 g, what are the mean weight and standard deviation of a pack of 10 000 sheets?

7. What are the mean thickness and the standard deviation of transformer cores each consisting of 50 layers of sheet metal and 49 insulating paper layers if the metal sheets have mean thickness 0.5 mm each with a standard deviation of 0.05 mm and the paper layers have mean 0.05 mm each with a standard deviation of 0.02 mm?

8. If the weight of certain (empty) containers has mean 2 lb and standard deviation 0.1 lb, and if the filling of the containers has mean weight 75 lb and standard deviation 0.8 lb, what are the mean weight and standard deviation of filled containers?

9. A 5-gear assembly is put together with spacers between the gears. The mean thickness of the gears is 5.020 cm with a standard deviation of 0.003 cm. The mean thickness of the spacers is 0.040 cm with a standard deviation of 0.002 cm. Find the mean and standard deviation of the assembled units consisting of 5 randomly selected gears and 4 randomly selected spacers.

10. Give an example of two different discrete distributions that have the same marginal distributions.

11. Show that the random variables with the densities

$$f(x, y) = x + y$$

and

$$g(x, y) = (x + \tfrac{1}{2})(y + \tfrac{1}{2})$$

if $0 \leqq x \leqq 1$, $0 \leqq y \leqq 1$ and $f(x, y) = 0$ and $g(x, y) = 0$ elsewhere, have the same marginal distribution.

12. Let X [cm] and Y [cm] be the diameter of a pin and hole, respectively. Suppose that (X, Y) has the density

$$f(x, y) = 2500 \qquad \text{if}$$

$$0.99 < x < 1.01,\ 1.00 < y < 1.02$$

and 0 otherwise. (a) Find the marginal distributions. (b) What is the probability that a pin chosen at random will fit a hole whose diameter is 1.00?

13. An electronic device consists of two components. Let X and Y [months] be the length of time until failure of the first and second component, respectively. Assume that (X, Y) has the probability density

$$f(x, y) = 0.01e^{-0.1(x+y)} \qquad \text{if } x > 0 \text{ and } y > 0$$

and 0 otherwise. (a) Are X and Y dependent or independent? (b) Find the densities of the marginal distributions. (c) What is the probability that the first component has a lifetime of 10 months or longer?

14. Prove (2).

15. Find $P(X > Y)$ when (X, Y) has the density

$$f(x, y) = 0.25e^{-0.5(x+y)} \text{ if } x \geqq 0, y \geqq 0$$

and 0 otherwise.

16. Let (X, Y) have the density

$$f(x, y) = k \text{ if } x^2 + y^2 < 1$$

and 0 otherwise. Determine k. Find the densities of the marginal distributions. Find the probability

$$P(X^2 + Y^2 < 1/4).$$

17. Let (X, Y) have the probability function

$$f(0, 0) = f(1, 1) = 1/8,$$

$$f(0, 1) = f(1, 0) = 3/8.$$

Are X and Y independent?

18. Using Theorem 1, obtain the formula for the mean of the hypergeometric distribution. Can you use Theorem 3 to obtain the variance of that distribution?

19. Using Theorems 1 and 3, obtain the formulas for the mean and the variance of the binomial distribution.

20. Prove the statement involving (18).

CHAPTER 24 REVIEW QUESTIONS AND PROBLEMS

1. Why did we begin the chapter with a section on handling data?
2. What are stem-and-leaf plots? Boxplots? Histograms? Compare their advantages.
3. What quantities measure the average size of data? The spread?
4. Why did we consider probability theory? What is its role in statistics?
5. What do we mean by an experiment? By a random variable related with it? What are outcomes? Events?
6. Give examples of experiments in which you have equally likely cases and others in which you don't.
7. State the definition of probability from memory.
8. What is the difference between the concepts of a permutation and a combination?
9. State the main theorems on probability. Illustrate them by simple examples.
10. What is the distribution of a random variable? The distribution function? The probability function? The density?
11. State the definitions of mean and variance of a random variable from memory.
12. If $P(A) = P(B)$ and $A \subseteq B$, can $A \neq B$?
13. If $E \neq S$ (= the sample space), can $P(E) = 1$?
14. What distributions correspond to sampling with replacement and without replacement?
15. When will an experiment involve a binomial distribution? A hypergeometric distribution?
16. When will the Poisson distribution be a good approximation of the binomial distribution?
17. What do you know about the approximation of the binomial distribution by the normal distribution?
18. Explain the use of the tables of the normal distribution. If you have a CAS, how would you proceed without the tables?
19. Can the probability function of a discrete random variable have infinitely many positive values?
20. State the most important facts about distributions of two random variables and their marginal distributions.
21. Make a stem-and-leaf plot, histogram, and boxplot of the data 22.5, 23.2, 22.1, 23.6, 23.3, 23.4, 24.0, 20.6, 23.3.
22. Do the same task as in Prob. 21, for the data 210, 213, 209, 218, 210, 215, 204, 211, 216, 213.
23. Find the mean, standard deviation, and variance in Prob. 21.
24. Find the mean, standard deviation, and variance in Prob. 22.
25. What are the outcomes of the sample space of *X: Tossing a coin until the first Head appears*?
26. What are the outcomes in the sample space of the experiment of simultaneously tossing three coins?
27. A box contains 50 screws, five of which are defective. Find the probability function of the random variable $X =$ *Number of defective screws in drawing two screws without replacement* and compute its values.
28. Find the values of the distribution function in Prob. 27.
29. Using a Venn diagram, show that $A \subseteq B$ if and only if $A \cup B = B$.
30. Using a Venn diagram, show that $A \subseteq B$ if and only if $A \cap B = A$.
31. If X has the density $f(x) = 0.5x$ $(0 \leqq x \leqq 2)$ and 0 otherwise, what are the mean and the variance of $X^* = -2X + 5$?
32. If 6 different inks are available, in how many ways can we select two colors for a printing job? Four colors?
33. Compute 5! by the Stirling formula and find the absolute and relative errors.
34. Two screws are randomly drawn without replacement from a box containing 7 right-handed and 3 left-handed screws. Let X be the number of left-handed screws drawn. Find $P(X = 0)$, $P(X = 1)$, $P(X = 2)$, $P(1 < X < 2)$, $P(0 < X < 5)$.
35. Find the mean and the variance of the distribution having the density $f(x) = \frac{1}{2}e^{-|x|}$.
36. Find the skewness of the distribution with density $f(x) = 2(1 - x)$ if $0 < x < 1$, $f(x) = 0$ otherwise.
37. Sketch the probability function $f(x) = x^2/30$ $(x = 1, 2, 3, 4)$ and the distribution function. Find μ.
38. Sketch $F(x) = 0$ if $x \leqq 0$, $F(x) = 0.2x$ if $0 < x \leqq 5$, $F(x) = 1$ if $x > 5$, and its density $f(x)$.
39. If the life of tires is normal with mean 25 000 km and variance 25 000 000 km^2, what is the probability that a given one of those tires will last at least 30 000 km? At least 35 000 km?
40. If the weight of bags of cement is normal with mean 50 kg and standard deviation 1 kg, what is the probability that 100 bags will be heavier than 5030 kg?

SUMMARY OF CHAPTER 24

Data Analysis. Probability Theory

A *random experiment,* briefly called **experiment,** is a process in which the result (**"outcome"**) depends on "chance" (effects of factors unknown to us). Examples are games of chance with dice or cards, measuring the hardness of steel, observing weather conditions, or recording the number of accidents in a city. (Thus the word "experiment" is used here in a much wider sense than in common language.) The outcomes are regarded as points (elements) of a set S, called the **sample space,** whose subsets are called **events.** For events E we define a **probability** $P(E)$ by the axioms (Sec. 24.3)

$$0 \leqq P(E) \leqq 1$$

(1)
$$P(S) = 1$$

$$P(E_1 \cup E_2 \cup \cdots) = P(E_1) + P(E_2) + \cdots \qquad (E_j \cap E_k = \varnothing).$$

These axioms are motivated by properties of frequency distributions of data (Sec. 24.1).

The complement E^c of E has the probability

(2)
$$P(E^c) = 1 - P(E).$$

The **conditional probability** of an event B under the condition that an event A happens is (Sec. 24.3)

(3)
$$P(B|A) = \frac{P(A \cap B)}{P(A)} \qquad [P(A) > 0].$$

Two events A and B are called **independent** if the probability of their simultaneous appearance in a trial equals the product of their probabilities, that is, if

(4)
$$P(A \cap B) = P(A)P(B).$$

With an experiment we associate a **random variable** X. This is a function defined on S whose values are real numbers; furthermore, X is such that the probability $P(X = a)$ with which X assumes any value a, and the probability $P(a < X \leqq b)$ with which X assumes any value in an interval $a < X \leqq b$ are defined (Sec. 24.5). The **probability distribution** of X is determined by the distribution function

(5)
$$F(x) = P(X \leqq x).$$

In applications there are two important kinds of random variables: those of the **discrete** type, which appear if we count (defective items, customers in a bank, etc.) and those of the **continuous** type, which appear if we measure (length, speed, temperature, weight, etc.).

A discrete random variable has a **probability function**

$$f(x) = P(X = x). \tag{6}$$

Its **mean** μ and **variance** σ^2 are (Sec. 24.6)

$$\mu = \sum_j x_j f(x_j) \quad \text{and} \quad \sigma^2 = \sum_j (x_j - \mu)^2 f(x_j) \tag{7}$$

where the x_j are the values for which X has a positive probability. Important discrete random variables and distributions are the **binomial, Poisson,** and **hypergeometric distributions** discussed in Sec. 24.7.

A continuous random variable has a **density**

$$f(x) = F'(x) \qquad \text{[see (5)]}. \tag{8}$$

Its mean and variance are (Sec. 24.6)

$$\mu = \int_{-\infty}^{\infty} x f(x)\, dx \quad \text{and} \quad \sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, dx. \tag{9}$$

Very important is the **normal distribution** (Sec. 24.8), whose density is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \tag{10}$$

and whose distribution function is (Sec. 24.8; Tables A7, A8 in App. 5)

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right). \tag{11}$$

A **two-dimensional random variable** (X, Y) occurs if we simultaneously observe two quantities (for example, height X and weight Y of adults). Its distribution function is (Sec. 24.9)

$$F(x, y) = P(X \leqq x, Y \leqq y). \tag{12}$$

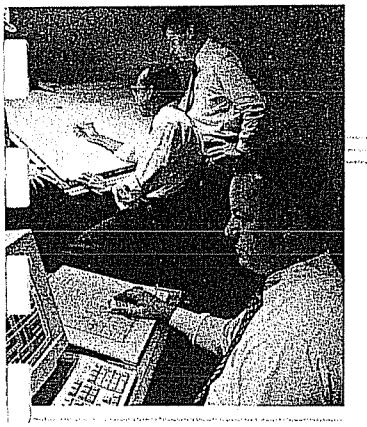
X and Y have the distribution functions (Sec. 24.9)

$$F_1(x) = P(X \leqq x, Y \text{ arbitrary}) \quad \text{and} \quad F_2(y) = P(x \text{ arbitrary}, Y \leqq y) \tag{13}$$

respectively; their distributions are called **marginal distributions.** If both X and Y are discrete, then (X, Y) has a probability function

$$f(x, y) = P(X = x, Y = y).$$

If both X and Y are continuous, then (X, Y) has a density $f(x, y)$.



CHAPTER 25

Mathematical Statistics

In probability theory we set up mathematical models of processes that are affected by "chance". In mathematical statistics or, briefly, **statistics,** we check these models against the observable reality. This is called **statistical inference.** It is done by **sampling,** that is, by drawing random samples, briefly called **samples.** These are sets of values from a much larger set of values that could be studied, called the **population.** An example is 10 diameters of screws drawn from a large lot of screws. Sampling is done in order to see whether a model of the population is accurate enough for practical purposes. If this is the case, the model can be used for predictions, decisions, and actions, for instance, in planning productions, buying equipment, investing in business projects, and so on.

Most important methods of statistical inference are **estimation of parameters** (Secs. 25.2), determination of **confidence intervals** (Sec. 25.3), and **hypothesis testing** (Secs. 25.4, 25.7, 25.8), with application to *quality control* (Sec. 25.5) and *acceptance sampling* (Sec. 25.6).

In the last section (25.9) we give an introduction to **regression** and **correlation analysis,** which concern experiments involving two variables.

Prerequisite: Chap. 24.

Sections that may be omitted in a shorter course: 25.5, 25.6, 25.8.

References, Answers to Problems, and Statistical Tables: App. 1 Part G, App. 2, App. 5.

25.1 Introduction. Random Sampling

Mathematical statistics consists of methods for designing and evaluating random experiments to obtain information about practical problems, such as exploring the relation between iron content and density of iron ore, the quality of raw material or manufactured products, the efficiency of air-conditioning systems, the performance of certain cars, the effect of advertising, the reactions of consumers to a new product, etc.

Random variables occur more frequently in engineering (and elsewhere) than one would think. For example, properties of mass-produced articles (screws, lightbulbs, etc.) always show **random variation,** due to small (uncontrollable!) differences in raw material or manufacturing processes. Thus the diameter of screws is a random variable X and we have *nondefective screws,* with diameter between given tolerance limits, and *defective screws,* with diameter outside those limits. We can ask for the distribution of X, for the percentage of defective screws to be expected, and for necessary improvements of the production process.

Samples are selected from populations—20 screws from a lot of 1000, 100 of 5000 voters, 8 beavers in a wildlife conservation project—because inspecting the entire population would be too expensive, time-consuming, impossible or even senseless (think of destructive testing of lightbulbs or dynamite). To obtain meaningful conclusions, samples must be **random selections.** Each of the 1000 screws must have the same chance of being sampled (of being drawn when we sample), at least approximately. Only then will the sample mean $\bar{x} = (x_1 + \cdots + x_{20})/20$ (Sec. 24.1) of a sample of size $n = 20$ (or any other n) be a good approximation of the population mean μ (Sec. 24.6); and the accuracy of the approximation will generally improve with increasing n, as we shall see. Similarly for other parameters (standard deviation, variance, etc.).

Independent sample values will be obtained in experiments with an infinite sample space S (Sec. 24.2), certainly for the normal distribution. This is also true in sampling with replacement. It is approximately true in drawing *small* samples from a large finite population (for instance, 5 or 10 of 1000 items). However, if we sample without replacement from a small population, the effect of dependence of sample values may be considerable.

Random numbers help in obtaining samples that are in fact random selections. This is sometimes not easy to accomplish because there are many subtle factors that can bias sampling (by personal interviews, by poorly working machines, by the choice of nontypical observation conditions, etc.). Random numbers can be obtained from a **random number generator** in Maple, Mathematica, or other systems listed on p. 991. (The numbers are not truly random, as they would be produced in flipping coins or rolling dice, but are calculated by a tricky formula that produces numbers that do have practically all the essential features of true randomness.)

EXAMPLE 1 **Random Numbers from a Random Number Generator**

To select a sample of size $n = 10$ from 80 given ball bearings, we number the bearings from 1 to 80. We then let the generator randomly produce 10 of the integers from 1 to 80 and include the bearings with the numbers obtained in our sample, for example.

44 55 53 03 52 61 67 78 39 54

or whatever.

Random numbers are also contained in (older) statistical tables. ■

Representing and processing data were considered in Sec. 24.1 in connection with frequency distributions. These are the empirical counterparts of probability distributions and helped motivating axioms and properties in probability theory. The new aspect in this chapter is **randomness:** the data are samples selected **randomly** from a population. Accordingly, we can immediately make the connection to Sec. 24.1, using stem-and-leaf plots, box plots, and histograms for representing samples graphically.

Also, we now call the mean $\bar{x}$ in (5), Sec. 24.1, the **sample mean**

$$\text{(1)} \qquad \bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j = \frac{1}{n}(x_1 + x_2 + \cdots + x_n).$$

We call n the **sample size,** the variance s^2 in (6), Sec. 24.1, the **sample variance**

$$\text{(2)} \qquad s^2 = \frac{1}{n-1}\sum_{j=1}^{n} (x_j - \bar{x})^2 = \frac{1}{n-1}\left[(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2\right],$$

and its positive square root s the **sample standard deviation.** $\bar{x}$, s^2, and s are called **parameters** *of a sample;* they will be needed throughout this chapter.

25.2 Point Estimation of Parameters

Beginning in this section, we shall discuss the most basic practical tasks in statistics and corresponding statistical methods to accomplish them. The first of them is point estimation of **parameters,** that is, of quantities appearing in distributions, such as p in the binomial distribution and μ and σ in the normal distribution.

A **point estimate** of a parameter is a number (point on the real line), which is computed from a given sample and serves as an approximation of the unknown exact value of the parameter of the population. An **interval estimate** is an interval (*"confidence interval"*) obtained from a sample; such estimates will be considered in the next section. Estimation of parameters is of great practical importance in many applications.

As an approximation of the mean μ of a population we may take the mean $\bar{x}$ of a corresponding sample. This gives the estimate $\hat{\mu} = \bar{x}$ for μ, that is,

$$\hat{\mu} = \bar{x} = \frac{1}{n}(x_1 + \cdots + x_n) \tag{1}$$

where n is the sample size. Similarly, an estimate $\hat{\sigma}^2$ for the variance of a population is the variance s^2 of a corresponding sample, that is,

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2. \tag{2}$$

Clearly, (1) and (2) are estimates of parameters for distributions in which μ or σ^2 appear explicity as parameters, such as the normal and Poisson distributions. For the binomial distribution, $p = \mu/n$ [see (3) in Sec. 24.7]. From (1) we thus obtain for p the estimate

$$\hat{p} = \frac{\bar{x}}{n}. \tag{3}$$

We mention that (1) is a special case of the so-called **method of moments.** In this method the parameters to be estimated are expressed in terms of the moments of the distribution (see Sec. 24.6). In the resulting formulas those moments of the distribution are replaced by the corresponding moments of the sample. This gives the estimates. Here the ***k*th moment of a sample** $x_1, \cdots, x_n$ is

$$m_k = \frac{1}{n}\sum_{j=1}^{n} x_j^k.$$

Maximum Likelihood Method

Another method for obtaining estimates is the so-called **maximum likelihood method** of R. A. Fisher [*Messenger Math.* **41** (1912), 155–160]. To explain it, we consider a discrete (or continuous) random variable X whose probability function (or density) $f(x)$ depends on a single parameter θ. We take a corresponding sample of n ***independent*** values $x_1, \cdots, x_n$. Then in the discrete case the probability that a sample of size n consists precisely of those n values is

$$l = f(x_1)f(x_2) \cdots f(x_n). \tag{4}$$

In the continuous case the probability that the sample consists of values in the small intervals $x_j \leqq x \leqq x_j + \Delta x$ $(j = 1, 2, \cdots, n)$ is

$$f(x_1)\Delta x\, f(x_2)\Delta x \cdots f(x_n)\Delta x = l(\Delta x)^n. \tag{5}$$

Since $f(x_j)$ depends on θ, the function l in (5) given by (4) depends on $x_1, \cdots, x_n$ and θ. We imagine $x_1, \cdots, x_n$ to be given and fixed. Then l is a function of θ, which is called the **likelihood function.** The basic idea of the maximum likelihood method is quite simple, as follows. We choose *that* approximation for the unknown value of θ for which l is as large as possible. If l is a differentiable function of θ, a necessary condition for l to have a maximum in an interval (not at the boundary) is

$$\frac{\partial l}{\partial \theta} = 0. \tag{6}$$

(We write a *partial* derivative, because l depends also on $x_1, \cdots, x_n$.) A solution of (6) depending on $x_1, \cdots, x_n$ is called a **maximum likelihood estimate** for θ. We may replace (6) by

$$\frac{\partial \ln l}{\partial \theta} = 0, \tag{7}$$

because $f(x_j) > 0$, a maximum of l is in general positive, and $\ln l$ is a monotone increasing function of l. This often simplifies calculations.

Several Parameters. If the distribution of X involves r parameters $\theta_1, \cdots, \theta_r$, then instead of (6) we have the r conditions $\partial l/\partial\theta_1 = 0, \cdots, \partial l/\partial\theta_r = 0$, and instead of (7) we have

$$\frac{\partial \ln l}{\partial \theta_1} = 0, \qquad \cdots, \qquad \frac{\partial \ln l}{\partial \theta_r} = 0. \tag{8}$$

EXAMPLE 1 Normal Distribution

Find maximum likelihood estimates for $\theta_1 = \mu$ and $\theta_2 = \sigma$ in the case of the normal distribution.

Solution. From (1), Sec. 24.8, and (4) we obtain the likelihood function

$$l = \left(\frac{1}{\sqrt{2\pi}}\right)^n \left(\frac{1}{\sigma}\right)^n e^{-h} \qquad \text{where} \qquad h = \frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j - \mu)^2.$$

Taking logarithms, we have

$$\ln l = -n \ln \sqrt{2\pi} - n \ln \sigma - h.$$

The first equation in (8) is $\partial(\ln l)/\partial\mu = 0$, written out

$$\frac{\partial \ln l}{\partial \mu} = -\frac{\partial h}{\partial \mu} = \frac{1}{\sigma^2} \sum_{j=1}^{n} (x_j - \mu) = 0, \qquad \text{hence} \qquad \sum_{j=1}^{n} x_j - n\mu = 0.$$

The solution is the desired estimate $\hat{\mu}$ for μ; we find

$$\hat{\mu} = \frac{1}{n} \sum_{j=1}^{n} x_j = \bar{x}.$$

The second equation in (8) is $\partial(\ln l)/\partial\sigma = 0$, written out

$$\frac{\partial \ln l}{\partial \sigma} = -\frac{n}{\sigma} - \frac{\partial h}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{j=1}^{n} (x_j - \mu)^2 = 0.$$

Replacing μ by $\hat{\mu}$ and solving for σ^2, we obtain the estimate

$$\tilde{\sigma}^2 = \frac{1}{n} \sum_{j=1}^{n} (x_j - \bar{x})^2$$

which we shall use in Sec. 25.7. Note that this differs from (2). We cannot discuss criteria for the goodness of estimates but want to mention that for small n, formula (2) is preferable. ■

PROBLEM SET 25.7

1. Find the maximum likelihood estimate for the parameter μ of a normal distribution with known variance $\sigma^2 = \sigma_0^2$.
2. Apply the maximum likelihood method to the normal distribution with $\mu = 0$.
3. **(Binomial distribution)** Derive a maximum likelihood estimate for p.
4. Extend Prob. 3 as follows. Suppose that m times n trials were made and in the first n trials A happened k_1 times, in the second n trials A happened k_2 times, $\cdots$, in the mth n trials A happened k_m times. Find a maximum likelihood estimate of p based on this information.
5. Suppose that in Prob. 4 we made 4 times 5 trials and A happened 2, 1, 4, 4 times, respectively. Estimate p.
6. Consider X = *Number of independent trials until an event A occurs.* Show that X has the probability function $f(x) = pq^{x-1}$, $x = 1, 2, \cdots$, where p is the probability of A in a single trial and $q = 1 - p$. Find the maximum likelihood estimate of p corresponding to a sample $x_1, \cdots, x_n$ of observed values of X.
7. In Prob. 6 find the maximum likelihood estimate of p corresponding to a single observation x of X.
8. In rolling a die, suppose that we get the first Six in the 7th trial and in doing it again we get it in the 6th trial. Estimate the probability p of getting a Six in rolling that die once.
9. **(Poisson distribution)** Apply the maximum likelihood method to the Poisson distribution.
10. **(Uniform distribution)** Show that in the case of the parameters a and b of the uniform distribution (see Sec. 24.6), the maximum likelihood estimate cannot be obtained by equating the first derivative to zero. How can we obtain maximum likelihood estimates in this case?
11. Find the maximum likelihood estimate of θ in the density $f(x) = \theta e^{-\theta x}$ if $x \geqq 0$ and $f(x) = 0$ if $x < 0$.
12. In Prob. 11, find the mean μ, substitute it in $f(x)$, find the maximum likelihood estimate of μ, and show that it is identical with the estimate for μ which can be obtained from that for θ in Prob. 11.
13. Compute $\hat{\theta}$ in Prob. 11 from the sample 1.8, 0.4, 0.8, 0.6, 1.4. Graph the sample distribution function $\hat{F}(x)$ and the distribution function $F(x)$ of the random variable, with $\theta = \hat{\theta}$, on the same axes. Do they agree reasonably well? (We consider goodness of fit systematically in Sec. 25.7.)

14. Do the same task as in Prob. 13 if the given sample is 0.5, 0.7, 0.1, 1.1, 0.1.

15. CAS EXPERIMENT. Maximum Likelihood Estimates. (MLEs). Find experimentally how much MLEs can differ depending on the sample size. *Hint.* Generate many samples of the same size n, e.g., of the standardized normal distribution, and record $\bar{x}$ and s^2. Then increase n.

25.3 Confidence Intervals

Confidence intervals[1] for an unknown parameter θ of some distribution (e.g., $\theta = \mu$) are intervals $\theta_1 \leqq \theta \leqq \theta_2$ that contain θ, not with certainty but with a high probability γ, which we can choose (95% and 99% are popular). Such an interval is calculated from a sample. $\gamma = 95\%$ means probability $1 - \gamma = 5\% = 1/20$ of being wrong—one of about 20 such intervals will not contain θ. Instead of writing $\theta_1 \leqq \theta \leqq \theta_2$, we denote this more distinctly by writing

$$\text{CONF}_\gamma \{\theta_1 \leqq \theta \leqq \theta_2\}. \tag{1}$$

Such a special symbol, CONF, seems worthwhile in order to avoid the misunderstanding that θ *must* lie between θ_1 and θ_2.

γ is called the **confidence level,** and θ_1 and θ_2 are called the **lower** and **upper confidence limits.** They depend on γ. The larger we choose γ, the smaller is the error probability $1 - \gamma$, but the longer is the confidence interval. If $\gamma \to 1$, then its length goes to infinity. The choice of γ depends on the kind of application. In taking no umbrella, a 5% chance of getting wet is not tragic. In a medical decision of life or death, a 5% chance of being wrong may be too large and a 1% chance of being wrong ($\gamma = 99\%$) may be more desirable.

Confidence intervals are more valuable than point estimates (Sec. 25.2). Indeed, we can take the midpoint of (1) as an approximation of θ and half the length of (1) as an "error bound" (not in the strict sense of numerics, but except for an error whose probability we know).

θ_1 and θ_2 in (1) are calculated from a sample $x_1, \cdots, x_n$. These are n observations of a random variable X. Now comes a **standard trick.** *We regard $x_1, \cdots, x_n$ as **single** observations of n random variables $X_1, \cdots, X_n$ (with the same distribution, namely, that of X).* Then $\theta_1 = \theta_1(x_1, \cdots, x_n)$ and $\theta_2 = \theta_2(x_1, \cdots, x_n)$ in (1) are observed values of two random variables $\Theta_1 = \Theta_1(X_1, \cdots, X_n)$ and $\Theta_2 = \Theta_2(X_1, \cdots, X_n)$. The condition (1) involving γ can now be written

$$P(\Theta_1 \leqq \theta \leqq \Theta_2) = \gamma. \tag{2}$$

Let us see what all this means in concrete practical cases.

In each case in this section we shall first state the steps of obtaining a confidence interval in the form of a table, then consider a typical example, and finally justify those steps theoretically.

---

[1]JERZY NEYMAN (1894–1981), American statistician, developed the theory of confidence intervals (*Annals of Mathematical Statistics* **6** (1935), 111–116).

Confidence Interval for μ of the Normal Distribution with Known σ^2

Table 25.1 Determination of a Confidence Interval for the Mean μ of a Normal Distribution with Known Variance σ^2

Step 1. Choose a confidence level γ (95%, 99%, or the like).

Step 2. Determine the corresponding c:

γ	0.90	0.95	0.99	0.999
c	1.645	1.960	2.576	3.291

Step 3. Compute the mean $\bar{x}$ of the sample $x_1, \cdots, x_n$.

Step 4. Compute $k = c\sigma/\sqrt{n}$. The confidence interval for μ is

$$\text{CONF}_\gamma \{\bar{x} - k \leqq \mu \leqq \bar{x} + k\}. \tag{3}$$

EXAMPLE 1 Confidence Interval for μ of the Normal Distribution with Known σ^2

Deterimine a 95% confidence interval for the mean of a normal distribution with variance $\sigma^2 = 9$, using a sample of $n = 100$ values with mean $\bar{x} = 5$.

Solution. *Step 1.* $\gamma = 0.95$ is required. *Step 2.* The corresponding c equals 1.960; see Table 25.1. *Step 3.* $\bar{x} = 5$ is given. *Step 4.* We need $k = 1.960 \cdot 3/\sqrt{100} = 0.588$. Hence $\bar{x} - k = 4.412$, $\bar{x} + k = 5.588$ and the confidence interval is $\text{CONF}_{0.95}\{4.412 \leqq \mu \leqq 5.588\}$.

This is sometimes written $\mu = 5 \pm 0.588$, but we shall not use this notation, which can be misleading. With your CAS you can determine this interval more directly. Similarly for the other examples in this section. ■

Theory for Table 25.1. The method in Table 25.1 follows from the basic

THEOREM 1 Sum of Independent Normal Random Variables

*Let $X_1, \cdots, X_n$ be **independent** normal random variables each of which has mean μ and variance σ^2. Then the following holds.*

(a) *The sum $X_1 + \cdots + X_n$ is normal with mean $n\mu$ and variance $n\sigma^2$.*

(b) The following random variable $\overline{X}$ *is normal with mean μ and variance σ^2/n.*

$$\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n) \tag{4}$$

(c) *The following random variable Z is normal with mean 0 and variance 1.*

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \tag{5}$$

PROOF The statements about the mean and variance in (a) follow from Theorems 1 and 3 in Sec. 24.9. From this and Theorem 2 in Sec. 24.6 we see that $\overline{X}$ has the mean $(1/n)n\mu = \mu$ and the variance $(1/n)^2 n\sigma^2 = \sigma^2/n$. This implies that Z has the mean 0 and variance 1, by Theorem 2(b) in Sec. 24.6. The normality of $X_1 + \cdots + X_n$ is proved in Ref. [G3] listed in App. 1. This implies the normality of (4) and (5). ■

Derivation of (3) in Table 25.1. Sampling from a normal distribution gives independent sample values (see Sec. 25.1), so that Theorem 1 applies. Hence we can choose γ and then determine c such that

$$(6) \qquad P(-c \leqq Z \leqq c) = P\left(-c \leqq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leqq c\right) = \Phi(c) - \Phi(-c) = \gamma.$$

For the value $\gamma = 0.95$ we obtain $z(D) = 1.960$ from Table A8 in App. 5, as used in Example 1. For $\gamma = 0.9$, 0.99. 0.999 we get the other values of c listed in Table 25.1. Finally, all we have to do is to convert the inequality in (6) into one for μ and insert observed values obtained from the sample. We multiply $-c \leqq Z \leqq c$ by -1 and then by $\sigma/\sqrt{n}$, writing $c\sigma/\sqrt{n} = k$ (as in Table 25.1),

$$P(-c \leqq Z \leqq c) = P(c \geqq -Z \geqq -c) = P\left(c \geqq \frac{\mu - \overline{X}}{\sigma/\sqrt{n}} \geqq -c\right)$$

$$= P(k \geqq \mu - \overline{X} \geqq -k) = \gamma.$$

Adding $\overline{X}$ gives $P(\overline{X} + k \geqq \mu \geqq \overline{X} - k) = \gamma$ or

$$(7) \qquad P(\overline{X} - k \leqq \mu \leqq \overline{X} + k) = \gamma.$$

Inserting the observed value $\bar{x}$ of $\overline{X}$ gives (3). Here we have regarded $x_1, \cdots, x_n$ as single observations of $X_1, \cdots, X_n$ (the standard trick!), so that $x_1 + \cdots + x_n$ is an observed value of $X_1 + \cdots + X_n$ and $\bar{x}$ is an observed value of $\overline{X}$. Note further that (7) is of the form (2) with $\Theta_1 = \overline{X} - k$ and $\Theta_2 = \overline{X} + k$. ▩

EXAMPLE 2 Sample Size Needed for a Confidence Interval of Prescribed Length

How large must n be in Example 1 if we want to obtain a 95% confidence interval of length $L = 0.4$?

Solution. The interval (3) has the length $L = 2k = 2c\sigma/\sqrt{n}$. Solving for n, we obtain

$$n = (2c\sigma/L)^2.$$

In the present case the answer is $n = (2 \cdot 1.960 \cdot 3/0.4)^2 \approx 870$.

Figure 525 shows how L decreases as n increases and that for $\gamma = 99\%$ the confidence interval is substantially longer than for $\gamma = 95\%$ (and the same sample size n). ▩

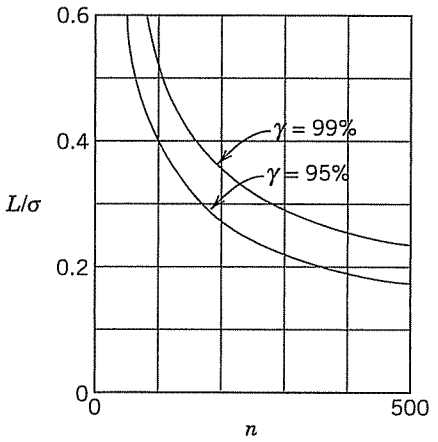


Fig. 525. Length of the confidence interval (3) (measured in multiples of σ) as a function of the sample size n for $\gamma = 95\%$ and $\gamma = 99\%$

Confidence Interval for μ of the Normal Distribution With Unknown σ^2

In practice σ^2 is frequently unknown. Then the method in Table 25.1 does not help and the whole theory changes, although the steps of determining a confidence interval for μ remain quite similar. They are shown in Table 25.2. We see that k differs from that in Table 25.1, namely, the sample standard deviation s has taken the place of the unknown standard deviation σ of the population. And c now depends on the sample size n and must be determined from Table A9 in App. 5 or from your CAS. That table lists values z for given values of the distribution function (Fig. 526)

$$F(z) = K_m \int_{-\infty}^{z} \left(1 + \frac{u^2}{m}\right)^{-(m+1)/2} du \tag{8}$$

of the ***t*-distribution**. Here, m $(= 1, 2, \cdots)$ is a parameter, called the **number of degrees of freedom** of the distribution (***abbreviated* d.f.).** In the present case, $m = n - 1$; see Table 25.2. The constant K_m is such that $F(\infty) = 1$. By integration it turns out that $K_m = \Gamma(\frac{1}{2}m + \frac{1}{2})/\left[\sqrt{m\pi}\,\Gamma(\frac{1}{2}m)\right]$, where Γ is the gamma function (see (24) in App. A3.1).

Table 25.2 Determination of a Confidence Interval for the Mean μ of a Normal Distribution with Unknown Variance σ^2

Step 1. Choose a confidence level γ (95%, 99%, or the like).

Step 2. Determine the solution c of the equation

$$F(c) = \tfrac{1}{2}(1 + \gamma) \tag{9}$$

from the table of the t-distribution with $n - 1$ degrees of freedom (Table A9 in App. 5; or use a CAS; n = sample size).

Step 3. Compute the mean $\bar{x}$ and the variance s^2 of the sample $x_1, \cdots, x_n$.

Step 4. Compute $k = cs/\sqrt{n}$. The confidence interval is

$$\text{CONF}_\gamma \{\bar{x} - k \leqq \mu \leqq \bar{x} + k\}. \tag{10}$$

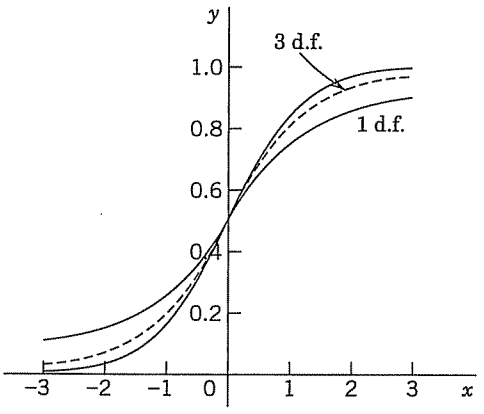


Fig. 526. Distribution functions of the t-distribution with 1 and 3 d.f. and of the standardized normal distribution (steepest curve)

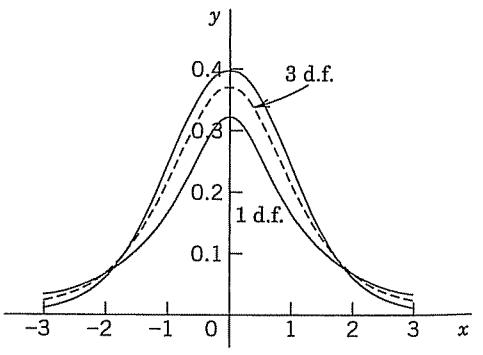


Fig. 527. Densities of the t-distribution with 1 and 3 d.f. and of the standardized normal distribution

Figure 527 compares the curve of the density of the t-distribution with that of the normal distribution. The latter is steeper. This illustrates that Table 25.1 (which uses more information, namely, the known value of σ^2) yields shorter confidence intervals than Table 25.2. This is confirmed in Fig. 528, which also gives an idea of the gain by increasing the sample size.

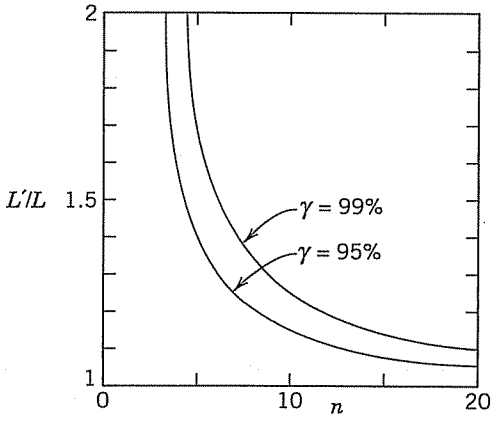


Fig. 528. Ratio of the lengths L' and L of the confidence intervals (10) and (3) with $\gamma = 95\%$ and $\gamma = 99\%$ as a function of the sample size n for equal s and σ

EXAMPLE 3 **Confidence Interval for μ of the Normal Distribution with Unknown σ^2**

Five independent measurements of the point of inflammation (flash point) of Diesel oil (D-2) gave the values (in °F) 144 147 146 142 144. Assuming normality, determine a 99% confidence interval for the mean.

Solution. *Step 1.* $\gamma = 0.99$ is required.

Step 2. $F(c) = \frac{1}{2}(1 + \gamma) = 0.995$, and Table A9 in App. 5 with $n - 1 = 4$ d.f. gives $c = 4.60$.

Step 3. $\bar{x} = 144.6$, $s^2 = 3.8$.

Step 4. $k = \sqrt{3.8} \cdot 4.60/\sqrt{5} = 4.01$. The confidence interval is $\text{CONF}_{0.99}\ \{140.5 \leqq \mu \leqq 148.7\}$.

If the variance σ^2 were known and equal to the sample variance s^2, thus $\sigma^2 = 3.8$, then Table 25.1 would give $k = c\sigma/\sqrt{n} = 2.576\sqrt{3.8}/\sqrt{5} = 2.25$ and $\text{CONF}_{0.99}\ \{142.35 \leqq \mu \leqq 146.85\}$. We see that the present interval is almost twice as long as that obtained from Table 25.1 (with $\sigma^2 = 3.8$). Hence for small samples the difference is considerable! See also Fig. 528. ▩

Theory for Table 25.2. For deriving (10) in Table 25.2 we need from Ref. [G3]

THEOREM 2 **Student's t-Distribution**

Let $X_1, \cdots, X_n$ be independent normal random variables with the same mean μ and the same variance σ^2. Then the random variable

(11) $$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

has a t-distribution [see (8)] *with $n - 1$ degrees of freedom* (d.f.); *here $\bar{X}$ is given by* (4) *and*

(12) $$S^2 = \frac{1}{n-1} \sum_{j=1}^{n} (X_j - \bar{X})^2.$$

Derivation of (10). This is similar to the derivation of (3). We choose a number γ between 0 and 1 and determine a number c from Table A9 in App. 5 with $n - 1$ d.f. (or from a CAS) such that

$$P(-c \leqq T \leqq c) = F(c) - F(-c) = \gamma. \tag{13}$$

Since the t-distribution is symmetric, we have

$$F(-c) = 1 - F(c),$$

and (13) assumes the form (9). Substituting (11) into (13) and transforming the result as before, we obtain

$$P(\overline{X} - K \leqq \mu \leqq \overline{X} + K) = \gamma \tag{14}$$

where

$$K = cS/\sqrt{n}.$$

By inserting the observed values $\bar{x}$ of $\overline{X}$ and s^2 of S^2 into (14) we finally obtain (10). ■

Confidence Interval for the Variance σ^2 of the Normal Distribution

Table 25.3 shows the steps, which are similar to those in Tables 25.1 and 25.2.

Table 25.3 Determination of a Confidence Interval for the Variance σ^2 of a Normal Distribution, Whose Mean Need Not Be Known

Step 1. Choose a confidence level γ (95%, 99%, or the like).

Step 2. Determine solutions c_1 and c_2 of the equations

$$F(c_1) = \tfrac{1}{2}(1 - \gamma), \qquad F(c_2) = \tfrac{1}{2}(1 + \gamma) \tag{15}$$

from the table of the chi-square distribution with $n - 1$ degrees of freedom (Table A10 in App. 5; or use a CAS; n = sample size).

Step 3. Compute $(n - 1)s^2$, where s^2 is the variance of the sample $x_1, \cdots, x_n$.

Step 4. Compute $k_1 = (n - 1)s^2/c_1$ and $k_2 = (n - 1)s^2/c_2$. The confidence interval is

$$\text{CONF}_\gamma \{k_2 \leqq \sigma^2 \leqq k_1\}. \tag{16}$$

EXAMPLE 4 Confidence Interval for the Variance of the Normal Distribution

Determine a 95% confidence interval (16) for the variance, using Table 25.3 and a sample (tensile strength of sheet steel in kg/mm^2, rounded to integer values)

89 84 87 81 89 86 91 90 78 89 87 99 83 89.

Solution. *Step 1.* $\gamma = 0.95$ is required.

Step 2. For $n - 1 = 13$ we find

$$c_1 = 5.01 \quad \text{and} \quad c_2 = 24.74.$$

Step 3. $13s^2 = 326.9$.

Step 4. $13s^2/c_1 = 65.25$, $13s^2/c_2 = 13.21$.

The confidence interval is

$$\text{CONF}_{0.95}\{13.21 \leqq \sigma^2 \leqq 65.25\}.$$

This is rather large, and for obtaining a more precise result, one would need a much larger sample. ▩

Theory for Table 25.3. In Table 25.1 we used the normal distribution, in Table 25.2 the *t*-distribution, and now we shall use the **χ^2-distribution** (*chi-square distribution*), whose distribution function is $F(z) = 0$ if $z < 0$ and

$$F(z) = C_m \int_0^z e^{-u/2} u^{(m-2)/2}\, du \qquad \text{if } z \geqq 0 \qquad \text{(Fig. 529).}$$

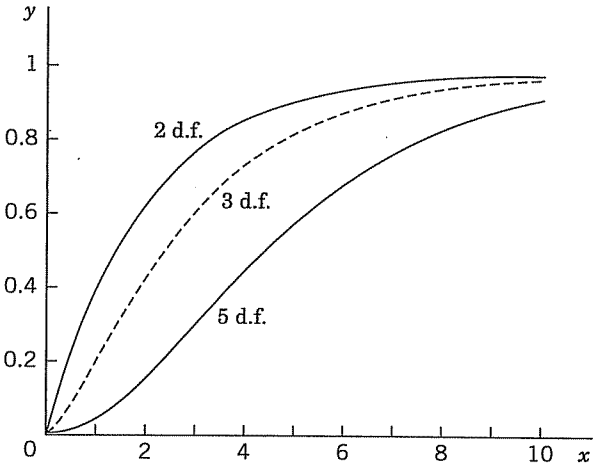


Fig. 529. Distribution function of the chi-square distribution with 2, 3, 5 d.f.

The parameter m $(= 1, 2, \cdots)$ is called the **number of degrees of freedom** (d.f.), and

$$C_m = 1/[2^{m/2}\Gamma(\tfrac{1}{2}m)].$$

Note that the distribution is not symmetric (see also Fig. 530).

For deriving (16) in Table 25.3 we need the following theorem.

THEOREM 3

Chi-Square Distribution

Under the assumptions in Theorem 2 the random variable

(17)
$$Y = (n-1)\frac{S^2}{\sigma^2}$$

with S^2 given by (12) *has a chi-square distribution with $n - 1$ degrees of freedom.*

Proof in Ref. [G3], listed in App. 1.

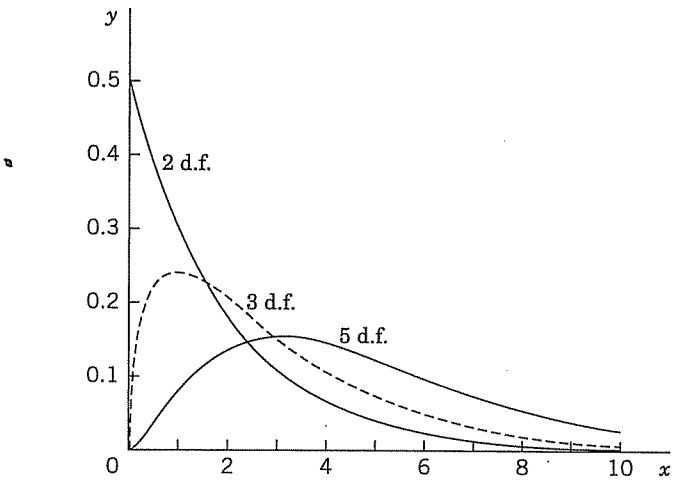


Fig. 530. Density of the chi-square distribution with 2, 3, 5 d.f.

Derivation of (16). This is similar to the derivation of (3) and (10). We choose a number γ between 0 and 1 and determine c_1 and c_2 from Table A10, App. 5, such that [see (15)]

$$P(Y \leqq c_1) = F(c_1) = \tfrac{1}{2}(1 - \gamma), \qquad P(Y \leqq c_2) = F(c_2) = \tfrac{1}{2}(1 + \gamma).$$

Subtraction yields

$$P(c_1 \leqq Y \leqq c_2) = P(Y \leqq c_2) - P(Y \leqq c_1) = F(c_2) - F(c_1) = \gamma.$$

Transforming $c_1 \leqq Y \leqq c_2$ with Y given by (17) into an inequality for σ^2, we obtain

$$\frac{n-1}{c_2} S^2 \leqq \sigma^2 \leqq \frac{n-1}{c_1} S^2.$$

By inserting the observed value s^2 of S^2 we obtain (16). ∎

Confidence Intervals for Parameters of Other Distributions

The methods in Tables 25.1–25.3 for confidence intervals for μ and σ^2 are designed for the normal distribution. We now show that they can also be applied to other distributions if we use large samples.

We know that if $X_1, \cdots, X_n$ are independent random variables with the same mean μ and the same variance σ^2, then their sum $Y_n = X_1 + \cdots + X_n$ has the following properties.

(A) Y_n has the mean $n\mu$ and the variance $n\sigma^2$ (by Theorems 1 and 3 in Sec. 24.9).

(B) If those variables are normal, then Y_n is normal (by Theorem 1).

If those random variables are not normal, then **(B)** is not applicable. However, for large n the random variable Y_n is still *approximately* normal. This follows from the central limit theorem, which is one of the most fundamental results in probability theory.

THEOREM 4

Central Limit Theorem

Let $X_1, \cdots, X_n, \cdots$ be independent random variables that have the same distribution function and therefore the same mean μ and the same variance σ^2. Let $Y_n = X_1 + \cdots + X_n$. Then the random variable

(18)
$$Z_n = \frac{Y_n - n\mu}{\sigma\sqrt{n}}$$

is **asymptotically normal** *with mean 0 and variance 1; that is, the distribution function $F_n(x)$ of Z_n satisfies*

$$\lim_{n\to\infty} F_n(x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2}\, du.$$

A proof can be found in Ref. [G3] listed in App. 1.

Hence when applying Tables 25.1–25.3 to a nonnormal distribution, we must use ***sufficiently large samples.*** As a rule of thumb, if the sample indicates that the skewness of the distribution (the asymmetry; see Team Project 16(d), Problem Set 24.6) is small, use at least $n = 20$ for the mean and at least $n = 50$ for the variance.

PROBLEM SET 25.3

1–7 MEAN (VARIANCE KNOWN)

1. Find a 95% confidence interval for the mean μ of a normal population with standard deviation 4.00 from the sample 30, 42, 40, 34, 48, 50.
2. Does the interval in Prob. 1 get longer or shorter if we take $\gamma = 0.99$ instead of 0.95? By what factor?
3. By what factor does the length of the interval in Prob. 1 change if we double the sample size?
4. Find a 90% confidence interval for the mean μ of a normal population with variance 0.25, using a sample of 100 values with mean 212.3.
5. What sample size would be needed for obtaining a 95% confidence interval (3) of length 2σ? Of length σ?
6. **(Use of Fig. 525)** Find a 95% confidence interval for a sample of 200 values with mean 120 from a normal distribution with variance 4, using Fig. 525.
7. What sample size is needed to obtain a 99% confidence interval of length 2.0 for the mean of a normal population with variance 25? Use Fig. 525. Check by calculation.

8–12 MEAN (VARIANCE UNKNOWN)

Find a 99% confidence interval for the mean of a normal population from the sample:

8. 425, 420, 425, 435
9. Length of 20 bolts with sample mean 20.2 cm and sample variance 0.04 cm^2
10. Knoop hardness of diamond 9500, 9800, 9750, 9200, 9400, 9550
11. Copper content (%) of brass 66, 66, 65, 64, 66, 67, 64, 65, 63, 64
12. Melting point (°C) of aluminum 660, 667, 654, 663, 662
13. Find a 95% confidence interval for the percentage of cars on a certain highway that have poorly adjusted brakes, using a random sample of 500 cars stopped at a roadblock on that highway, 87 of which had poorly adjusted brakes.
14. Find a 99% confidence interval for p in the binomial distribution from a classical result by K. Pearson, who in 24000 trials of tossing a coin obtained 12012 Heads. Do you think that the coin was fair?

15–20 VARIANCE

Find a 95% confidence interval for the variance of a normal population from the sample:

15. A sample of 30 values with variance 0.0007

16. The sample in Prob. 9

17. The sample in Prob. 11

18. Carbon monoxide emission (grams per mile) of a certain type of passenger car (cruising at 55 mph): 17.3, 17.8, 18.0, 17.7, 18.2, 17.4, 17.6, 18.1

19. Mean energy (keV) of delayed neutron group (Group 3, half-life 6.2 sec.) for uranium U^{235} fission: 435, 451, 430, 444, 438

20. Ultimate tensile strength (k psi) of alloy steel (Maraging H) at room temperature: 251, 255, 258, 253, 253, 252, 250, 252, 255, 256

21. If X is normal with mean 27 and variance 16, what distributions do $-X$, $3X$, and $5X - 2$ have?

22. If X_1 and X_2 are independent normal random variables with mean 23 and 4 and variance 3 and 1, respectively, what distribution does $4X_1 - X_2$ have? *Hint.* Use Team Project 14(g) in Sec. 24.8.

23. A machine fills boxes weighing Y lb with X lb of salt, where X and Y are normal with mean 100 lb and 5 lb and standard deviation 1 lb and 0.5 lb, respectively. What percent of filled boxes weighing between 104 lb and 106 lb are to be expected?

24. If the weight X of bags of cement is normally distributed with a mean of 40 kg and a standard deviation of 2 kg, how many bags can a delivery truck carry so that the probability of the total load exceeding 2000 kg will be 5%?

25. CAS EXPERIMENT. Confidence Intervals. Obtain 100 samples of size 10 of the standardized normal distribution. Calculate from them and graph the corresponding 95% confidence intervals for the mean and count how many of them do not contain 0. Does the result support the theory? Repeat the whole experiment, compare and comment.

25.4 Testing of Hypotheses. Decisions

The ideas of confidence intervals and of tests[2] are the two most important ideas in modern statistics. In a statistical **test** we make inference from sample to population through testing a **hypothesis,** resulting from experience or observations, from a theory or a quality requirement, and so on. In many cases the result of a test is used as a basis for a **decision,** for instance, to buy (or not to buy) a certain model of car, depending on a test of the fuel efficiency (miles/gal) (and other tests, of course), to apply some medication, depending on a test of its effect; to proceed with a marketing strategy, depending on a test of consumer reactions, etc.

Let us explain such a test in terms of a typical example and introduce the corresponding standard notions of statistical testing.

EXAMPLE 1 Test of a Hypothesis. Alternative. Significance Level α

We want to buy 100 coils of a certain kind of wire, provided we can verify the manufacturer's claim that the wire has a breaking limit $\mu = \mu_0 = 200$ lb (or more). This is a test of the **hypothesis** (also called *null hypothesis*) $\mu = \mu_0 = 200$. We shall not buy the wire if the (statistical) test shows that actually $\mu = \mu_1 < \mu_0$, the wire is weaker, the claim does not hold. μ_1 is called the **alternative** (or *alternative hypothesis*) of the test. We shall **accept** the hypothesis if the test suggests that it is true, except for a small error probability α, called the **significance level** of the test. Otherwise we **reject** the hypothesis. Hence α is the probability of rejecting a hypothesis although it is true. The choice of α is up to us. 5% and 1% are popular values.

For the test we need a sample. We randomly select 25 coils of the wire, cut a piece from each coil, and determine the breaking limit experimentally. Suppose that this sample of $n = 25$ values of the breaking limit has the mean $\bar{x} = 197$ lb (somewhat less than the claim!) and the standard deviation $s = 6$ lb.

[2]Beginning around 1930, a systematic theory of tests was developed by NEYMAN (see Sec. 25.3) and EGON SHARPE PEARSON (1895–1980), English statistician, the son of Karl Pearson (see the footnote on p. 1066).

At this point we could only speculate whether this difference $197 - 200 = -3$ is due to randomness, is a chance effect, or whether it is **significant,** due to the actually inferior quality of the wire. To continue beyond speculation requires probability theory, as follows.

We assume that the breaking limit is normally distributed. (This assumption could be tested by the method in Sec. 25.7. Or we could remember the central limit theorem (Sec. 25.3) and take a still larger sample.) Then

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

in (11), Sec. 25.3, with $\mu = \mu_0$ has a ***t*-distribution** with $n - 1$ degrees of freedom ($n - 1 = 24$ for our sample). Also $\bar{x} = 197$ and $s = 6$ are observed values of $\bar{X}$ and S to be used later. We can now choose a significance level, say, $\alpha = 5\%$. From Table A9 in App. 5 or from a CAS we then obtain a critical value c such that $P(T \leqq c) = \alpha = 5\%$. For $P(T \leqq \tilde{c}) = 1 - \alpha = 95\%$ the table gives $\tilde{c} = 1.71$, so that $c = -\tilde{c} = -1.71$ because of the symmetry of the distribution (Fig. 531).

We now reason as follows—this is the ***crucial idea*** of the test. If the hypothesis is true, we have a chance of only α (= 5%) that we observe a value t of T (calculated from a sample) that will fall between $-\infty$ and -1.71. Hence if we nevertheless do observe such a t, we assert that the hypothesis cannot be true and we reject it. Then we accept the alternative. If, however, $t \geqq c$, we accept the hypothesis.

A simple calculation finally gives $t = (197 - 200)/(6/\sqrt{25}) = -2.5$ as an observed value of T. Since $-2.5 < -1.71$, we reject the hypothesis (the manufacturer's claim) and accept the alternative $\mu = \mu_1 < 200$, the wire seems to be weaker than claimed. ■

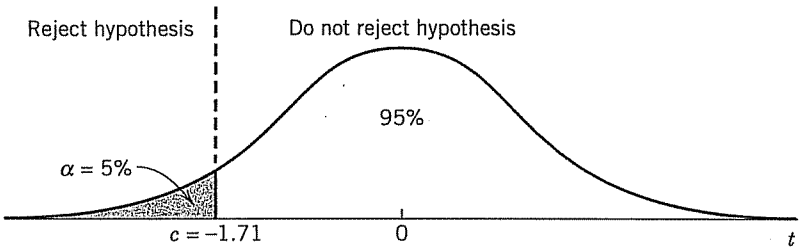


Fig. 531. *t*-distribution in Example 1

This example illustrates the ***steps of a test:***

1. Formulate the **hypothesis** $\theta = \theta_0$ to be tested. ($\theta_0 = \mu_0$ in the example.)
2. Formulate an **alternative** $\theta = \theta_1$. ($\theta_1 = \mu_1$ in the example.)
3. Choose a **significance level** α (5%, 1%, 0.1%).
4. Use a random variable $\hat{\Theta} = g(X_1, \cdots, X_n)$ whose distribution depends on the hypothesis and on the alternative, and this distribution is known in both cases. Determine a critical value c from the distribution of $\hat{\Theta}$, assuming the hypothesis to be true. (In the example, $\hat{\Theta} = T$, and c is, obtained from $P(T \leqq c) = \alpha$.)
5. Use a sample $x_1, \cdots, x_n$ to determine an observed value $\hat{\theta} = g(x_1, \cdots, x_n)$ of $\hat{\Theta}$. (t in the example.)
6. Accept or reject the hypothesis, depending on the size of $\hat{\theta}$ relative to c. ($t < c$ in the example, rejection of the hypothesis.)

Two important facts require further discussion and careful attention. The first is the choice of an alternative. In the example, $\mu_1 < \mu_0$, but other applications may require $\mu_1 > \mu_0$ or $\mu_1 \neq \mu_0$. The second fact has to do with errors. We know that α (the significance level of the test) is the probability of *rejecting* a *true* hypothesis. And we shall discuss the probability β of *accepting* a *false* hypothesis.

One-Sided and Two-Sided Alternatives (Fig. 532)

Let θ be an unknown parameter in a distribution, and suppose that we want to test the hypothesis $\theta = \theta_0$. Then there are three main kinds of alternatives, namely,

(1) $\theta > \theta_0$

(2) $\theta < \theta_0$

(3) $\theta \neq \theta_0$.

(1) and (2) are **one-sided alternatives,** and (3) is a **two-sided alternative.**

We call **rejection region** (or **critical region**) the region such that we reject the hypothesis if the observed value in the test falls in this region. In ① the critical c lies to the right of θ_0 because so does the alternative. Hence the rejection region extends to the right. This is called a **right-sided test.** In ② the critical c lies to the left of θ_0 (as in Example 1), the rejection region extends to the left, and we have a **left-sided test** (Fig. 532, middle part). These are **one-sided tests.** In ③ we have two rejection regions. This is called a **two-sided test** (Fig. 532, lower part).

All three kinds of alternatives occur in practical problems. For example, (1) may arise if θ_0 is the maximum tolerable inaccuracy of a voltmeter or some other instrument. Alternative (2) may occur in testing strength of material, as in Example 1. Finally, θ_0 in (3) may be the diameter of axle-shafts, and shafts that are too thin or too thick are equally undesirable, so that we have to watch for deviations in both directions.

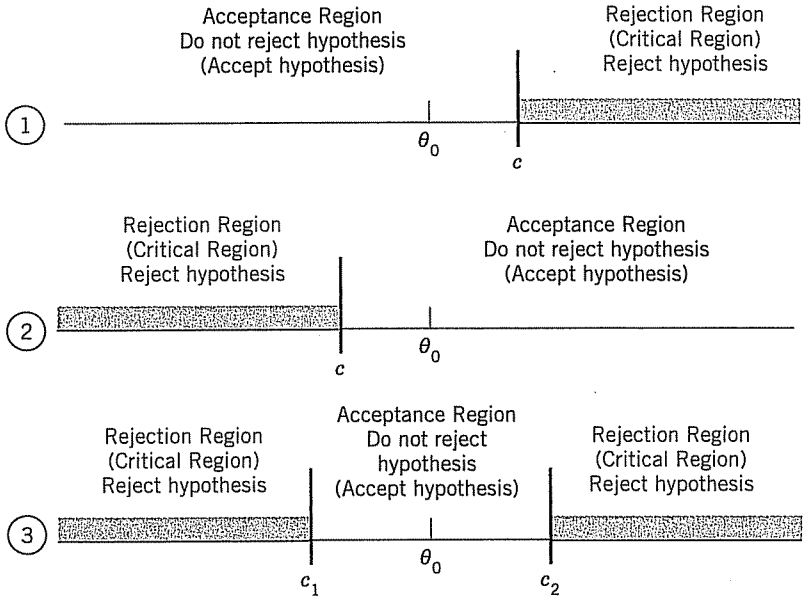


Fig. 532. Test in the case of alternative (1) (upper part of the figure), alternative (2) (middle part), and alternative (3)

Errors in Tests

Tests always involve **risks of making false decisions:**

(I) Rejecting a true hypothesis **(Type I error).**
α = Probability of making a Type I error.

(II) Accepting a false hypothesis **(Type II error).**
β = Probability of making a Type II error.

Clearly, we cannot avoid these errors because no absolutely certain conclusions about populations can be drawn from samples. But we show that there are ways and means of choosing suitable levels of risks, that is, of values α and β. The choice of α depends on the nature of the problem (e.g., a small risk $\alpha = 1\%$ is used if it is a matter of life or death).

Let us discuss this systematically for a test of a hypothesis $\theta = \theta_0$ against an alternative that is a single number θ_1, for simplicity. We let $\theta_1 > \theta_0$, so that we have a right-sided test. For a left-sided or a two-sided test the discussion is quite similar.

We choose a critical $c > \theta_0$ (as in the upper part of Fig. 532, by methods discussed below). From a given sample $x_1, \cdots, x_n$ we then compute a value

$$\hat{\theta} = g(x_1, \cdots, x_n)$$

with a suitable g (whose choice will be a main point of our further discussion; for instance, take $g = (x_1 + \cdots + x_n)/n$ in the case in which θ is the mean). If $\hat{\theta} > c$, we reject the hypothesis. If $\hat{\theta} \leqq c$, we accept it. Here, the value $\hat{\theta}$ can be regarded as an observed value of the random variable

$$\hat{\Theta} = g(X_1, \cdots, X_n) \tag{4}$$

because x_j may be regarded as an observed value of X_j, $j = 1, \cdots, n$. In this test there are two possibilities of making an error, as follows.

Type I Error (see Table 25.4). The hypothesis is true but is rejected (hence the alternative is accepted) because Θ assumes a value $\hat{\theta} > c$. Obviously, the probability of making such an error equals

$$P(\hat{\Theta} > c)_{\theta=\theta_0} = \alpha. \tag{5}$$

α is called the **significance level** of the test, as mentioned before.

Type II Error (see Table 25.4). The hypothesis is false but is accepted because $\hat{\Theta}$ assumes a value $\hat{\theta} \leqq c$. The probability of making such an error is denoted by β; thus

$$P(\hat{\Theta} \leqq c)_{\theta=\theta_1} = \beta. \tag{6}$$

$\eta = 1 - \beta$ is called the **power** of the test. Obviously, the power η is the probability of avoiding a Type II error.

Table 25.4 Type I and Type II Errors in Testing a Hypothesis $\theta = \theta_0$ Against an Alternative $\theta = \theta_1$

		Unknown Truth	
		$\theta = \theta_0$	$\theta = \theta_1$
Accepted	$\theta = \theta_0$	True decision $P = 1 - \alpha$	Type II error $P = \beta$
	$\theta = \theta_1$	Type 1 error $P = \alpha$	True decision $P = 1 - \beta$

Formulas (5) and (6) show that both α and β depend on c, and we would like to choose c so that these probabilities of making errors are as small as possible. But the important Figure 533 shows that these are conflicting requirements because to let α decrease we must shift c to the right, but then β increases. In practice we first choose α (5%, sometimes 1%), then determine c, and finally compute β. If β is large so that the power $\eta = 1 - \beta$ is small, we should repeat the test, choosing a larger sample, for reasons that will appear shortly.

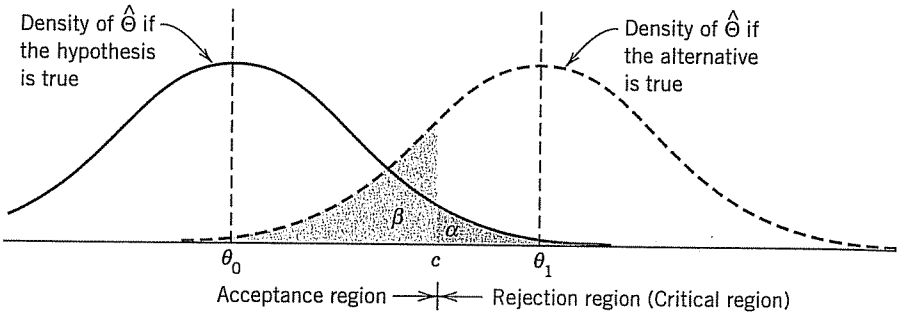


Fig. 533. Illustration of Type I and II errors in testing a hypothesis $\theta = \theta_0$ against an alternative $\theta = \theta_1$ ($> \theta_0$, right-sided test)

If the alternative is not a single number but is of the form (1)–(3), then β becomes a function of θ. This function $\beta(\theta)$ is called the **operating characteristic** (OC) of the test and its curve the **OC curve.** Clearly, in this case $\eta = 1 - \beta$ also depends on θ. This function $\eta(\theta)$ is called the **power function** of the test. (Examples will follow.)

Of course, from a test that leads to the acceptance of a certain hypothesis θ_0, it does ***not*** follow that this is the only possible hypothesis or the best possible hypothesis. Hence the terms **"not reject"** or **"fail to reject"** are perhaps better than the term **"accept."**

Test for μ of the Normal Distribution with Known σ^2

The following example explains the three kinds of hypotheses.

EXAMPLE 2 **Test for the Mean of the Normal Distribution with Known Variance**

Let X be a normal random variable with variance $\sigma^2 = 9$. Using a sample of size $n = 10$ with mean $\bar{x}$, test the hypothesis $\mu = \mu_0 = 24$ against the three kinds of alternatives, namely,

$$\text{(a)} \quad \mu > \mu_0 \qquad \text{(b)} \quad \mu < \mu_0 \qquad \text{(c)} \quad \mu \neq \mu_0.$$

Solution. We choose the significance level $\alpha = 0.05$. An estimate of the mean will be obtained from

$$\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n).$$

If the hypothesis is true, $\bar{X}$ is normal with mean $\mu = 24$ and variance $\sigma^2/n = 0.9$, see Theorem 1, Sec. 25.3. Hence we may obtain the critical value c from Table A8 in App. 5.

Case (a). ***Right-Sided Test.*** We determine c from $P(\bar{X} > c)_{\mu=24} = \alpha = 0.05$, that is,

$$P(\bar{X} \leqq c)_{\mu=24} = \Phi\left(\frac{c - 24}{\sqrt{0.9}}\right) = 1 - \alpha = 0.95.$$

Table A8 in App. 5 gives $(c - 24)/\sqrt{0.9} = 1.645$, and $c = 25.56$, which is greater than μ_0, as in the upper part of Fig. 532. If $\bar{x} \leqq 25.56$, the hypothesis is accepted. If $\bar{x} > 25.56$, it is rejected. The power function of the test is (Fig. 534)

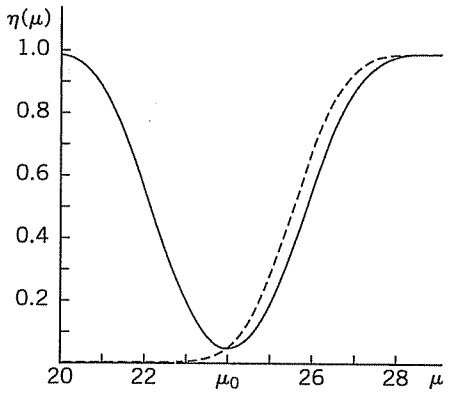


Fig. 534. Power function $\eta(\mu)$ in Example 2, case (a) (dashed) and case (c)

$$\eta(\mu) = P(\overline{X} > 25.56)_\mu = 1 - P(\overline{X} \leqq 25.56)_\mu$$

(7)
$$= 1 - \Phi\left(\frac{25.56 - \mu}{\sqrt{0.9}}\right) = 1 - \Phi(26.94 - 1.05\mu)$$

Case (b). ***Left-Sided Test.*** The critical value c is obtained from the equation

$$P(\overline{X} \leqq c)_{\mu=24} = \Phi\left(\frac{c - 24}{\sqrt{0.9}}\right) = \alpha = 0.05.$$

Table A8 in App. 5 yields $c = 24 - 1.56 = 22.44$. If $\bar{x} \geqq 22.44$, we accept the hypothesis. If $\bar{x} < 22.44$, we reject it. The power function of the test is

(8)
$$\eta(\mu) = P(\overline{X} \leqq 22.44)_\mu = \Phi\left(\frac{22.44 - \mu}{\sqrt{0.9}}\right) = \Phi(23.65 - 1.05\mu).$$

Case (c). ***Two-Sided Test.*** Since the normal distribution is symmetric, we choose c_1 and c_2 equidistant from $\mu = 24$, say, $c_1 = 24 - k$ and $c_2 = 24 + k$, and determine k from

$$P(24 - k \leqq \overline{X} \leqq 24 + k)_{\mu=24} = \Phi\left(\frac{k}{\sqrt{0.9}}\right) - \Phi\left(-\frac{k}{\sqrt{0.9}}\right) = 1 - \alpha = 0.95.$$

Table A8 in App. 5 gives $k/\sqrt{0.9} = 1.960$, hence $k = 1.86$. This gives the values $c_1 = 24 - 1.86 = 22.14$ and $c_2 = 24 + 1.86 = 25.86$. If $\bar{x}$ is not smaller than c_1 and not greater than c_2, we accept the hypothesis. Otherwise we reject it. The power function of the test is (Fig. 534)

$$\eta(\mu) = P(\overline{X} < 22.14)_\mu + P(\overline{X} > 25.86)_\mu = P(\overline{X} < 22.14)_\mu + 1 - P(\overline{X} \leqq 25.86)_\mu$$

(9)
$$= 1 + \Phi\left(\frac{22.14 - \mu}{\sqrt{0.9}}\right) - \Phi\left(\frac{25.86 - \mu}{\sqrt{0.9}}\right)$$

$$= 1 + \Phi(23.34 - 1.05\mu) - \Phi(27.26 - 1.05\mu).$$

Consequently, the operating characteristic $\beta(\mu) = 1 - \eta(\mu)$ (see before) is (Fig. 535)

$$\beta(\mu) = \Phi(27.26 - 1.05\mu) - \Phi(23.34 - 1.05\mu).$$

If we take a larger sample, say, of size $n = 100$ (instead of 10), then $\sigma^2/n = 0.09$ (instead of 0.9) and the critical values are $c_1 = 23.41$ and $c_2 = 24.59$, as can be readily verified. Then the operating characteristic of the test is

$$\beta(\mu) = \Phi\left(\frac{24.59 - \mu}{\sqrt{0.09}}\right) - \Phi\left(\frac{23.41 - \mu}{\sqrt{0.09}}\right)$$

$$= \Phi(81.97 - 3.33\mu) - \Phi(78.03 - 3.33\mu).$$

Figure 535 shows that the corresponding OC curve is steeper than that for $n = 10$. This means that the increase of n has led to an improvement of the test. In any practical case, n is chosen as small as possible but so large that the test brings out deviations between μ and μ_0 that are of practical interest. For instance, if deviations of ± 2 units are of interest, we see from Fig. 535 that $n = 10$ is much too small because when $\mu = 24 - 2 = 22$ or $\mu = 24 + 2 = 26$ β is almost 50%. On the other hand, we see that $n = 100$ is sufficient for that purpose. ▩

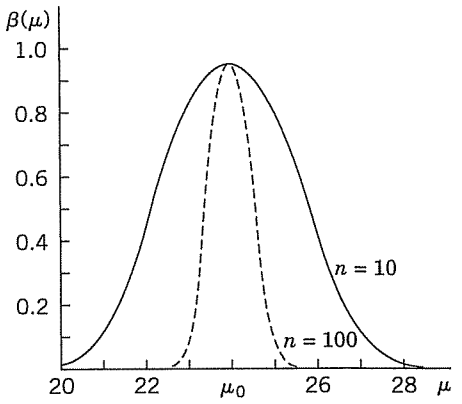


Fig. 535. Curves of the operating characteristic (OC curves) in Example 2, case (*c*), for two different sample sizes n

Test for μ When σ^2 is Unknown, and for σ^2

EXAMPLE 3 **Test for the Mean of the Normal Distribution with Unknown Variance**

The tensile strength of a sample of $n = 16$ manila ropes (diameter 3 in.) was measured. The sample mean was $\bar{x} = 4482$ kg, and the sample standard deviation was $s = 115$ kg (N. C. Wiley, 41st Annual Meeting of the American Society for Testing Materials). Assuming that the tensile strength is a normal random variable, test the hypothesis $\mu_0 = 4500$ kg against the alternative $\mu_1 = 4400$ kg. Here μ_0 may be a value given by the manufacturer, while μ_1 may result from previous experience.

Solution. We choose the significance level $\alpha = 5\%$. If the hypothesis is true, it follows from Theorem 2 in Sec. 25.3, that the random variable

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} = \frac{\overline{X} - 4500}{S/4}$$

has a t-distribution with $n - 1 = 15$ d.f. The test is left-sided. The critical value c is obtained from $P(T < c)_{\mu_0} = \alpha = 0.05$. Table A9 in App. 5 gives $c = -1.75$. As an observed value of T we obtain from the sample $t = (4482 - 4500)/(115/4) = -0.626$. We see that $t > c$ and accept the hypothesis. For obtaining numeric values of the power of the test, we would need tables called noncentral Student t-tables; we shall not discuss this question here. ▩

EXAMPLE 4 **Test for the Variance of the Normal Distribution**

Using a sample of size $n = 15$ and sample variance $s^2 = 13$ from a normal population, test the hypothesis $\sigma^2 = {\sigma_0}^2 = 10$ against the alternative $\sigma^2 = {\sigma_1}^2 = 20$.

Solution. We choose the significance level $\alpha = 5\%$. If the hypothesis is true, then

$$Y = (n - 1)\frac{S^2}{{\sigma_0}^2} = 14\,\frac{S^2}{10} = 1.4S^2$$

has a chi-square distribution with $n - 1 = 14$ d.f. by Theorem 3, Sec. 25.3. From

$$P(Y > c) = \alpha = 0.05, \qquad \text{that is,} \qquad P(Y \leqq c) = 0.95,$$

and Table A10 in App. 5 with 14 degrees of freedom we obtain $c = 23.68$. This is the critical value of Y. Hence

to $S^2 = {\sigma_0}^2 Y/(n-1) = 0.714Y$ there corresponds the critical value $c^* = 0.714 \cdot 23.68 = 16.91$. Since $s^2 < c^*$, we accept the hypothesis.

If the alternative is true, the random variable $Y_1 = 14S^2/{\sigma_1}^2 = 0.7S^2$ has a chi-square distribution with 14 d.f. Hence our test has the power

$$\eta = P(S^2 > c^*)_{\sigma^2=20} = P(Y_1 > 0.7c^*)_{\sigma^2=20} = 1 - P(Y_1 \leqq 11.84)_{\sigma^2=20}.$$

From a more extensive table of the chi-square distribution (e.g. in Ref. [G3] or [G8]) or from your CAS, you see that $\eta \approx 62\%$. Hence the Type II risk is very large, namely, 38%. To make this risk smaller, we would have to increase the sample size. ▩

Comparison of Means and Variances

EXAMPLE 5 **Comparison of the Means of Two Normal Distributions**

Using a sample $x_1, \cdots, x_{n_1}$ from a normal distribution with unknown mean μ_x and a sample $y_1, \cdots, y_{n_2}$ from another normal distribution with unknown mean μ_y, we want to test the hypothesis that the means are equal, $\mu_x = \mu_y$, against an alternative, say, $\mu_x > \mu_y$. The variances need not be known but are assumed to be equal.[3]

Two cases of comparing means are of practical importance:

Case A. *The samples have the* ***same size.*** *Furthermore, each value of the first sample corresponds to precisely one value of the other,* because corresponding values result from the same person or thing **(paired comparison)**—for example, two measurements of the same thing by two different methods or two measurements from the two eyes of the same person. More generally, they may result from pairs of *similar* individuals or things, for example, identical twins, pairs of used front tires from the same car, etc. Then we should form the differences of corresponding values and test the hypothesis that the population corresponding to the differences has mean 0, using the method in Example 3. If we have a choice, this method is better than the following.

Case B. *The two samples are independent and not necessarily of the same size.* Then we may proceed as follows. Suppose that the alternative is $\mu_x > \mu_y$. We choose a significance level α. Then we compute the sample means $\bar{x}$ and $\bar{y}$ as well as $(n_1 - 1){s_x}^2$ and $(n_2 - 1){s_y}^2$, where ${s_x}^2$ and ${s_y}^2$ are the sample variances. Using Table A9 in App. 5 with $n_1 + n_2 - 2$ degrees of freedom, we now determine c from

$$P(T \leqq c) = 1 - \alpha. \tag{10}$$

We finally compute

$$t_0 = \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}} \, \frac{\bar{x} - \bar{y}}{\sqrt{(n_1 - 1){s_x}^2 + (n_2 - 1){s_y}^2}}. \tag{11}$$

It can be shown that this is an observed value of a random variable that has a t-distribution with $n_1 + n_2 - 2$ degrees of freedom, provided the hypothesis is true. If $t_0 \leqq c$, the hypothesis is accepted. If $t_0 > c$, it is rejected.

If the alternative is $\mu_x \neq \mu_y$, then (10) must be replaced by

$$P(T \leqq c_1) = 0.5\alpha, \qquad P(T \leqq c_2) = 1 - 0.5\alpha. \tag{10*}$$

Note that for samples of equal size $n_1 = n_2 = n$, formula (11) reduces to

$$t_0 = \sqrt{n} \, \frac{\bar{x} - \bar{y}}{\sqrt{{s_x}^2 + {s_y}^2}}. \tag{12}$$

[3]This assumption of equality of variances can be tested, as shown in the next example. If the test shows that they differ significantly, choose two samples of the same size $n_1 = n_2 = n$ (not too small, > 30, say), use the test in Example 2 together with the fact that (12) is an observed value of an approximately standardized normal random variable.

To illustrate the computations, let us consider the two samples $(x_1, \cdots, x_{n_1})$ and $(y_1, \cdots, y_{n_2})$ given by

105	108	86	103	103	107	124	105

and

89	92	84	97	103	107	111	97

showing the relative output of tin plate workers under two different working conditions [J. J. B. Worth, *Journal of Industrial Engineering* **9**, 249–253). Assuming that the corresponding populations are normal and have the same variance, let us test the hypothesis $\mu_x = \mu_y$ against the alternative $\mu_x \neq \mu_y$. (Equality of variances will be tested in the next example.)

Solution. We find

$$\bar{x} = 105.125, \qquad \bar{y} = 97.500, \qquad s_x^2 = 106.125, \qquad s_y^2 = 84.000.$$

We choose the significance level $\alpha = 5\%$. From (10*) with $0.5\alpha = 2.5\%$, $1 - 0.5\alpha = 97.5\%$ and Table A9 in App. 5 with 14 degrees of freedom we obtain $c_1 = -2.14$ and $c_2 = 2.14$. Formula (12) with $n = 8$ gives the value

$$t_0 = \sqrt{8} \cdot 7.625/\sqrt{190.125} = 1.56.$$

Since $c_1 \leqq t_0 \leqq c_2$, we ***accept the hypothesis*** $\mu_x = \mu_y$ that under both conditions the mean output is the same.

Case A applies to the example because the two first sample values correspond to a certain type of work, the next two were obtained in another kind of work, etc. So we may use the differences

16	16	2	6	0	0	13	8

of corresponding sample values and the method in Example 3 to test the hypothesis $\mu = 0$, where μ is the mean of the population corresponding to the differences. As a logical alternative we take $\mu \neq 0$. The sample mean is $\bar{d} = 7.625$, and the sample variance is $s^2 = 45.696$. Hence

$$t = \sqrt{8}\,(7.625 - 0)/\sqrt{45.696} = 3.19.$$

From $P(T \leqq c_1) = 2.5\%$, $P(T \leqq c_2) = 97.5\%$ and Table A9 in App. 5 with $n - 1 = 7$ degrees of freedom we obtain $c_1 = -2.36$, $c_2 = 2.36$ and ***reject the hypothesis*** because $t = 3.19$ does not lie between c_1 and c_2. Hence our present test, in which we used more information (but the same samples), shows that the difference in output is significant. ▩

EXAMPLE 6 Comparison of the Variance of Two Normal Distributions

Using the two samples in the last example, test the hypothesis $\sigma_x^2 = \sigma_y^2$; assume that the corresponding populations are normal and the nature of the experiment suggests the alternative $\sigma_x^2 > \sigma_y^2$.

Solution. We find $s_x^2 = 106.125$, $s_y^2 = 84.000$. We choose the significance level $\alpha = 5\%$. Using $P(V \leqq c) = 1 - \alpha = 95\%$ and Table A11 in App. 5, with $(n_1 - 1, n_2 - 1) = (7, 7)$ degrees of freedom, we determine $c = 3.79$. We finally compute $v_0 = s_x^2/s_y^2 = 1.26$. Since $v_0 \leqq c$, we accept the hypothesis. If $v_0 > c$, we would reject it.

This test is justified by the fact that v_0 is an observed value of a random variable that has a so-called ***F*-distribution** with $(n_1 - 1, n_2 - 1)$ degrees of freedom, provided the hypothesis is true. (Proof in Ref. [G3] listed in App. 1.) The F-distribution with (m, n) degrees of freedom was introduced by R. A. Fisher[4] and has the distribution function $F(z) = 0$ if $z < 0$ and

$$\textbf{(13)} \qquad F(z) = K_{mn} \int_0^z t^{(m-2)/2}(mt + n)^{-(m+n)/2}\,dt \qquad (z \geqq 0),$$

where $K_{mn} = m^{m/2}n^{n/2}\Gamma(\frac{1}{2}m + \frac{1}{2}n)/\Gamma(\frac{1}{2}m)\Gamma(\frac{1}{2}n)$. (For Γ see App. A3.1.) ▩

[4]After the pioneering work of the English statistician and biologist, KARL PEARSON (1857–1936), the founder of the English school of statistics, and WILLIAM SEALY GOSSET (1876–1937), who discovered the t-distribution (and published under the name 'Student'), the English statistician Sir RONALD AYLMER FISHER (1890–1962), professor of eugenics in London (1933–1943) and professor of genetics in Cambridge, England (1943–1957) and Adelaide, Australia (1957–1962), had great influence on the further development of modern statistics.

This long section contained the basic ideas and concepts of testing, along with typical applications and you may perhaps want to review it quickly before going on, because the next sections concern an adaption of these ideas to tasks of great practical importance and resulting tests in connection with quality control, acceptance (or rejection) of goods produced, and so on.

PROBLEM SET 25.4

1. Test $\mu = 0$ against $\mu > 0$, assuming normality and using the sample 1, −1, 1, 3, −8, 6, 0 (deviations of the azimuth [multiples of 0.01 radian] in some revolution of a satellite). Choose $\alpha = 5\%$.
2. In one of his classical experiments Buffon obtained 2048 heads in tossing a coin 4040 times. Was the coin fair?
3. Do the same test as in Prob. 2, using a result by K. Pearson, who obtained 6 019 heads in 12 000 trials.
4. Assuming normality and known variance $\sigma^2 = 4$, test the hypothesis $\mu = 30.0$ against the alternative (a) $\mu = 28.5$, (b) $\mu = 30.7$, using a sample of size 10 with mean $\bar{x} = 28.5$ and choosing $\alpha = 5\%$.
5. How does the result in Prob. 4(a) change if we use a smaller sample, say, of size 4, the other data ($\bar{x} = 28.5$, $\alpha = 5\%$, etc.) remaining as before?
6. Detemine the power of the test in Prob. 4(a).
7. What is the rejection region in Prob. 4 in the case of a two-sided test with $\alpha = 5\%$?
8. Using the sample 0.80, 0.81, 0.81, 0.82, 0.81, 0.82, 0.80, 0.82, 0.81, 0.81 (length of nails in inches), test the hypothesis $\mu = 0.80$ in. (the length indicated on the box) against the alternative $\mu \neq 0.80$ in. (Assume normality, choose $\alpha = 5\%$.)
9. A firm sells oil in cans containing 1000 g oil per can and is interested to know whether the mean weight differs significantly from 1000 g at the 5% level, in which case the filling machine has to be adjusted. Set up a hypothesis and an alternative and perform the test, assuming normality and using a sample of 20 fillings with mean 996 g and standard deviation 5 g.
10. If a sample of 50 tires of a certain kind has a mean life of 32 000 mi and a standard deviation of 4000 mi, can the manufacturer claim that the true mean life of such tires is greater than 30 000 mi? Set up and test a corresponding hypothesis at a 5% level, assuming normality.
11. If simultaneous measurements of electric voltage by two different types of voltmeter yield the differences (in volts) 0.8, 0.2, −0.3, 0.1, 0.0, 0.5, 0.7, 0.2, can we assert at the 5% level that there is no significant difference in the calibration of the two types of instruments? (Assume normality.)
12. If a standard medication cures about 70% of patients with a certain disease and a new medication cured 148 of the first 200 patients on whom it was tried, can we conclude that the new medication is better? (Choose $\alpha = 5\%$.)
13. Suppose that in the past the standard deviation of weights of certain 25.0-oz packages filled by a machine was 0.4 oz. Test the hypothesis H_0: $\sigma = 0.4$ against the alternative H_1: $\sigma > 0.4$ (an undesirable increase), using a sample of 10 packages with standard deviation 0.5 oz and assuming normality. (Choose $\alpha = 5\%$.)
14. Suppose that in operating battery-powered electrical equipment, it is less expensive to replace all batteries at fixed intervals than to replace each battery individually when it breaks down, provided the standard deviation of the lifetime is less than a certain limit, say, less than 5 hours. Set up and apply a suitable test, using a sample of 28 values of lifetimes with standard deviation $s = 3.5$ hours and assuming normality; choose $\alpha = 5\%$.
15. Brand A gasoline was used in 9 automobiles of the same model under identical conditions. The corresponding sample of 9 values (miles per gallon) had mean 20.2 and standard deviation 0.5. Under the same conditions, high-power brand B gasoline gave a sample of 10 values with mean 21.8 and standard deviation 0.6. Is the mileage of B significantly better than that of A? (Test at the 5% level; assume normality.)
16. The two samples 70, 80, 30, 70, 60, 80 and 140, 120, 130, 120, 120, 130, 120 are values of the differences of temperatures (°C) of iron at two stages of casting, taken from two different crucibles. Is the variance of the first population larger than that of the second? (Assume normality. Choose $\alpha = 5\%$.)
17. Using samples of sizes 10 and 16 with variances $s_x^2 = 50$ and $s_y^2 = 30$ and assuming normality of the corresponding populations, test the hypothesis H_0: $\sigma_x^2 = \sigma_y^2$ against the alternative $\sigma_x^2 > \sigma_y^2$. Choose $\alpha = 5\%$.
18. Assuming normality and equal variance and using independent samples with $n_1 = 9$, $\bar{x} = 12$, $s_x = 2$, $n_2 = 9$, $\bar{y} = 15$, $s_y = 2$, test H_0: $\mu_x = \mu_y$ against $\mu_x \neq \mu_y$; choose $\alpha = 5\%$.

19. Show that for a normal distribution the two types of errors in a test of a hypothesis H_0: $\mu = \mu_0$ against an alternative H_1: $\mu = \mu_1$ can be made as small as one pleases (not zero) by taking the sample sufficiently large.

20. CAS EXPERIMENT. Tests of Means and Variances. (a) Obtain 100 samples of size 10 each from the normal distribution with mean 100 and variance 25. For each sample test the hypothesis $\mu_0 = 100$ against the alternative $\mu_1 > 100$ at the level of $\alpha = 10\%$. Record the number of rejections of the hypothesis. Do the whole experiment once more and compare.

(b) Set up a similar experiment for the variance of a normal distribution and perform it 100 times.

25.5 Quality Control

The ideas on testing can be adapted and extended in various ways to serve basic practical needs in engineering and other fields. We show this in the remaining sections for some of the most important tasks solvable by statistical methods. As a first such area of problems, we discuss industrial quality control, a highly successful method used in various industries.

No production process is so perfect that all the products are completely alike. There is always a small variation that is caused by a great number of small, uncontrollable factors and must therefore be regarded as a chance variation. It is important to make sure that the products have required values (for example, length, strength, or whatever property may be essential in a particular case). For this purpose one makes a test of the hypothesis that the products have the required property, say, $\mu = \mu_0$, where μ_0 is a required value. If this is done after an entire lot has been produced (for example, a lot of 100 000 screws), the test will tell us how good or how bad the products are, but it it obviously too late to alter undesirable results. It is much better to test during the production run. This is done at regular intervals of time (for example, every hour or half-hour) and is called **quality control.** Each time a sample of the same size is taken, in practice 3 to 10 times. If the hypothesis is rejected, we stop the production and look for the cause of the trouble.

If we stop the production process even though it is progressing properly, we make a Type I error. If we do not stop the process even though something is not in order, we make a Type II error (see Sec. 25.4). The result of each test is marked in graphical form on what is called a **control chart.** This was proposed by W. A. Shewhart in 1924 and makes quality control particularly effective.

Control Chart for the Mean

An illustration and example of a control chart is given in the upper part of Fig. 536. This control chart for the mean shows the **lower control limit** LCL, the **center control line** CL, and the **upper control limit** UCL. The two **control limits** correspond to the critical values c_1 and c_2 in case (c) of Example 2 in Sec. 25.4. As soon as a sample mean falls outside the range between the control limits, we reject the hypothesis and assert that the production process is "out of control"; that is, we assert that there has been a shift in process level. Action is called for whenever a point exceeds the limits.

If we choose control limits that are too loose, we shall not detect process shifts. On the other hand, if we choose control limits that are too tight, we shall be unable to run the process because of frequent searches for nonexistent trouble. The usual significance level

is $\alpha = 1\%$. From Theorem 1 in Sec. 25.3 and Table A8 in App. 5 we see that in the case of the normal distribution the corresponding control limits for the mean are

$$\text{(1)} \qquad \text{LCL} = \mu_0 - 2.58\,\frac{\sigma}{\sqrt{n}}, \qquad \text{UCL} = \mu_0 + 2.58\,\frac{\sigma}{\sqrt{n}}.$$

Here σ is assumed to be known. If σ is unknown, we may compute the standard deviations of the first 20 or 30 samples and take their arithmetic mean as an approximation of σ. The broken line connecting the means in Fig. 536 is merely to display the results.

Additional, more subtle controls are often used in industry. For instance, one observes the motions of the sample means above and below the centerline, which should happen frequently. Accordingly, long runs (conventionally of length 7 or more) of means all above (or all below) the centerline could indicate trouble.

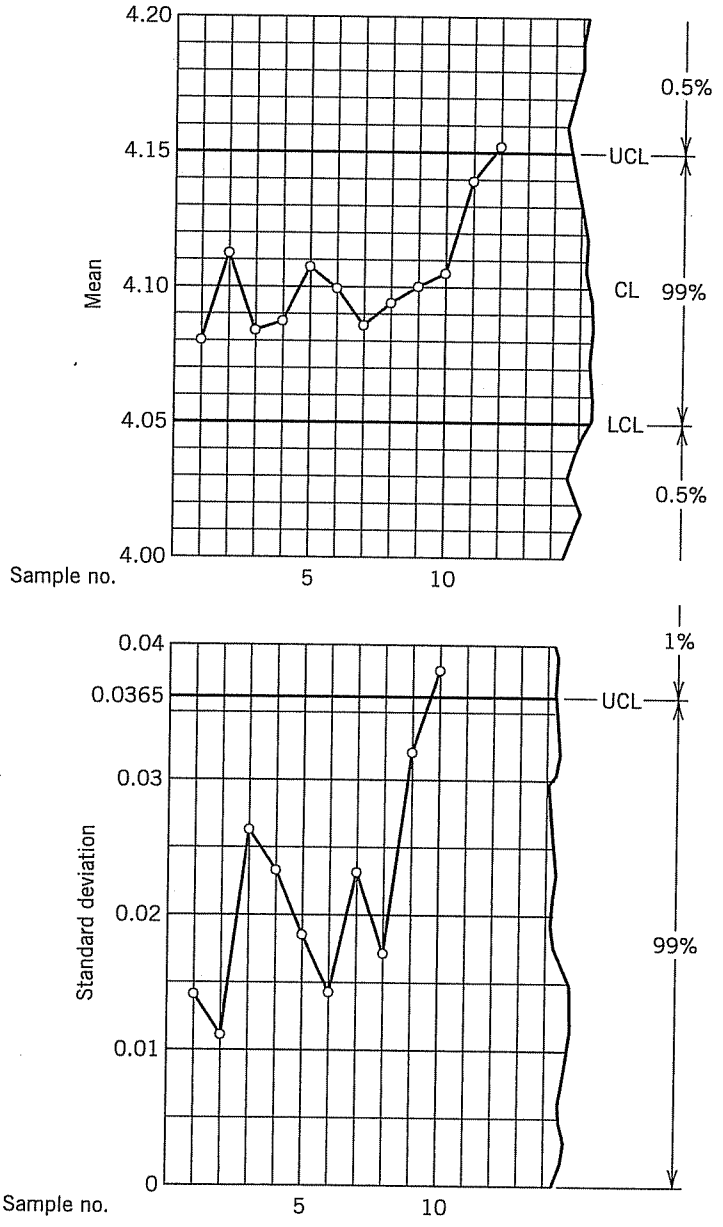


Fig. 536. Control charts for the mean (upper part of figure) and the standard deviation in the case of the samples on p. 1070

Table 25.5 Twelve Samples of Five Values Each (Diameter of Small Cylinders, Measured in Millimeters)

Sample Number	Sample Values					$\bar{x}$	s	R
1	4.06	4.08	4.08	4.08	4.10	4.080	0.014	0.04
2	4.10	4.10	4.12	4.12	4.12	4.112	0.011	0.02
3	4.06	4.06	4.08	4.10	4.12	4.084	0.026	0.06
4	4.06	4.08	4.08	4.10	4.12	4.088	0.023	0.06
5	4.08	4.10	4.12	4.12	4.12	4.108	0.018	0.04
6	4.08	4.10	4.10	4.10	4.12	4.100	0.014	0.04
7	4.06	4.08	4.08	4.10	4.12	4.088	0.023	0.06
8	4.08	4.08	4.10	4.10	4.12	4.096	0.017	0.04
9	4.06	4.08	4.10	4.12	4.14	4.100	0.032	0.08
10	4.06	4.08	4.10	4.12	4.16	4.104	0.038	0.10
11	4.12	4.14	4.14	4.14	4.16	4.140	0.014	0.04
12	4.14	4.14	4.16	4.16	4.16	4.152	0.011	0.02

Control Chart for the Variance

In addition to the mean, one often controls the variance, the standard deviation, or the range. To set up a control chart for the variance in the case of a normal distribution, we may employ the method in Example 4 of Sec. 25.4 for determining control limits. It is customary to use only one control limit, namely, an upper control limit. Now from Example 4 of Sec. 25.4 we have $S^2 = \sigma_0{}^2 Y/(n-1)$, where because of our normality assumption the random variable Y has a chi-square distribution with $n-1$ degrees of freedom. Hence the desired control limit is

$$\text{UCL} = \frac{\sigma^2 c}{n-1} \tag{2}$$

where c is obtained from the equation

$$P(Y > c) = \alpha, \qquad \text{that is,} \qquad P(Y \leqq c) = 1 - \alpha$$

and the table of the chi-square distribution (Table A10 in App. 5) with $n-1$ degrees of freedom (or from your CAS); here α (5% or 1%, say) is the probability that in a properly running process an observed value s^2 of S^2 is greater than the upper control limit.

If we wanted a control chart for the variance with both an upper control limit UCL and a lower control limit LCL, these limits would be

$$\text{LCL} = \frac{\sigma^2 c_1}{n-1} \qquad \text{and} \qquad \text{UCL} = \frac{\sigma^2 c_2}{n-1}. \tag{3}$$

where c_1 and c_2 are obtained from Table A10 with $n-1$ d.f. and the equations

$$P(Y \leqq c_1) = \frac{\alpha}{2} \qquad \text{and} \qquad P(Y \leqq c_2) = 1 - \frac{\alpha}{2}. \tag{4}$$

Control Chart for the Standard Deviation

To set up a control chart for the standard deviation, we need an upper control limit

$$\text{UCL} = \frac{\sigma\sqrt{c}}{\sqrt{n-1}} \tag{5}$$

obtained from (2). For example, in Table 25.5 we have $n = 5$. Assuming that the corresponding population is normal with standard deviation $\sigma = 0.02$ and choosing $\alpha = 1\ \%$, we obtain from the equation

$$P(Y \leqq c) = 1 - \alpha = 99\%$$

and Table A10 in App. 5 with 4 degrees of freedom the critical value $c = 13.28$ and from (5) the corresponding value

$$\text{UCL} = \frac{0.02\sqrt{13.28}}{\sqrt{4}} = 0.0365,$$

which is shown in the lower part of Fig. 536.

A control chart for the standard deviation with both an upper and a lower control limit is obtained from (3).

Control Chart for the Range

Instead of the variance or standard deviation, one often controls the **range** R (= largest sample value minus smallest sample value). It can be shown that in the case of the normal distribution, the standard deviation σ is proportional to the expectation of the random variable R^* for which R is an observed value, say, $\sigma = \lambda_n E(R^*)$, where the factor of proportionality λ_n depends on the sample size n and has the values

n	2	3	4	5	6	7	8	9	10
$\lambda_n = \sigma/E(R^*)$	0.89	0.59	0.49	0.43	0.40	0.37	0.35	0.34	0.32

n	12	14	16	18	20	30	40	50
$\lambda_n = \sigma/E(R^*)$	0.31	0.29	0.28	0.28	0.27	0.25	0.23	0.22

Since R depends on two sample values only, it gives less information about a sample than s does. Clearly, the larger the sample size n is, the more information we lose in using R instead of s. A practical rule is to use s when n is larger than 10.

PROBLEM SET 25.5

1. Suppose a machine for filling cans with lubricating oil is set so that it will generate fillings which form a normal population with mean 1 gal and standard deviation 0.03 gal. Set up a control chart of the type shown in Fig. 536 for controlling the mean (that is, find LCL and UCL), assuming that the sample size is 6.
2. **(Three-sigma control chart)** Show that in Prob. 1, the requirement of the significance level $\alpha = 0.3\%$ leads to $\text{LCL} = \mu - 3\sigma/\sqrt{n}$ and $\text{UCL} = \mu + 3\sigma/\sqrt{n}$, and find the corresponding numeric values.
3. What sample size should we choose in Prob. 1 if we want LCL and UCL somewhat closer together, say, $\text{UCL} - \text{LCL} = 0.05$, without changing the significance level?

4. How does the meaning of the control limits (1) change if we apply a control chart with these limits in the case of a population that is not normal?

5. How should we change the sample size in controlling the mean of a normal population if we want the difference

$$\text{UCL} - \text{LCL}$$

to decrease to half its original value?

6. What LCL and UCL should we use instead of (1) if instead of $\bar{x}$ we use the sum $x_1 + \cdots + x_n$ of the sample values? Determine these limits in the case of Fig. 536.

7. Ten samples of size 2 were taken from a production lot of bolts. The values (length in mm) are as shown. Assuming that the population is normal with mean 27.5 and variance 0.024 and using (1), set up a control chart for the mean and graph the sample means on the chart.

Sample No.	1	2	3	4	5	6	7	8	9	10
Length	27.4	27.4	27.5	27.3	27.9	27.6	27.6	27.8	27.5	27.3
	27.6	27.4	27.7	27.4	27.5	27.5	27.4	27.3	27.4	27.7

8. Graph the means of the following 10 samples (thickness of washers, coded values) on a control chart for means, assuming that the population is normal with mean 5 and standard deviation 1.55.

Time	8:00	8:30	9:00	9:30	10:00	10:30	11:00	11:30	12:00	12:30
Sample Values	3	3	5	7	7	4	5	6	5	5
	4	6	2	5	3	4	6	4	5	2
	8	6	5	4	6	3	4	6	6	5
	4	8	6	4	5	6	6	4	4	3

9. Graph the ranges of the samples in Prob. 8 on a control chart for ranges.

10. What effect on UCL − LCL does it have if we double the sample size? If we switch from $\alpha = 1\%$ to $\alpha = 5\%$?

11. Since the presence of a point outside control limits for the mean indicates trouble ("the process is out of control"), how often would we be making the mistake of looking for nonexistent trouble if we used (a) 1-sigma limits, (b) 2-sigma limits? (Assume normality.)

12. Graph $\lambda_n = \sigma/E(R^*)$ as a function of n. Why is λ_n a monotone decreasing function of n?

13. **(Number of defectives)** Find formulas for the UCL, CL, and LCL (corresponding to 3σ-limits) in the case of a control chart for the number of defectives, assuming that in a state of statistical control the fraction of defectives is p.

14. How would progressive tool wear in an automatic lathe operation be indicated by a control chart of the mean? Answer the same question for a sudden change in the position of the tool in that operation.

15. **(Number of defects per unit)** A so-called *c-chart* or *defects-per-unit chart* is used for the control of the number X of defects per unit (for instance, the number of defects per 10 meters of paper, the number of missing rivets in an airplane wing, etc.) (a) Set up formulas for CL and LCL, UCL corresponding to

$$\mu \pm 3\sigma,$$

assuming that X has a Poisson distribution. (b) Compute CL, LCL, and UCL in a control process of the number of imperfections in sheet glass; assume that this number is 2.5 per sheet on the average when the process is under control.

16. **(Attribute control charts).** Twenty samples of size 100 were taken from a production of containers. The numbers of defectives (leaking containers) in those samples (in the order observed) were

3 7 6 1 4 5 4 9 7 0 5 6 13 4

9 0 2 1 12 8.

From previous experience it was known that the average fraction defective is $p = 5\%$ provided that the process of production is running properly. Using the binomial distribution, set up a *fraction defective chart* (also called a ***p*-chart**), that is, choose the LCL = 0 and determine the UCL for the fraction defective (in percent) by the use of 3-sigma limits, where σ^2 is the variance of the random variable

$$\overline{X} = \textit{Fraction defective in a sample of size } 100.$$

Is the process under control?

17. **CAS PROJECT. Control Charts. (a)** Obtain 100 samples of 4 values each from the normal distribution with mean 8.0 and variance 0.16 and their means, variances, and ranges.

(b) Use these samples for making up a control chart for the mean.

(c) Use them on a control chart for the standard deviation.

(d) Make up a control chart for the range.

(e) Describe quantitative properties of the samples that you can see from those charts (e.g., whether the corresponding process is under control, whether the quantities observed vary randomly, etc.).

25.6 Acceptance Sampling

Acceptance sampling is usually done when products leave the factory (or in some cases even within the factory). The standard situation in acceptance sampling is that a **producer** supplies to a **consumer** (a buyer or wholesaler) a lot of N items (a carton of screws, for instance). The decision to **accept** or **reject** the lot is made by determining the number x of **defectives** (= defective items) in a sample of size n from the lot. The lot is accepted if $x \leqq c$, where c is called the **acceptance number,** giving the allowable number of defectives. If $x > c$, the consumer rejects the lot. Clearly, producer and consumer must agree on a certain **sampling plan** giving n and c.

From the hypergeometric distribution we see that the event A: *"Accept the lot"* has probability (see Sec. 24.7)

$$P(A) = P(X \leqq c) = \sum_{x=0}^{c} \binom{M}{x} \binom{N-M}{n-x} \Big/ \binom{N}{n} \tag{1}$$

where M is the number of defectives in a lot of N items. In terms of the **fraction defective** $\theta = M/N$ we can write (1) as

$$P(A;\ \theta) = \sum_{x=0}^{c} \binom{N\theta}{x} \binom{N-N\theta}{n-x} \Big/ \binom{N}{n}. \tag{2}$$

$P(A;\ \theta)$ can assume $n + 1$ values corresponding to $\theta = 0,\ 1/N,\ 2/N,\ \cdots,\ N/N;$ here, n and c are fixed. A monotone smooth curve through these points is called the **operating characteristic curve (OC curve)** of the sampling plan considered.

EXAMPLE 1 **Sampling Plan**

Suppose that certain tool bits are packaged 20 to a box, and the following sampling plan is used. A sample of two tool bits is drawn, and the corresponding box is accepted if and only if both bits in the sample are good. In this case, $N = 20$, $n = 2$, $c = 0$, and (2) takes the form (a factor 2 drops out)

$$P(A;\ \theta) = \binom{20\theta}{0} \binom{20 - 20\theta}{2} \Big/ \binom{20}{2}$$

$$= \frac{(20 - 20\theta)(19 - 20\theta)}{380}.$$

The values of $P(A,\ \theta)$ for $\theta = 0,\ 1/20,\ 2/20,\ \cdots,\ 20/20$ and the resulting OC curve are shown in Fig. 537 on p. 1074. (Verify!) ■

In most practical cases θ will be small (less than 10%). Then if we take small samples compared to N, we can approximate (2) by the Poisson distribution (Sec. 24.7); thus

$$P(A;\ \theta) \sim e^{-\mu} \sum_{x=0}^{c} \frac{\mu^x}{x!} \qquad (\mu = n\theta). \tag{3}$$

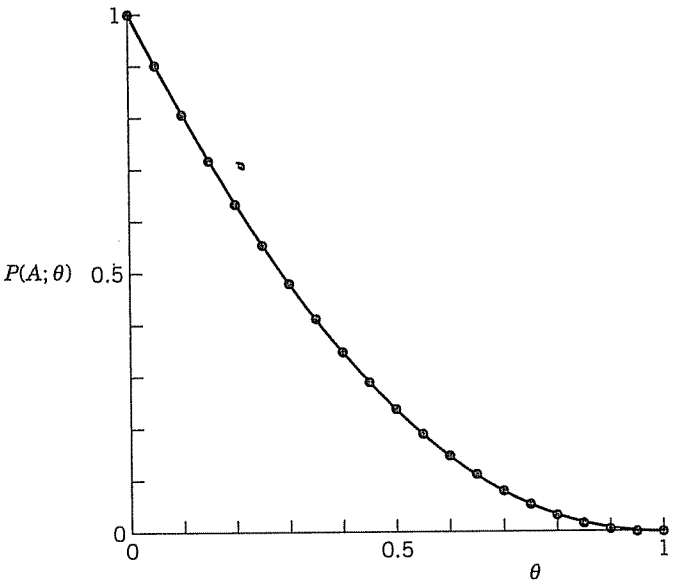


Fig. 537. OC curve of the sampling plan with $n = 2$ and $c = 0$ for lots of size $N = 20$

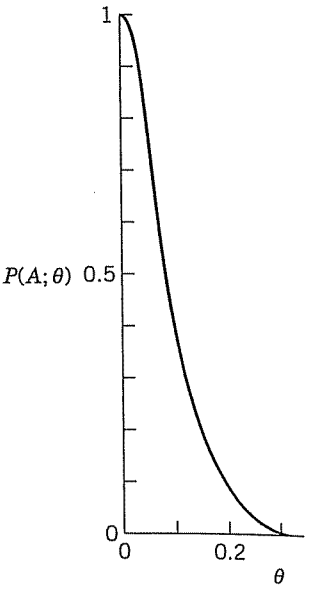


Fig. 538. OC curve in Example 2

EXAMPLE 2 Sampling Plan. Poisson Distribution

Suppose that for large lots the following sampling plan is used. A sample of size $n = 20$ is taken. If it contains not more than one defective, the lot is accepted. If the sample contains two or more defectives, the lot is rejected. In this plan, we obtain from (3)

$$P(A;\ \theta) \sim e^{-20\,\theta}(1 + 20\,\theta).$$

The corresponding OC curve is shown in Fig. 538. ■

Errors in Acceptance Sampling

We show how acceptance sampling fits into general test theory (Sec. 25.4) and what this means from a practical point of view. The producer wants the probability α of rejecting an **acceptable lot** (a lot for which θ does not exceed a certain number θ_0 on which the two parties agree) to be small. θ_0 is called the **acceptable quality level** (AQL). Similarly,

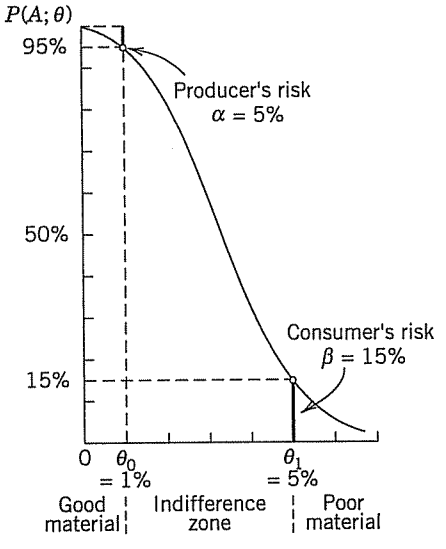


Fig. 539. OC curve, producer's and consumer's risks

the consumer (the buyer) wants the probability β of accepting an **unacceptable lot** (a lot for which θ is greater than or equal to some θ_1) to be small. θ_1 is called the **lot tolerance percent defective** (LTPD) or the **rejectable quality level** (RQL). α is called **producer's risk.** It corresponds to a Type I error in Sec. 25.4. β is called **consumer's risk** and corresponds to a Type II error. Figure 539 shows an example. We see that the points $(\theta_0, 1 - \alpha)$ and (θ_1, β) lie on the OC curve. It can be shown that for large lots we can choose θ_0, θ_1 $(> \theta_0)$, α, β and then determine n and c such that the OC curve runs very close to those prescribed points. Table 25.6 shows the analogy between acceptance sampling and hypothesis testing in Sec. 25.4.

Table 25.6 Acceptance Sampling and Hypothesis Testing

Acceptance Sampling	Hypothesis Testing
Acceptable quality level (AQL) $\theta = \theta_0$	Hypothesis $\theta = \theta_0$
Lot tolerance percent defectives (LTPD) $\theta = \theta_1$	Alternative $\theta = \theta_1$
Allowable number of defectives c	Critical value c
Producer's risk α of rejecting a lot with $\theta \leqq \theta_0$	Probability α of making a Type I error (significance level)
Consumer's risk β of accepting a lot with $\theta \geqq \theta_1$	Probability β of making a Type II error

Rectification

Rectification of a *rejected* lot means that the lot is inspected item by item and all defectives are removed and replaced by nondefective items. (This may be too expensive if the lot is cheap; in this case the lot may be sold at a cut-rate price or scrapped.) If a production turns out $100\theta\%$ defectives, then in K lots of size N each, $KN\theta$ of the KN items are defectives. Now $KP(A;\ \theta)$ of these lots are accepted. These contain $KPN\theta$ defectives, whereas the rejected and rectified lots contain no defectives, because of the rectification. Hence after the rectification the fraction defective in all K lots equals $KPN\theta/KN$. This is called the **average outgoing quality** (AOQ); thus

(4) $$\text{AOQ}(\theta) = \theta P(A;\ \theta).$$

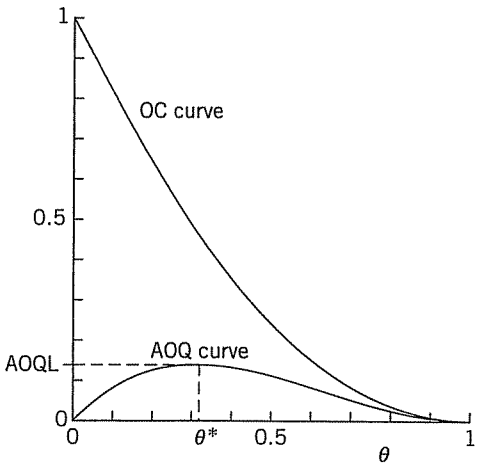


Fig. 540. OC curve and AOQ curve for the sampling plan in Fig. 537

Figure 540 on p. 1075 shows an example. Since AOQ(0) = 0 and $P(A;\ 1) = 0$, the AOQ curve has a maximum at some $\theta = \theta^*$, giving the **average outgoing quality limit** (AOQL). This is the worst average quality that may be expected to be accepted under rectification.

PROBLEM SET 25.6

1. Lots of knives are inspected by a sampling plan that uses a sample of size 20 and the acceptance number $c = 1$. What are probabilities of accepting a lot with 1%, 2%, 10% defectives (dull blades)? Use Table A6 in App. 5. Graph the OC curve.
2. What happens in Prob. 1 if the sample size is increased to 50? First guess. Then calculate. Graph the OC curve and compare.
3. How will the probabilities in Prob. 1 with $n = 20$ change (up or down) if we decrease c to zero? First guess.
4. What are the producer's and consumer's risks in Prob. 1 if the AQL is 1.5% and the RQL is 7.5%?
5. Large lots of batteries are inspected according to the following plan. $n = 30$ batteries are randomly drawn from a lot and tested. If this sample contains at most $c = 1$ defective battery, the lot is accepted. Otherwise it is rejected. Graph the OC curve of the plan, using the Poisson approximation.
6. Graph the AOQ curve in Prob. 5. Determine the AOQL, assuming that rectification is applied.
7. Do the work required in Prob. 5 if $n = 50$ and $c = 0$.
8. Find the binomial approximation of the hypergeometric distribution in Example 1 and compare the approximate and the accurate values.
9. In Example 1, what are the producer's and consumer's risks if the AQL is 0.1 and the RQL is 0.6?
10. Calculate $P(A;\ \theta)$ in Example 1 if the sample size is increased from $n = 2$ to $n = 3$, the other data remaining as before. Compute $P(A;\ 0.10)$ and $P(A;\ 0.20)$ and compare with Example 1.
11. Samples of 5 screws are drawn from a lot with fraction defective θ. The lot is accepted if the sample contains (a) no defective screws, (b) at most 1 defective screw. Using the binomial distribution, find, graph, and compare the OC curves.
12. Find the risks in the single sampling plan with $n = 5$ and $c = 0$, assuming that the AQL is $\theta_0 = 1\%$ and the RQL is $\theta_1 = 15\%$.
13. Why is it impossible for an OC curve to have a vertical portion separating good from poor quality?
14. If in a single sampling plan for large lots of spark plugs, the sample size is 100 and we want the AQL to be 5% and the producer's risk 2%, what acceptance number c should we choose? (Use the normal approximation.)
15. What is the consumer's risk in Prob. 14 if we want the RQL to be 12%?
16. Graph and compare sampling plans with $c = 1$ and increasing values of n, say, $n = 2, 3, 4$. (Use the binomial distribution.)
17. Samples of 3 fuses are drawn from lots and a lot is accepted if in the corresponding sample we find no more than 1 defective fuse. Criticize this sampling plan. In particular, find the probability of accepting a lot that is 50% defective. (Use the binomial distribution.)
18. Graph the OC curve and the AOQ curve for the single sampling plan for large lots with $n = 5$ and $c = 0$, and find the AOQL.

25.7 Goodness of Fit. χ^2-Test

To test for **goodness of fit** means that we wish to test that a certain function $F(x)$ is the distribution function of a distribution from which we have a sample $x_1, \cdots, x_n$. Then we test whether the **sample distribution function** $\tilde{F}(x)$ defined by

$$\tilde{F}(x) = \textit{Sum of the relative frequencies of all sample values } x_j \textit{ not exceeding } x$$

fits $F(x)$ "sufficiently well." If this is so, we shall accept the hypothesis that $F(x)$ is the distribution function of the population; if not, we shall reject the hypothesis.

This test is of considerable practical importance, and it differs in character from the tests for parameters (μ, σ^2, etc.) considered so far.

To test in that fashion, we have to know how much $\tilde{F}(x)$ can differ from $F(x)$ if the hypothesis is true. Hence we must first introduce a quantity that measures the deviation of $\tilde{F}(x)$ from $F(x)$, and we must know the probability distribution of this quantity under the assumption that the hypothesis is true. Then we proceed as follows. We determine a number c such that if the hypothesis is true, a deviation greater than c has a small preassigned probability. If, nevertheless, a deviation greater than c occurs, we have reason to doubt that the hypothesis is true and we reject it. On the other hand, if the deviation does not exceed c, so that $\tilde{F}(x)$ approximates $F(x)$ sufficiently well, we accept the hypothesis. Of course, if we accept the hypothesis, this means that we have insufficient evidence to reject it, and this does not exclude the possibility that there are other functions that would not be rejected in the test. In this respect the situation is quite similar to that in Sec. 25.4.

Table 25.7 shows a test of that type, which was introduced by R. A. Fisher. This test is justified by the fact that if the hypothesis is true, then ${\chi_0}^2$ is an observed value of a random variable whose distribution function approaches that of the chi-square distribution with $K - 1$ degrees of freedom (or $K - r - 1$ degrees of freedom if r parameters are estimated) as n approaches infinity. The requirement that at least five sample values lie in each interval in Table 25.7 results from the fact that for finite n that random variable has only *approximately* a chi-square distribution. A proof can be found in Ref. [G3] listed in App. 1. If the sample is so small that the requirement cannot be satisfied, one may continue with the test, but then use the result with caution.

Table 25.7 Chi-square Test for the Hypothesis That $F(x)$ is the Distribution Function of a Population from Which a Sample $x_1, \cdots, x_n$ is Taken

Step 1. Subdivide the x-axis into K intervals $I_1, I_2, \cdots, I_K$ such that each interval contains at least 5 values of the given sample $x_1, \cdots, x_n$. Determine the number b_j of sample values in the interval I_j, where $j = 1, \cdots, K$. If a sample value lies at a common boundary point of two intervals, add 0.5 to each of the two corresponding b_j.

Step 2. Using $F(x)$, compute the probability p_j that the random variable X under consideration assumes any value in the interval I_j, where $j = 1, \cdots, K$. Compute

$$e_j = np_j.$$

(This is the number of sample values theoretically expected in I_j if the hypothesis is true.)

Step 3. Compute the deviation

(1)
$$\chi_0^2 = \sum_{j=1}^{K} \frac{(b_j - e_j)^2}{e_j}.$$

Step 4. Choose a significance level (5%, 1%, or the like).

Step 5. Determine the solution c of the equation

$$P(\chi^2 \leqq c) = 1 - \alpha$$

from the table of the chi-sqare distribution with $K - 1$ degrees of freedom (Table A10 in App. 5). If r parameters of $F(x)$ are unknown and their maximum likelihood estimates (Sec. 25.2) are used, then use $K - r - 1$ degrees of freedom (instead of $K - 1$). If ${\chi_0}^2 \leqq c$, accept the hypothesis. If ${\chi_0}^2 > c$, reject the hypothesis.

Table 25.8 Sample of 100 Values of the Splitting Tensile Strength (lb/in.^{2}) of Concrete Cylinders

320	380	340	410	380	340	360	350	320	370
350	340	350	360	370	350	380	370	300	420
370	390	390	440	330	390	330	360	400	370
320	350	360	340	340	350	350	390	380	340
400	360	350	390	400	350	360	340	370	420
420	400	350	370	330	320	390	380	400	370
390	330	360	380	350	330	360	300	360	360
360	390	350	370	370	350	390	370	370	340
370	400	360	350	380	380	360	340	330	370
340	360	390	400	370	410	360	400	340	360

D. L. IVEY, Splitting tensile tests on structural lightweight aggregate concrete. Texas Transportation Institute, College Station, Texas.

EXAMPLE 1 Test of Normality

Test whether the population from which the sample in Table 25.8 was taken is normal.

Solution. Table 25.8 shows the values (column by column) in the order obtained in the experiment. Table 25.9 gives the frequency distribution and Fig. 541 the histogram. It is hard to guess the outcome of the test—does the histogram resemble a normal density curve sufficiently well or not?

The maximum likelihood estimates for μ and σ^2 are $\hat{\mu} = \bar{x} = 364.7$ and $\tilde{\sigma}^2 = 712.9$. The computation in Table 25.10 yields $\chi_0^2 = 2.942$. It is very interesting that the interval $375 \cdots 385$ contributes over 50% of χ_0^2. From the histogram we see that the corresponding frequency looks much too small. The second largest contribution comes from $395 \cdots 405$, and the histogram shows that the frequency seems somewhat too large, which is perhaps not obvious from inspection.

Table 25.9 Frequency Table of the Sample in Table 25.8

1 Tensile Strength x [lb/in.^{2}]	2 Absolute Frequency	3 Relative Frequency $\tilde{f}(x)$	4 Cumulative Absolute Frequency	5 Cumulative Relative Frequency $\tilde{F}(x)$
300	2	0.02	2	0.02
310	0	0.00	2	0.02
320	4	0.04	6	0.06
330	6	0.06	12	0.12
340	11	0.11	23	0.23
350	14	0.14	37	0.37
360	16	0.16	53	0.53
370	15	0.15	68	0.68
380	8	0.08	76	0.76
390	10	0.10	86	0.86
400	8	0.08	94	0.94
410	2	0.02	96	0.96
420	3	0.03	99	0.99
430	0	0.00	99	0.99
440	1	0.01	100	1.00

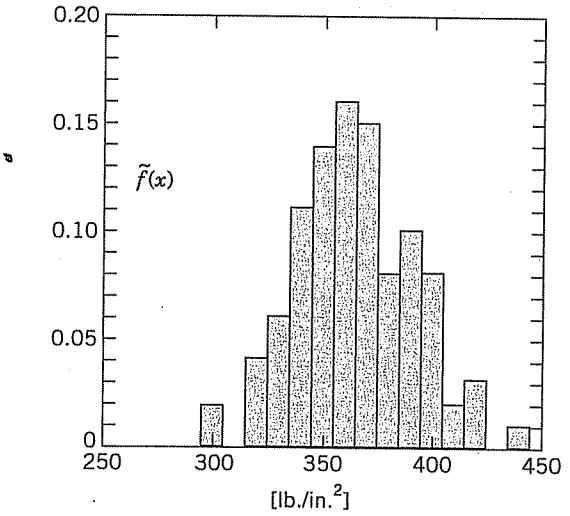


Fig. 541. Frequency histogram of the sample in Table 25.8

We choose $\alpha = 5\%$. Since $K = 10$ and we estimated $r = 2$ parameters we have to use Table A10 in App. 5 with $K - r - 1 = 7$ degrees of freedom. We find $c = 14.07$ as the solution of $P(\chi^2 \leqq c) = 95\%$. Since $\chi_0^2 < c$, we accept the hypothesis that the population is normal. ▩

Table 25.10 Computations in Example 1

x_j	$\frac{x_j - 364.7}{26.7}$	$\Phi\left(\frac{x_j - 364.7}{26.7}\right)$	e_j	b_j	Term in (1)
$-\infty \cdots 325$	$-\infty \cdots -1.49$	$0.0000 \cdots 0.0681$	6.81	6	0.096
$325 \cdots 335$	$-1.49 \cdots -1.11$	$0.0681 \cdots 0.1335$	6.54	6	0.045
$335 \cdots 345$	$-1.11 \cdots -0.74$	$0.1335 \cdots 0.2296$	9.61	11	0.201
$345 \cdots 355$	$-0.74 \cdots -0.36$	$0.2296 \cdots 0.3594$	12.98	14	0.080
$355 \cdots 365$	$-0.36 \cdots 0.01$	$0.3594 \cdots 0.4960$	13.66	16	0.401
$365 \cdots 375$	$0.01 \cdots 0.39$	$0.4960 \cdots 0.6517$	15.57	15	0.021
$375 \cdots 385$	$0.39 \cdots 0.76$	$0.6517 \cdots 0.7764$	12.47	8	1.602
$385 \cdots 395$	$0.76 \cdots 1.13$	$0.7764 \cdots 0.8708$	9.44	10	0.033
$395 \cdots 405$	$1.13 \cdots 1.51$	$0.8708 \cdots 0.9345$	6.37	8	0.417
$405 \cdots \infty$	$1.51 \cdots \infty$	$0.9345 \cdots 1.0000$	6.55	6	0.046

$\chi_0^2 = 2.942$

PROBLEM SET 25.7

1. If 100 flips of a coin result in 30 heads and 70 tails, can we assert on the 5% level that the coin is fair?
2. If in 10 flips of a coin we get the same ratio as in Prob. 1 (3 heads and 7 tails), is the conclusion the same as in Prob. 1? First conjecture, then compute.
3. What would be the smallest number of heads in Prob. 1 under which the hypothesis "Fair coin" is still accepted (with $\alpha = 5\%$)?
4. If in rolling a die 180 times we get 39, 22, 41, 26, 20, 32, can we claim on the 5% level that the die is fair?
5. Solve Prob. 4 if the sample is 25, 31, 33, 27, 29, 35.
6. A manufacturer claims that in a process of producing kitchen knives, only 2.5% of the knives are dull. Test the claim against the alternative that more than 2.5% of the knives are dull, using a sample of 400 knives containing 17 dull ones. (Use $\alpha = 5\%$.)
7. Between 1 P.M. and 2 P.M. on five consecutive days (Monday through Friday) a certain service station has 92, 60, 66, 62, and 90 customers, respectively. Test the hypothesis that the expected number of customers during that hour is the same on those days. (Use $\alpha = 5\%$.)

8. Test for normality at the 1% level using a sample of $n = 79$ (rounded) values x (tensile strength [kg/mm^2] of steel sheets of 0.3 mm thickness). $a = a(x)$ = absolute frequency. (Take the first two values together, also the last three, to get $K = 5$.)

x	57	58	59	60	61	62	63	64
a	4	10	17	27	8	9	3	1

9. In a sample of 100 patients having a certain disease 45 are men and 55 women. Does this support the claim that the disease is equally common among men and women? Choose $\alpha = 5\%$.

10. In Prob. 9 find the smallest number (>50) of women that leads to the rejection of the hypothesis on the levels 5%, 1%, 0.5%.

11. Verify the calculations in Example 1 of the text.

12. Does the random variable X = *Number of accidents per week in a certain foundry* have a Poisson distribution if within 50 weeks, 33 were accident-free, 1 accident occurred in 11 of the 50 weeks, 2 in 6 of the weeks and more than 2 accidents in no week? (Choose $\alpha = 5\%$.)

13. Using the given sample, test that the corresponding population has a Poisson distribution. x is the number of alpha particles per 7.5-sec intervals observed by E. Rutherford and H. Geiger in one of their classical experiments in 1910, and $a(x)$ is the absolute frequency (= number of time periods during which exactly x particles were observed). (Use $\alpha = 5\%$.)

x	0	1	2	3	4	5	6
a	57	203	383	525	532	408	273
x	7	8	9	10	11	12	≧13
a	139	45	27	10	4	2	0

14. Can we assert that the traffic on the three lanes of an expressway (in one direction) is about the same on each lane if a count gives 910, 850, 720 cars on the right, middle, and left lanes, respectively, during a particular time interval? (Use $\alpha = 5\%$.)

15. If it is known that 25% of certain steel rods produced by a standard process will break when subjected to a load of 5000 lb, can we claim that a new process yields the same breakage rate if we find that in a sample of 80 rods produced by the new process, 27 rods broke when subjected to that load? (Use $\alpha = 5\%$.)

16. Three samples of 200 rivets each were taken from a large production of each of three machines. The numbers of defective rivets in the samples were 7, 8, and 12. Is this difference significant? (Use $\alpha = 5\%$.)

17. In a table of properly rounded function values, even and odd last decimals should appear about equally often. Test this for the 90 values of $J_1(x)$ in Table A1 in App. 5.

18. Are the 5 tellers in a certain bank equally time-efficient if during the same time interval on a certain day they serve 120, 95, 110, 108, 102 customers? (Use $\alpha = 5\%$.)

19. CAS EXPERIMENT. Random Number Generator. Check your generator experimentally by imitating results of n trials of rolling a fair die, with a convenient n (e.g., 60 or 300 or the like). Do this many times and see whether you can notice any "nonrandomness" features, for example, too few Sixes, too many even numbers, etc., or whether your generator seems to work properly. Design and perform other kinds of checks.

20. TEAM PROJECT. Difficulty with Random Selection. 77 students were asked to choose 3 of the integers 11, 12, 13, · · · , 30 completely arbitrarily. The amazing result was as follows.

Number	11	12	13	14	15	16	17	18	19	20
Frequ.	11	10	20	8	13	9	21	9	16	8
Number	21	22	23	24	25	26	27	28	29	30
Frequ.	12	8	15	10	10	9	12	8	13	9

If the selection were completely random, the following hypotheses should be true.

(a) The 20 numbers are equally likely.

(b) The 10 even numbers together are as likely as the 10 odd numbers together.

(c) The 6 prime numbers together have probability 0.3 and the 14 other numbers together have probability 0.7.

Test these hypotheses, using $\alpha = 5\%$. Design further experiments that illustrate the difficulties of random selection.

25.8 Nonparametric Tests

Nonparametric tests, also called **distribution-free tests,** are valid for any distribution. Hence they are used in cases when the kind of distribution is unknown, or is known but such that no tests specifically designed for it are available. In this section we shall explain the basic idea of these tests, which are based on **"order statistics"** and are rather simple.

If there is a choice, then tests designed for a specific distribution generally give better results than do nonparametric tests. For instance, this applies to the tests in Sec. 25.4 for the normal distribution.

We shall discuss two tests in terms of typical examples. In deriving the distributions used in the test, it is essential that the distributions from which we sample are continuous. (Nonparametric tests can also be derived for discrete distributions, but this is slightly more complicated.)

EXAMPLE 1 **Sign Test for the Median**

A **median** of the population is a solution $x = \tilde{\mu}$ of the equation $F(x) = 0.5$, where F is the distribution function of the population.

Suppose that eight radio operators were tested, first in rooms without air-conditioning and then in air-conditioned rooms over the same period of time, and the difference of errors (unconditioned minus conditioned) were

$$9 \quad 4 \quad 0 \quad 6 \quad 4 \quad 0 \quad 7 \quad 11.$$

Test the hypothesis $\tilde{\mu} = 0$ (that is, air-conditioning has no effect) against the alternative $\tilde{\mu} > 0$ (that is, inferior performance in unconditioned rooms).

Solution. We choose the significance level $\alpha = 5\%$. If the hypothesis is true, the probability p of a positive difference is the same as that of a negative difference. Hence in this case, $p = 0.5$, and the random variable

$$X = \textit{Number of positive values among } n \textit{ values}$$

has a binomial distribution with $p = 0.5$. Our sample has eight values. We omit the values 0, which do not contribute to the decision. Then six values are left, all of which are positive. Since

$$P(X = 6) = \binom{6}{6} (0.5)^6(0.5)^0$$
$$= 0.0156$$
$$= 1.56\%$$

we do have observed an event whose probability is very small if the hypothesis is true; in fact $1.56\% < \alpha = 5\%$. Hence we assert that the alternative $\tilde{\mu} > 0$ is true. That is, the number of errors made in unconditioned rooms is significantly higher, so that installation of air conditioning should be considered. ■

EXAMPLE 2 **Test for Arbitrary Trend**

A certain machine is used for cutting lengths of wire. Five successive pieces had the lengths

$$29 \quad 31 \quad 28 \quad 30 \quad 32.$$

Using this sample, test the hypothesis that there is **no trend,** that is, the machine does not have the tendency to produce longer and longer pieces or shorter and shorter pieces. Assume that the type of machine suggests the alternative that there is *positive trend,* that is, there is the tendency of successive pieces to get longer.

Solution. We count the number of **transpositions** in the sample, that is, the number of times a larger value precedes a smaller value:

29 precedes 28 (1 transposition),

31 precedes 28 and 30 (2 transpositions).

The remaining three sample values follow in ascending order. Hence in the sample there are $1 + 2 = 3$ transpositions. We now consider the random variable

$$T = \textit{Number of transpositions.}$$

If the hypothesis is true (no trend), then each of the $5! = 120$ permutations of five elements 1 2 3 4 5 has the same probability (1/120). We arrange these permutations according to their number of transpositions:

$T = 0$	$T = 1$	$T = 2$	$T = 3$
1 2 3 4 5	1 2 3 5 4	1 2 4 5 3	1 2 5 4 3
	1 2 4 3 5	1 2 5 3 4	1 3 4 5 2
	1 3 2 4 5	1 3 2 5 4	1 3 5 2 4
	2 1 3 4 5	1 3 4 2 5	1 4 2 5 3
		1 4 2 3 5	1 4 3 2 5
		2 1 3 5 4	1 5 2 3 4
		2 1 4 3 5	2 1 4 5 3
		2 3 1 4 5	2 1 5 3 4 etc.
		3 1 2 4 5	2 3 1 5 4
			2 3 4 1 5
			2 4 1 3 5
			3 1 2 5 4
			3 1 4 2 5
			3 2 1 4 5
			4 1 2 3 5

From this we obtain

$$P(T \leqq 3) = \tfrac{1}{120} + \tfrac{4}{120} + \tfrac{9}{120} + \tfrac{15}{120} = \tfrac{29}{120} = 24\%.$$

We accept the hypothesis because we have observed an event that has a relatively large probability (certainly much more than 5%) if the hypothesis is true.

Values of the distribution function of T in the case of no trend are shown in Table A12, App. 5. For instance, if $n = 3$, then $F(0) = 0.167$, $F(1) = 0.500$, $F(2) = 1 - 0.167$. If $n = 4$, then $F(0) = 0.042$, $F(1) = 0.167$, $F(2) = 0.375$, $F(3) = 1 - 0.375$, $F(4) = 1 - 0.167$, and so on.

Our method and those values refer to *continuous* distributions. Theoretically, we may then expect that all the values of a sample are different. Practically, some sample values may still be equal, because of rounding: If m values are equal, add $m(m - 1)/4$ (= mean value of the transpositions in the case of the permutations of m elements), that is, $\frac{1}{2}$ for each pair of equal values, $\frac{3}{2}$ for each triple, etc. ■

PROBLEM SET 25.8

1. What would change in Example 1, had we observed only 5 positive values? Only 4?

2. Does a process of producing plastic pipes of length $\mu = 2$ meters need adjustment if in a sample, 4 pipes have the exact length and 15 are shorter and 3 longer than 2 meters? (Use the normal approximation of the binomial distribution.)

3. Do the computations in Prob. 2 without the use of the DeMoivre–Laplace limit theorem (in Sec. 24.8).

4. Test whether a thermostatic switch is properly set to 20°C against the alternative that its setting is too low. Use a sample of 9 values, 8 of which are less than 20°C and 1 is greater than 20°C.

5. Are air filters of type A better than type B filters if in 10 trials, A gave cleaner air than B in 7 cases, B gave cleaner air than A in 1 case, whereas in 2 of the trials the results for A and B were practically the same?

6. In a clinical experiment, each of 10 patients were given two different sedatives A and B. The following table shows the effect (increase of sleeping time, measured in hours). Using the sign test, find out whether the difference is significant.

A	1.9	0.8	1.1	0.1	−0.1	4.4	5.5	1.6	4.6	3.4
B	0.7	−1.6	−0.2	−1.2	−0.1	3.4	3.7	0.8	0.0	2.0
Difference	1.2	2.4	1.3	1.3	0.0	1.0	1.8	0.8	4.6	1.4

7. Assuming that the populations corresponding to the samples in Prob. 6 are normal, apply a suitable test for the normal distribution.

8. Thirty new employees were grouped into 15 pairs of similar intelligence and experience and were then instructed in data processing by an old method (A) applied to one (randomly selected) person of each pair, and by a new presumably better method (B) applied to

the other person of each pair. Test for equality of methods against the alternative that (B) is better than (A), using the following scores obtained after the end of the training period.

A	60	70	80	85	75	40	70	45	95	80	90	60	80	75	65
B	65	85	85	80	95	65	100	60	90	85	100	75	90	60	80

9. Assuming normality, solve Prob. 8 by a suitable test from Sec. 25.4.

10. Set up a sign test for the lower quartile q_{25} (defined by the condition $F(q_{25}) = 0.25$).

11. How would you proceed in the sign test if the hypothesis is $\tilde{\mu} = \tilde{\mu}_0$ (any number) instead of $\tilde{\mu} = 0$?

12. Check the table in Example 2 of the text.

13. Apply the test in Example 2 to the following data (x = disulfide content of a certain type of wool, measured in percent of the content in unreduced fibers; y = saturation water content of the wool, measured in percent). Test for no trend against negative trend.

x	10	15	30	40	50	55	80	100
y	50	46	43	42	36	39	37	33

14. Test the hypothesis that for a certain type of voltmeter, readings are independent of temperature T [°C] against the alternative that they tend to increase with T. Use a sample of values obtained by applying a constant voltage:

Temperature T [°C]	10	20	30	40	50
Reading V [volts]	99.5	101.1	100.4	100.8	101.6

15. In a swine-feeding experiment, the following gains in weight [kg] of 10 animals (ordered according to increasing amounts of food given per day) were recorded:

20 17 19 18 23 16 25 28 24 22.

Test for no trend against positive trend.

16. Apply the test explained in Example 2 to the following data (x = diastolic blood pressure [mm Hg], y = weight of heart [in grams] of 10 patients who died of cerebral hemorrhage).

x	121	120	95	123	140	112	92	100	102	91
y	521	465	352	455	490	388	301	395	375	418

17. Does an increase in temperature cause an increase of the yield of a chemical reaction from which the following sample was taken?

Temperature [°C]	10	20	30	40	60	80
Yield [kg/min]	0.6	1.1	0.9	1.6	1.2	2.0

18. Does the amount of fertilizer increase the yield of wheat X [kg/plot]? Use a sample of values ordered according to increasing amounts of fertilizer:

41.4 43.3 39.6 43.0 44.1 45.6 44.5 46.7.

25.9 Regression. Fitting Straight Lines. Correlation

So far we were concerned with random experiments in which we observed a single quantity (random variable) and got samples whose values were single numbers. In this section we discuss experiments in which we observe or measure two quantities simultaneously, so that we get samples of *pairs* of values (x_1, y_1), (x_2, y_2), $\cdots$, (x_n, y_n). Most applications involve one of two kinds of experiments, as follows.

1. In **regression analysis** one of the two variables, call it x, can be regarded as an ordinary variable because we can measure it without substantial error or we can even give it values we want. x is called the **independent variable,** or sometimes the **controlled variable** because we can control it (set it at values we choose). The other variable, Y, is a random variable, and we are interested in the dependence of Y on x. Typical examples are the dependence of the blood pressure Y on the age x of a person or, as we shall now say, the regression of Y on x, the regression of the gain of weight Y of certain animals on the daily ration of food x, the regression of the heat conductivity Y of cork on the specific weight x of the cork, etc.

2. In **correlation analysis** both quantities are random variables and we are interested in relations between them. Examples are the relation (one says "correlation") between wear X and wear Y of the front tires of cars, between grades X and Y of students in mathematics and in physics, respectively, between the hardness X of steel plates in the center and the hardness Y near the edges of the plates, etc.

Regression Analysis

In regression analysis the dependence of Y on x is a dependence of the mean μ of Y on x, so that $\mu = \mu(x)$ is a function in the ordinary sense. The curve of $\mu(x)$ is called the **regression curve** of Y on x.

In this section we discuss the simplest case, namely, that of a straight **regression line**

$$\mu(x) = \kappa_0 + \kappa_1 x. \tag{1}$$

Then we may want to graph the sample values as n points in the xY-plane, fit a straight line through them, and use it for estimating $\mu(x)$ at values of x that interest us, so that we know what values of Y we can expect for those x. Fitting that line by eye would not be good because it would be subjective; that is, different persons' results would come out differently, particularly if the points are scattered. So we need a mathematical method that gives a unique result depending only on the n points. A widely used procedure is the method of least squares by Gauss and Legendre. For our task we may formulate it as follows.

Least Squares Principle

The straight line should be fitted through the given points so that the sum of the squares of the distances of those points from the straight line is minimum, where the distance is measured in the vertical direction (the y-direction). (Formulas below.)

To get uniqueness of the straight line, we need some extra condition. To see this, take the sample (0, 1), (0, −1). Then all the lines $y = k_1 x$ with any k_1 satisfy the principle. (Can you see it?) The following assumption will imply uniqueness, as we shall find out.

General Assumption (A1)

The x-values $x_1, \cdots, x_n$ in our sample $(x_1, y_1), \cdots, (x_n, y_n)$ are not all equal.

From a given sample $(x_1, y_1), \cdots, (x_n, y_n)$ we shall now determine a straight line by least squares. We write the line as

$$y = k_0 + k_1 x \tag{2}$$

and call it the **sample regression line** because it will be the counterpart of the population regression line (1).

Now a sample point (x_j, y_j) has the vertical distance (distance measured in the y-direction) from (2) given by

$$|y_j - (k_0 + k_1 x_j)| \qquad \text{(see Fig. 542).}$$

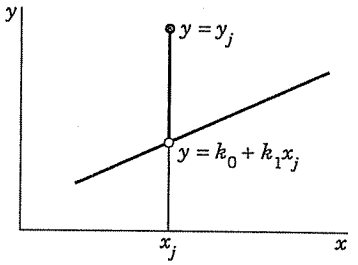


Fig. 542. Vertical distance of a point (x_j, y_j) from a straight line $y = k_0 + k_1 x$

Hence the sum of the squares of these distances is

$$q = \sum_{j=1}^{n} (y_j - k_0 - k_1 x_j)^2. \tag{3}$$

In the method of least squares we now have to determine k_0 and k_1 such that q is minimum. From calculus we know that a necessary condition for this is

$$\frac{\partial q}{\partial k_0} = 0 \quad \text{and} \quad \frac{\partial q}{\partial k_1} = 0. \tag{4}$$

We shall see that from this condition we obtain for the sample regression line the formula

$$y - \bar{y} = k_1(x - \bar{x}). \tag{5}$$

Here $\bar{x}$ and $\bar{y}$ are the means of the x- and the y-values in our sample, that is,

$$\text{(a)} \quad \bar{x} = \frac{1}{n}(x_1 + \cdots + x_n)$$
$$\text{(b)} \quad \bar{y} = \frac{1}{n}(y_1 + \cdots + y_n). \tag{6}$$

The slope k_1 in (5) is called the **regression coefficient** of the sample and is given by

$$k_1 = \frac{s_{xy}}{s_x^2}. \tag{7}$$

Here the **"sample covariance"** s_{xy} is

$$s_{xy} = \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \bar{x})(y_j - \bar{y}) = \frac{1}{n-1} \left[\sum_{j=1}^{n} x_j y_j - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right) \left(\sum_{j=1}^{n} y_j \right) \right] \tag{8}$$

and s_x^2 is given by

$$s_x^2 = \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \bar{x})^2 = \frac{1}{n-1} \left[\sum_{j=1}^{n} x_j^2 - \frac{1}{n} \left(\sum_{j=1}^{n} x_j \right)^2 \right]. \tag{9a}$$

From (5) we see that the sample regression line passes through the point $(\bar{x}, \bar{y})$, by which it is determined, together with the regression coefficient (7). We may call s_x^2 the *variance* of the x-values, but we should keep in mind that x is an ordinary variable, not a random variable.

We shall soon also need

$$\textbf{(9b)} \qquad s_y^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_j - \bar{y})^2 = \frac{1}{n-1}\left[\sum_{j=1}^{n} y_j^2 - \frac{1}{n}\left(\sum_{j=1}^{n} y_j\right)^2\right].$$

Derivation of (5) and (7). Differentiating (3) and using (4), we first obtain

$$\frac{\partial q}{\partial k_0} = -2\sum (y_j - k_0 - k_1 x_j) = 0,$$

$$\frac{\partial q}{\partial k_1} = -2\sum x_j(y_j - k_0 - k_1 x_j) = 0$$

where we sum over j from 1 to n. We now divide by 2, write each of the two sums as three sums, and take the sums containing y_j and $x_j y_j$ over to the right. Then we get the **"normal equations"**

$$\textbf{(10)} \qquad \begin{aligned} k_0 n + k_1 \sum x_j &= \sum y_j \\ k_0 \sum x_j + k_1 \sum x_j^2 &= \sum x_j y_j. \end{aligned}$$

This is a linear system of two equations in the two unknowns k_0 and k_1. Its coefficient determinant is [see (9)]

$$\begin{vmatrix} n & \sum x_j \\ \sum x_j & \sum x_j^2 \end{vmatrix} = n\sum x_j^2 - \left(\sum x_j\right)^2 = n(n-1)s_x^2 = n\sum (x_j - \bar{x})^2$$

and is not zero because of Assumption (A1). Hence the system has a unique solution. Dividing the first equation of (10) by n and using (6), we get $k_0 = \bar{y} - k_1\bar{x}$. Together with $y = k_0 + k_1 x$ in (2) this gives (5). To get (7), we solve the system (10) by Cramer's rule (Sec. 7.6) or elimination, finding

$$(11) \qquad k_1 = \frac{n\sum x_j y_j - \sum x_i \sum y_j}{n(n-1)s_x^2}.$$

This gives (7)–(9) and completes the derivation. [The equality of the two expressions in (8) and in (9) may be shown by the student; see Prob. 14]. ∎

EXAMPLE 1 Regression Line

The decrease of volume y [%] of leather for certain fixed values of high pressure x [atmospheres] was measured. The results are shown in the first two columns of Table 25.11. Find the regression line of y on x.

Solution. We see that $n = 4$ and obtain the values $\bar{x} = 28\,000/4 = 7000$, $\bar{y} = 19.0/4 = 4.75$, and from (9) and (8)

Table 25.11 Regression of the Decrease of Volume y [%] of Leather on the Pressure x [Atmospheres]

Given Values		Auxiliary Values	
x_j	y_j	x_j^2	$x_j y_j$
4 000	2.3	16 000 000	9 200
6 000	4.1	36 000 000	24 600
8 000	5.7	64 000 000	45 600
10 000	6.9	100 000 000	69 000
28 000	19.0	216 000 000	148 400

$$s_x^2 = \frac{1}{3}\left(216\,000\,000 - \frac{28\,000^2}{4}\right) = \frac{20\,000\,000}{3}$$

$$s_{xy} = \frac{1}{3}\left(148\,400 - \frac{28\,000 \cdot 19}{4}\right) = \frac{15\,400}{3}.$$

Hence $k_1 = 15\,400/20\,000\,000 = 0.000\,77$ from (7), and the regression line is

$$y - 4.75 = 0.000\,77(x - 7000) \qquad \text{or} \qquad y = 0.000\,77x - 0.64.$$

Note that $y(0) = -0.64$, which is physically meaningless, but typically indicates that a linear relation is merely an approximation valid on some restricted interval. ▩

Confidence Intervals in Regression Analysis

If we want to get confidence intervals, we have to make assumptions about the distribution of Y (which we have not made so far; least squares is a "geometric principle," nowhere involving probabilities!). We assume normality and independence in sampling:

Assumption (A2)

For each fixed x the random variable Y is normal with mean (1), *that is,*

$$\mu(x) = \kappa_0 + \kappa_1 x \tag{12}$$

and variance σ^2 independent of x.

Assumption (A3)

The n performances of the experiment by which we obtain a sample

$$(x_1, y_1), \qquad (x_2, y_2), \qquad \cdots, \qquad (x_n, y_n)$$

are independent.

κ_1 in (12) is called the **regression coefficient** of the population because it can be shown that under Assumptions (A1)–(A3) the maximum likelihood estimate of κ_1 is the sample regression coefficient k_1 given by (11).

Under Assumptions (A1)–(A3) we may now obtain a confidence interval for κ_1, as shown in Table 25.12.

Table 25.12 Determination of a Confidence Interval for κ_1 in (1) under Assumptions (A1)–(A3)

Step 1. Choose a confidence level γ (95%, 99%, or the like).

Step 2. Determine the solution c of the equation

$$F(c) = \tfrac{1}{2}(1 + \gamma) \tag{13}$$

from the table of the t-distribution with $n - 2$ degrees of freedom (Table A9 in App. 5; n = sample size).

Step 3. Using a sample $(x_1, y_1), \cdots, (x_n, y_n)$, compute $(n-1)s_x^2$ from (9a), $(n-1)s_{xy}$ from (8), k_1 from (7),

$$(n-1)s_y^2 = \sum_{j=1}^{n} y_j^2 - \frac{1}{n}\left(\sum_{j=1}^{n} y_j\right)^2 \tag{14}$$

[as in (9b)], and

$$q_0 = (n-1)(s_y^2 - k_1^2 s_x^2). \tag{15}$$

Step 4. Compute

$$K = c\sqrt{\frac{q_0}{(n-2)(n-1)s_x^2}}.$$

The confidence interval is

$$\mathrm{CONF}_\gamma\{k_1 - K \leqq \kappa_1 \leqq k_1 + K\}. \tag{16}$$

EXAMPLE 2 Confidence Interval for the Regression Coefficient

Using the sample in Table 25.11, determine a confidence interval for κ_1 by the method in Table 25.12.

Solution. ***Step 1.*** We choose $\gamma = 0.95$.

Step 2. Equation (13) takes the form $F(c) = 0.975$, and Table A9 in App. 5 with $n - 2 = 2$ degrees of freedom gives $c = 4.30$.

Step 3. From Example 1 we have $3s_x^2 = 20\,000\,000$ and $k_1 = 0.00077$. From Table 25.11 we compute

$$\begin{aligned} 3s_y^2 &= 102.2 - \frac{19^2}{4} \\ &= 11.95, \\ q_0 &= 11.95 - 20\,000\,000 \cdot 0.00077^2 \\ &= 0.092. \end{aligned}$$

Step 4. We thus obtain

$$\begin{aligned} K &= 4.30\sqrt{0.092/(2 \cdot 20\,000\,000)} \\ &= 0.000\,206 \end{aligned}$$

and

$$\mathrm{CONF}_{0.95}\{0.00056 \leqq \kappa_1 \leqq 0.00098\}. \cdot$$

Correlation Analysis

We shall now give an introduction to the basic facts in correlation analysis; for proofs see Ref. [G2] or [G8] in App. 1.

Correlation analysis is concerned with the relation between X and Y in a two-dimensional random variable (X, Y) (Sec. 24.9). A sample consists of n ordered pairs of values $(x_1, y_1), \cdots, (x_n, y_n)$, as before. The interrelation between the x and y values in the sample is measured by the sample covariance s_{xy} in (8) or by the sample **correlation coefficient**

(17)
$$r = \frac{s_{xy}}{s_x s_y}$$

with s_x and s_y given in (9). Here r has the advantage that it does not change under a multiplication of the x and y values by a factor (in going from feet to inches, etc.).

THEOREM 1

Sample Correlation Coefficient

The sample correlation coefficient r satisfies $-1 \leqq r \leqq 1$. *In particular,* $r = \pm 1$ *if and only if the sample values lie on a straight line.* (See Fig. 543.)

The theoretical counterpart of r is the **correlation coefficient** ρ of X and Y,

(18)
$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

where $\mu_X = E(X)$, $\mu_Y = E(Y)$, ${\sigma_X}^2 = E([X - \mu_X]^2)$, ${\sigma_Y}^2 = E([Y - \mu_Y]^2)$ (the means and variances of the marginal distributions of X and Y; see Sec. 24.9), and σ_{XY} is the

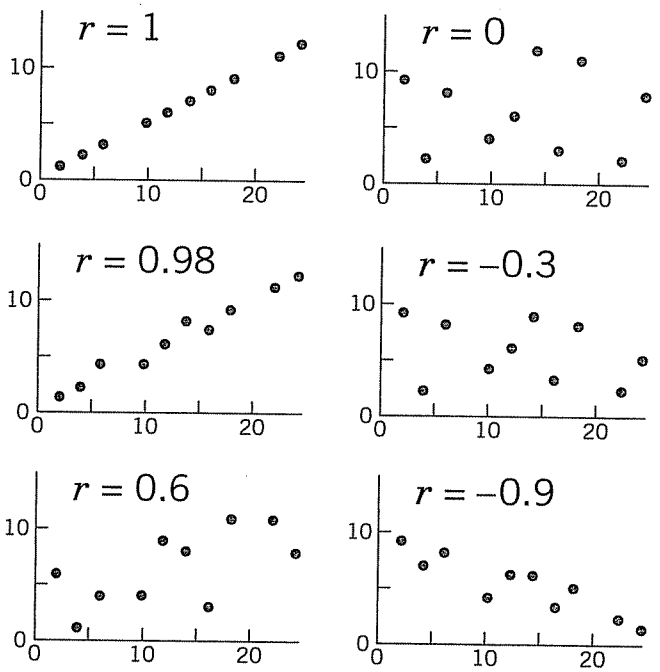


Fig. 543. Samples with various values of the correlation coefficient r

covariance of X and Y given by (see Sec. 24.9)

$$\sigma_{XY} = E([X - \mu_X][Y - \mu_Y]) = E(XY) - E(X)E(Y). \tag{19}$$

The analog of Theorem 1 is

THEOREM 2

Correlation Coefficient

The correlation coefficient ρ satisfies $-1 \leqq \rho \leqq 1$. In particular, $\rho = \pm 1$ if and only if X and Y are **linearly related,** *that is, $Y = \gamma X + \delta$, $X = \gamma^* Y + \delta^*$.*

X and Y are called **uncorrelated** if $\rho = 0$.

THEOREM 3

Independence. Normal Distribution

(a) *Independent X and Y* (see Sec. 24.9) *are uncorrelated.*

(b) *If (X, Y) is normal* (see below), *then uncorrelated X and Y are independent.*

Here the two-dimensional normal distribution can be introduced by taking two independent standardized normal random variables X^*, Y^*, whose joint distribution thus has the density

$$f^*(x^*, y^*) = \frac{1}{2\pi} e^{-(x^{*2}+y^{*2})/2} \tag{20}$$

(representing a surface of revolution over the x^*y^*-plane with a bell-shaped curve as cross section) and setting

$$X = \mu_X + \sigma_X X^*$$
$$Y = \mu_Y + \rho\sigma_Y X^* + \sqrt{1 - \rho^2}\, \sigma_Y Y^*.$$

This gives the general **two-dimensional normal distribution** with the density

$$f(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1 - \rho^2}} e^{-h(x,y)/2} \tag{21a}$$

where

$$h(x, y) = \frac{1}{1 - \rho^2}\left[\left(\frac{x - \mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x - \mu_X}{\sigma_X}\right)\left(\frac{y - \mu_Y}{\sigma_Y}\right) + \left(\frac{y - \mu_Y}{\sigma_Y}\right)^2\right]. \tag{21b}$$

In Theorem 3(b), normality is important, as we can see from the following example.

EXAMPLE 3 **Uncorrelated but Dependent Random Variables**

If X assumes -1, 0, 1 with probability 1/3 and $Y = X^2$, then $E(X) = 0$ and in (3)

$$\sigma_{XY} = E(XY) = E(X^3) = (-1)^3 \cdot \frac{1}{3} + 0^3 \cdot \frac{1}{3} + 1^3 \cdot \frac{1}{3} = 0,$$

so that $\rho = 0$ and X and Y are uncorrelated. But they are certainly not independent since they are even functionally related. ■

Test for the Correlation Coefficient ρ

Table 25.13 shows a test for ρ in the case of the two-dimensional normal distribution. t is an observed value of a random variable that has a t-distribution with $n - 2$ degrees of freedom. This was shown by R. A. Fisher (*Biometrika* **10** (1915), 507–521).

Table 25.13 Test of the Hypothesis $\rho = 0$ Against the Alternative $\rho > 0$ in the Case of the Two-Dimensional Normal Distribution

Step 1. Choose a significance level α (5%, 1%, or the like).

Step 2. Determine the solution c of the equation

$$P(T \leqq c) = 1 - \alpha$$

from the t-distribution (Table A9 in App. 5) with $n - 2$ degrees of freedom.

Step 3. Compute r from (17), using a sample $(x_1, y_1), \cdots, (x_n, y_n)$.

Step 4. Compute

$$t = r\left(\sqrt{\frac{n-2}{1-r^2}}\right).$$

If $t \leqq c$, accept the hypothesis. If $t > c$, reject the hypothesis.

EXAMPLE 4 Test for the Correlation Coefficient ρ

Test the hypothesis $\rho = 0$ (independence of X and Y, because of Theorem 3) against the alternative $\rho > 0$, using the data in the lower left corner of Fig. 543, where $r = 0.6$ (manual soldering errors on 10 two-sided circuit boards done by 10 workers; x = front, y = back of the boards).

Solution. We choose $\alpha = 5\%$; thus $1 - \alpha = 95\%$. Since $n = 10$, $n - 2 = 8$, the table gives $c = 1.86$. Also, $t = 0.6\sqrt{8/0.64} = 2.12 > c$. We reject the hypothesis and assert that there is a **positive correlation.** A worker making few (many) errors on the front side also tends to make few (many) errors on the reverse side of the board. ■

PROBLEM SET 25.9

1–10 SAMPLE REGRESSION LINE

Find and sketch or graph the sample regression line of y and x and the given data as points on the same axes.

1. $(-1, 1)$, $(0, 1.7)$, $(1, 3)$
2. $(3, 3.5)$, $(5, 2)$, $(7, 4.5)$, $(9, 3)$
3. $(2, 12)$, $(5, 24)$, $(9, 33)$, $(14, 50)$
4. $(11, 22)$, $(15, 18)$, $(17, 16)$, $(20, 9)$, $(22, 10)$

5. Speed x [mph] of a car

Stopping distance y [ft]	150	195	240	295

Also find the stopping distance at 35 mph.

6. x = Deformation of a certain steel [mm], y = Brinell hardness [kg/mm^2]

x	6	9	11	13	22	26	28	33	35
y	68	67	65	53	44	40	37	34	32

7. x = Revolutions per minute, y = Power of a Diesel engine [hp]

x	400	500	600	700	750
y	580	1030	1420	1880	2100

8.

Humidity of air x [%]	10	20	30	40
Expansion of gelatin y [%]	0.8	1.6	2.3	2.8

9.

Voltage x [V]	40	40	80	80	110	110
Current y [A]	5.1	4.8	10.0	10.3	13.0	12.7

Also find the resistance R [Ω] by **Ohms' law** (Sec. 2.9].

10.

Force x [lb]	2	4	6	8
Extension y [in] of a spring	4.1	7.8	12.3	15.8

Also find the spring modulus by **Hooke's law** (Sec. 2.4).

11–13 CONFIDENCE INTERVALS

Find a 95% confidence interval for the regression coefficient κ_1, assuming that (A2) and (A3) hold and using the sample:

11. In Prob. 6

12. In Prob. 7

13. In Prob. 8

14. Derive the second expression for s_x^2 in (9a) from the first one.

15. CAS EXPERIMENT. Moving Data. Take a sample, for instance, that in Prob. 6, and investigate and graph the effect of changing y-values (a) for small x, (b) for large x, (c) in the middle of the sample.

CHAPTER 25 REVIEW QUESTIONS AND PROBLEMS

1. What is a sample? Why do we take samples?

2. What is the role of probability theory in statistics?

3. Will you get better results by taking larger samples? Explain.

4. Do several samples from a certain population have the same mean? The same variance?

5. What is a parameter? How can we estimate it? Give an example.

6. What is a statistical test? What errors occur in testing?

7. How do we test in quality control?

8. What is the χ^2-test? Give a simple example from memory.

9. What are nonparametric tests? When would you apply them?

10. In what tests did we use the t-distribution? The χ^2-distribution?

11. What are one-sided and two-sided tests? Give typical examples.

12. List some areas of application of statistical tests.

13. What do we mean by "goodness of fit"?

14. Acceptance sampling uses principles of testing. Explain.

15. What is the power of a test? What can you do if the power is low?

16. Explain the idea of a maximum likelihood estimate from memory.

17. How does the length of a confidence interval depend on the sample size? On the confidence level?

18. Couldn't we make the error in interval estimation zero simply by choosing the confidence level 1?

19. What is the least squares principle? Give applications.

20. What is the difference between regression and correlation analysis?

21. Find the maximum likelihood estimates of mean and variance of a normal distribution using the sample 5, 4, 6, 5, 3, 5, 7, 4, 6, 5, 8, 6.

22. Determine a 95% confidence interval for the mean μ of a normal population with variance $\sigma^2 = 16$, using a sample of size 400 with mean 53.

23. What will happen to the length of the interval in Prob. 22 if we reduce the sample size to 100?

24. Determine a 99% confidence interval for the mean of a normal population with standard deviation 2.2, using the sample 28, 24, 31, 27, 22.

25. What confidence interval do we obtain in Prob. 24 if we assume the variance to be unknown?

26. Assuming normality, find a 95% confidence interval for the variance from the sample 145.3, 145.1, 145.4, 146.2.

27–29 Find a 95% confidence interval for the mean μ, assuming normality and using the sample:

27. Nitrogen content [%] of steel 0.74, 0.75, 0.73, 0.75, 0.74, 0.72

28. Diameters of 10 gaskets with mean 4.37 cm and standard deviation 0.157 cm

29. Density [g/cm^3] of coke 1.40, 1.45, 1.39, 1.44, 1.38

30. What sample size should we use in Prob. 28 if we want to obtain a confidence interval of length 0.1, assuming that the standard deviation of the samples is (about) the same?

31–32 Find a 99% confidence interval for the variance σ^2, assuming normality and using the sample:

31. Rockwell hardness of tool bits 64.9, 64.1, 63.8, 64.0

32. A sample of size $n = 128$ with variance $s^2 = 1.921$

33. Using a sample of 10 values with mean 14.5 from a normal population with variance $\sigma^2 = 0.25$, test the hypothesis $\mu_0 = 15.0$ against the alternative $\mu_1 = 14.4$ on the 5% level.

34. In Prob. 33, change the alternative to $\mu \neq 15.0$ and test as before.

35. Find the power in Prob. 33.

36. Using a sample of 15 values with mean 36.2 and variance 0.9, test the hypothesis $\mu_0 = 35.0$ against the alternative $\mu_1 = 37.0$, assuming normality and taking $\alpha = 1\%$.

37. Using a sample of 20 values with variance 8.25 from a normal population, test the hyothesis ${\sigma_0}^2 = 5.0$ against the alternative ${\sigma_1}^2 = 8.1$, choosing $\alpha = 5\%$.

38. A firm sells paint in cans containing 1 kg of paint per can and is interested to know whether the mean weight differs significantly from 1 kg, in which case the filling machine must be adjusted. Set up a hypothesis and an alternative and perform the test, assuming normality and using a sample of 20 fillings having a mean of 991 g and a standard deviation of 8 g. (Choose $\alpha = 5\%$.)

39. Using samples of sizes 10 and 5 with variances ${s_x}^2 = 50$ and ${s_y}^2 = 20$ and assuming normality of the corresponding populations, test the hypothesis H_0: ${\sigma_x}^2 = {\sigma_y}^2$ against the alternative ${\sigma_x}^2 > {\sigma_y}^2$. Choose $\alpha = 5\%$.

40. Assume the thickness X of washers to be normal with mean 2.75 mm and variance 0.00024 mm^2. Set up a control chart for μ, choosing $\alpha = 1\%$, and graph the means of the five samples (2.74, 2.76), (2.74, 2.74), (2.79, 2.81), (2.78, 2.76), (2.71, 2.75) on the chart.

41. What effect on UCL − LCL in a control chart for the mean does it have if we double the sample size? If we switch from $\alpha = 1\%$ to $\alpha = 5\%$?

42. The following samples of screws (length in inches) were taken from an ongoing production. Assuming that the population is normal with mean 3.500 and variance 0.0004, set up a control chart for the mean, choosing $\alpha = 1\%$, and graph the sample means on the chart.

Sample No.	1	2	3	4	5	6	7	8
Length	3.49	3.48	3.52	3.50	3.51	3.49	3.52	3.53
	3.50	3.47	3.49	3.51	3.48	3.50	3.50	3.49

43. A purchaser checks gaskets by a single sampling plan that uses a sample size of 40 and an acceptance number of 1. Use Table A6 in App. 5 to compute the probability of acceptance of lots containing the following percentages of defective gaskets $\frac{1}{4}\%$, $\frac{1}{2}\%$, 1%, 2%, 5%, 10%. Graph the OC curve. (Use the Poisson approximation.)

44. Does an automatic cutter have the tendency of cutting longer and longer pieces of wire if the lengths of subsequent pieces [in.] were 10.1, 9.8, 9.9, 10.2, 10.6, 10.5?

45. Find the least squares regression line to the data (−2, 1), (0, 1), (2, 3), (4, 4), (6, 5).

SUMMARY OF CHAPTER 25
Mathematical Statistics

We recall from Chap. 24 that with an experiment in which we observe some quantity (number of defectives, height of persons, etc.) there is associated a random variable X whose probability distribution is given by a distribution function

(1) $$F(x) = P(X \leqq x)$$ (Sec. 24.5)

which for each x gives the probability that X assumes any value not exceeding x.

In statistics we take random samples $x_1, \cdots, x_n$ of size n by performing that experiment n times (Sec. 25.1) and draw conclusions from properties of samples about properties of the distribution of the corresponding X. We do this by calculating *point estimates* or *confidence intervals* or by performing a *test* for **parameters** (μ and σ^2 in the normal distribution, p in the binomial distribution, etc.) or by a test for distribution functions.

A **point estimate** (Sec. 25.2) is an approximate value for a parameter in the distribution of X obtained from a sample. Notably, the **sample mean** (Sec. 25.1)

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j = \frac{1}{n}(x_1 + \cdots + x_n) \tag{2}$$

is an estimate of the mean μ of X, and the **sample variance** (Sec. 25.1)

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2 = \frac{1}{n-1}[(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2] \tag{3}$$

is an estimate of the variance σ^2 of X. Point estimation can be done by the basic **maximum likelihood method** (Sec. 25.2).

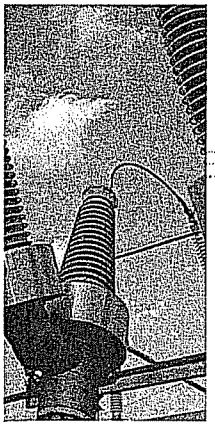
Confidence intervals (Sec. 25.3) are intervals $\theta_1 \leqq \theta \leqq \theta_2$ with endpoints calculated from a sample such that with a high probability γ we obtain an interval that contains the unknown true value of the parameter θ in the distribution of X. Here, γ is chosen at the beginning, usually 95% or 99%. We denote such an interval by $\text{CONF}_\gamma\{\theta_1 \leqq \theta \leqq \theta_2\}$.

In a **test** for a parameter we test a *hypothesis* $\theta = \theta_0$ against an *alternative* $\theta = \theta_1$ and then, on the basis of a sample, accept the hypothesis, or we reject it in favor of the alternative (Sec. 25.4). Like any conclusion about X from samples, this may involve errors leading to a false decision. There is a small probability α (which we can choose, 5% or 1%, for instance) that we reject a true hypothesis, and there is a probability β (which we can compute and decrease by taking larger samples) that we accept a false hypothesis. α is called the **significance level** and $1 - \beta$ the **power** of the test. Among many other engineering applications, testing is used in **quality control** (Sec. 25.5) and **acceptance sampling** (Sec. 25.6).

If not merely a parameter but the kind of distribution of X is unknown, we can use the **chi-square test** (Sec. 25.7) for testing the hypothesis that some function $F(x)$ is the unknown distribution function of X. This is done by determining the discrepancy between $F(x)$ and the distribution function $\tilde{F}(x)$ of a given sample.

"Distribution-free" or **nonparametric tests** are tests that apply to any distribution, since they are based on combinatorial ideas. These tests are usually very simple. Two of them are discussed in Sec. 25.8.

The last section deals with samples of **pairs of values,** which arise in an experiment when we simultaneously observe two quantities. In **regression analysis,** one of the quantities, x, is an ordinary variable and the other, Y, is a random variable whose mean μ depends on x, say, $\mu(x) = \kappa_0 + \kappa_1 x$. In **correlation analysis** the relation between X and Y in a two-dimensional random variable (X, Y) is investigated, notably in terms of the **correlation coefficient** ρ.



APPENDIX 1

References

Software *see* at the beginning of Chaps. 19 and 24.

General References

[GR1] Abramowitz, M. and I. A. Stegun (eds.), *Handbook of Mathematical Functions.* 10th printing, with corrections. Washington, DC: National Bureau of Standards. 1972 (also New York: Dover, 1965).

[GR2] Cajori, F., *History of Mathematics.* 5th ed. Reprinted. Providence, RI: American Mathematical Society, 2002.

[GR3] Courant, R. and D. Hilbert, *Methods of Mathematical Physics.* 2 vols. Hoboken, NJ: Wiley, 2003.

[GR4] Courant, R., *Differential and Integral Calculus.* 2 vols. Hoboken, NJ: Wiley, 2003.

[GR5] Graham, R. L. et al., *Concrete Mathematics.* 2nd ed. Reading, MA: Addison-Wesley, 1994.

[GR6] Ito, K. (ed.), *Encyclopedic Dictionary of Mathematics.* 4 vols. 2nd ed. Cambridge, MA: MIT Press, 1993.

[GR7] Kreyszig, E., *Introductory Functional Analysis with Applications.* New York: Wiley, 1989.

[GR8] Kreyszig, E., *Differential Geometry.* Mineola, NY: Dover, 1991.

[GR9] Kreyszig, E. *Introduction to Differential Geometry and Riemannian Geometry.* Toronto: University of Toronto Press, 1975.

[GR10] Szegö, G., *Orthogonal Polynomials.* 4th ed. Reprinted. New York: American Mathematical Society, 2003.

[GR11] Thomas, G. et al., *Thomas' Calculus, Early Transcendentals Update.* 10th ed. Reading, MA: Addison-Wesley, 2003.

Part A. Ordinary Differential Equations (ODEs) (Chaps. 1–6)

See also Part E: Numeric Analysis

[A1] Arnold, V. I., *Ordinary Differential Equations.* 3rd ed. New York: Springer, 1997.

[A2] Bhatia, N. P. and G. P. Szego, *Stability Theory of Dynamical Systems.* New York: Springer, 2002.

[A3] Birkhoff, G. and G.-C. Rota, *Ordinary Differential Equations.* 4th ed. New York: Wiley, 1989.

[A4] Brauer, F. and J. A. Nohel, *Qualitative Theory of Ordinary Differential Equations.* Mineola, NY: Dover, 1994.

[A5] Churchill, R. V., *Operational Mathematics.* 3rd ed. New York: McGraw-Hill, 1972.

[A6] Coddington, E. A. and R. Carlson, *Linear Ordinary Differential Equations.* Philadelphia: SIAM, 1997.

[A7] Coddington, E. A. and N. Levinson, *Theory of Ordinary Differential Equations.* Malabar, FL: Krieger, 1984.

[A8] Dong, T.-R. et al., *Qualitative Theory of Differential Equations.* Providence, RI: American Mathematical Society, 1992.

[A9] Erdélyi, A. et al., *Tables of Integral Transforms.* 2 vols. New York: McGraw-Hill, 1954.

[A10] Hartman, P., *Ordinary Differential Equations.* 2nd ed. Philadelphia: SIAM, 2002.

[A11] Ince, E. L., *Ordinary Differential Equations.* New York: Dover, 1956.

[A12] Schiff, J. L., *The Laplace Transform: Theory and Applications.* New York: Springer, 1999.

[A13] Watson, G. N., *A Treatise on the Theory of Bessel Functions.* 2nd ed. Reprinted. New York: Cambridge University Press, 1995.

[A14] Widder, D. V., *The Laplace Transform.* Princeton, NJ: Princeton University Press, 1941.

[A15] Zwillinger, D., *Handbook of Differential Equations.* 3rd ed. New York: Academic Press, 1997.

Part B. Linear Algebra, Vector Calculus (Chaps. 7–10)

For books on *numeric* linear algebra, *see also* Part E: Numeric Analysis.

[B1] Bellman, R., *Introduction to Matrix Analysis.* 2nd ed. Philadelphia: SIAM, 1997.

[B2] Chatelin, F., *Eigenvalues of Matrices.* New York: Wiley-Interscience, 1993.

[B3] Gantmacher, F. R., *The Theory of Matrices.* 2 vols. Providence, RI: American Mathematical Society, 2000.

[B4] Gohberg, I. P. et al., *Invariant Subspaces of Matrices with Applications.* New York: Wiley, 1986.

[B5] Greub, W. H., *Linear Algebra.* 4th ed. New York: Springer, 1996.

[B6] Herstein, I. N., *Abstract Algebra.* 3rd ed. New York: Wiley, 1996.

[B7] John, A. W., *Matrices and Tensors in Physics.* 3rd ed. New York: Wiley, 1995.

[B8] Lang, S., *Linear Algebra.* 3rd ed. New York: Springer, 1996.

[B9] Nef, W., *Linear Algebra.* 2nd ed. New York: Dover, 1988.

[B10] Parlett, B., *The Symmetric Eigenvalue Problem.* Philadelphia: SIAM, 1997.

Part C. Fourier Analysis and PDEs (Chaps. 10–11)

For books on *numerics* for PDEs *see also* Part E: Numeric Analysis.

[C1] Antimirov, M. Ya., *Applied Integral Transforms.* Providence, RI: American Mathematical Society, 1993.

[C2] Bracewell, R., *The Fourier Transform and Its Applications.* 3rd ed. New York: McGraw-Hill, 2000.

[C3] Carslaw, H. S. and J. C. Jaeger, *Conduction of Heat in Solids.* 2nd ed. Reprinted. Oxford: Clarendon, 1986.

[C4] Churchill, R. V. and J. W. Brown, *Fourier Series and Boundary Value Problems.* 6th ed. New York: McGraw-Hill, 2000.

[C5] DuChateau, P. and D. Zachmann, *Applied Partial Differential Equations.* Mineola, NY: Dover, 2001.

[C6] Hanna, J. R. and J. H. Rowland, *Fourier Series, Transforms, and Boundary Value Problems.* 2nd ed. New York: Wiley, 1990.

[C7] Jerri, A. J., *The Gibbs Phenomenon in Fourier Analysis, Splines, and Wavelet Approximations.* Boston: Kluwer, 1998.

[C8] John, F., *Partial Differential Equations.* New York: Springer, 1995.

[C9] Tolstov, G. P., *Fourier Series.* New York: Dover, 1976.

[C10] Widder, D. V., *The Heat Equation.* New York: Academic Press, 1975.

[C11] Zauderer, E., *Partial Differential Equations of Applied Mathematics.* 2nd ed. New York: Wiley, 1989.

[C12] Zygmund, A. and R. Fefferman, *Trigonometric Series.* 3rd ed. New York: Cambridge University Press, 2003.

Part D. Complex Analysis (Chaps. 13–18)

[D1] Ahlfors, L. V., *Complex Analysis.* 3rd ed. New York: McGraw-Hill, 1979.

[D2] Bieberbach, L., *Conformal Mapping.* Providence, RI: American Mathematical Society, 2000.

[D3] Henrici, P., *Applied and Computational Complex Analysis.* 3 vols. New York: Wiley, 1993.

[D4] Hille, E., *Analytic Function Theory.* 2 vols. 2nd ed. Providence, RI: American Mathematical Society, 1997.

[D5] Knopp, K., *Elements of the Theory of Functions.* New York: Dover, 1952.

[D6] Knopp, K., *Theory of Functions.* 2 parts. New York: Dover, 1945, 1947.

[D7] Krantz, S. G., *Complex Analysis: The Geometric Viewpoint.* Washington, DC: The Mathematical Association of America, 1990.

[D8] Lang, S., *Complex Analysis.* 3rd ed. New York: Springer, 1993.

[D9] Narasimhan, R., *Compact Riemann Surfaces.* New York: Springer, 1996.

[D10] Nehari, Z., *Conformal Mapping.* Mineola, NY: Dover, 1975.

[D11] Springer, G., *Introduction to Riemann Surfaces.* Providence, RI: American Mathematical Society, 2002.

Part E. Numeric Analysis (Chaps. 19–21)

[E1] Ames, W. F., *Numerical Methods for Partial Differential Equations.* 3rd ed. New York: Academic Press, 1992.

[E2] Anderson, E., et al., *LAPACK User's Guide.* 3rd ed. Philadelphia: SIAM, 1999.

[E3] Bank, R. E., *PLTMG. A Software Package for Solving Elliptic Partial Differential Equations: Users' Guide 7.0.* Philadelphia: SIAM, 1994.

[E4] Constanda, C., *Solution Techniques for Elementary Partial Differential Equations.* Boca Raton, FL: CRC Press, 2002.

[E5] Dahlquist, G. and A. Björck, *Numerical Methods.* Mineola, NY: Dover, 2003.

[E6] DeBoor, C., *A Practical Guide to Splines.* Reprinted. New York: Springer, 1991.

[E7] Dongarra, J. J. et al., *LINPACK Users Guide.* Philadelphia: SIAM, 1978. (See also at the beginning of Chap. 19.)

[E8] Garbow, B. S. et al., *Matrix Eigensystem Routines: EISPACK Guide Extension.* Reprinted. New York: Springer, 1990.

[E9] Golub, G. H. and C. F. Van Loan, *Matrix Computations.* 3rd ed. Baltimore, MD: Johns Hopkins University Press, 1996.

[E10] Higham, N. J., *Accuracy and Stability of Numerical Algorithms.* 2nd ed. Philadelphia: SIAM, 2002.

[E11] IMSL (International Mathematical and Statistical Libraries), *FORTRAN Numerical Library.* Houston, TX: Visual Numerics, 2002. (See also at the beginning of Chap. 19.)

[E12] IMSL, *IMSL for Java.* Houston, TX: Visual Numerics, 2002.

[E13] IMSL, *C Library.* Houston, TX: Visual Numerics, 2002.

[E14] Kelley, C. T., *Iterative Methods for Linear and Nonlinear Equations.* Philadelphia: SIAM, 1995.

[E15] Knabner, P. and L. Angerman, *Numerical Methods for Partial Differential Equations.* New York: Springer, 2003.

[E16] Knuth, D. E., *The Art of Computer Programming.* 3 vols. 3rd ed. Reading, MA: Addison-Wesley, 2005.

[E17] Kreyszig, E., *Introductory Functional Analysis with Applications.* New York: Wiley, 1989.

[E18] Kreyszig, E., On methods of Fourier analysis in multigrid theory. *Lecture Notes in Pure and Applied Mathematics* 157. New York: Dekker, 1994, pp. 225–242.

[E19] Kreyszig, E., Basic ideas in modern numerical analysis and their origins. *Proceedings of the Annual Conference of the Canadian Society for the History and Philosophy of Mathematics.* 1997, pp. 34–45.

[E20] Kreyszig, E., and J. Todd, *QR* in two dimensions. *Elemente der Mathematik* 31 (1976), pp. 109–114.

[E21] Mortensen, M. E., *Geometric Modeling.* 2nd ed. New York: Wiley, 1997.

[E22] Morton, K. W., and D. F. Mayers, *Numerical Solution of Partial Differential Equations: An Introduction.* New York: Cambridge University Press, 1994.

[E23] Ortega, J. M., *Introduction to Parallel and Vector Solution of Linear Systems.* New York: Plenum Press, 1988.

[E24] Overton, M. L., *Numerical Computing with IEEE Floating Point Arithmetic.* Philadelphia: SIAM, 2001.

[E25] Press, W. H. et al., *Numerical Recipes in C: The Art of Scientific Computing.* 2nd ed. New York: Cambridge University Press, 1992.

[E26] Shampine, L. F., *Numerical Solutions of Ordinary Differential Equations.* New York: Chapman and Hall, 1994.

[E27] Varga, R. S., *Matrix Iterative Analysis.* 2nd ed. New York: Springer, 2000.

[E28] Varga, R. S., *Geršgorin and His Circles.* New York: Springer, 2004.

[E29] Wilkinson, J. H., *The Algebraic Eigenvalue Problem.* Oxford: Oxford University Press, 1988.

Part F. Optimization, Graphs (Chaps. 22–23)

[F1] Bondy, J. A., *Graph Theory with Applications.* Hoboken, NJ: Wiley-Interscience, 2003.

[F2] Cook, W. J. et al., *Combinatorial Optimization.* New York: Wiley, 1993.

[F3] Diestel, R., *Graph Theory.* 2nd ed. New York: Springer, 2000.

[F4] Diwekar, U. M., *Introduction to Applied Optimization.* Boston: Kluwer, 2003.

[F5] Gass, S. L., *Linear Programming. Method and Applications.* 3rd ed. New York: McGraw-Hill, 1969.

[F6] Gross, J. T., *Handbook of Graph Theory and Applications.* Boca Raton, FL: CRC Press, 1999.

[F7] Goodrich, M. T., and R. Tamassia, *Algorithm Design: Foundations, Analysis, and Internet Examples.* Hoboken, NJ: Wiley, 2002.

[F8] Harary, F., *Graph Theory.* Reprinted. Reading, MA: Addison-Wesley, 2000.

[F9] Merris, R., *Graph Theory.* Hoboken, NJ: Wiley-Interscience, 2000.

[F10] Ralston, A., and P. Rabinowitz, *A First Course in Numerical Analysis.* 2nd ed. Mineola, NY: Dover, 2001.

[F11] Thulasiraman, K., and M. N. S. Swamy, *Graph Theory and Algorithms.* New York: Wiley-Interscience, 1992.

[F12] Tucker, A., *Applied Combinatorics.* 4th ed. Hoboken, NJ: Wiley, 2001.

Part G. Probability and Statistics (Chaps. 24–25)

[G1] American Society for Testing Materials, *Manual on Presentation of Data and Control Chart Analysis.* 7th ed. Philadelphia: ASTM, 2002.

[G2] Anderson, T. W., *An Introduction to Multivariate Statistical Analysis.* 3rd ed. Hoboken, NJ: Wiley, 2003.

[G3] Cramér, H., *Mathematical Methods of Statistics.* Reprinted. Princeton, NJ: Princeton University Press, 1999.

[G4] Dodge, Y., *The Oxford Dictionary of Statistical Terms.* 6th ed. Oxford: Oxford University Press, 2003.

[G5] Gibbons, J. D., *Nonparametric Statistical Inference.* 4th ed. New York: Dekker, 2003.

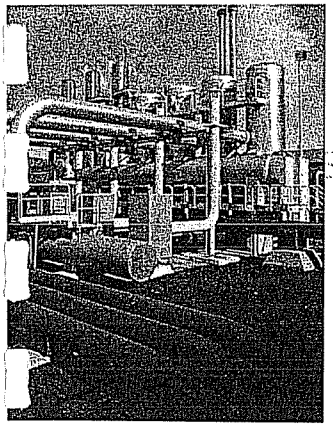
[G6] Grant, E. L. and R. S. Leavenworth, *Statistical Quality Control.* 7th ed. New York: McGraw-Hill, 1996.

[G7] IMSL, *Fortran Numerical Library.* Houston, TX: Visual Numerics, 2002.

[G8] Kreyszig, E., *Introductory Mathematical Statistics. Principles and Methods.* New York: Wiley, 1970.

[G9] O'Hagan, T. et al., *Kendall's Advanced Theory of Statistics 3-Volume Set.* Kent, U.K.: Hodder Arnold, 2004.

[G10] Rohatgi, V. K. and A. K. MD. E. Saleh, *An Introduction to Probability and Statistics.* 2nd ed. Hoboken, NJ: Wiley-Interscience, 2001.



APPENDIX 2

Answers to Odd-Numbered Problems

Problem Set 1.1, page 8

1. $(\cos \pi x)/\pi + c$ **3.** $e^{x^2/2} + c$ **5.** First order
7. Second order **9.** Third order
11. $y = \frac{1}{2} \tan (2x + n\pi)$, $n = 0, \pm 1, \pm 2, \cdots$
13. $y = e^{-x^2}$ **15.** (A) No. (B) No. Only $y = 0$.
17. $y'' = g$, $y' = gt$, $y = gt^2/2$
19. $y'' = k$, $y' = kt + 6$, $y = \frac{1}{2}kt^2 + 6t$, $y(60) = 1800k + 360 = 3000$, $k = 1.47$, $y'(60) = 1.47 \cdot 60 + 6 = 94$ [m/sec] $= 210$ [mph]
21. $e^{kH} = \frac{1}{2}$, $H = (\ln \frac{1}{2})/k = (10^{11} \ln 2)/1.4 = 1570$ [years]

Problem Set 1.2, page 11

11. $y = -(2/\pi) \cos \frac{1}{2}\pi x + c$ **15.** $y = x(1 - \ln x) + c$
17. Verify the general solution $y^2 + t^2 = c$. Circle of radius $3\sqrt{2}$
19. $mv' = mg - bv^2$, $v' = 9.8 - v^2$, $v(0) = 10$. $v' = 0$ gives the limit $\sqrt{9.8} = 3.1$ [meter/sec].

Problem Set 1.3, page 18

3. $\cos 2y\, dy = 2\, dx$, $y = \frac{1}{2} \arcsin (4x + c)$ **5.** $y^2 + 36x^2 = c$, ellipses
7. $dy/y = \cot \pi x\, dx$, $y = c(\sin \pi x)^{1/\pi}$ **9.** $y = \tan (c - e^{-\pi x}/\pi)$
11. $r = r_0 e^{-t^2}$ **13.** $I = I_0 e^{-Rt/L}$
15. $y = e^x/\sqrt{2x + 5}$ **17.** $y = 4 \ln x$
19. $y = \sqrt{\ln (x^2 - 2x + e)}$
21. $y' = (y - b)/(x - a)$, $y - b = c(x - a)$
23. $y_0 e^k = 2y_0$, $e^k = 2$ (1 week), $e^{2k} = 2^2$ (2 weeks), $e^{4k} = 2^4$
25. $y = y_0 e^{kt} = y_0 e^{-0.0001213t} = y_0 e^{-0.0001213 \cdot 4000} = 0.62y_0$; 62%; cf. Example 2.
27. $y' = -ky$, $y = y_0 e^{-kt}$, $e^{-5k} = 0.5$, $k = -(\ln 0.5)/5 = 0.139$, $t = -(\ln 0.05)/0.139 = 22$ [min]
29. $T(0) = 10$, $T = 23 - 13e^{kt}$, $T(2) = 23 - 13e^{2k} = 18$, $k = -0.478$, $T = 22.8$ gives $t = [\ln (-0.2/-13)]/(-0.478) = 8.73$ [min].
31. $h = gt^2/2$, $t = \sqrt{2h/g}$, $v = gt = g\sqrt{2h/g} = \sqrt{2gh}$
33. $y' = 0 - (2/800)y$, $y = 200e^{-0.0025t}$, $t = 300$ [min], $y(300) = 94.5$ [lb]
35. (A) is related to the error function and (C) concerns the Fresnel integral $C(x)$; see App. 3.1. (D) $y' = 2xy + 1$, $y(0) = 0$

Problem Set 1.4, page 25

1. Exact. $x^4 + y^4 = c$

3. Exact. $u = \cos \pi x \sinh y + k(y)$, $u_y = \cos \pi x \cosh y + k'$, $k' = 0$.
Ans. $\cos \pi x \sinh y = c$

5. Exact. $9x^2 + 4y^2 = c$

7. Exact, $M_\theta = N_r = -2e^{-2\theta}$, $u = re^{-2\theta} + k(\theta)$, $u_\theta = -2re^{-2\theta} + k'$, $k' = 0$.
Ans. $re^{-2\theta} = c$, $r = ce^{2\theta}$

9. Exact. $u = y/x + \sin 2x + k(y)$, $u_y = 1/x + k' = 1/x - 2 \sin 2y$.
Ans. $y/x + \sin 2x + \cos 2y = c$

11. Not exact. $F = 1/x^2$ by Theorem 1. $-y/x^2\, dx + 1/x\, dy = d(y/x) = 0$. *Ans.* $y = cx$

13. $-3y^2/x^4\, dx + 2y/x^3\, dy = d(y^2/x^3) = 0$. $y = cx^{3/2}$ (semicubical parabolas)

15. Exact, $u = e^{2x} \cos y + k(y)$, $u_y = -e^{2x} \sin y + k'$, $k' = 0$. *Ans.* $e^{2x} \cos y = c$, $c = 1$

17. Not exact. Try R. $F = e^{-x}$, $e^{-x}(\cos \omega x + \omega \sin \omega x)\, dx + dy = 0$, $u = y + l(x)$, $u_x = l' = e^{-x}(\cos \omega x + \omega \sin \omega x)$, $u = y + l = y - e^{-x} \cos \omega x = c$, $c = 0$

19. $u = e^x + k(y)$, $u_y = k' = -1 + e^y$, $k = -y + e^y$. *Ans.* $e^x - y + e^y = c$

21. $B = C$, $\frac{1}{2}Ax^2 + Cxy + \frac{1}{2}Dy^2 = c$

Problem Set 1.5, page 32

3. $y = ce^{-3.5x} + 0.8$

5. $y = 2.6e^{-1.25x} + 4$

7. $y = x + c$ (if $k = 0$). $y = ce^{-kx} + e^{2kx}/3k$ if $k \neq 0$

9. Separate. $y - 2.5 = c \cosh^4 1.5x$

11. $y = 2xe^{\cos 2x}$

13. $y = \sin 2x + c/\sin^2 2x$, $c = 1$

15. $y = e^{1/x}(x^2 + c)$, $c = 4.1$

17. $y = (c + \frac{1}{2} \cosh 10x)/x^3$. Note $(x^3y)' = 5 \sinh 10x$.

19. $y = 1/u$, $u = ce^{-5.7x} - \dfrac{6.5}{5.7}$

21. $u = y^{-2} = e^{x^2}(1 + ce^{2x})$, $c = 3$, $u(0) = 4$

23. Separate variables. $y^2 = 1 - ce^{\cos x}$, $c = -1/e$

25. $y' = Ry + k$, $y = ce^{Rt} - k/R$, $c = y_0 + k/R$. $y_0 = 1000$, $R = 0.06$, $t = 65 - 25 = 40$, $k = 1000$, $y = \$178\,076.12$. Start at 45 gives $y_0[(1 + 1/0.06)e^{0.06 \cdot 20} - 1/0.06] = 41.988732y_0 = 178\,076.12$, $y_0 = k = \$4241.05$.

27. $y' = 175(0.0001 - y/450)$, $y(0) = 450 \cdot 0.0004 = 0.18$,
$y = 0.135e^{-0.3889t} + 0.045 = 0.18/2$,
$e^{-0.3889t} = (0.09 - 0.045)/0.135 = 1/3$,
$t = (\ln 3)/0.3889 = 2.82$. *Ans.* About 3 years

29. $y' = A - ky$, $y(0) = 0$, $y = A(1 - e^{-kt})/k$

31. $y' = By^2 - Ay = By(y - A/B)$, $A > 0$, $B > 0$. Constant solutions $y = 0$, $y = A/B$.
$y' > 0$ if $y > A/B$ (unlimited growth), $y' < 0$ if $0 < y < A/B$ (extinction).
$y = A/(ce^{At} + B)$, $y(0) > A/B$ if $c < 0$, $y(0) < A/B$ if $c > 0$.

33. $y' = y - y^2 - 0.2y$, $y = 1/(1.25 - 0.75e^{-0.8t})$, limit 0.8, limit 1

35. $y' = y - 0.25y^2 - 0.1y = 0.25y(3.6 - y)$. Equilibrium harvest 3.6,
$y = 18/(5 + ce^{-0.9t})$

37. $(y_1 + y_2)' + p(y_1 + y_2) = (y_1' + py_1) + (y_2' + py_2) = 0 + 0 = 0$

39. $(y_1 + y_2)' + p(y_1 + y_2) = (y_1' + py_1) + (y_2' + py_2) = r + 0 = r$

41. Solution of $cy_1' + pcy_1 = c(y_1' + py_1) = cr$

43. CAS Experiment (a) $y = x \sin(1/x) + cx$. $c = 0$ if $y(2/\pi) = 2/\pi$. y is undefined at $x = 0$, the point at which the "waves" of $\sin(1/x)$ accumulate; the factor x makes them smaller and smaller. Experiment with various x-intervals.
(b) $y = x^n[\sin(1/x) + c]$. $y(2/\pi) = (2/\pi)^n$. n need not be an integer. Try $n = \frac{1}{2}$. Try $n = -1$ and see how the "waves" near 0 become larger and larger.

45. $y = uy^*$, $y' + py = u'y^* + uy^{*\prime} + puy^* = u'y^* + u(y^{*\prime} + py^*) = u'y^* + u \cdot 0 = r$, $u' = r/y^* = re^{\int p\,dx}$, $u = \int e^{\int p\,dx} r\,dx + c$. Thus, $y = uy_h$ gives (4). We shall see that this method extends to higher-order ODEs (Secs. 2.10 and 3.3).

Problem Set 1.6, page 36

1. $y' = 4$, $\tilde{y}' = -1/4$, $\tilde{y} = -x/4 + c^*$

3. $y/x = c$, $y'/x = y/x^2$, $y' = y/x$, $\tilde{y}' = -x/\tilde{y}$, $\tilde{y}^2 + x^2 = c^*$, circles

5. $2xy + x^2y' = 0$, $y' = -2y/x$, $\tilde{y}' = x/(2\tilde{y})$, $\tilde{y}^2 - x^2/2 = c^*$, hyperbolas

7. $ye^{-x^2/2} = c$, $y' = xy$, $\tilde{y}' = -1/(x\tilde{y})$, $\tilde{y}\tilde{y}' = -1/x$, $\tilde{y}^2/2 = -\ln|x| + c^{**}$, $x = c^*e^{-\tilde{y}^2/2}$, bell-shaped curves (with x and $\tilde{y}$ interchanged)

9. $y' = -4x/y$, $\tilde{y}' = \tilde{y}/4x$, $4\ln|\tilde{y}| = \ln|x| + c^{**}$, $x = c^*\tilde{y}^4$, parabolas

11. $xe^{-y/4} = c$, $y' = 4/x$, $\tilde{y}' = -x/4$, $\tilde{y} = -x^2/8 + c^*$

13. Use $dy/dx = 1/(dx/dy)$. $(y - 2x)e^x = c$, $(y' - 2 + y - 2x)e^x = 0$, $y' = 2 - y + 2x$, $dx/d\tilde{y} = -2 + \tilde{y} - 2x$ is linear, $dx/d\tilde{y} + 2x = \tilde{y} - 2$, $x = c^*e^{-2y} + y/2 - 5/4$

15. $u = c$, $u_x dx + u_y dy = 0$, $y' = -u_x/u_y$. Trajectories $\tilde{y}' = u_{\tilde{y}}/u_x$. Now $v = c^*$, $v_x dx + v_y dy = 0$, $y' = -v_x/v_y$. This agrees with the trajectory ODE in u if $u_x = v_y$ (equal denominators) and $u_y = -v_x$ (equal numerators). But these are just the Cauchy-Riemann equations.

17. $2x + 2yy' = 0$, $y' = -x/y$. Trajectories $\tilde{y}' = \tilde{y}/x$, $\ln|\tilde{y}| = \ln|x| + c^{**}$, $\tilde{y} = c^*x$.

19. $y' = -4x/9y$. Trajectories $\tilde{y}' = 9\tilde{y}/4x$. $\tilde{y} = c^*x^{9/4}$ ($c^* > 0$). Sketch or graph these curves.

Problem Set 1.7, page 41

1. In $|x - x_0| < a$; just take b in $\alpha = b/K$ large, namely, $b = aK$.

3. No. At a common point (x_1, y_1) they would both satisfy the "initial condition" $y(x_1) = y_1$, violating uniqueness.

5. $y' = f(x, y) = r(x) - p(x)y$; hence $\partial f/\partial y = -p(x)$ is continuous and is thus bounded in the closed interval $|x - x_0| \leqq a$.

7. R has sides $2a$ and $2b$ and center $(1, 1)$ since $y(1) = 1$. In R, $f = 2y^2 \leqq 2(b + 1)^2 = K$, $\alpha = b/K = b/(2(b + 1)^2)$. $d\alpha/db = 0$ gives $b = 1$, and $\alpha_{opt} = b/K = 1/8$. Solution by $dy/y^2 = 2\,dx$, etc., $y = 1/(3 - 2x)$.

9. $|1 + y^2| \leqq K = 1 + b^2$, $\alpha = b/K$, $d\alpha/db = 0$, $b = 1$, $\alpha = 1/2$.

Chapter 1 Review Questions and Problems, page 42

11. $dy/(y^2 + \frac{1}{4}) = 4\,dx$, $2\arctan 2y = 4x + c^*$, $y = \frac{1}{2}\tan(2x + c)$

13. Logistic ODE. $y = 1/u$, $y' = -u'/u^2 = 4/u - 1/u^2$, $u = c^*e^{-4x} + \frac{1}{4}$

15. $dy/(y^2 + 1) = x^2\,dx$, $\arctan y = x^3/3 + c$, $y = \tan(x^3/3 + c)$

17. Bernoulli. $y' + xy = x/y$, $u = y^2$, $u' = 2yy' = 2x - 2xu$ linear, $u = e^{-x^2}(\int e^{x^2} 2x\,dx + c) = 1 + ce^{-x^2}$, $y = \sqrt{u}$. Or write $yy' = -x(y^2 - 1)$ and separate.

19. Linear, $y = e^{\cos x}(\int e^{-\cos x} \sin x\, dx + c) = ce^{\cos x} + 1$. Or by separation.

21. Not exact. Use Theorem 1, Sec. 1.4; $R = 2/x$, $F = x^2$; the resulting exact ODE is $3x^2 \sin 2y\, dx + 2x^3 \cos 2y\, dy = d(x^3 \sin 2y)$, $x^3 \sin 2y = c$. Or by separation, $\cot 2y\, dy = -3/(2x)\, dx$, etc., $\sin 2y = cx^{-3}$.

23. Exact. $u = \int M\, dx = \sin xy - x^2 + k$, $u_y = x \cos xy + k' = N$, $k = y^2$, $\sin xy - x^2 + y^2 = c$.

25. Not exact. $R^* = 1$ in Theorem 2, Sec. 1.4, $F^* = e^y$. Exact is $e^y \sin (y - x)\, dx + e^y[\cos (y - x) - \sin (y - x)]\, dy = 0$. $u = \int M\, dx = e^y \cos (y - x) + k$, $u_y = e^y(\cos (y - x) - \sin (y - x)) + k' = N$, $e^y \cos (y - x) = c$.

27. Separation. $y^2 + x^2 = 25$

29. Separation. $y = \tan (x + c)$, $c = -\frac{1}{4}\pi$

31. Exact. $u = x^2y^2 + \cos x + 2y = c$, $c = u(0, 1) = 3$

33. $y' = x/y$. Trajectories $\tilde{y}' = -\tilde{y}/x$, $\tilde{y} = c^*/x$ by separation. Hyperbolas.

35. $y = y_0e^{kt}$, $e^{4k} = 0.9$, $k = \frac{1}{4} \ln 0.9$, $e^{kt} = 0.5$, $t = (\ln 0.5)/k = (\ln 0.5)/[(\ln 0.9)/4] = 26.3$ [days]

37. $e^{kt} = 0.01$, $t = (\ln 0.01)/k = 175$ [days]

39. $y' = -4x/y$. Trajectories $\tilde{y} = c_1x^{1/4}$ or $x = c_2\tilde{y}^4$

41. Logistic ODE $y' = Ay - By^2$, $y = 1/u$, $u' + Au = +B$, $u = ce^{-At} + B/A$

43. A = amount of incident light. A thin layer of thickness Δx absorbs $\Delta A = -kA\Delta x$ ($-k$ = constant of proportionality). Thus $\Delta A/\Delta x = -kA$. Let $\Delta x \to 0$. Then $A' = -kA$, $A = A_0e^{-kx}$ = amount of light in a thick layer at depth x from the surface of incidence.

Problem Set 2.1, page 52

1. $y = 2.5e^{4x} + 0.5e^{-4x}$ **3.** $y = e^{-x} \cos x$ **5.** $y = 4x^2 + 7/x^2$

7. Yes **9.** Yes if $a \neq 0$ **11.** No

13. No **15.** $F(x, z, z') = 0$ **17.** $y = c_1e^{kx} + c_2$

19. $y\, dz/dy = 4z$, $y = (c_1x + c_2)^{-1/3}$

21. $(dz/dy)z = -z^3 \sin y$, $-1/z = -dx/dy = \cos y + \tilde{c}_1$, $x = -\sin y + c_1y + c_2$

23. $y''y' = 2$, $y = \frac{4}{3}(t + 1)^{3/2} - \frac{1}{3}$, $y(3) = \frac{31}{3}$, $y'(3) = 4$

25. $y'' = ky'$, $z' = kz$, $z = c_1e^{kx} = y'$, $c_1 = 1$, $y = (e^{kx} - 1)/k$

Problem Set 2.2, page 59

1. $y = c_1e^{7x} + c_2e^{-x}$ **3.** $y = (c_1 + c_2x)e^{2.5x}$

5. $y = c_1e^{0.9x} + c_2e^{-1.1x}$ **7.** $y = e^{0.5x}(A \cos 1.5x + B \sin 1.5x)$

9. $y = c_1e^{3.5x} + c_2e^{-1.5x}$ **11.** $y = A \cos 3\pi x + B \sin 3\pi x$

13. $y = c_1e^{12x} + c_2e^{-12x}$ **15.** $y'' - 3y' + 2y = 0$

17. $y'' - 2\sqrt{3}\, y' + 3y = 0$ **19.** $y'' - 16y = 0$

21. $y = 4e^{3x} - 2e^{-x}$ **23.** $y = e^{-2x}(2 \cos x - \sin x)$

25. $y = 2 + e^{-\pi x}$ **27.** $y = (2 - 4x)e^{-0.25x}$

29. $y = e^{-0.1x}(3.2 \cos 0.2x + 1.6 \sin 0.2x)$ **31.** $y = 4e^{5x} - 4e^{-5x}$

33. $y_1 = e^{-x}$, $y_2 = 0.001e^x + e^{-x}$

35. Write $E = e^{-ax/2}$, $c = \cos \omega x$, $s = \sin \omega x$. Note that $E' = -\frac{1}{2}aE$, $c' = -\omega s$, $s' = \omega c$. Substitute, drop E, collect c-terms, then s-terms, and use $\omega^2 = b - \frac{1}{4}a^2$, to get $c(b - \frac{1}{2}a^2 + \frac{1}{4}a^2 - \omega^2) + s(-a\omega + \frac{1}{2}a\omega + \frac{1}{2}a\omega) = 0 + 0 = 0$.

Problem Set 2.3, page 61

1. 0, 0, $-2 \cos x$
3. $-0.8x^3 + 6x^2 + 0.4$, 0, $e^{0.4x}$
5. $-12x^3 + 9x^2 + 8x - 2$, $-28 \sin 4x - 4 \cos 4x$, 0
7. $y = (c_1 + c_2x)e^{-2x}$
9. $y = e^{-3x}(A \cos 2x + B \sin 2x)$
11. $y = c_1e^{-3.1x} + c_2e^{-x}$
13. $y = A \cos 4.2\omega x + B \sin 4.2\omega x$

Problem Set 2.4, page 68

1. $y = y_0 \cos \omega_0 t + (v_0/\omega_0) \sin \omega_0 t$. At integer t (if $\omega_0 = \pi$), because of periodicity.
3. $mL\theta'' = -mg \sin \theta \approx -mg\theta$ (tangential component of $W = mg$), $\theta'' + \omega_0{}^2\theta = 0$, $\omega_0/(2\pi) = \sqrt{g/L}/(2\pi)$.
5. No, because the frequency depends only on k/m.
7. (i) Greater by a factor $\sqrt{3}$. (ii) Lower
9. $\omega^* = [\omega_0{}^2 - c^2/(4m^2)]^{1/2} = \omega_0[1 - c^2/(4mk)]^{1/2} \approx \omega_0(1 - c^2/8mk) = 2.9583$
11. $2\pi/\omega^*$ since Eq. (10) and $y' = 0$ give $\tan(\omega^* t - \delta) = -\alpha/\omega^*$; tan is periodic with period π/ω^*.
13. Case (II) of (5) with $c = \sqrt{4mk} = \sqrt{4 \cdot 500 \cdot 4500} = 3000$ [kg/sec], where 500 kg is the mass per wheel.
15. $y = [y_0 + (v_0 + \alpha y_0)t]e^{-\alpha t}$, $y = [1 + (v_0 + 1)t]e^{-t}$; (ii) $v_0 = -2, -3/2, -4/3, -5/4, -6/5$
17. $y = 0$ gives $c_1 = -c_2e^{-2\beta t}$, which has one or no positive zero, depending on the initial conditions.

Problem Set 2.5, page 72

1. $c_1x^3 + c_2x^{-2}$
3. $(c_1 + c_2 \ln |x|)x^4$
5. $x[A \cos(\ln |x|) + B \sin(\ln |x|)]$
7. $c_1x^{1.4} + c_2x^{1.6}$
9. $c_1x^{0.1} + c_2x^{0.9}$
11. $3x^2 - 2x^3$
13. $x^{-0.5}[2 \cos(10 \ln |x|) - \sin(10 \ln |x|)]$
15. $2x^{-3} + 10$

Problem Set 2.6, page 77

1. $y'' - 0.25y = 0$, $W = -1$
3. $y'' - 2ky' + k^2y = 0$, $W = e^{2kx}$
5. $x^2y'' + 0.5xy' + 0.0625y = 0$, $W = x^{-0.5}$
7. $x^2y'' + xy' + 4y = 0$, $W = 2/x$
9. $x^2y'' - 0.75y = 0$, $W = -2$
11. $y'' - 6.25y = 0$, $W = 2.5$
13. $y'' + 2y' + 1.64y = 0$, $W = 0.8e^{-2x}$
15. $y'' + 5y' + 6.34y = 0$, $W = 0.3e^{-5x}$
17. $y'' + 7.6\pi y' + 14.44\pi^2y = 0$, $W = e^{-7.6\pi x}$

Problem Set 2.7, page 83

1. $c_1e^{-x} + c_2e^{-2x} + 2.5e^{2x}$
3. $c_1e^{4x} + c_2e^{-4x} + 2.4xe^{4x} - 4e^x$
5. $c_1e^{2x} + c_2e^{-3x} - x^3 - 3x - 0.5$
7. $e^{-3x}(A \cos 8x + B \sin 8x) + e^x(\cos 4x + \frac{1}{2} \sin 4x)$
9. $c_1e^{-0.4x} + c_2e^{0.4x} + 20xe^{0.4x} - 20xe^{-0.4x}$
11. $c_1 \cos 1.2x + c_2 \sin 1.2x + 10x \sin 1.2x$
13. $e^{-2x}(A \cos x + B \sin x) + 5x^2 - 8x + 4.4 - 1.6 \cos 2x + 0.2 \sin 2x$
15. $4x \sin 2x$

17. $e^{-0.1x}(1.5 \cos 0.5x - \sin 0.5x) + 2e^{0.5x}$

19. $2e^{-3x} + 3e^{4x} - 12x^3 + 3x^2 - 6.5x$

Problem Set 2.8, page 90

1. $-0.4 \cos 3t + 7.2 \sin 3t$

3. $-12.8 \cos 4.5t + 3.6 \sin 4.5t$

5. $0.16 \cos 2t + 0.12 \sin 2t$

7. $\frac{7}{45} \cos 3t - \frac{1}{45} \sin 3t$

9. $c_1e^{-t/2} + c_2e^{-3t/2} - \frac{32}{5} \cos t - \frac{4}{5} \sin t$

11. $(c_1 + c_2t)e^{-3t/2} - \frac{4}{3} \cos 3t - \sin 3t$

13. $e^{-1.5t}(A \cos t + B \sin t) + 4 + 0.8 \cos 2t - 6.4 \sin 2t$

15. $0.32e^{-t} \cos 5t + 0.68 \cos 3t + 0.24 \sin 3t$

17. $5e^{-4t} - 4e^{-2t} - 0.3 \cos 2t + 0.1 \sin 2t$

19. $e^{-1.5t}(0.2 \cos t - 1.1 \sin t) + 0.8 \cos t + 0.4 \sin t$

Problem Set 2.9, page 97

1. $LI' + RI = E$, $I = (E/R) + ce^{-Rt/L} = 2.4 + ce^{-50t}$

3. $RI' + I/C = 0$, $I = ce^{-t/(RC)}$

5. $I = 5(\cos t - \cos 10t)/99$

7. I_0 is maximum when $S = 0$; thus $C = 1/(\omega^2 L)$.

9. $R > R_{\text{crit}} = 2\sqrt{L/C}$ is Case I, etc.

11. 0

13. $c_1e^{-20t} + c_2e^{-10t} + 16.5 \sin 10t + 5.5 \cos 10t$

15. $E' = -e^{-4t}(7.605 \cos \frac{1}{2}t + 1.95 \sin \frac{1}{2}t)$, $I = e^{-0.1t}(A \cos \frac{1}{2}t + B \sin \frac{1}{2}t) - e^{-4t} \cos \frac{1}{2}t$

17. $E(0) = 600$, $I'(0) = 600$, $I = e^{-3t}(-100 \cos 4t + 75 \sin 4t) + 100 \cos t$

19. (*b*) $R = 2\ \Omega$, $L = 1$ H, $C = 1/12$ F, $E = 4.4 \sin 10t$ V

Problem Set 2.10, page 101

1. $A \cos x + B \sin x - x \cos x + (\sin x) \ln |\sin x|$

3. $c_1x + c_2x^2 - x \cos x$

5. $(\cos x)(c_1 + \sin x - \ln |\sec x + \tan x|) + (\sin x)(c_2 - \cos x)$
$= (c_1 - \ln |\sec x + \tan x|) \cos x + c_2 \sin x$

7. $(c_1 + \frac{3}{2}x) \sin x + (c_2 + \ln |\cos x|) \cos x$

9. $(c_1 + c_2x)e^x + x^2 + 4x + 6 - e^x(\ln |x| + 1)$

11. $c_1 \cos 2x + c_2 \sin 2x + \frac{1}{8}x \cosh 2x$

13. $c_1x + c_2x^2 - x \sin x$

15. $A \cos x + B \sin x + y_{p1} + y_{p2}$, y_{p1} as y_p in Example 1, $y_{p2} = \frac{10}{24} \sin 5x$

17. $u'' + u = 0$ by substitution of $y = ux^{-1/2}$. $y_1 = x^{-1/2} \cos x$, $y_2 = x^{-1/2} \sin x$, $y_p = -\frac{1}{2}x^{1/2} \cos x + \frac{1}{2}x^{-1/2} \sin x$ from (2) with the ODE in standard form.

Chapter 2 Review Questions and Problems, page 102

9. $c_1e^{4x} + c_2e^{-2x} - 1.1 \cos 6x - 0.3 \sin 6x$

11. $e^{-4x}(A \cos 3x + B \sin 3x) - \frac{3}{4} \cos 3x + \frac{1}{2} \sin 3x$

13. $y_1 = x^3$, $y_2 = x^{-4}$, $r = x^{-5}$, $W = -7x^{-2}$, $y_p = -\frac{1}{42}x^{-3} - \frac{1}{7}x^{-3} = -\frac{1}{6}x^{-3}$
15. $y_1 = e^x$, $y_2 = xe^x$, $W = e^{2x}$, $y_p = e^x/(2x)$
17. $y_1 = e^x \cos x$, $y_2 = e^x \sin x$, $W = e^{2x}$, $y_p = -xe^x \cos x + e^x(\sin x) \ln|\sin x|$
19. $y = 4e^{2x} + 2e^{-7x}$ **21.** $y = 9x^{-4} + 6x^6$
23. $y = e^{-2x} - 2e^{-3x} + 18x^2 - 30x + 19$ **25.** $y = \frac{1}{5}x^3 + 4x^2 - 5x^{-2}$
27. $y = -16 \cos 2t + 12 \sin 2t + 16(\cos 0.5t - \sin 1.5t)$.
Resonance for $\omega/(2\pi) = 2/(2\pi) = 1/\pi$
29. $\omega = 3.1$ is close to $\omega_0 = \sqrt{k/m} = 3$, $y = 25(\cos 3t - \cos 3.1t)$.
31. $R = 9\ \Omega$, $L = 0.5$ H, $C = 0.025$ F, $E = 17 \sin 6t$ V, hence $0.5I'' + 9I' + 40I = 102 \cos 6t$, $I = -8.16e^{-8t} + 7.5e^{-10t} + 0.66 \cos 6t + 1.62 \sin 6t$
33. $E' = 220 \cdot 314 \cos 314t$, $I = e^{-50t}(A \cos 150t + B \sin 150t) + 0.847001 \sin 314t - 1.985219 \cos 314t$

Problem Set 3.1, page 111

7. Linearly independent
9. Linearly dependent
11. $x|x| = x^2$ if $x > 0$, linearly dependent
13. Linearly independent
15. Linearly independent
17. Linearly independent
19. Linearly dependent

Problem Set 3.2, page 115

1. $y''' - 6y'' + 11y' - 6y = 0$ **3.** $y^{iv} - y = 0$
5. $y^{iv} + 4y'' = 0$ **7.** $c_1 + c_2 \cos x + c_3 \sin x$
9. $c_1e^x + (c_2 + c_3x)e^{-x}$ **11.** $c_1e^x + c_2e^{(1+\sqrt{7})x} + c_3e^{(1-\sqrt{7})x}$
13. $e^{0.25x} + 4.3e^{-0.7x} + 12.1 \cos 0.1x - 0.6 \sin 0.1x$
15. $2.4 + e^{-1.6x}(\cos 1.5x - 2 \sin 1.5x)$
17. $y = \cosh 5x - \cos 4x$
19. $y = c_1x^{-2} + c_2x + c_3x^2$. $W = 12/x^2$

Problem Set 3.3, page 122

1. $(c_1 + c_2x)e^{2x} + c_3e^{-2x} - 0.04e^{-3x} + x^2 + x + 1$
3. $c_1 \cos \frac{1}{2}x + c_2 \sin \frac{1}{2}x + x(c_3 \cos \frac{1}{2}x + c_4 \sin \frac{1}{2}x) - \frac{1}{2}e^{-x} \sin \frac{1}{2}x$
5. $c_1x^{0.5} + c_2x + c_3x^{1.5} + 0.1x^{5.5}$
7. $c_1 \cos x + c_2 \sin x + c_3 \cos 3x + c_4 \sin 3x + 0.2 \cosh 2x$
9. $y = (4 - x^2)e^{3x} - 0.5 \cos 3x + 0.5 \sin 3x$
11. $x^{-2} - x^2 + 5x^4 + x(\ln x + 1)$
13. $3 + 9e^{-2x} \cos 9x - (1.6 - 1.5x)e^x$

Chapter 3 Review Questions and Problems, page 122

7. $c_1 + c_2x^{1/2} + c_3x^{-1/2}$ **9.** $c_1e^{-0.5x} + c_2e^{0.5x} + c_3e^{-1.5x}$
11. $c_1x^2(\frac{1}{2} \ln x - \frac{3}{4}) + c_2x^2 + c_3x + c_4 + \frac{1}{7}x^7$
13. $c_1e^{-x} + e^{x/2}(c_2 \cos(\frac{1}{2}\sqrt{3}x) + c_3 \sin(\frac{1}{2}\sqrt{3}x)) + 8e^{x/2}$
15. $(c_1 + c_2x)e^x + c_3e^{-x} + 0.25x^2e^x$ **17.** $-0.5x^{-1} + 1.5x^{-5}$
19. $\cos 7x + e^{3x} - 0.02 \cosh x$

Problem Set 4.1, page 135

1. Yes **5.** $y_1' = 0.02(-y_1 + y_2)$, $y_2' = 0.02(y_1 - 2y_2 + y_3)$, $y_3' = 0.02(y_2 - y_3)$
7. $c_1 = 1$, $c_2 = -5$
9. 3 and 0
11. $y_1' = y_2$, $y_2' = 4y_1$, $y_1 = c_1e^{-2t} + c_2e^{2t} = y$, $y_2 = y_1'$
13. $y_1' = y_2$, $y_2' = y_2$, eigenvalues 0, 1, $y_1 = c_1 + c_2e^t$, $y_2 = y_1' = y'$
15. $y_1' = y_2$, $y_2' = 0.109375y_1 + 0.75y_2$ (divide by 64), $y_1 = c_1e^{-0.125t} + c_2e^{0.875t}$

Problem Set 4.3, page 146

1. $y_1 = c_1e^{-6t} + c_2e^{6t}$, $y_2 = -2c_1e^{-6t} + 2c_2e^{6t}$
3. $y_1 = c_1e^{2t} + c_2$, $y_2 = c_1e^{2t} - c_2$
5. $y_1 = c_1e^{4it} + c_2e^{-4it} = (c_1 + c_2)\cos 4t + i(c_1 - c_2)\sin 4t$
$= A\cos 4t + B\sin 4t$, $y_2 = ic_1e^{4it} - ic_2e^{-4it}$
$= (ic_1 - ic_2)\cos 4t + i(ic_1 + ic_2)\sin 4t = B\cos 4t - A\sin 4t$, $A = c_1 + c_2$, $B = i(c_1 - c_2)$
7. $y_1 = 2c_1 + c_2e^{-6t}$, $y_2 = -c_1 + c_3e^{-6t}$, $y_3 = -c_1 + 2(c_2 + c_3)e^{-6t}$
9. $y_1 = c_1e^{1.8t} + 2c_2e^{-0.9t} + 2c_3e^{-1.8t}$, $y_2 = 2c_1e^{1.8t} + c_2e^{-0.9t} - 2c_3e^{-1.8t}$, $y_3 = 2c_1e^{1.8t} - 2c_2e^{-0.9t} + c_3e^{-1.8t}$
11. $y_1 = 10 + 6e^{2t}$, $y_2 = -5 + 3e^{2t}$
13. $y_1 = 2.4e^{-t} - 2e^{2.5t}$, $y_2 = 1.8e^{-t} + 2e^{2.5t}$
15. $y_1 = 2e^{14.5t} + 10$, $y_2 = 5e^{14.5t} - 4$
17. $y_2 = y_1' + y_1$, $y_2' = y_1'' + y_1' = -y_1 - y_2 = -y_1 - (y_1' + y_1)$, $y_1'' + 2y_1' + 2y_1 = 0$, $y_1 = e^{-t}(A\cos t + B\sin t)$, $y_2 = y_1' + y_1 = e^{-t}(B\cos t - A\sin t)$. Note that $r^2 = {y_1}^2 + {y_2}^2 = e^{-2t}(A^2 + B^2)$.
19. $I_1 = 4c_1e^{-200t} + c_2e^{-50t}$, $I_2 = -c_1e^{-200t} - 4c_2e^{-50t}$

Problem Set 4.4, page 150

1. Saddle point, unstable, $y_1 = c_1e^{-4t} + c_2e^{4t}$, $y_2 = -2c_1e^{-4t} + 2c_2e^{4t}$
3. Unstable node, $y_1 = c_1e^t + c_2e^{3t}$, $y_2 = -c_1e^t + c_2e^{3t}$
5. Stable and attractive node, $y_1 = c_1e^{-3t} + c_2e^{-5t}$, $y_2 = c_1e^{-3t} - c_2e^{-5t}$
7. Center, stable, $y_1 = A\cos 4t + B\sin 4t$, $y_2 = -2B\cos 4t + 2A\sin 4t$
9. Saddle point, unstable, $y_1 = c_1e^{3t} + c_2e^{-t}$, $y_2 = c_1e^{3t} - c_2e^{-t}$
11. $y_1 = y = c_1e^{kt} + c_2e^{-kt}$, $y_2 = y'$, hyperbolas $k^2{y_1}^2 - {y_2}^2 = const$
13. $y = e^{-2t}(A\cos t + B\sin t)$, stable and attractive spirals
17. For instance, (a) -2, (b) -1, (c) $-\frac{1}{2}$, (d) 1, (e) 4.

Problem Set 4.5, page 158

1. (0, 0), $y_1' = y_2$, $y_2' = 3y_1$, saddle point; (0, -1), $y_1 = \tilde{y}_1$, $y_2 = -1 + \tilde{y}_2$, $\tilde{y}_1' = -\tilde{y}_2$, $\tilde{y}_2' = 3\tilde{y}_1$, center
3. (0, 0), $y_1' = 4y_2$, $y_2' = 2y_1$, saddle point; (2, 0), $y_1 = 2 + \tilde{y}_1$, $y_2 = \tilde{y}_2$, $\tilde{y}_1' = 4\tilde{y}_2$, $\tilde{y}_2' = -2\tilde{y}_1$, center
5. (0, 0), $y_1' = -y_1 + y_2$, $y_2' = -y_1 - y_2$, stable and attractive spiral point; (-2, 2), $y_1 = -2 + \tilde{y}_1$, $y_2 = 2 + \tilde{y}_2$, $\tilde{y}_1' = -\tilde{y}_1 - 3\tilde{y}_2$, $\tilde{y}_2' = -\tilde{y}_1 - \tilde{y}_2$, saddle point

7. $y_1' = y_2$, $\tilde{y}_2' = -y_1(1 - 4y_1)$, $(0, 0)$, $y_1' = y_2$, $y_2' = -y_1$, center; $(\frac{1}{4}, 0)$, $y_1 = \frac{1}{4} + \tilde{y}_1$, $y_2 = \tilde{y}_2$, $\tilde{y}_1' = \tilde{y}_2$, $\tilde{y}_2' = (-\frac{1}{4} - \tilde{y}_1)(-4\tilde{y}_1)$, $\tilde{y}_2' = \tilde{y}_1$, saddle
9. $(\frac{1}{2}\pi \pm 2n\pi, 0)$ saddle points; $(-\frac{1}{2}\pi \pm 2n\pi, 0)$ centers. Use $-\cos(\pm\frac{1}{2}\pi + \tilde{y}_1) = \sin(\pm\tilde{y}_1) \approx \pm\tilde{y}_1$.
11. $y_1' = y_2$, $y_2' = -y_1(2 + y_1)(2 - y_1)$. $(0, 0)$, $y_2' = -4y_1$, center; $(-2, 0)$, $\tilde{y}_2' = 8\tilde{y}_1$, saddle point; $(2, 0)$, $\tilde{y}_2' = 8\tilde{y}_1$ saddle point
13. $y''/y' + 2y'/y = 0$, $\ln y' + 2 \ln y = c$, $y'y^2 = y_2{y_1}^2 = const$
15. $y = A \cos t + B \sin t$, radius $\sqrt{A^2 + B^2}$

Problem Set 4.6, page 162

3. $y_1 = A \cos 4t + B \sin 4t + \frac{29}{16}$, $y_2 = B \cos 4t - A \sin 4t - \frac{9}{4}t$
5. $y_1 = c_1e^{4t} + c_2e^{-3t} + 4$, $y_2 = c_1e^{4t} - 2.5c_2e^{-3t} - 10$
7. $y_1 = 2c_1e^{-9t} + c_2e^{-4t} - 90t + 28$, $y_2 = c_1e^{-9t} + c_2e^{-4t} - 126t + 14$
9. $y_1 = c_1e^{t} + 4c_2e^{2t} - 3t - 4 - 2e^{-t}$, $y_2 = -c_1e^{t} - 5c_2e^{2t} + 5t + 7.5 + e^{-t}$
11. $y_1 = 3 \cos 2t - \sin 2t + t + 1$, $y_2 = \cos 2t + 3 \sin 2t + 2t - \frac{1}{2}$
13. $y_1 = 4e^{-t} - 4e^{t} + e^{2t}$, $y_2 = -4e^{-t} + t$
15. $y_1 = 7 - 2e^{2t} + e^{3t} - 4e^{-3t}$, $y_2 = -e^{2t} + 3e^{-3t}$
17. $I_1' + 2.5(I_1 - I_2) = 845 \sin t$, $2.5(I_2' - I_1') + 25I_2 = 0$,
$I_1 = (95 + 162.5t)e^{-5t} - 95 \cos t + 312.5 \sin t$,
$I_2 = (-30 - 162.5t)e^{-5t} + 30 \cos t + 12.5 \sin t$
19. $I_1' + 2(I_1 - I_2) = 200$, $2(I_2 - I_1) + 8I_2 + 2\int I_2\,dt = 0$.
$I_1 = 2c_1e^{\lambda_1 t} + 2c_2e^{\lambda_2 t} + 100$,
$I_2 = (1.1 + \sqrt{0.41})c_1e^{\lambda_1 t} + (1.1 - \sqrt{0.41})c_2e^{\lambda_2 t}$, $\lambda_1 = -0.9 + \sqrt{0.41}$,
$\lambda_2 = -0.9 - \sqrt{0.41}$

Chapter 4 Review Questions and Problems, page 163

11. $y_1 = c_1e^{8t} + c_2e^{-8t}$, $y_2 = 2c_1e^{8t} - 2c_2e^{-8t}$. Saddle point
13. $y_1 = c_1e^{t} + c_2e^{-6t}$, $y_2 = c_1e^{t} - 6c_2e^{-6t}$. Saddle point
15. $y_1 = c_1e^{7.5t} + c_2e^{-3t}$, $y_2 = -c_1e^{7.5t} + 0.75c_2e^{-3t}$. Saddle point
17. $y_1 = c_1e^{5t} + c_2e^{t}$, $y_2 = c_1e^{5t} - c_2e^{t}$. Unstable node
19. $y_1 = e^{-t}(A \cos 2t + B \sin 2t)$, $y_2 = e^{-t}(B \cos 2t - A \sin 2t)$. Stable and attractive spiral point
21. $y_1 = c_1e^{t} + c_2e^{-t} + e^{2t} + e^{-2t}$, $y_2 = -c_2e^{-t} - 1.5e^{-2t}$
23. $y_1 = c_1e^{t} + c_2e^{-2t} - 6e^{-t} - 5$, $y_2 = -c_1e^{t} - 2c_2e^{-2t} + 10e^{-t} + 6$
25. $y_1 = c_1e^{3t} + c_2e^{-t} + t^2 - 2t + 2$, $y_2 = c_1e^{3t} - c_2e^{-t} - t^2 + 2t - 2$
27. A saddle point at $(0, 0)$
29. $I_1 = 4e^{-40t} - e^{-10t}$, $I_2 = -e^{-40t} + 4e^{-10t}$
31. $(n\pi, 0)$ center for even n and saddle point for odd n
33. Saddle points at $(0, 0)$ and $(\frac{1}{2}, \frac{1}{2})$, centers at $(0, \frac{1}{2})$ and $(\frac{1}{2}, 0)$

Problem Set 5.1, page 170

1. $a_0(1 + x + \frac{1}{2}x^2 + \cdots) = a_0e^{x}$
3. $a_0(1 - 2x^2 + \frac{2}{3}x^4 - + \cdots) + a_1(x - \frac{2}{3}x^3 + \frac{2}{15}x^5 - + \cdots)$
$= a_0 \cos 2x + \frac{1}{2}a_1 \sin 2x$

5. $a_0(1 + \frac{1}{2}x)$

7. $a_0 + a_0x + (\frac{1}{2}a_0 + \frac{1}{2})x^2 + \cdots = a_0e^x + e^x - x - 1 = ce^x - x - 1,\ c = a_0 + 1$

9. $a_0 + a_1x + \frac{1}{2}a_1x^2 + \cdots = a_0 - a_1 + a_1e^x$

11. $s = \frac{5}{4} - 4x + 8x^2 - \frac{32}{3}x^3 + \frac{32}{3}x^4 - \frac{128}{15}x^5,\quad s(0.2) = 0.69900$

13. $s = \frac{1}{2} + \frac{1}{4}x - \frac{1}{48}x^3 + \frac{1}{480}x^5,\quad s(1) = 0.73125$

15. $s = 1 + x - x^2 - \frac{5}{6}x^3 + \frac{2}{3}x^4 + \frac{11}{24}x^5,\quad s(\frac{1}{2}) = \frac{923}{768}$

Problem Set 5.2, page 176

1. $|c|$

3. 2 (as function of $t = (x - 3)^2$). *Ans.* $\sqrt{2}$

5. 0

7. 2

9. 1

11. π

13. $\displaystyle\sum_{s=3}^{\infty} \frac{(-1)^{s-1}}{5(s-2)}\, x^s;\ R = 1$

15. $\displaystyle\sum_{s=5}^{\infty} \frac{(s-4)^2}{(s-3)!}\, x^s;\ R = \infty$

17. $a_0(1 - \frac{1}{12}x^4 - \frac{1}{60}x^5 - \cdots) + a_1(x + \frac{1}{2}x^2 + \frac{1}{6}x^3 + \frac{1}{24}x^4 - \frac{1}{24}x^5 - \cdots)$

19. $a_0 + a_1(x - \frac{2}{3}x^3 + \frac{2}{5}x^5 - \frac{4}{21}x^7 + \frac{2}{27}x^9 - \frac{4}{165}x^{11} + - \cdots)$

21. $a_0(1 - \frac{1}{2}x^2 - \frac{1}{24}x^4 + \frac{13}{720}x^6 + \cdots) + a_1(x - \frac{1}{6}x^3 - \frac{1}{24}x^5 + \frac{5}{1008}x^7 + \cdots)$

23. $a_0(1 + x^2 + x^3 + x^4 + x^5 + x^6 + \cdots) + a_1x$

Problem Set 5.3, page 180

3. $P_6(x) = \frac{1}{16}(231x^6 - 315x^4 + 105x^2 - 5)$,
$P_7(x) = \frac{1}{16}(429x^7 - 693x^5 + 315x^3 - 35x)$

7. Set $x = az$. $y = c_1P_n(x/a) + c_2Q_n(x/a)$

15. $P_1^{\,1} = \sqrt{1 - x^2}$, $P_2^{\,1} = 3x\sqrt{1 - x^2}$, $P_2^{\,2} = 3(1 - x^2)$,
$P_4^{\,2} = (1 - x^2)(105x^2 - 15)/2$

Problem Set 5.4, page 187

1. $\displaystyle y_1 = 1 + \frac{x^2}{3!} + \frac{x^4}{5!} + \cdots = \frac{\sinh x}{x},\quad y_2 = \frac{1}{x} + \frac{x}{2!} + \frac{x^3}{4!} + \cdots = \frac{\cosh x}{x}$

3. $\displaystyle y_1 = 1 - \frac{x^2}{12} + \frac{1}{384}x^4 - \cdots,\ y_2 = 9y_1 \ln x - \frac{144}{x^4} - \frac{36}{x^2} + \frac{x^2}{2} - \frac{25x^4}{1024} + \cdots$

5. $r(r - 1) + 4r + 2 = 0,\ r_1 = -1,\ r_2 = -2;\ y_1 = \dfrac{1}{x} - \dfrac{x}{6} + \dfrac{x^3}{120} - + \cdots,$

$\displaystyle y_2 = \frac{1}{x^2} - \frac{1}{2} + \frac{x^2}{24} - \frac{x^4}{720} + - \cdots$

7. Euler-Cauchy equation with $t = x + 3$, $y_1 = (x + 3)^5$, $y_2 = y_1 \ln (x + 3)$

9. $b_0 = 1,\ c_0 = 0,\ r^2 = 0,\ y_1 = e^{-x},\ y_2 = e^{-x} \ln x$

11. $y_1 = 1/(x + 1),\ y_2 = 1/x$

13. $b_0 = \frac{1}{2},\ c_0 = 0,\ r_1 = \frac{1}{2},\ r_2 = 0,\ y_1 = x^{1/2}(1 + 2x + 2x^2 + \frac{4}{3}x^3 + \cdots)$,
$y_2 = 1 + 2x + 2x^2 + \cdots$

15. $y_1 = (x - 4)^7,\ y_2 = (x - 4)^{-5}$ (Euler-Cauchy with $t = x - 4$)

17. $y_1 = x + x^3 - \frac{1}{12}x^4 + \frac{3}{10}x^5 - \frac{1}{20}x^6 + \cdots,\ y_2 = 1 + 3x^2 - \frac{1}{6}x^3 + \frac{3}{2}x^4 - \frac{1}{5}x^5$
$+ \frac{11}{36}x^6 - + \cdots$

19. $y = c_1F(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}; x) + c_2\sqrt{x}F(1, 1, \frac{3}{2}; x)$
21. $y = A(1 - 4x + \frac{8}{3}x^2) + B\sqrt{x}F(-\frac{3}{2}, \frac{3}{2}, \frac{3}{2}; x)$
23. $y = c_1F(2, -2, -\frac{1}{2}; t - 2) + c_2(t - 2)^{3/2}F(\frac{7}{2}, -\frac{1}{2}, \frac{5}{2}; t - 2)$

Problem Set 5.5, page 197

1. Use (7b) in Sec. 5.2.
3. 0.77958 (exact 0.76520), 0.19674 (0.22389), −0.27651 (−0.26005), −0.39788 (−0.39715), −0.17038 (−0.17760), 0.15680 (0.15065), 0.30086 (0.30008), 0.16833 (0.17165)
5. $y = c_1J_\nu(\lambda x) + c_2J_{-\nu}(\lambda x)$, $\nu \neq 0, \pm 1, \cdots$
7. $y = c_1J_\nu(\sqrt{x}) + c_2J_{-\nu}(\sqrt{x})$, $\nu \neq 0, \pm 1, \cdots$
9. $y = c_1xJ_1(2x)$, J_1, J_{-1} linearly dependent
11. $y = x^{-\nu}[c_1J_\nu(x) + c_2J_{-\nu}(x)]$, $\nu \neq 0, \pm 1, \cdots$
13. $y = c_1J_\nu(x^3) + c_2J_{-\nu}(x^3)$, $\nu \neq 0, \pm 1, \cdots$
15. $y = c_1\sqrt{x}J_1(2\sqrt{x})$, J_1, J_{-1} linearly dependent
17. $y = x^{1/4}J_1(\frac{2}{3}x^{1/4})$, J_1, J_{-1} linearly dependent
19. $y = x^{2/5}(c_1J_{8/5}(4x^{1/4}) + c_2J_{-8/5}(4x^{1/4}))$
21. Use (24b) with $\nu = 0$, (24a) with $\nu = 1$, (24d) with $\nu = 2$, respectively.
23. $J_n(x_1) = J_n(x_2) = 0$ implies ${x_1}^{-n}J_n(x_1) = {x_2}^{-n}J_n(x_2) = 0$ and $[x^{-n}J_n(x)]' = 0$ somewhere between x_1 and x_2 by Rolle's theorem. Now use (24b) to get $J_{n+1}(x) = 0$ there. Conversely, $J_{n+1}(x_3) = J_{n+1}(x_4) = 0$, thus ${x_3}^{n+1}J_{n+1}(x_3) = {x_4}^{n+1}J_{n+1}(x_4) = 0$ implies $J_n(x) = 0$ in between by Rolle's theorem and (24a) with $\nu = n + 1$.
25. Integrate the formulas in (24).
27. Use (24a) with $\nu = 1$, partial integration, (24b) with $\nu = 0$, partial integration.
33. CAS Experiment. (b) $x_0 = 1$, $x_1 = 2.5$, $x_2 = 20$, approximately. It increases with n. (c) (14) is exact. (d) It oscillates. (e) Formula (24b) with $\nu = 0$

Problem Set 5.6, page 202

1. $y = c_1J_5(x) + c_2Y_5(x)$
3. $y = c_1J_0(\sqrt{x}) + c_2Y_0(\sqrt{x})$
5. $y = c_1J_2(x^2) + c_2Y_2(x^2)$
7. $y = x^{-5}(c_1J_5(x) + c_2Y_5(x))$
9. $y = x^3(c_1J_3(x^3) + c_2Y_3(x^3))$
11. Set $H^{(1)} = kH^{(2)}$, use (10).
13. Set $x = is$ in (1), Sec. 5.5, to get the present ODE (12) in terms of s. Use (20), Sec. 5.5.

Problem Set 5.7, page 209

3. Set $x = ct + k$.
5. $x = \cos\theta$, $dx = -\sin\theta\, d\theta$, etc.
7. $\lambda_m = (m\pi/5)^2$, $m = 1, 2, \cdots$; $y_m = \sin(m\pi x/5)$
9. $\lambda_m = [(2m + 1)\pi/2L]^2$, $m = 0, 1, \cdots$; $y_m(x) = \sin[(2m + 1)\pi x/2L]$
11. $\lambda_m = m^2$, $m = 0, 1, \cdots$; $y_0 = 1$, $y_m = \cos mx, \sin mx$, $m = 1, 2, \cdots$
13. $k = k_m$ from $\tan k = -k$. $\lambda_m = {k_m}^2$, $m = 1, 2, \cdots$; $y_m = \sin k_m x$
15. $\lambda_m = m^2$, $m = 1, 2, \cdots$; $y_m = x\sin(m \ln|x|)$
17. $p = e^{8x}$, $q = 0$, $r = e^{8x}$, $\lambda_m = m^2$; $y_m = e^{-4x}\sin mx$, $m = 1, 2, \cdots$
19. $\lambda_m = (m\pi)^2$, $y_m = x\cos m\pi x, x\sin m\pi x$, $m = 0, 1, \cdots$

Problem Set 5.8, page 216

1. $1.6P_4(x) - 0.6P_0(x)$ **3.** $\frac{2}{5}P_3(x) - \frac{2}{3}P_2(x) + \frac{8}{5}P_1(x) - \frac{4}{3}P_0(x)$

7. $-0.4775P_1(x) - 0.6908P_3(x) + 1.844P_5(x) - 0.8234P_7(x) + 0.1544P_9(x) + \cdots$, $m_0 = 9$. Rounding seems to have considerable influence in Probs. 6–15.

9. $0.3799P_2(x) + 1.673P_4(x) - 1.397P_6(x) + 0.3968P_8(x) + \cdots$, $m_0 = 8$

11. $1.175P_0(x) + 1.104P_1(x) + 0.3575P_2(x) + 0.0700P_3(x) - \cdots$, $m_0 = 3$ or 4

13. $0.7855P_0(x) - 0.3550P_2(x) + 0.0900P_4(x) - \cdots$, $m_0 = 4$

15. $0.1212P_0(x) - 0.7955P_2(x) + 0.9600P_4(x) - 0.3360P_6(x) + \cdots$, $m_0 = 8$

17. (c) $a_m = (2/J_1{}^2(\alpha_{0,m}))(J_1(\alpha_{0,m})/\alpha_{0,m}) = 2/(\alpha_{0,m}J_1(\alpha_{0,m}))$

Chapter 5 Review Questions and Problems, page 217

11. e^{3x}, e^{-3x}, or $\cosh 3x$, $\sinh 3x$ **13.** e^x, $1 + x$

15. e^{-x^2}, xe^{-x^2} **17.** e^{-x}, $e^{-x}\ln x$

19. $1/(1 - x^2)$, $x/(1 - x^2)$ or $1/(1 - x)$, $1/(1 + x)$

21. $y = c_1J_{\sqrt{2}}(6x) + c_2J_{-\sqrt{2}}(6x)$ **23.** $y = c_1J_1(x^2) + c_2Y_1(x^2)$

25. $y = \sqrt{x}[c_1J_{1/4}(\frac{1}{2}kx^2) + c_2J_{-1/4}(\frac{1}{2}kx^2)]$

27. $\lambda_m = (2m\pi)^2$, $y_0 = 1$, $y_m = \cos 2m\pi x$, $\sin 2m\pi x$, $m = 1, 2, \cdots$

29. $y = c_1J_1(kx) + c_2Y_1(kx)$, $c_2 = 0$, $y(1) = c_1J_1(k) = 0$, $k = k_m = \alpha_{1,m}$ (the positive zeros of J_1), $y_m = J_1(\alpha_{1,m}x)$

31. $1.813P_0(x) + 2.923P_1(x) + 1.759P_2(x) + 0.663P_3(x) + 0.185P_4(x) + \cdots$

33. $0.693P_0(x) - 0.285P_2(x) + 0.144P_4(x) - 0.091P_6(x) + \cdots$

35. $0.25P_0(x) + 0.5P_1(x) + 0.3125P_2(x) - 0.0938P_4(x) + 0.0508P_6(x) + \cdots$

Problem Set 6.1, page 226

1. $\dfrac{2}{s^3} - \dfrac{2}{s^2}$ **3.** $\dfrac{s}{s^2 + 4\pi^2}$ **5.** $\dfrac{s - 2}{(s - 2)^2 - 1}$

7. $\dfrac{s\cos\theta - \omega\sin\theta}{s^2 + \omega^2}$ **9.** $\dfrac{e^{3a}}{s + 2b}$ **11.** $\dfrac{1}{s^2 + 4}$

13. $\dfrac{k}{s}(1 - e^{-bs})$ **15.** $\dfrac{1 - (1 + 2s)e^{-2s}}{2s^2}$ **17.** $\dfrac{1 - e^{-bs}}{s^2} - \dfrac{be^{-bs}}{s}$

19. $\dfrac{(1 - e^{-s})^2}{s}$ **21.** $\mathscr{L}(f_1) = \mathscr{L}(1) - \mathscr{L}(f) = \dfrac{1}{s} - \dfrac{1}{s}(1 - e^{-2s}) = e^{-2s}/s$

23. Set $ct = p$. Then $\mathscr{L}(f(ct)) = \displaystyle\int_0^\infty e^{-st}f(ct)\,dt = \int_0^\infty e^{-(s/c)p}f(p)\,dp/c = F(s/c)/c.$

29. $4\cos\pi t - 3\sin\pi t$ **31.** $1 - \frac{3}{2}t^2 + \frac{1}{2}t^4$ **33.** $\sin\dfrac{n\pi t}{L}$

35. $2 - 2e^{-4t}$ **37.** $(e^{\sqrt{3}t} - e^{-\sqrt{5}t})/(\sqrt{3} + \sqrt{5})$

39. $\dfrac{1}{\sqrt{5}}\sin\sqrt{5}t - e^{-5t}$ **41.** $\dfrac{3.8}{(s - 2.4)^2}$ **43.** $\dfrac{5\omega}{(s + a)^2 + \omega^2}$

45. $\dfrac{a(s + k) + b}{(s + k)^2 + 1}$ **47.** $3.5t^2e^t$ **49.** $\sqrt{2}t^2e^{-t\sqrt{2}}$

51. $3e^{-2t}\sin 5t$ **53.** $e^{-5\pi t}\sinh\pi t$

Problem Set 6.2, page 232

1. $\dfrac{1}{(s-k)^2}$ **3.** $\dfrac{2\omega^2}{s(s^2+4\omega^2)}$ **5.** $\dfrac{2a^2}{s(s^2-4a^2)}$

7. $\dfrac{\pi s}{(s^2+\frac{1}{4}\pi^2)^2}$

9. Use shifting. Use $\cos^2\alpha = \frac{1}{2} + \frac{1}{2}\cos 2\alpha$; use $\cos^2\alpha + \sin^2\alpha = 1$.
Ans. $(2s^2+1)/[2s(s^2+1)]$

11. $(s+\frac{1}{2})Y = -1 + 17\cdot 2/(s^2+4)$, $y = 7e^{-t/2} + 2\sin 2t - 8\cos 2t$

13. $(s^2-\frac{1}{4})Y = 4s$, $y = 4\cosh\frac{1}{2}t$

15. $(s^2+2s+2)Y = s - 3 + 2\cdot 1$, $Y = (s+1-2)/[(s+1)^2+1]$,
$y = e^{-t}(\cos t - 2\sin t)$

17. $(s^2+7s+12)Y = 3.5s - 10 + 24.5 + 21/(s-3)$, $y = \frac{1}{2}e^{3t} + \frac{5}{2}e^{-4t} + \frac{1}{2}e^{-3t}$

19. $(s+1.5)^2Y = s + 31.5 + 3 + 54/s^4 + 64/s$,
$Y = 1/(s+1.5) + 1/(s+1.5)^2 + 24/s^4 - 32/s^3 + 32/s^2$,
$y = (1+t)e^{-1.5t} + 4t^3 - 16t^2 + 32t$

21. $t = \tilde{t} + 2$, $\tilde{Y} = 4/(s-6)$, $\tilde{y} = 4e^{6t}$, $y = 4e^{6(t-2)}$

23. $t = \tilde{t} + 1$, $(s-1)(s+4)\tilde{Y} = 4s + 17 + 6/(s-2)$, $y = 3e^{t-1} + e^{2(t-1)}$

25. (b) In the proof, integrate from 0 to a and then from a to ∞ and see what happens.
(c) Find $\mathscr{L}(f)$ and $\mathscr{L}(f')$ by integration and substitute them into (1*).

27. $2 - 2e^{-t/2}$ **29.** $\dfrac{1}{k^2}(e^{kt}-1) - \dfrac{t}{k}$ **31.** $\cosh\sqrt{5}t - 1$

33. $\frac{1}{8}\sinh 2t - \frac{1}{4}t$

Problem Set 6.3, page 240

3. $(1 - e^{2-2s})/(s-1)$

5. $\left(\dfrac{2}{s^3} + \dfrac{2}{s^2} + \dfrac{1}{s}\right)e^{-s} - \left(\dfrac{2}{s^3} + \dfrac{4}{s^2} + \dfrac{4}{s}\right)e^{-2s}$

7. $\dfrac{s}{s^2+\pi^2}(-e^{-s} - e^{-4s})$ **9.** $\left(\dfrac{1}{s^2} + \dfrac{5}{s}\right)e^{-5s} - \left(\dfrac{1}{s^2} + \dfrac{10}{s}\right)e^{-10s}$

11. $\dfrac{-20s}{s^2+\pi^2}(e^{-3s} + e^{-6s})$ **13.** $\dfrac{1}{s-\pi}(e^{-2s+2\pi} - e^{-4s+4\pi})$

15. 0 if $t < 4$, $t - 4$ if $t > 4$ **17.** $\sin t$ if $2\pi < t < 8\pi$, 0 elsewhere

19. 0 if $t < 2$, $(t-2)^4/24$ if $t > 2$ **21.** $u(t-3)\cosh(2t-6)$

23. $e^{-t}\sin t$ **25.** $e^{-2t}\cos 3t + 9\cos 2t + 8\sin 2t$

27. $\sin 3t + \sin t$ if $0 < t < \pi$ and $\frac{4}{3}\sin 3t$ if $t > \pi$

29. $t - \sin t$ if $0 < t < 1$, $\cos(t-1) + \sin(t-1) - \sin t$ if $t > 1$

31. $e^t - \sin t + u(t-2\pi)(\sin t - \frac{1}{2}\sin 2t)$

33. $t = 1 + \tilde{t}$, $\tilde{y}'' + 4\tilde{y} = 8(1+\tilde{t})^2(1 - u(\tilde{t}-4))$, $\cos 2t + 2t^2 - 1$ if $t < 5$,
$\cos 2t + 49\cos(2t-10) + 10\sin(2t-10)$ if $t > 5$

35. $Rq' + q/C = 0$, $Q = \mathscr{L}(q)$, $q(0) = CV_0$, $i = q'(t)$, $R(sQ - CV_0) + Q/C = 0$,
$q = CV_0e^{-t/(RC)}$

37. $10I + \dfrac{100}{s}I = \dfrac{100}{s^2}e^{-2s}$, $I = e^{-2s}\left(\dfrac{1}{s} - \dfrac{1}{s+10}\right)$, $i = 0$ if $t < 2$ and
$1 - e^{-10(t-2)}$ if $t > 2$

39. $i = e^{-20t} + 20t - 1 + u(t-2)[-20t + 1 + 39e^{-20(t-2)}]$

41. $0.1i' + 25i = 490e^{-5t}[1 - u(t-1)]$, $i = 20(e^{-5t} - e^{-250t}) + 20u(t-1)[-e^{-5t} + e^{-250t+245}]$

43. $i = (10 \sin 10t + 100 \sin t)(u(t-\pi) - u(t-3\pi))$

45. $i' + 2i + 2\int_0^t i(\tau)\,d\tau = 1 - u(t-2)$, $I = (1 - e^{-2s})/(s^2 + 2s + 2)$,
$i = e^{-t} \sin t - u(t-2)\, e^{-t+2} \sin (t-2)$

47. $i = 27 \cos t + 6 \sin t - e^{-t}(27 \cos 3t + 11 \sin 3t)$
$+ u(t - 2\pi)\,[-27 \cos t - 6 \sin t + e^{-(t-2\pi)}(27 \cos 3t + 11 \sin 3t)]$

Problem Set 6.4, page 247

1. $y = 10 \cos t$ if $0 < t < 2\pi$ and $10 \cos t + \sin t$ if $t > 2\pi$

3. $y = 5.5e^t + 4.5e^{-t} + 5(e^{t-1/2} - e^{-t+1/2})u(t - \frac{1}{2}) - 50(e^{t-1} - e^{-t+1})u(t-1)$

5. $y = 0.1[e^t + e^{-2t}(-\cos t + 7 \sin t)]$
$+ 0.1u(t-10)[-e^t + e^{-2t+30}(\cos (t-10) - 7 \sin (t-10))]$

7. $y = 1 + \frac{1}{3}e^{-t} \sin 3t + u(t-4)[-1 + e^{-t+4}(\cos (3t-12) + \frac{1}{3} \sin (3t-12))]$
$- \frac{10}{3}u(t-5)e^{-t+5} \sin (3t-15)$

9. $y = 5t - 2 - 50u(t-\pi)e^{-t+\pi} \sin 2t$. Straight line, sharply deformed between π and about 8

11. $y = (0.4t + 1.52)e^t + 0.48e^{-4t} + 1.6u(t-2)[-e^t + e^{-4t+10}]$

Problem Set 6.5, page 253

1. t

3. $e^t - t - 1$

5. $\dfrac{1}{\omega} \sin \omega t$

7. $\dfrac{1}{2k}(e^{kt} - e^{-kt}) = \dfrac{1}{k} \sinh kt$

9. $\frac{1}{8}(e^{3t} - e^{-5t})$

11. $\frac{1}{2}(\frac{1}{2} - \frac{1}{2} \cos 2t) = \frac{1}{2} \sin^2 t$

13. $t - \sin t$

15. $\frac{1}{9}(\cosh 3t - 1)$

19. $Y = 3/((s^2+4)(s^2+9))$, $y = 0.3 \sin 2t - 0.2 \sin 3t$

21. $(s^2+9)Y = 4 + 8(1 + e^{-\pi s})/(s^2+1)$, $y = \sin t + \sin 3t$ if $t < \pi$, $\frac{4}{3} \sin 3t$ if $t > \pi$

23. 0 if $0 < t < 1$, $\frac{5}{2}\int_1^t \sin (2(\tau - 1))\,d\tau = -\frac{5}{4} \cos (2t-2) + \frac{5}{4}$ if $t > 1$

25. $y = 2e^{-2t} - e^{-4t} + (e^{-2t+2} - e^{-4t+4})u(t-1) + (e^{-2t+4} - e^{-4t+8})u(t-2)$

27. $y - 1 * y = 1$, $y = e^t$

29. $y - y * \sin t = \cos t$, $Y = 1/s$, $y = 1$

31. $Y(1 + 1/s^2) = 1/s$, $y = \cos t$

33. $Y(1 + 2/(s-1)) = (s-1)^{-2}$, $y = \sinh t$

Problem Set 6.6, page 257

1. $\dfrac{4}{(s-1)^2}$

3. $\dfrac{2\omega s}{(s^2+\omega^2)^2}$

5. $\dfrac{2s+4}{(s^2+4s+5)^2}$

7. $\dfrac{24s^2 + 128}{(s^2-16)^3}$

9. $\dfrac{2\omega(3s^2 - \omega^2)}{(s^2+\omega^2)^3}$

11. $\dfrac{2s \cos k + (s^2-1) \sin k}{(s^2+1)^2}$

13. $6te^{-t}$
15. $te^{-2t}\sin t$
17. t^2e^{kt}
19. $\ln s - \ln(s-1);\ (e^t-1)/t$

Problem Set 6.7, page 262

1. $y_1 = -e^{-t}\sin t,\quad y_2 = e^{-t}\cos t$
3. $y_1 = 2e^{-4t} - 4e^{2t},\quad y_2 = e^{-4t} - 8e^{2t}$
5. $y_1 = 2e^{-t} + 4e^{-2t} + \frac{1}{2}t - \frac{1}{4},\quad y_2 = -3e^{-t} - 4e^{-2t} - \frac{1}{2}t + \frac{1}{4}$
7. $y_1 = e^{-t}(2\cos 2t + 6\sin 2t) + t^2,\quad y_2 = 10e^{-t}\sin 2t - t^2$
9. $y_1 = 4\cos 5t + 6\sin 5t - 2\cos t - 25\sin t,\quad y_2 = 2\cos 5t - 10\sin 5t + 20\sin t$
11. $y_1 = -\cos t + \sin t + 1 + u(t-1)[-1 + \cos(t-1) - \sin(t-1)]$
$y_2 = \cos t + \sin t - 1 + u(t-1)[1 - \cos(t-1) - \sin(t-1)]$
13. $y_1 = 2u(t-2)(e^{4t} - e^{t+6}),\quad y_2 = e^{2t} + u(t-2)(e^{4t} - 3e^{2t+4} + 2e^{t+6})$
15. $y_1 = -e^{-2t} + e^t + \frac{1}{3}u(t-1)(-e^{-2t+3} + e^t),\quad y_2 = -e^{-2t} + 4e^t$
$+ \frac{1}{3}u(t-1)(-e^{-2t+3} + e^t)$
17. $y_1 = 3\sin 2t + 8e^{-3t},\quad y_2 = -3\sin 2t + 5e^{-3t}$
19. $y_1 = e^t - e^{-t},\quad y_2 = e^t,\quad y_3 = e^{-t}$
25. $4i_1 + 8(i_1 - i_2) + 2i_1' = 390\cos t,\quad 8i_2 + 8(i_2 - i_1) + 4i_2' = 0,\quad i_1 = -26e^{-2t}$
$- 16e^{-8t} + 42\cos t + 15\sin t,\quad i_2 = -26e^{-2t} + 8e^{-8t} + 18\cos t + 12\sin t$

Chapter 6 Review Questions and Problems, page 267

11. $\dfrac{1}{(s-3)^2}$
13. $\dfrac{2}{s(s^2+4)}$
15. $e^{-\pi s}\left(\dfrac{\pi}{s} + \dfrac{1}{s^2}\right)$
17. $\dfrac{s}{(s-1)(s^2+4)}$
19. $\dfrac{2s^2}{s^4-1}$
21. $\dfrac{a-b}{(s-a)(s-b)}$
23. $10\cos t\sqrt{2}$
25. $3e^{-2t}\sin 4t$
27. $u(t-2)(5 + 4(t-2))$
29. $te^{-2t}\sin t$
31. $(t^2-1)u(t-1)$
33. $\dfrac{\pi}{\omega^3}(\omega t - \sin \omega t)$
35. $20\sin t + u(t-1)[1 - \cos(t-1)]$
37. $10\cos 2t - \frac{1}{2}\sin 2t + 4u(t-5)\sin(2t-10)$
39. $e^{-t}(7\cos 3t + 2\sin 3t)$
41. $e^{-t} + u(t-\pi)[1.2\cos t - 3.6\sin t + 2e^{-t+\pi} - 0.8e^{2t-2\pi}]$
43. $u(t-1)(t-1)e^{2t-2} + 4u(t-2)(2-t)e^{2t-4}$
45. $y_1 = e^t + \frac{1}{2}e^{-t} - \frac{1}{2}\cos t - \frac{1}{2}\sin t,\quad y_2 = -e^t + \frac{1}{2}e^{-t} + \frac{1}{2}\cos t + \frac{1}{2}\sin t$
47. $y_1 = \frac{1}{2}e^{-t}\sin 2t,\quad y_2 = e^{-t}(\cos 2t - \frac{1}{2}\sin 2t)$
49. $y_1 = e^{2t},\quad y_2 = e^{2t} + e^t$
51. $I = (1 - e^{-2s})/[s(s+10)],\quad i = 0.1(1 - e^{-10t}) + 0.1u(t-2)[-1 + e^{-10t+20}]$
53. $I = e^{-2t}(76\cos 4t - 42\sin 4t) - 76\cos 20t + 16\sin 20t$
55. $i_1' + 10(i_1 - i_2) = 100t^2,\quad 30i_2' + 10(i_2' - i_1') + 100i_2 = 0,$
$i_1 = (\frac{4}{5} + 4t)e^{-5t} + 10t^2 - \frac{4}{5},\quad i_2 = (\frac{4}{5} + 2t)e^{-5t} + 2t - \frac{4}{5}$

Problem Set 7.1, page 277

1. $\begin{bmatrix} 6 & 3 \\ -2 & 11 \\ -7 & 6 \end{bmatrix}$, same, $\begin{bmatrix} 36 & -6 \\ -36 & 18 \\ -54 & 0 \end{bmatrix}$, $\begin{bmatrix} -36 & 6 \\ 36 & -18 \\ 54 & 0 \end{bmatrix}$

3. Undef., $\begin{bmatrix} -6 & 1 \\ 6 & -3 \\ 9 & 0 \end{bmatrix}$, $\begin{bmatrix} 6 & -1 \\ -6 & 3 \\ -9 & 0 \end{bmatrix}$, undef.

5. $\begin{bmatrix} -48 & -2 \\ 38 & -44 \\ 67 & -15 \end{bmatrix}$, $\begin{bmatrix} 36 & 0 & 48 \\ -12 & 24 & 24 \\ 72 & 60 & -48 \end{bmatrix}$, same, $\begin{bmatrix} -0.3 & -5.0 & -3.4 \\ -4.9 & 1.8 & 3.8 \\ -3.6 & 3.5 & 0.4 \end{bmatrix}$

7. $\begin{bmatrix} 66 \\ 0 \\ -33 \end{bmatrix}$, $\begin{bmatrix} 0 \\ 3.2 \\ -4.2 \end{bmatrix}$, same, $\begin{bmatrix} 6.5 \\ -0.8 \\ -2.2 \end{bmatrix}$

9.
$$\begin{aligned} -5x_2 &= -3 \\ -5x_1 + 2x_2 &= 4 \\ -3x_1 + 4x_2 &= 0 \end{aligned}$$

Problem Set 7.2, page 286

1. $\begin{bmatrix} 24 \\ 49 \\ -43 \end{bmatrix}$, undef., $\begin{bmatrix} 2 \\ 38 \\ -22 \end{bmatrix}$, $\begin{bmatrix} 54 & 10 & -46 \\ 74 & 19 & -29 \\ -74 & -5 & 51 \end{bmatrix}$

3. $\begin{bmatrix} 54 & 10 & -46 \\ 74 & 19 & -29 \\ -74 & -5 & 51 \end{bmatrix}$, $\begin{bmatrix} 134 & -50 & -18 \\ 94 & -29 & -1 \\ -134 & 63 & 19 \end{bmatrix}$, $\begin{bmatrix} 44 & 64 & -72 \\ 64 & 110 & -114 \\ -72 & -114 & 126 \end{bmatrix}$,

$\begin{bmatrix} 236 & -92 & -12 \\ -92 & 38 & 6 \\ -12 & 6 & 6 \end{bmatrix}$

5. [20 −3 −7], [−62 34 2], $\begin{bmatrix} 565 \\ 525 \\ 790 \end{bmatrix}$, same

7. $\begin{bmatrix} 15 & 0 & 40 \\ 3 & 0 & 8 \\ 6 & 0 & 16 \end{bmatrix}$, 31, $\begin{bmatrix} -310 & 170 & 10 \\ -62 & 34 & 2 \\ -124 & 68 & 4 \end{bmatrix}$, same

9. $\begin{bmatrix} 337 & 8 & -160 \\ 252 & 49 & -68 \\ -308 & 52 & 233 \end{bmatrix}$, same, $\begin{bmatrix} 257 & 68 & -188 \\ 232 & 97 & -96 \\ -248 & -16 & 265 \end{bmatrix}$

11. $\begin{bmatrix} 324 & 32 & -320 \\ 244 & 38 & -322 \\ -244 & -10 & 366 \end{bmatrix}, \begin{bmatrix} 216 & -104 & -104 \\ 280 & -132 & -68 \\ -280 & 140 & 76 \end{bmatrix}, \begin{bmatrix} 7060 & 960 & -5120 \\ 7548 & 1246 & -5434 \\ -8140 & -1090 & 6150 \end{bmatrix}$

$\begin{bmatrix} 4324 & 1520 & -4816 \\ 3636 & 1242 & -4518 \\ -3700 & -1046 & 5002 \end{bmatrix}$

13. 83, 166, 593, 0

19. (d) $\mathbf{AB} = (\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T = \mathbf{BA}$; etc. (e) *Ans.* If $\mathbf{AB} = -\mathbf{BA}$.

21. Triangular are $\mathbf{U}_1 + \mathbf{U}_2$, $\mathbf{U}_1\mathbf{U}_2$, $\mathbf{U}_1^2$, $\mathbf{L}_1 + \mathbf{L}_2$, $\mathbf{L}_1\mathbf{L}_2$, $\mathbf{L}_1^2$.

23. $[0.8 \quad 1.2]^T$, $[0.76 \quad 1.24]^T$, $[0.752 \quad 1.248]^T$

27. $\mathbf{p} = [110 \quad 45 \quad 80]^T$, $\mathbf{v} = [92000 \quad 86300]^T$

Problem Set 7.3, page 295

1. $x = 2.5, y = -4.2$

3. $x = 0.2, y = 1.6$

5. $x = 0, y = -2, z = 9$

7. $x = 4, y = 0, z = -2$

9. $x = 3y + 2$, y arb., $z = -y + 6$

11. $y = 2x + 3z + 1$, x, z arb.

13. $w = 1, y = 2z - x$, x, z arb.

15. $w = 3, x = 0, y = -2, z = 8$

17. $I_1 = (R_1 + R_2)E_0/(R_1R_2)$, $I_2 = E_0/R_1$, $I_3 = E_0/R_2$ [Amps]

19. $I_1 - I_2 - I_3 = 0$, $(3 + 2 + 5)I_1 + 10I_2 = 95 + 35$, $10I_2 - 5I_3 = 35$, $I_1 = 8$, $I_2 = 5$, $I_3 = 3$ Amps

21. $x_1 + x_4 = 500$, $x_1 + x_2 = 800$, $x_2 + x_3 = 1100$, $x_3 + x_4 = 800$, $x_1 = 500 - x_4$, $x_2 = 300 + x_4$, $x_3 = 800 - x_4$, x_4 arbitrary

Problem Set 7.4, page 301

1. 1, $[1 \quad -2]$; $[1 \quad 0 \quad -3]^T$

3. 3, $[1 \quad 4 \quad 0 \quad 7]$, $[0 \quad -2 \quad 1 \quad 3]$, $[0 \quad 0 \quad 5 \quad 105]$; $[-2 \quad 4 \quad 5]^T$, $[0 \quad 1 \quad 5]^T$, $[0 \quad 0 \quad 1]^T$

5. 2, $[3 \quad 0 \quad 5]$, $[0 \quad 3 \quad 4]$; $[3 \quad 0 \quad 5]^T$, $[0 \quad 3 \quad 4]^T$

7. 2, $[8 \quad 0 \quad 4]$, $[0 \quad 2 \quad 0]$; $[8 \quad 0 \quad 4 \quad 0]^T$, $[0 \quad 2 \quad 0 \quad 4]^T$

9. 3, $[1 \quad 0 \quad 3 \quad 0]$, $[0 \quad 5 \quad 8 \quad -37]$, $[0 \quad 0 \quad -74 \quad 296]$; same transposed

11. 4, $[1 \quad 0 \quad 0 \quad 0]$, $[0 \quad 1 \quad 0 \quad 0]$, $[0 \quad 0 \quad 1 \quad 0]$, $[0 \quad 0 \quad 0 \quad 1]$; same transposed

13. No

15. No

17. Yes

19. Yes

21. (c) 1

27. 2, $[1 \quad -1 \quad 0]$, $[0 \quad 0 \quad 1]$

29. No

31. 1, $[-\frac{1}{4} \quad \frac{1}{3} \quad 1]$

33. No

35. 1, $[5 \quad \frac{5}{2} \quad \frac{5}{3} \quad \frac{5}{4} \quad 1]$

Problem Set 7.7, page 314

5. 107

7. $\cos(\alpha + \beta)$

9. -66.88

11. 0

13. $u^3 + v^3 + w^3 - 3uvw$

15. 4

19. $x = -1.2, y = 0.8, z = 3.1$

21. 1

23. 3

Problem Set 7.8, page 322

1. $\begin{bmatrix} 1.80 & -2.32 \\ -0.25 & 0.60 \end{bmatrix}$ **3.** $\begin{bmatrix} \cos 2\theta & -\sin 2\theta \\ \sin 2\theta & \cos 2\theta \end{bmatrix}$ **5.** $\begin{bmatrix} 2 & 0 & -1 \\ 5 & 1 & 0 \\ 0 & 1 & 3 \end{bmatrix}$

7. $\begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 3 & -4 & 1 \end{bmatrix}$ **9.** $\mathbf{A}^{-1} = \mathbf{A}$ **11.** No inverse

15. $(\mathbf{A}^2)^{-1} = (\mathbf{A}^{-1})^2 = \begin{bmatrix} 4 & -1 & -5 \\ 15 & 1 & -5 \\ 5 & 4 & 9 \end{bmatrix}$

19. $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$, $(\mathbf{A}\mathbf{A}^{-1})^{-1} = (\mathbf{A}^{-1})^{-1}\mathbf{A}^{-1} = \mathbf{I}$. Multiply by $\mathbf{A}$ from the right.
21. $\det \mathbf{A} = -1$. $C_{12} = C_{21} = C_{33} = -1$, the other C_{jk} are zero.
23. $\det \mathbf{A} = 1$. $C_{11} = 1$, $C_{12} = -2$, $C_{22} = 1$, $C_{13} = 3$, $C_{23} = -4$, $C_{33} = 1$

Problem Set 7.9, page 329

1. Yes, 2, $[3 \quad 5 \quad 0]^T$, $[2 \quad 0 \quad -5]^T$ **3.** No
5. Yes, 2, $[0 \quad 0 \quad 0 \quad 1 \quad 0]^T$, $[0 \quad 0 \quad 0 \quad 0 \quad 1]^T$
7. Yes, 1, $\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$ **9.** No
11. Yes, 2, xe^{-x}, e^{-x}
13. $[1 \quad 0]^T$, $[0 \quad 1]^T$; $[1 \quad 1]^T$, $[-1 \quad 1]^T$; $[1 \quad 0]^T$, $[0 \quad -1]^T$
15. $x_1 = -0.6y_1 + 0.4y_2$
$x_2 = -0.8y_1 + 0.2y_2$
17. $x_1 = 2y_1 + y_2$
$x_2 = 5y_1 + 3y_2$
19. $x_1 = 5y_1 + 3y_2 - 3y_3$
$x_2 = 3y_1 + 2y_2 - 2y_3$
$x_3 = 2y_1 - y_2 + 2y_3$
21. $\sqrt{56}$ **23.** $16\sqrt{5}$
25. 2 **29.** $4v_1 - 3v_2 = 0$, $\mathbf{v} = \pm [\frac{3}{5} \quad \frac{4}{5}]^T$

Chapter 7 Review Questions and Problems, page 330

11. $x = 4$, $y = 7$ **13.** $x = y + 6$, $z = y$, y arbitrary
15. $x = \frac{1}{4}$, $y = -\frac{1}{2}$, $z = \frac{3}{2}$ **17.** $x = 7$, $y = -3$
19. $x = 2z$, $y = 4$, z arbitrary **21.** 0
23. 638, 0, 0 **25.** $\begin{bmatrix} 8.0 & -3.6 & 1.2 \\ -3.6 & 2.6 & 2.4 \\ 1.2 & 2.4 & 9.0 \end{bmatrix}$
27. 14, 14, $\begin{bmatrix} 12 & 0 & 6 \\ 28 & 0 & 14 \\ 4 & 0 & 2 \end{bmatrix}$ **29.** $[-20 \quad 9 \quad -3]$, $\begin{bmatrix} -20 \\ 9 \\ -3 \end{bmatrix}$

31. 2, 2

33. 2, 2

35. 2, 2

37. $\frac{1}{51}\begin{bmatrix} 4 & 3 \\ -5 & 9 \end{bmatrix}$

39. $\frac{1}{45}\begin{bmatrix} -5 & 10 & 5 \\ 11 & 5 & -2 \\ 23 & -10 & 4 \end{bmatrix}$

41. $\frac{1}{12}\begin{bmatrix} 72 & -72 & 132 \\ 31 & -32 & 59 \\ -19 & 20 & -35 \end{bmatrix}$

43. $I_1 = 33$ A, $I_2 = 11$ A, $I_3 = 22$ A

45. $I_1 = 12$ A, $I_2 = 18$ A, $I_3 = 6$ A

Problem Set 8.1, page 338

1. -2, $[1 \quad 0]^T$; 0.4, $[0 \quad 1]^T$

3. 4, $2x_1 + (-4 - 4)x_2 = 0$, say, $x_1 = 4$, $x_2 = 1$; -4, $[0 \quad 1]^T$

5. -4, $[2 \quad 9]^T$; 3, $[1 \quad 1]^T$ **7.** $0.8 + 0.6i$, $[1 \quad -i]^T$; $0.8 - 0.6i$, $[1 \quad i]^T$

9. 5, $[1 \quad 2]^T$; 0, $[-2 \quad 1]^T$ **11.** 4, $[1 \quad 0 \quad 0]^T$; 0, $[0 \quad 1 \quad 0]^T$; -1, $[0 \quad 0 \quad 1]^T$

13. $-(\lambda^3 - 18\lambda^2 + 99\lambda - 162)/(\lambda - 3) = -(\lambda^2 - 15\lambda + 54)$; 3, $[2 \quad -2 \quad 1]^T$; 6, $[1 \quad 2 \quad 2]^T$; 9, $[2 \quad 1 \quad -2]^T$

15. 1, $[-3 \quad 2 \quad 10]^T$; 4, $[0 \quad 1 \quad 2]^T$; 2, $[0 \quad 0 \quad 1]^T$

17. $-(\lambda^3 - 7\lambda^2 - 5\lambda + 75)/(\lambda + 3) = -(\lambda^2 - 10\lambda + 25)$; -3, $[1 \quad 2 \quad -1]^T$; 5, $[3 \quad 0 \quad 1]^T$, $[-2 \quad 1 \quad 0]^T$

19. $-(\lambda - 9)^3$; 9, $[2 \quad -2 \quad 1]^T$; defect 2

21. $\lambda(\lambda^3 - 8\lambda^2 - 16\lambda + 128)/(\lambda - 4) = \lambda(\lambda^2 - 4\lambda - 32)$; 4, $[-1 \quad 3 \quad 1 \quad 1]^T$; -4, $[1 \quad 1 \quad -1 \quad -1]^T$; 0, $[1 \quad 1 \quad 1 \quad 1]^T$; 8, $[1 \quad -3 \quad 1 \quad -3]^T$

23. 2, $[8 \quad 8 \quad -16 \quad 1]^T$; 1, $[0 \quad 7 \quad 0 \quad 4]^T$; 3, $[0 \quad 0 \quad 9 \quad 2]^T$, -6, $[0 \quad 0 \quad 0 \quad 1]^T$

25. $(\lambda + 1)^2(\lambda^2 + 2\lambda - 15)$; -1, $[1 \quad 0 \quad 0 \quad 0]^T$, $[0 \quad 1 \quad 0 \quad 0]^T$; -5, $[-3 \quad -3 \quad 1 \quad 1]^T$, 3, $[3 \quad -3 \quad 1 \quad -1]^T$

29. Use that real entries imply real coefficients of the characteristic polynomial.

Problem Set 8.2, page 343

1. $\begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$; -1, $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$; 1, $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$; any point $(x, 0)$ on the x-axis is mapped onto $(-x, 0)$, so that $[1 \quad 0]^T$ is an eigenvector corresponding to $\lambda = -1$.

3. (x, y) maps onto $(x, 0)$. $\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$; 1, $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$; 0, $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$. A point on the x-axis maps onto itself, a point on the y-axis maps onto the origin.

5. (x, y) maps onto $(5x, 5y)$. 2×2 diagonal matrix with entries 5.

7. -2, $[1 \quad -1]^T$, $-45°$; 8, $[1 \quad 1]^T$, $45°$

9. 2, $[3 \quad -1]^T$, $-18.4°$; 7, $[1 \quad 3]^T$, $71.6°$

11. 1, $[-1/\sqrt{6} \quad 1]$, $112.2°$; 8, $[1 \quad 1/\sqrt{6}]$, $22.2°$

13. 1, $[1 \quad 1]^T$, $45°$; -5, $[1 \quad -1]^T$, $-45°$

15. $c[15 \quad 24 \quad 50]^T$, $c > 0$

17. $\mathbf{x} = (\mathbf{I} - \mathbf{A})^{-1}\mathbf{y} = [0.73 \quad 0.59 \quad 1.04]^T$ (rounded)

19. $[1 \quad 1 \quad 1]^T$ **21.** 1.8 **23.** 2.1

Problem Set 8.3, page 348

3. No

5. $\mathbf{A}^{-1} = (-\mathbf{A}^{\mathsf{T}})^{-1} = -(\mathbf{A}^{-1})^{\mathsf{T}}$

7. No since $\det \mathbf{A} = \det(\mathbf{A}^{\mathsf{T}}) = \det(-\mathbf{A}) = (-1)^3 \det \mathbf{A} = -\det \mathbf{A} = 0$.

9. Orthogonal, $0.96 \pm 0.28i$

11. Neither, 2, 2, defect 1

13. Symmetric, 9, 18, 18

15. Orthogonal, 1, i, $-i$

17. Symmetric, $a + 2b$, $a - b$, $a - b$

Problem Set 8.4, page 355

1. $[1 \quad 2]^{\mathsf{T}}$, $[2 \quad -1]^{\mathsf{T}}$; $\mathbf{X} = \begin{bmatrix} 1 & 2 \\ 2 & -1 \end{bmatrix}$, $\mathbf{D} = \begin{bmatrix} 7 & 0 \\ 0 & 2 \end{bmatrix}$

3. $[1 \quad -1]^{\mathsf{T}}$, $[1 \quad 1]^{\mathsf{T}}$, $\mathbf{D} = \begin{bmatrix} 4 & 0 \\ 0 & 6 \end{bmatrix}$

5. $[2 \quad -1]^{\mathsf{T}}$, $[2 \quad 1]^{\mathsf{T}}$, diag $(-2, 4)$

7. $[1 \quad 0 \quad 0]^{\mathsf{T}}$, $[1 \quad -2 \quad 1]^{\mathsf{T}}$, $[0 \quad 1 \quad 0]^{\mathsf{T}}$, diag $(1, 2, 3)$

9. $[0 \quad 3 \quad 2]^{\mathsf{T}}$, $[5 \quad 3 \quad 0]^{\mathsf{T}}$, $[1 \quad 0 \quad 2]^{\mathsf{T}}$, diag $(45, 9, -27)$

13. $\begin{bmatrix} 16 & -7 \\ 42 & -19 \end{bmatrix}$; -5, $\begin{bmatrix} 1 \\ 3 \end{bmatrix}$; 2, $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$; $\mathbf{x} = \begin{bmatrix} -2 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \end{bmatrix}$

15. $\begin{bmatrix} -30 & -72 \\ \frac{40}{3} & 32 \end{bmatrix}$; 2, $\begin{bmatrix} 9 \\ -4 \end{bmatrix}$; 0, $\begin{bmatrix} 12 \\ -5 \end{bmatrix}$; $\mathbf{x} = \begin{bmatrix} -3 \\ 3 \end{bmatrix}, \begin{bmatrix} -3 \\ 6 \end{bmatrix}$

17. $\begin{bmatrix} -95 & 18 & -144 \\ 24 & -2 & 36 \\ 66 & -12 & 100 \end{bmatrix}$; 4, $\begin{bmatrix} 4 \\ -2 \\ -3 \end{bmatrix}$; -2, $\begin{bmatrix} 6 \\ -1 \\ -4 \end{bmatrix}$; 1, $\begin{bmatrix} 3 \\ 0 \\ -2 \end{bmatrix}$;

$\mathbf{x} = \begin{bmatrix} -2 \\ -4 \\ -6 \end{bmatrix}, \begin{bmatrix} 0 \\ -2 \\ -4 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ -2 \end{bmatrix}$

19. $\mathbf{C} = \begin{bmatrix} 1 & 12 \\ 12 & -6 \end{bmatrix}$, $10y_1^2 - 15y_2^2 = 5$, $\mathbf{x} = \begin{bmatrix} 0.8 & 0.6 \\ 0.6 & -0.8 \end{bmatrix}\mathbf{y}$, hyperbola

21. $\mathbf{C} = \begin{bmatrix} 3 & -4 \\ -4 & -3 \end{bmatrix}$, $5y_1^2 - 5y_2^2 = 0$, $\mathbf{x} = \begin{bmatrix} 2/\sqrt{5} & 1/\sqrt{5} \\ -1/\sqrt{5} & 2/\sqrt{5} \end{bmatrix}\mathbf{y}$, straight lines

23. $\mathbf{C} = \begin{bmatrix} 4 & \sqrt{3} \\ \sqrt{3} & 2 \end{bmatrix}$, $y_1^2 + 5y_2^2 = 10$, $\mathbf{x} = \begin{bmatrix} 1/2 & \sqrt{3}/2 \\ -\sqrt{3}/2 & 1/2 \end{bmatrix}\mathbf{y}$, ellipse

25. $\mathbf{C} = \begin{bmatrix} 1 & -6 \\ -6 & 1 \end{bmatrix}$, $7y_1^2 - 5y_2^2 = 35$, $\mathbf{x} = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}\mathbf{y}$, hyperbola

27. $\mathbf{C} = \begin{bmatrix} 12 & 16 \\ 16 & 12 \end{bmatrix}$, $28y_1^2 - 4y_2^2 = 112$, $\mathbf{x} = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}\mathbf{y}$, hyperbola

Problem Set 8.5, page 361

3. $(\overline{\mathbf{ABC}})^{\mathrm{T}} = \overline{\mathbf{C}}^{\mathrm{T}}\overline{\mathbf{B}}^{\mathrm{T}}\overline{\mathbf{A}}^{\mathrm{T}} = \mathbf{C}^{-1}(-\mathbf{B})\mathbf{A}$
5. Hermitian, $3 + \sqrt{2}$, $[-i \quad 1 - \sqrt{2}]^{\mathrm{T}}$; $3 - \sqrt{2}$, $[-i \quad 1 + \sqrt{2}]^{\mathrm{T}}$
7. Hermitian, unitary, 1, $[1 \quad i - i\sqrt{2}]^{\mathrm{T}}$; -1, $[1 \quad i + i\sqrt{2}]^{\mathrm{T}}$
9. Skew-Hermitian, $5i$, $[1 \quad 0 \quad 0]^{\mathrm{T}}$, $[0 \quad 1 \quad 1]^{\mathrm{T}}$; $-5i$, $[0 \quad 1 \quad -1]^{\mathrm{T}}$
11. Skew-Hermitian, unitary, i, $[1 \quad 0 \quad 1]^{\mathrm{T}}$, $[0 \quad 1 \quad 0]^{\mathrm{T}}$; $-i$, $[1 \quad 0 \quad -1]^{\mathrm{T}}$
13. Skew-Hermitian, $-66i$ **15.** Hermitian, 10

Chapter 8 Review Questions and Problems, page 362

9. $\begin{bmatrix} 3 & 2 \\ -4 & -3 \end{bmatrix} \mathbf{A} \begin{bmatrix} 3 & 2 \\ -4 & -3 \end{bmatrix} = \begin{bmatrix} 5 & 0 \\ 0 & -7 \end{bmatrix}$

11. $\begin{bmatrix} -1/3 & 2/3 \\ 2/3 & -1/3 \end{bmatrix} \mathbf{A} \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} 6 & 0 \\ 0 & -9 \end{bmatrix}$

13. $\begin{bmatrix} \frac{2}{9} & \frac{1}{9} & 0 \\ \frac{1}{9} & 0 & \frac{2}{9} \\ 0 & -\frac{2}{9} & -\frac{1}{9} \end{bmatrix} \mathbf{A} \begin{bmatrix} 4 & 1 & 2 \\ 1 & -2 & -4 \\ -2 & 4 & -1 \end{bmatrix} = \begin{bmatrix} 6 & 0 & 0 \\ 0 & -3 & 0 \\ 0 & 0 & 0 \end{bmatrix}$

15. $\begin{bmatrix} -1.0 & 0 \\ 4.8 & 5.0 \end{bmatrix}$, 5, -1 **17.** $\begin{bmatrix} -2 & 0 & -4.50 \\ 1 & 0 & -2.75 \\ 0 & 0 & 4 \end{bmatrix}$, 4, 0, -2

19. $1.1y_1^2 + y_2^2 = 1$, ellipse

Problem Set 9.1, page 370

1. 2, -4, 0; $\sqrt{20}$; $[1/\sqrt{5}, -2/\sqrt{5}, 0]$
3. -1, 0, 5; $\sqrt{26}$; $[-1/\sqrt{26}, 0, 5/\sqrt{26}]$
5. -8, -6, 0; 10; $[-0.8, -0.6, 0]$
7. (7, 5, 0); $\sqrt{10}$
9. $(\frac{1}{4}, \frac{3}{2}, \frac{9}{4})$; $\sqrt{37/8}$
11. $(0, 1, \frac{1}{2})$; $\sqrt{37/2}$
13. $[4, -2, 0]$, $[-2, 1, 0]$, $[-1, \frac{1}{2}, 0]$
15. $[10, -5, -15]$
17. $[28, -14, -14]$
19. $[-2, 1, 8]$, $[6, -3, -24]$
23. (5.5, 5.5, 0), $(\frac{1}{8}, \frac{1}{2}, \frac{15}{8})$
25. $[0, 0, 9]$; 9
27. $[-8, -2, 4]$; $\sqrt{84}$
29. $\mathbf{v} = [0, 0, -9]$
31. $[-9, 0, 0]$, $[0, -2, 0]$, $[0, 0, -11]$. Yes.
33. $|\mathbf{p} + \mathbf{q} + \mathbf{u}| \leqq 6$. Nothing

35. $\left[\frac{25}{\sqrt{2}}, \frac{25}{\sqrt{2}}\right] - \left[-\frac{20}{\sqrt{2}}, \frac{20}{\sqrt{2}}\right] = \left[\frac{45}{\sqrt{2}}, \frac{5}{\sqrt{2}}\right]$ **37.** $|\mathbf{w}|/(2 \sin \alpha)$

Problem Set 9.2, page 376

1. 4 **3.** $\sqrt{241}$
5. $[12, -8, 4]$, $[-18, -9, -36]$
7. 17
9. -4, 4
11. -24
15. Use (1) and $|\cos \gamma| \leqq 1$.
17. $|\mathbf{a} + \mathbf{b}|^2 + |\mathbf{a} - \mathbf{b}|^2 = \mathbf{a}\bullet\mathbf{a} + 2\mathbf{a}\bullet\mathbf{b} + \mathbf{b}\bullet\mathbf{b} + (\mathbf{a}\bullet\mathbf{a} - 2\mathbf{a}\bullet\mathbf{b} + \mathbf{b}\bullet\mathbf{b}) = 2|\mathbf{a}|^2 + 2|\mathbf{b}|^2$
19. 0
21. 15
23. Orthogonality. Yes

25. 2, 2, 0, −2 **27.** 79.11° **29.** 82.45°
31. 54.74° **33.** 54.79°, 79.11°, 46.10° **35.** 63.43°, 116.57°
37. 3 **39.** 1.4
41. If $|\mathbf{a}| = |\mathbf{b}|$ or if $\mathbf{a}$ and $\mathbf{b}$ are orthogonal

Problem Set 9.3, page 383

1. [0, 0, −10], [0, 0, 10] **3.** [−4, −8, 26] **5.** [0, 0, −60]
7. −20, −20 **9.** 240 **11.** [19, −21, 24], $\sqrt{1378}$
13. [10, −5, −1] **15.** 2 **17.** 30, −30
19. −20, −20 **25.** $[-2, 2, 0] \times [4, 4, 0] = -16\mathbf{k}$, 16
27. $[1, -1, 2] \times [1, 2, 3] = [-7, -1, 3]$, $\sqrt{59}$
29. $[0, 10, 0] \times [4, 3, 0] = [0, 0, -40]$, speed 40
31. $|[7, 0, 0] \times [1, 1, 0]| = 7$ **33.** $\frac{1}{2}\sqrt{3}$
35. [18, 14, 26]; $9x + 7y + 13z = c$, $9 \cdot 4 + 7 \cdot 8 + 13 \cdot 0 = 92 = c$
37. 16 **39.** $c = 2.5$

Problem Set 9.4, page 389

1. Hyperbolas **3.** Hyperbolas **5.** Circles
7. Ellipses; 288, 100, 409; elliptic ring between the ellipses

$$\frac{x^2}{(2/3)^2} + \frac{y^2}{(1/2)^2} = 1 \quad \text{and} \quad \frac{x^2}{(4/3)^2} + y^2 = 1.$$

9. Ellipsoids **11.** Cones **13.** Planes
23. $[8x, 0, yz]$, $[0, 0, xz]$, $[0, 18z, xy]$; $[0, z, y]$, $[z, 0, x]$, $[y, x, 0]$

Problem Set 9.5, page 398

1. $[4 + 3\cos t, \quad 6 + 3\sin t]$ **3.** $[2 - t, \quad 0, \quad 4 + t]$
5. $[3, \quad -2 + 3\cos t, \quad 3\sin t]$ **7.** $[a + 3t, \quad b - 2t, \quad c + 5t]$
9. $[\sqrt{2}\cos t, \quad \sin t, \quad \sin t]$ **11.** Helix on $(x - 2)^2 + (y - 6)^2 = r^2$
13. Circle $(x - 2)^2 + (y + 2)^2 = 1$, $z = 5$ **15.** $x^4 + y^4 = 1$
17. Hyperbola $xy = 1$
23. $\mathbf{r}' = [-5\sin t, \quad 5\cos t, \quad 0]$, $\mathbf{u} = [-\sin t, \quad \cos t, \quad 0]$, $\mathbf{q} = [4 - 3w, \quad 3 + 4w, \quad 0]$
25. $\mathbf{r}' = [\sinh t, \quad \cosh t]$, $\mathbf{u} = (\cosh 2t)^{-1/2}[\sinh t, \quad \cosh t]$, $\mathbf{q} = [\frac{5}{3} + 4w, \quad \frac{4}{3} + 5w]$
27. $\sqrt{\mathbf{r}' \cdot \mathbf{r}'} = \cosh t$, $\ell = \sinh 1 = 1.175$
29. Start from $\mathbf{r}(t) = [t, \quad f(t)]$.
33. $\mathbf{v} = \mathbf{r}' = [1, \quad 2t, \quad 0]$, $|\mathbf{v}| = \sqrt{1 + 4t^2}$, $\mathbf{a} = [0, \quad 2, \quad 0]$
35. $\mathbf{v}(0) = 2\omega R\mathbf{i}$, $\mathbf{a}(0) = -\omega^2 R\mathbf{j}$
37. 1 year $= 365 \cdot 86400$ sec, $R = 30 \cdot 365 \cdot 86400/2\pi = 151 \cdot 10^6$ [km]. $|\mathbf{a}| = \omega^2 R = |\mathbf{v}|^2/R = 5.98 \cdot 10^{-6}$ [km/sec^2]
39. $R = 3960 + 80$ mi $= 2.133 \cdot 10^7$ ft, $g = |\mathbf{a}| = \omega^2 R = |\mathbf{v}|^2/R$, $|\mathbf{v}| = \sqrt{gR} = \sqrt{6.61 \cdot 10^8} = 25700$ [ft/sec] $= 17500$ [mph]
43. $\mathbf{r}(t) = [t, \quad y(t), \quad 0]$, $\mathbf{r}' = [1, \quad y', \quad 0]$, $\mathbf{r}' \cdot \mathbf{r}' = 1 + y'^2$, $\mathbf{r}'' = [0, \quad y'', \quad 0]$, etc.
47. $3/(1 + 9t^2 + 9t^4)$

Problem Set 9.6, page 403

1. $w' = 2\sqrt{2}(\sinh 4t)/(\cosh 4t)^{1/2}$
3. $w' = (\cosh t)^{\sinh t - 1}((\cosh^2 t) \ln(\cosh t) + \sinh^2 t)$
5. $w' = 3(2t^4 + t^8)^2(8t^3 + 8t^7)$ **7.** $e^{4u} \sin^2 2v$, $\frac{1}{2}e^{4u} \sin 4v$
9. $-2(u^2 + v^2)^{-3}u$, $-2(u^2 + v^2)^{-3}v$

Problem Set 9.7, page 409

1. $[2x, \quad 2y]$ **3.** $[1/y, \quad -x/y^2]$ **5.** $[y + 2, \quad x - 2]$
7. $[6, \quad 4, \quad 4]$ **9.** $[-1.25, \quad 0]$ **11.** $[0, \quad -e]$
13. $[-4, \quad 2]$ **15.** $[-18, \quad 24]$ **17.** $[48, \quad -36]$
19. $[6, \quad 4]$ **21.** $[-6, \quad -12]$
23. $[-0.0015, \quad 0, \quad -0.0020]$ **27.** $[a, \quad b, \quad c]$
29. $[8, \quad 6, \quad 0]$ **31.** $[108, \quad 108, \quad 108]$ **33.** $\sqrt{2/3}$
35. 7/3 **37.** $2e^2/\sqrt{13}$
39. $\frac{3}{2}x^2 + \frac{5}{2}y^2 - 2z^2$ **41.** $x^4 + y^3 - 3z^2$

Problem Set 9.8, page 413

1. $3(x + y)^2$ **3.** $2(x + xz + z)$ **5.** $(y + x + 1)\cos xy$
7. $9x^2y^2z^2$
9. $[v_1, v_2, v_3] = \mathbf{r}' = [x', y', z'] = [y, 0, 0]$, $z' = 0$, $z = c_3$, $y' = 0$, $y = c_2$, $x' = y = c_2$, $x = c_2t + c_1$. Hence as t increases from 0 to 1, this "shear flow" transforms the cube into a parallelepiped of volume 1.
11. div $(\mathbf{w} \times \mathbf{r}) = 0$ because v_1, v_2, v_3 do not depend on x, y, z, respectively.
13. (b) $(fv_1)_x + (fv_2)_y + (fv_3)_z = f[(v_1)_x + (v_2)_y + (v_3)_z] + f_xv_1 + f_yv_2 + f_zv_3$, etc.
(c) Use (b) with $\mathbf{v} = \nabla g$.
15. $4(x + y)/(y - x)^3$ **17.** 0 **19.** $e^{xyz}(y^2z^2 + x^2z^2 + x^2y^2)$

Problem Set 9.9, page 416

1. $[0, \quad 0, \quad 4x - 1]$ **3.** $[0, \quad 0, \quad 2e^x \sin y]$ **5.** $[0, \quad 0, \quad -4y/(x^2 + y^2)]$
9. curl $\mathbf{v} = [-2z, \quad 0, \quad 0]$, incompressible, $\mathbf{v} = \mathbf{r}' = [x', \quad y', \quad z'] = [0, \quad z^2, \quad 0]$, $x = c_1$, $z = c_3$, $y' = z^2 = c_3^2$, $y = c_3^2t + c_2$
11. curl $\mathbf{v} = [0, \quad 0, \quad -2]$, incompressible, $x' = y$, $y' = -x$, $z' = 0$, $z = c_3$, $y\,dy + x\,dx = 0$, $x^2 + y^2 = c$
13. Irrotational, div $\mathbf{v} = 1$, compressible, $\mathbf{r} = [c_1e^t, \quad c_2e^{-t}, \quad c_3e^t]$
17. **0, 0,** $[xy - zx, \quad yz - xy, \quad zx - yz]$
19. **0, 0,** $0, -2yz^2 - 2zx^2 - 2xy^2$

Chapter 9 Review Questions and Problems, page 416

11. $[-1, \quad 9, \quad 24]$
13. **0,** $[-43, \quad 54, \quad 3]$, $[43, \quad -54, \quad -3]$
15. $[0, \quad 0, \quad -740]$, $[0, \quad 0, \quad -740]$
17. $[-24, \quad 3, \quad -398]$, $[114, \quad 95, \quad -76]$
19. $-495, -495$
21. 90°, 95.4°
23. If $\mathbf{u} \times \mathbf{v} = \mathbf{0}$. Always
25. $[v_1, v_2, -3]$
27. 3.4
29. If $\gamma > \frac{1}{2}\pi, \frac{1}{2}\pi$

33. 45/6
35. No
37. **0**, $2y^2 + (z + x)^2$
39. $[-1, 1, -1]$, $[-2z, -2x, -2y]$
41. 0, $2x^2 + 4y^2 + 2z^2 + 4xz$
43. $488/\sqrt{3323}$ **45. 0**

Problem Set 10.1, page 425

1. $\mathbf{F}(\mathbf{r}(t)) = [125t^6, t^3, 0]$, $16448/7 = 2350$ **3.** $0 + 160$
5. $\mathbf{F}(\mathbf{r}(t)) = [\cosh t \sinh^2 t, \cosh^2 t \sinh t]$, 93.09
7. $\mathbf{F}(\mathbf{r}(t)) = [t, \cos t, \sin t]$, 6π
9. $\mathbf{F}(\mathbf{r}(t)) = [\cosh \frac{1}{2}t, \sinh \frac{1}{4}t, e^{t/8}]$, 0.6857
11. $\mathbf{F}(\mathbf{r}(t)) = [e^t, e^{t^2}, e^{t^2}]$, $e^2 + 2e^4 - 3$
15. 17/3
17. $[36\pi, \frac{4}{3}(8\pi)^3, 36\pi]$

Problem Set 10.2, page 432

1. $\sin xy$, 1
3. $-\frac{1}{2}e^{-(x^2+y^2)}$, 0
5. $e^{xz} + y$, -2
7. $x^2y + \cosh z$, 392
11. $\sinh ac$
13. No
15. $ce^a - ae^b$
17. $\frac{1}{2}a^2bc^2$
19. No

Problem Set 10.3, page 438

3. $\int_0^1 [x - x^3 - (x^2 - x^5)]\, dx = \frac{1}{12}$
5. $\frac{1}{2}\cosh 6 - \cosh 3 + \frac{1}{2}$
7. $\int_0^4 \frac{1}{2}(e^{3x} - e^{-x})\, dx = \frac{1}{6}e^{12} + \frac{1}{2}e^{-4} - \frac{2}{3}$
9. $\int_0^{\pi/2} (e^{\sin y} \cos y - \cos y)\, dy = e - 2$
11. $\int_{-1}^1 (2x^2 + \frac{2}{3})\, dx = \frac{8}{3}$
13. $\int_0^1 \int_0^x \sqrt{1 - x^2}\, dy\, dx = \frac{1}{3}$
15. $\bar{x} = \frac{2}{3}b$, $\bar{y} = \frac{1}{3}h$
17. $I_x = bh^3/12$, $I_y = b^3h/4$
19. $I_x = (a + b)h^3/24$, $I_y = h(a^4 - b^4)/(48(a - b))$

Problem Set 10.4, page 444

1. $2x^3y - 2xy^3$, $81 - 36 = 45$
3. $3x^2 + 3y^2$, $1875\pi/2 = 2945$
5. $e^{x-y} - e^{x+y}$, $-\frac{1}{3}e^3 + \frac{1}{2}e^2 + e^{-1} - \frac{1}{6}$
7. $2x - 2y$, $-56/15$
9. 0 (why?)
11. Integrand 4. *Ans.* 40π
13. y from 0 to $\frac{1}{2}x$, x from 0 to 2. *Ans.* $\cosh 2 - \frac{1}{2}\sinh 2$
15. y from 1 to $5 - x^2$. *Ans.* 56
19. $4e^4 - 4$

Problem Set 10.5, page 448

1. Straight lines, **k**
3. $x^2/a^2 + y^2/b^2 = 1$, ellipses, straight lines, $[-b \cos v, a \sin v, 0]$
5. $z = (c/a)\sqrt{x^2 + y^2}$, circles, straight lines, $[-acu \cos v, -acu \sin v, a^2u]$
7. $x^2/9 + y^2/16 = z$, ellipses, parabolas, $[-8u^2 \cos v, -6u^2 \sin v, 12u]$
9. $x^2/4 + y^2/9 + z^2/16 = 1$, ellipses, $[12 \cos^2 v \cos u, 8 \cos^2 v \sin u, 6 \cos v \sin v]$
13. $[10u, 10v, 1.6 - 4u + 2v]$, $[40, -20, 100]$
15. $[-2 + \cos v \cos u, \cos v \sin u, 2 + \sin v]$,
$[\cos^2 v \cos u, \cos^2 v \sin u, \cos v \sin v]$

17. $[u, \quad v, \quad 3v^2]$, $[0, \quad -6v, \quad 1]$
19. $[u\cos v, \quad 3u\sin v, \quad 3u]$, $[-9u\cos v, \quad -3u\sin v, \quad 3u]$
21. Because $\mathbf{r}_u$ and $\mathbf{r}_v$ are tangent to the coordinate curves $v = const$ and $u = const$, respectively.
23. $[\tilde{u}, \quad \tilde{v}, \quad \tilde{u}^2 + \tilde{v}^2]$, $\tilde{\mathbf{N}} = [-2\tilde{u}, \quad -2\tilde{v}, \quad 1]$

Problem Set 10.6, page 456

1. -64
3. -18
5. -128π
7. 2π
9. $\frac{1}{3}a^3$
11. $17h/4$
15. $140\sqrt{6}/3$
17. $128\pi\sqrt{2}/3 = 189.6$
19. $\frac{1}{96}\pi^2(37^{3/2} - 5^{3/2}) = 22.00$
25. $2\pi h$
27. $\pi h^4/\sqrt{2}$
29. $\pi h + 2\pi h^3/3$

Problem Set 10.7, page 463

1. $8a^3b^3c^3/27$
3. 6
5. $42\frac{2}{3}\pi$
7. 234π
9. $2a^5/3$
11. $ha^4\pi/2$
13. $\pi h^5/10$
17. 108π
19. 216π
21. 0
23. 8
25. 384π

Problem Set 10.8, page 468

1. Integrals $4 \cdot 1 \cdot 1$ $(x = 1)$, $4 \cdot 1 \cdot 1$ $(y = 1)$, $-8 \cdot 1 \cdot 1$ $(z = 1)$, 0 $(x = y = z = 0)$
3. 2 (volume integral of $6y^2$), 2 (surface integral over $x = 1$). Others 0
5. Volume integral of $6y^2 - 6x^2$ is 0. 2 $(x = 1)$, -2 $(y = 1)$, others 0.
7. $\mathbf{F} = [x, \quad y, \quad z]$, div $\mathbf{F} = 3$, In (2), Sec. 10.7, $\mathbf{F}\bullet\mathbf{n} = |\mathbf{F}||\mathbf{n}|\cos\phi$
$= \sqrt{x^2 + y^2 + z^2}\cos\phi = r\cos\phi$.
9. $\mathbf{F} = [x, \quad 0, \quad 0]$, div $\mathbf{F} = 1$, use (2*), Sec. 10.7, etc.

Problem Set 10.9, page 473

1. $[0, \quad 8z, \quad 16]\bullet[0, \quad -1, \quad 1]$, ± 12
3. $[-e^z, \quad -e^x, \quad e^y]\bullet[-1, \quad -1, \quad 1]$, $\pm(e^2 - 1)$
5. S: $[u, \quad v, \quad v^2]$, (curl $\mathbf{F}$)$\bullet\mathbf{N} = -4ve^{2v^2}$, $\pm(4 - 4e^2)$
7. (curl $\mathbf{F}$)$\bullet\mathbf{n} = 3/2$, $\pm 3a^2/2$
9. The sides contribute a, $3a^2/2$, $-a$, 0.
11. curl $\mathbf{F} = [0, \quad 0, \quad 6]$, 24π
13. (curl $\mathbf{F}$)$\bullet\mathbf{n} = 2x - 2y$, $1/3$
15. $-\pi/4$
17. (curl $\mathbf{F}$)$\bullet\mathbf{N} = \pi(\cos\pi x + \sin\pi y)$, 2
19. $\mathbf{F}\bullet\mathbf{r} = [-\sin\theta, \quad \cos\theta]\bullet[-\sin\theta, \quad \cos\theta] = 1, 2\pi, 0$

Chapter 10 Review Questions and Problems, page 473

11. Exact, $-542/3$
13. Not exact, $e^4 - 7$
15. By Green, 1152π
17. By Stokes, $\pm 18\pi$
19. By Stokes, $\pm 12\pi$
21. 4/5, 8/15
23. 0, $4a/3\pi$
25. 8/7, 118/49
27. Direct, 5
29. By Gauss, 100π
31. By Gauss, $40abc$
33. Direct, πh
35. Direct, $5(e^2 - 1)$

Problem Set 11.1, page 485

3. $2\pi/n$, $2\pi/n$, k, k, k/n, k/n

13. $\frac{1}{2} + \frac{2}{\pi}\left(\cos x - \frac{1}{3}\cos 3x + \frac{1}{5}\cos 5x - + \cdots\right)$

15. $\frac{\pi}{2} + \frac{4}{\pi}\left(\cos x + \frac{1}{9}\cos 3x + \frac{1}{25}\cos 5x + \cdots\right)$

17. $\frac{\pi}{4} - \frac{2}{\pi}\left(\cos x + \frac{1}{9}\cos 3x + \frac{1}{25}\cos 5x + \cdots\right)$
$+ \sin x - \frac{1}{2}\sin 2x + \frac{1}{3}\sin 3x - + \cdots$

19. $-\frac{4}{\pi}\left(\cos x + \frac{1}{9}\cos 3x + \frac{1}{25}\cos 5x + \cdots\right)$
$+ 2\left(\sin x + \frac{1}{3}\sin 3x + \frac{1}{5}\sin 5x + \cdots\right)$

21. $\frac{1}{3}\pi^2 - 4\left(\cos x - \frac{1}{4}\cos 2x + \frac{1}{9}\cos 3x - + \cdots\right)$

23. $\frac{1}{6}\pi^2 - \frac{4}{\pi}\cos x - \frac{1}{2}\cos 2x + \frac{4}{27\pi}\cos 3x + \frac{1}{8}\cos 4x - \cdots$

29. $f' = 2x$, $f'' = 2$, $j_1 = 0$, $j_1' = -4\pi$, $j_1'' = 0$, $a_n = \frac{1}{n\pi}\left(-\frac{1}{n}\right)(-4\pi)\cos n\pi$, etc.

Problem Set 11.2, page 490

1. $\frac{4}{\pi}\left(\sin\frac{\pi x}{2} + \frac{1}{3}\sin\frac{3\pi x}{2} + \frac{1}{5}\sin\frac{5\pi x}{2} + \cdots\right)$

3. $\frac{1}{3} - \frac{4}{\pi^2}\left(\cos \pi x - \frac{1}{4}\cos 2\pi x + \frac{1}{9}\cos 3\pi x - + \cdots\right)$

5. Rectifier, $\frac{2}{\pi} - \frac{4}{\pi}\left(\frac{1}{1\cdot 3}\cos 2\pi x + \frac{1}{3\cdot 5}\cos 4\pi x + \frac{1}{5\cdot 7}\cos 6\pi x + \cdots\right)$

7. Rectifier, $\frac{1}{2} - \frac{4}{\pi^2}\left(\cos \pi x + \frac{1}{9}\cos 3\pi x + \frac{1}{25}\cos 5\pi x + \cdots\right)$

9. $\frac{2}{3} + \frac{4}{\pi^2}\left(\cos \pi x - \frac{1}{4}\cos 2\pi x + \frac{1}{9}\cos 3\pi x - \frac{1}{16}\cos 4\pi x + - \cdots\right)$

11. $\frac{3}{4} - \frac{4}{\pi^2}\left(\cos\frac{\pi x}{2} + \frac{1}{2}\cos \pi x + \frac{1}{9}\cos\frac{3\pi x}{2} + \frac{1}{25}\cos\frac{5\pi x}{2} + \frac{1}{18}\cos 3\pi x\right.$
$\left. + \cdots\right)$

13. $\frac{3}{8} + \frac{1}{2}\cos 2x + \frac{1}{8}\cos 4x$

15. Translate by $\frac{1}{2}$. **17.** Set $x = 0$.

Problem Set 11.3, page 496

1. Even, odd, neither, even, neither, odd

3. Odd

5. Neither

7. Odd

9. Odd

11. $\frac{\pi}{2} + \frac{4}{\pi}\left(\cos x + \frac{1}{9}\cos 3x + \frac{1}{25}\cos 5x + \cdots\right)$

13. $\frac{4}{\pi}\left(\sin x - \frac{1}{9}\sin 3x + \frac{1}{25}\sin 5x - + \cdots\right)$

15. $1 - \frac{4}{\pi}\left(\sin\frac{\pi x}{2} + \frac{1}{3}\sin\frac{3\pi x}{2} + \frac{1}{5}\sin\frac{5\pi x}{2} + \cdots\right)$

17. (a) 1, (b) $\frac{4}{\pi}\left(\sin\frac{\pi x}{2} + \frac{1}{3}\sin\frac{3\pi x}{2} + \frac{1}{5}\sin\frac{5\pi x}{2} + \cdots\right)$

19. (a) $1 + \frac{8}{\pi^2}\left(\cos\frac{\pi x}{2} + \frac{1}{9}\cos\frac{3\pi x}{2} + \frac{1}{25}\cos\frac{5\pi x}{2} + \cdots\right)$

(b) $\frac{4}{\pi}\left(\sin\frac{\pi x}{2} + \frac{1}{2}\sin\pi x + \frac{1}{3}\sin\frac{3\pi x}{2} + \frac{1}{4}\sin 2\pi x + \cdots\right)$

21. (a) $\frac{3}{2} - \frac{2}{\pi}\left(\cos\frac{\pi x}{2} - \frac{1}{3}\cos\frac{3\pi x}{2} + \frac{1}{5}\cos\frac{5\pi x}{2} - \frac{1}{7}\cos\frac{7\pi x}{2} + - \cdots\right)$

(b) $\frac{6}{\pi}\left(\sin\frac{\pi x}{2} - \frac{1}{3}\sin\pi x + \frac{1}{3}\sin\frac{3\pi x}{2} + \frac{1}{5}\sin\frac{5\pi x}{2} - \frac{1}{9}\sin 3\pi x + \cdots\right)$

23. (a) $\frac{L}{2} - \frac{4L}{\pi^2}\left(\cos\frac{\pi x}{L} + \frac{1}{9}\cos\frac{3\pi x}{L} + \frac{1}{25}\cos\frac{5\pi x}{L} + \cdots\right)$

(b) $\frac{2L}{\pi}\left(\sin\frac{\pi x}{L} - \frac{1}{2}\sin\frac{2\pi x}{L} + \frac{1}{3}\sin\frac{3\pi x}{L} - + \cdots\right)$

25. (a) $\frac{\pi}{2} + \frac{4}{\pi}\left(\cos x + \frac{1}{9}\cos 3x + \frac{1}{25}\cos 5x + \cdots\right)$

(b) $2\left(\sin x + \frac{1}{2}\sin 2x + \frac{1}{3}\sin 3x + \cdots\right)$

Problem Set 11.4, page 499

3. Use (5).

7. $-\frac{2i}{\pi}\sum_{n=-\infty}^{\infty}\frac{1}{2n+1}e^{(2n+1)ix}$

9. $i\sum_{\substack{n=-\infty\\ n\neq 0}}^{\infty}\frac{(-1)^n}{n}e^{inx}$

11. $\frac{\pi^2}{3} + 2\sum_{\substack{n=-\infty\\ n\neq 0}}^{\infty}\frac{(-1)^n}{n^2}e^{inx}$

13. $\pi + i\sum_{\substack{n=-\infty\\ n\neq 0}}^{\infty}\frac{1}{n}e^{inx}$

Problem Set 11.5, page 501

3. $(0.05n)^2$ in D_n changes to $(0.02n)^2$, which gives $C_5 = 0.5100$, leaving the other coefficients almost unaffected.

5. $y = c_1 \cos \omega t + c_2 \sin \omega t + A(\omega) \cos t$, $A(\omega) = 1/(\omega^2 - 1) < 0$ if $\omega^2 < 1$ (phase shift!) and > 0 if $\omega^2 > 1$

7. $y = c_1 \cos \omega t + c_2 \sin \omega t + \sum_{n=1}^{N} \frac{a_n}{\omega^2 - n^2} \cos nt$

9. $y = c_1 \cos \omega t + c_2 \sin \omega t + \frac{\pi}{2\omega^2} + \frac{4}{\pi}\left(\frac{1}{\omega^2 - 1} \cos t + \frac{1/9}{\omega^2 - 9} \cos 3t + \cdots\right)$

11. $y = c_1 \cos \omega t + c_2 \sin \omega t + \frac{1}{2\omega^2} - \frac{1}{1 \cdot 3(\omega^2 - 4)} \cos 2t - \frac{1}{3 \cdot 5(\omega^2 - 16)} \cos 4t - \cdots$

13. The situation is the same as in Fig. 53 in Sec. 2.8.

15. $y = -\frac{3c}{64 + 9c^2} \cos 3t - \frac{8}{64 + 9c^2} \sin 3t$

17. $y = \sum_{n=1}^{N} \left(-\frac{ncb_n}{D_n} \cos nt + \frac{(1 - n^2)b_n}{D_n} \sin nt\right)$, $D_n = (1 - n^2)^2 + n^2c^2$

19. $I(t) = \sum_{n=1}^{\infty} (A_n \cos nt + B_n \sin nt)$, $A_n = (-1)^{n+1} \frac{240(10 - n^2)}{n^2 D_n}$,

$B_n = (-1)^{n+1} \frac{2400}{nD_n}$, $D_n = (10 - n^2)^2 + 100n^2$

Problem Set 11.6, page 505

1. $F = 2\left(\sin x - \frac{1}{2} \sin 2x + \cdots + \frac{(-1)^{N+1}}{N} \sin Nx\right)$, $E^* = 8.1, 5.0, 3.6, 2.8, 2.3$

3. $F = \frac{\pi}{2} - \frac{4}{\pi}\left(\cos x + \frac{1}{9} \cos 3x + \frac{1}{25} \cos 5x + \cdots\right)$, $E^* = 0.0748, 0.0748, 0.0119, 0.0119, 0.0037$

5. $F = \frac{2}{\pi} - \frac{4}{\pi}\left(\frac{1}{1 \cdot 3} \cos 2x + \frac{1}{3 \cdot 5} \cos 4x + \frac{1}{5 \cdot 7} \cos 6x + \cdots\right)$,
$E^* = 0.5951, 0.0292, 0.0292, 0.0066, 0.0066$

7. $F = \frac{4}{\pi}\left(\sin x + \frac{1}{3} \sin 3x + \frac{1}{5} \sin 5x + \cdots\right)$, $E^* = 1.1902, 1.1902, 0.6243, 0.6243, 0.4206$ (0.1272 when $N = 20$)

9. $\frac{8}{\pi}\left(\sin x + \frac{1}{27} \sin 3x + \frac{1}{125} \sin 5x + \cdots\right)$, $E^* = 0.0295, 0.0295, 0.0015, 0.0015, 0.00023$

Problem Set 11.7, page 512

1. $f(x) = \pi e^{-x}$ $(x > 0)$ gives $A = \displaystyle\int_0^\infty e^{-v} \cos wv\, dv = \frac{1}{1 + w^2}$, $B = \dfrac{w}{1 + w^2}$ (see Example 3), etc.

3. $f(x) = \frac{1}{2}\pi e^{-x}$ gives $A = 1/(1 + w^2)$.

5. Use $f = (\pi/2) \cos v$ and (11) in App. 3.1 to get $A = (\cos(\pi w/2))/(1 - w^2)$.

7. $\dfrac{2}{\pi}\displaystyle\int_0^\infty \frac{\sin aw \cos xw}{w}\, dw$

9. $\dfrac{2}{\pi}\displaystyle\int_0^\infty \frac{\cos w + w \sin w - 1}{w^2} \cos xw\, dw$

11. $\dfrac{2}{\pi}\displaystyle\int_0^\infty \frac{\cos \pi w + 1}{1 - w^2} \cos xw\, dw$

15. $\dfrac{2}{\pi}\displaystyle\int_0^\infty \frac{\sin \pi w}{1 - w^2} \sin xw\, dw$

17. $\dfrac{2}{\pi}\displaystyle\int_0^\infty \frac{\pi w - \sin \pi w}{w^2} \sin xw\, dw$

19. $\dfrac{2}{\pi}\displaystyle\int_0^\infty \frac{wa - \sin wa}{w^2} \sin xw\, dw$

Problem Set 11.8, page 517

1. $\sqrt{\dfrac{2}{\pi}}\left(\dfrac{\sin 2w - 2 \sin w}{w}\right)$

5. $\sqrt{\pi/2}\, e^{-x}$ $(x > 0)$

7. $\sqrt{\pi/2} \cos w$ if $0 < w < \pi/2$ and 0 if $w > \pi/2$

9. Yes, no

11. $\sqrt{(2/\pi)}\, w/(w^2 + \pi^2)$

13. $\mathcal{F}_s(xe^{-x^2/2}) = \mathcal{F}_s(-(e^{-x^2/2})') = w\mathcal{F}_c(e^{-x^2/2}) = we^{-w^2/2}$

17. $\mathcal{F}_c(f') = \mathcal{F}_c(-af) = -a\mathcal{F}_c(f) = -\sqrt{\dfrac{2}{\pi}}\dfrac{a^2}{a^2 + w^2} = w\mathcal{F}_s(f) - \sqrt{\dfrac{2}{\pi}} \cdot 1,$

$\mathcal{F}_s(f) = \sqrt{\dfrac{2}{\pi}}\dfrac{w}{a^2 + w^2}$

19. In (5) for $f(ax)$ set $ax = v$.

Problem Set 11.9, page 528

3. $ik(e^{-ibw} - 1)/(\sqrt{2\pi}w)$

5. $\sqrt{(2/\pi)}k\,(\sin w)/w$

7. $[(1 + iw)e^{-iw} - 1]/(\sqrt{2\pi}w^2)$

9. $\sqrt{(2/\pi)}\, i(\cos w - 1)/w$

11. $\frac{1}{2}e^{-w^2/2}$

13. $(e^{ibw} - e^{-ibw})/(iw\sqrt{2\pi}) = \sqrt{2/\pi}(\sin bw)/w$

Chapter 11 Review Questions and Problems, page 532

11. $\dfrac{4k}{\pi}\left(\sin \pi x + \frac{1}{3}\sin 3\pi x + \frac{1}{5}\sin 5\pi x + \cdots\right)$

13. $4\left(\sin\dfrac{x}{2} - \frac{1}{2}\sin x + \frac{1}{3}\sin\dfrac{3x}{2} - \frac{1}{4}\sin 4x + \frac{1}{5}\sin\dfrac{5x}{2} - + \cdots\right)$

15. $\dfrac{8}{\pi^2}\left(\sin\dfrac{\pi x}{2} - \frac{1}{9}\sin\dfrac{3\pi x}{2} + \frac{1}{25}\sin\dfrac{5\pi x}{2} - + \cdots\right)$

17. $\dfrac{2}{\pi} - \dfrac{4}{\pi}\left(\dfrac{1}{1 \cdot 3}\cos 16\pi x + \dfrac{1}{3 \cdot 5}\cos 32\pi x + \dfrac{1}{5 \cdot 7}\cos 48\pi x + \cdots\right)$

19. $\frac{\pi^2}{12} - \cos 2x + \frac{1}{4}\cos 4x - \frac{1}{9}\cos 6x + \frac{1}{16}\cos 8x - + \cdots$

21. $\pi/4$ by Prob. 11 **23.** $\pi^2/8$ by Prob. 15

25. $\frac{1}{2}[f(x) + f(-x)]$, $\frac{1}{2}[f(x) - f(-x)]$

27. $\pi - \frac{8}{\pi}\left(\cos\frac{x}{2} + \frac{1}{9}\cos\frac{3x}{2} + \frac{1}{25}\cos\frac{5x}{2} + \cdots\right)$

29. 8.105, 4.963, 3.567, 2.781, 2.279, 1.929, 1.673, 1.477

31. $y = C_1\cos\omega t + C_2\sin\omega t + \frac{\pi^2}{3\omega^2} - 4\left(\frac{\cos t}{\omega^2-1} - \frac{1}{4}\cdot\frac{\cos 2t}{\omega^2-4} + \frac{1}{9}\cdot\frac{\cos 3t}{\omega^2-9} - \frac{1}{16}\cdot\frac{\cos 4t}{\omega^2-16} + - \cdots\right)$

33. $\frac{1}{\pi}\int_0^\infty \frac{(\cos w + w\sin w - 1)\cos wx + (\sin w - w\cos w)\sin wx}{w^2}\,dw$

35. $\frac{2}{\pi}\int_0^\infty \frac{w - \sin w\cos w}{w^2}\sin wx\,dw$

37. $\frac{4}{\pi}\int_0^\infty \frac{\sin 2w - 2w\cos 2w}{w^3}\cos wx\,dw$ **39.** $\sqrt{\frac{8}{\pi}}\cdot\frac{1}{w^2+4}$

Problem Set 12.1, page 537

1. $u = c_1(x)\cos 4y + c_2(x)\sin 4y$
3. $u = c_1(x) + c_2(x)y$
5. $u = c(x)e^{-y} + e^{xy}/(x+1)$
7. $u = c(x)\exp(\frac{1}{2}y^2\cosh x)$
9. $u = c_1(x)y + c_2(x)y^{-2}$
11. $u = c(x)e^y + h(y)$
15. $c = 1/4$
17. Any c
19. $\pi/4$
21. Any c and ω
27. $u = 110 - (110/\ln 100)\ln(x^2+y^2)$
29. $u = c_1x + c_2(y)$

Problem Set 12.3, page 546

1. $k\cos 2\pi t\sin 2\pi x$

3. $\frac{8k}{\pi^3}\left(\cos\pi t\sin\pi x + \frac{1}{27}\cos 3\pi t\sin 3\pi x + \frac{1}{125}\cos 5\pi t\sin 5\pi x + \cdots\right)$

5. $\frac{4}{5\pi^2}\left(\cos\pi t\sin\pi x - \frac{1}{9}\cos 3\pi t\sin 3\pi x + \frac{1}{25}\cos 5\pi t\sin 5\pi x - + \cdots\right)$

7. $\frac{2}{\pi^2}\left((\sqrt{2}-1)\cos\pi t\sin\pi x + \frac{1}{2}\cos 2\pi t\sin 2\pi x + \frac{1}{9}(\sqrt{2}+1)\cos 3\pi t\sin 3\pi x - \cdots\right)$

9. $\frac{2}{\pi^2}\left((2-\sqrt{2})\cos \pi t \sin \pi x - \frac{1}{9}(2+\sqrt{2})\cos 3\pi t \sin 3\pi x + \frac{1}{25}(2+\sqrt{2})\cos 5\pi t \sin 5\pi x + \cdots\right)$

17. $u = \frac{8L^2}{\pi^3}\left(\cos\left[c\left(\frac{\pi}{L}\right)^2 t\right]\sin\frac{\pi x}{L} + \frac{1}{3^3}\cos\left[c\left(\frac{3\pi}{L}\right)^2 t\right]\sin\frac{3\pi x}{L} + \cdots\right)$

19. (a) $u(0, t) = 0$, (b) $u(L, t) = 0$, (c) $u_x(0, t) = 0$, (d) $u_x(L, t) = 0$. $C = -A$, $D = -B$ from (a), (b). Insert this. The coefficient determinant resulting from (c), (d) must be zero to have a nontrivial solution. This gives (22).

Problem Set 12.4, page 552

3. $c^2 = 300/[0.9/(2 \cdot 9.80)] = 80.83^2$ [m^2/sec^2]

11. Hyperbolic, $u = f_1(x) + f_2(x + y)$

13. Elliptic, $u = f_1(y + 3ix) + f_2(y - 3ix)$

15. Parabolic, $u = xf_1(x - y) + f_2(x - y)$

17. Parabolic, $u = xf_1(2x + y) + f_2(2x + y)$

19. Hyperbolic, $u = (1/y)f_1(xy) + f_2(y)$

Problem Set 12.5, page 560

5. $u = \sin 0.4\pi x\, e^{-1.752 \cdot 16\pi^2 t/100}$

7. $u = \frac{2}{\pi}\left(\frac{2}{\pi}\sin 0.1\pi x\, e^{-0.01752\pi^2 t} + \frac{1}{2}\sin 0.2\pi x\, e^{-0.01752(2\pi)^2 t} - \cdots\right)$

9. $u = \frac{20\sqrt{2}}{\pi^2}\left(\sin 0.1\pi x\, e^{-0.01752\pi^2 t} + \frac{1}{9}\sin 0.3\pi x\, e^{-0.01752(3\pi)^2 t} - - \cdots\right)$

11. $u = u_{\mathrm{I}} + u_{\mathrm{II}}$, where $u_{\mathrm{II}} = u - u_{\mathrm{I}}$ satisfies the boundary conditions of the text, so that $u_{\mathrm{II}} = \sum_{n=1}^{\infty} B_n \sin\frac{n\pi x}{L} e^{-(cn\pi/L)^2 t}$, $B_n = \frac{2}{L}\int_0^L [f(x) - u_{\mathrm{I}}(x)]\sin\frac{n\pi x}{L}\,dx$

13. $F = A\cos px + B\sin px$, $F'(0) = Bp = 0$, $B = 0$, $F'(L) = -Ap\sin pL = 0$, $p = n\pi/L$, etc.

15. $u = 1$

17. $u = \frac{2\pi^2}{3} + 4\left(\cos x\, e^{-t} - \frac{1}{4}\cos 2x\, e^{-4t} + \frac{1}{9}\cos 3x\, e^{-9t} - + \cdots\right)$

19. $u = \frac{\pi^2}{12} + \cos 2x\, e^{-4t} + \frac{1}{4}\cos 4x\, e^{-16t} + \frac{1}{9}\cos 6x\, e^{-36t} + \cdots$

23. $-\frac{K\pi}{L}\sum_{n=1}^{\infty} nB_n e^{-\lambda_n^2 t}$ **25.** $w = e^{-\beta t}$

27. $v_t - c^2 v_{xx} = 0$, $w'' = -Ne^{-\alpha x}/c^2$, $w = \frac{N}{c^2\alpha^2}\left[-e^{-\alpha x} - \frac{1}{L}(1 - e^{-\alpha L})x + 1\right]$, so that $w(0) = w(L) = 0$.

29. $u = (\sin\frac{1}{2}\pi x \sinh\frac{1}{2}\pi y)/\sinh\pi$

31. $u = \frac{80}{\pi}\sum_{n=1}^{\infty}\frac{1}{(2n-1)\sinh(2n-1)\pi}\sin\frac{(2n-1)\pi x}{24}\sinh\frac{(2n-1)\pi y}{24}$

33. $u = A_0 x + \sum_{n=1}^{\infty} A_n \dfrac{\sinh (n\pi x/24)}{\sinh n\pi} \cos \dfrac{n\pi y}{24}$,

$A_0 = \dfrac{1}{24^2} \displaystyle\int_0^{24} f(y)\, dy, \quad A_n = \dfrac{1}{12} \int_0^{24} f(y) \cos \dfrac{n\pi y}{24}\, dy$

35. $\displaystyle\sum_{n=1}^{\infty} A_n \sin \dfrac{n\pi x}{a} \sinh \dfrac{n\pi(b-y)}{a}$, $A_n = \dfrac{2}{a \sinh (n\pi b/a)} \displaystyle\int_0^a f(x) \sin \dfrac{n\pi x}{a}\, dx$

Problem Set 12.6, page 568

1. $A = \dfrac{2 \sin ap}{\pi p}$, $B = 0$, $u = \dfrac{2}{\pi} \displaystyle\int_0^{\infty} \dfrac{\sin ap}{p} \cos px\, e^{-c^2p^2t}\, dp$

3. $A = e^{-p}$, $B = 0$, $u = \displaystyle\int_0^{\infty} \cos px\, e^{-p-c^2p^2t}\, dp$

5. Set $\pi v = s$. $A = 1$ if $0 < p/\pi < 1$, $B = 0$, $u = \displaystyle\int_0^{\pi} \cos px\, e^{-c^2p^2t}\, dp$

7. $A = 2[\cos p + p \sin p - 1)/(\pi p^2)]$, $B = 0$, $u = \displaystyle\int_0^{\infty} A \cos px\, e^{-c^2p^2t}\, dp$

Problem Set 12.8, page 578

1. (a), (b) It is multiplied by $\sqrt{2}$. (c) Half

3. $B_{mn} = 16/(mn\pi^2)$ if m, n odd, 0 otherwise

5. $B_{mn} = (-1)^{n+1}8/(mn\pi^2)$ if m odd, 0 if m even

7. $B_{mn} = (-1)^{m+n}4/(mn\pi^2)$

11. $k \cos \sqrt{29}\, t \sin 2x \sin 5y$

13. $\dfrac{6.4}{\pi^2} \displaystyle\sum_{\substack{m=1 \\ m,\, n \text{ odd}}}^{\infty} \sum_{n=1}^{\infty} \dfrac{1}{m^3n^3} \cos (t\sqrt{m^2 + n^2}) \sin mx \sin ny$

17. $c\pi\sqrt{260}$ (corresponding eigenfunctions $F_{4,16}$ and $F_{16,14}$), etc.

19. $B_{mn} = 0$ (m or n even), $B_{mn} = 16k/(mn\pi^2)$ (m, n odd)

21. $B_{mn} = (-1)^{m+n}144a^3b^3/(m^3n^3\pi^6)$

23. $\cos \left(\pi t \sqrt{\dfrac{9}{a^2} + \dfrac{16}{b^2}}\right) \sin \dfrac{3\pi x}{a} \sin \dfrac{4\pi y}{b}$

Problem Set 12.9, page 585

7. $30r \cos \theta + 10r^3 \cos 3\theta$

9. $55 + \dfrac{220}{\pi} \left(r \cos \theta - \dfrac{1}{3} r^3 \cos 3\theta + \dfrac{1}{5} r^5 \cos 5\theta - + \cdots\right)$

11. $\dfrac{\pi}{2} - \dfrac{4}{\pi} \left(r \cos \theta + \dfrac{1}{9} r^3 \cos 3\theta + \dfrac{1}{25} r^5 \cos 5\theta + \cdots\right)$

15. Solve the problem in the disk $r < a$ subject to u_0 (given) on the upper semicircle and $-u_0$ on the lower semicircle.

$u = \dfrac{4u_0}{\pi} \left(\dfrac{r}{a} \sin \theta + \dfrac{1}{3a^3} r^3 \sin 3\theta + \dfrac{1}{5a^5} r^5 \sin 5\theta + \cdots\right)$

17. Increase by a factor $\sqrt{2}$

19. $T = 6.826\rho R^2 f_1^2$

21. No

23. Differentiation brings in a factor $1/\lambda_m = R/(c\alpha_m)$.

Problem Set 12.10, page 593

11. $v = F(r)G(t)$, $F'' + k^2F = 0$, $\dot{G} + c^2k^2G = 0$, $F_n = \sin(n\pi r/R)$,

$$G_n = B_n \exp(-c^2n^2\pi^2t/R^2),\quad B_n = \frac{2}{R}\int_0^R rf(r)\sin\frac{n\pi r}{R}\,dr$$

13. $u = 100$

15. $u = \frac{8}{5}r^3P_3(\cos\phi) - \frac{3}{5}rP_1(\cos\phi)$

17. $64r^4P_4(\cos\phi)$

21. Analog of Example 1 in the text with 55 replaced by 50

23. $v = r(\cos\theta)/r^2 = x/(x^2 + y^2)$, $v = xy/(x^2 + y^2)^2$

Problem Set 12.11, page 596

5. $W = \dfrac{c(s)}{x^s} + \dfrac{x}{s^2(s+1)}$, $W(0, s) = 0$, $c(s) = 0$, $w(x, t) = x(t - 1 + e^{-t})$

7. $w = f(x)g(t)$, $xf'g + f\dot{g} = xt$, take $f(x) = x$ to get $g = ce^{-t} + t - 1$ and $c = 1$ from $w(x, 0) = x(c - 1) = 0$.

9. Set $x^2/(4c^2\tau) = z^2$. Use z as a new variable of integration. Use $\operatorname{erf}(\infty) = 1$.

Chapter 12 Review Questions and Problems, page 597

19. $u = c_1(y)e^x + c_2(y)e^{-2x}$

21. $u = g(x)(1 - e^{-y}) + f(x)$

23. $u = \cos t \sin x - \frac{1}{2}\cos 2t \sin 2x$

25. $u = \frac{3}{4}\cos t \sin x - \frac{1}{4}\cos 3t \sin 3x$

27. $u = \sin(0.02\pi x)\, e^{-0.004572t}$

29. $u = \dfrac{200}{\pi^2}\left(\sin\dfrac{\pi x}{50}\, e^{-0.004572t} - \dfrac{1}{9}\sin\dfrac{3\pi x}{50}\, e^{-0.04115t} + \cdots\right)$

31. $u = 100\cos 4x\, e^{-16t}$

33. $u = \dfrac{\pi}{2} - \dfrac{16}{\pi}\left(\dfrac{1}{4}\cos 2x\, e^{-4t} + \dfrac{1}{36}\cos 6x\, e^{-36t} + \dfrac{1}{100}\cos 10x\, e^{-100t} + \cdots\right)$

37. $u = f_1(y) + f_2(x + y)$

39. $u = f_1(y - 2ix) + f_2(y + 2ix)$

41. $u = xf_1(y - x) + f_2(y - x)$

49. $u = (u_1 - u_0)(\ln r)/\ln(r_1/r_0) + (u_0 \ln r_1 - u_1 \ln r_0)/\ln(r_1/r_0)$

Problem Set 13.1, page 606

5. $x - iy = -(x + iy)$, $x = 0$

7. 484

9. $-5/169$

11. $-7/13 - (22/13)i$

13. $-273 + 136i$

15. $-7/17 - (11/17)i$

17. $x/(x^2 + y^2)$

19. $(x^2 - y^2)/(x^2 + y^2)^2$

Problem Set 13.2, page 611

1. $3\sqrt{2}(\cos(-\frac{1}{4}\pi) + i\sin(-\frac{1}{4}\pi))$

3. $5(\cos\pi + i\sin\pi) = 5\cos\pi$

5. $\cos\frac{1}{2}\pi + i\sin\frac{1}{2}\pi$

7. $\frac{1}{3}\sqrt{61}(\cos \arctan \frac{6}{5} + i \sin \arctan \frac{6}{5})$ **9.** $-3\pi/4$
11. $\arctan(\pm 3/4)$ **13.** $\pm\pi/4$ **15.** $3\pi/4$
17. $2.94020 + 0.59601i$ **19.** $0.54030 - 0.84147i$
21. $\cos(-\frac{1}{4}\pi) + i \sin(-\frac{1}{4}\pi)$, $\cos \frac{3}{4}\pi + i \sin \frac{3}{4}\pi$
23. $\pm(1 \pm i)/\sqrt{2}$ **25.** -1, $\cos \frac{1}{5}\pi \pm i \sin \frac{1}{5}\pi$, $\cos \frac{3}{5}\pi \pm i \sin \frac{3}{5}\pi$
27. $4 + 3i$, $4 - 8i$ **29.** $\frac{5}{2} - i$, $2 + \frac{1}{4}i$
35. $|z_1 + z_2|^2 = (z_1 + z_2)(\overline{z_1 + z_2}) = (z_1 + z_2)(\bar{z}_1 + \bar{z}_2)$. Multiply out and use $\operatorname{Re} z_1\bar{z}_2 \leqq |z_1\bar{z}_2|$ (Prob. 32):
$z_1\bar{z}_1 + z_1\bar{z}_2 + z_2\bar{z}_1 + z_2\bar{z}_2 = |z_1|^2 + 2 \operatorname{Re} z_1\bar{z}_2 + |z_2|^2 \leqq |z_1|^2 + 2|z_1||z_2| + |z_2|^2 = (|z_1| + |z_2|)^2$.
Take the square root to get (6).

Problem Set 13.3, page 617

1. Circle of radius $\frac{4}{3}$, center $3 + 2i$
3. Set obtained from an open disk of radius 1 by omitting its center $z = 1$
5. Hyperbola $xy = 1$ **7.** y-axis
9. The region above $y = x$
13. $f = 1 - 1/(z + 1) = 1 - (x + 1 - iy)/[(x + 1)^2 + y^2]$; $0.9 - 0.1i$
15. $(x^2 - y^2 - 2ixy)/(x^2 + y^2)^2$, $-i/2$ **17.** Yes since $r^2(\sin 2\theta)/r \to 0$
19. Yes **21.** $6z^2(z^3 + i)$
23. $2i(1 - z)^{-3}$

Problem Set 13.4, page 623

1. Yes **3.** No **5.** Yes
7. No **9.** Yes for $z \neq 0$
11. $r_x = x/r = \cos\theta$, $r_y = \sin\theta$, $\theta_x = -(\sin\theta)/r$, $\theta_y = (\cos\theta)/r$,
(a) $0 = u_x - v_y = u_r \cos\theta + u_\theta(-\sin\theta)/r - v_r \sin\theta - v_\theta(\cos\theta)/r$.
(b) $0 = u_y + v_x = u_r \sin\theta + u_\theta(\cos\theta)/r + v_r \cos\theta + v_\theta(-\sin\theta)/r$.
Multiply (a) by $\cos\theta$, (b) by $\sin\theta$, and add. Etc.
13. $z^2/2$ **15.** $\ln|z| + i \operatorname{Arg} z$ **17.** z^3
19. No **21.** No **23.** $c = 1$, $\cos x \sinh y$
27. Use (4), (5), and (1).

Problem Set 13.5, page 626

3. $-1.13120 + 2.47173i$, $e = 2.71828$ **5.** $-i$, 1
7. $e^{0.8}(\cos 5 - i \sin 5)$, 2.22554 **9.** $e^{-2x} \cos 2y$, $-e^{-2x} \sin 2y$
11. $\exp(x^2 - y^2) \cos 2xy$, $\exp(x^2 - y^2) \sin 2xy$
13. $e^{i\pi/4}$, $e^{5\pi i/4}$
15. $\sqrt[n]{r} \exp[i(\theta + 2k\pi)/n]$, $k = 0, \cdots, n - 1$
17. $9e^{\pi i}$ **19.** $z = \ln 2 + \pi i + 2n\pi i$ $(n = 0, \pm 1, \cdots)$
21. $z = \ln 5 - \arctan \frac{3}{4}i \pm 2n\pi i$ $(n = 0, 1, \cdots)$

Problem Set 13.6, page 629

3. Use (11), then (5) for e^{iy}, and simplify. **5.** Use (11) and simplify.
7. $\cos 1 \cosh 1 - i \sin 1 \sinh 1 = 0.83373 - 0.98890i$

9. 74.203, 74.210
11. $-3.7245 - 0.51182i$
13. -1
15. $\cosh 4 = 27.308$
17. $z = \pm(2n + 1)\pi i/2$
19. $z = \frac{1}{2}(2n + 1)\pi - (-1)^n 1.4436i$
21. $z = \pm n\pi i$
25. Insert the definitions on the left, multiply out, simplify.

Problem Set 13.7, page 633

1. $\ln 10 + \pi i$
3. $\frac{1}{2}\ln 8 - \frac{1}{4}\pi i$
5. $\ln 5 + (\arctan \frac{4}{3} - \pi)i = 1.609 - 2.214i$
7. $0.9273i$
9. $\frac{1}{2}\ln 2 - \frac{1}{4}\pi i$
11. $\pm(2n + 1)\pi i, \quad n = 0, 1, \cdots$
13. $\ln 6 \pm (2n + 1)\pi i, \quad n = 0, 1, \cdots$
15. $(\pi - 1 \pm 2n\pi)i, \quad n = 0, 1, \cdots$
17. $\ln (i^2) = (\pm 2n + 1)\pi i$, $2 \ln i = \pm(4n + 1)\pi i, \quad n = 0, 1, \cdots$
19. $e^{0.3}(\cos 0.7 + i \sin 0.7) = 1.032 + 0.870i$
21. $e^2(1 + i)/\sqrt{2}$
23. $64(\cos (\ln 4) + i \sin (\ln 4))$
25. $2.8079 + 1.3179i$
27. $(1 + i)/\sqrt{2}$

Chapter 13 Review Questions and Problems, page 634

17. $-32 - 24i$
19. $-\frac{1}{2} - \frac{1}{2}i$
21. $5 - 3i$
23. $6\sqrt{2}e^{3\pi i/4}$
25. $12e^{-\pi i/2}$
27. $\pm(2 + 2i)$
29. $(\pm 1 \pm i)/\sqrt{2}$
31. $f(z) = 1/z$
33. $f(z) = (1 + i)z^2$
35. $f(z) = e^{z^2}$
37. $(-x^2 + y^2)/2$
39. No
41. 0
43. $0.6435i$
45. -1.5431

Problem Set 14.1, page 645

1. Straight segment from $1 + 3i$ to $4 + 12i$
3. Circle of radius 3, center $4 + i$
5. Semicircle, radius 1, center 0
7. Ellipse, half-axes 6 and 5
9. Parabola $y = \frac{1}{2}x^3$ from $-1 - \frac{1}{2}i$ to $2 + 4i$
11. e^{-it} $(0 \leqq t \leqq 2\pi)$
13. $t + i/t$ $(1 \leqq t \leqq 4)$
15. $t + (4 - 4t^2)i$ $(-1 \leqq t \leqq 1)$
17. $-a - ib + re^{-it}$ $(0 \leqq t \leqq 2\pi)$
19. $\frac{1}{2} + \frac{1}{2}i$
21. 0
23. $\pi i + \frac{1}{2}i \sinh 2\pi$
25. $i/2$
27. $-1 + i \tanh \frac{1}{4}\pi = -1 + 0.6558i$
29. $2 \sinh \frac{1}{2}$

Problem Set 14.2, page 653

1. πi, no
3. 0, yes
5. 0, yes
7. 0, yes
9. 0, no
11. 0, yes
15. Yes, by the deformation principle
19. πi by path deformation
21. πi
23. $2\pi i$
25. 0
27. (a) 0, (b) π
29. 0

Problem Set 14.3, page 657

1. -4π **3.** 4π **5.** $8\pi i$
7. 0 **9.** $-\pi i$ **11.** πi
13. π **15.** $2\pi i \operatorname{Ln} 4 = 8.710i$
17. $\pi i \cosh^2(1 + i) = \pi(-0.2828 + 1.6489i)$

Problem Set 14.4, page 661

1. $27\pi i/4$ **3.** $-2\pi i e^{\pi/2}$ **5.** $\pi i a^3/3$
7. $2\pi i$ if $|a| < 2$, 0 if $|a| > 2$ **9.** $\pi i(\cos \frac{1}{2} - \sin \frac{1}{2})$
11. $2\pi^2 i$ **13.** $\pi i e^{a/2}/24$ if $|a - 2 - i| < 3$, 0 if $|a - 2 - i| > 3$

Chapter 14 Review Questions and Problems, page 662

17. $-6\pi i$ **19.** 0 **21.** $-\pi i$
23. $\frac{1}{4} i \sin 8$ **25.** 0 **27.** π **29.** 0

Problem Set 15.1, page 672

1. Bounded, divergent, ± 1 **3.** Bounded, convergent, 0
5. Unbounded
7. Bounded, divergent, $\pm 1/\sqrt{2} \pm i$, 0, 1, -2
9. Convergent, 0
13. $|z_n - l| < \frac{1}{2}\epsilon$, $|z_n^* - l^*| < \frac{1}{2}\epsilon$ $(n > N(\epsilon))$, hence $|z_n + z_n^* - (l + l^*)| < \frac{1}{2}\epsilon + \frac{1}{2}\epsilon$
17. Convergent **19.** Divergent
21. Conditionally convergent **23.** Divergent by Theorem 3
27. $n = |100 + 75i| = 125$ (why?); $|100 + 75i|^{125}/125! = 125^{125}/[\sqrt{250\pi}\,(125/e)^{125}]$
$= e^{125}/\sqrt{250\pi} = 6.91 \cdot 10^{52}$

Problem Set 15.2, page 677

1. $\sum a_n z^{2n} = \sum a_n (z^2)^n$, $|z^2| < R = \lim |a_n/a_{n+1}|$, hence $|z| < \sqrt{R}$.
3. $-i$, 1 **5.** -1, e by (6) and $(1 + 1/n)^n \to e$.
7. 0, $|b/a|$ **9.** 0, 1 **11.** 0, 1
13. $3 - 2i$, 1 **15.** i, $1/\sqrt[4]{2}$ **17.** 0, $\sqrt{2}$

Problem Set 15.3, page 682

1. 3 **3.** $\sqrt{2}$ **5.** $\sqrt{5/3}$
7. $1/\sqrt{7}$ **9.** 1

Problem Set 15.4, page 690

1. $1 - 2z + 2z^2 - \frac{4}{3}z^3 + \frac{2}{3}z^4 - + \cdots$, $R = \infty$
3. $e^{-2i}(1 + (z + 2i) + \frac{1}{2}(z + 2i)^2 + \frac{1}{6}(z + 2i)^3 + \frac{1}{24}(z + 2i)^4 + \cdots$, $R = \infty$
5. $1 - \frac{1}{2}(z - \frac{1}{2}\pi)^2 + \frac{1}{24}(z - \frac{1}{2}\pi)^4 - \frac{1}{720}(z - \frac{1}{2}\pi)^6 + - \cdots$, $R = \infty$
7. $\frac{1}{2} + \frac{1}{2}i + \frac{1}{2}i(z - i) + (-\frac{1}{4} + \frac{1}{4}i)(z - i)^2 - \frac{1}{4}(z - i)^3 + - \cdots$, $R = \sqrt{2}$

9. $1 - \frac{1}{2}z^2 + \frac{1}{8}z^4 - \frac{1}{48}z^6 + \frac{1}{384}z^8 - + \cdots, \quad R = \infty$

11. $4(z-1) + 10(z-1)^2 + 16(z-1)^3 + 14(z-1)^4 + 6(z-1)^5 + (z-1)^6$

13. $(2/\sqrt{\pi})(z - z^3/3 + z^5(2!5) - z^7/(3!7) + \cdots), \quad R = \infty$

15. $z^3/(1!3) - z^7/(3!7) + z^{11}/(5!11) - + \cdots, \quad R = \infty$

19. $z + \frac{1}{3}z^3 + \frac{2}{15}z^5 + \frac{17}{315}z^7 + \cdots, \quad R = \frac{1}{2}\pi$

Problem Set 15.5, page 697

1. Use Theorem 1.

3. $R = 1/\sqrt{\pi} > 0.56$

5. $|z^n| \leqq 1$ and $\Sigma\, 1/n^2$ converges.

7. $|\tanh^n |z|| \leqq 1,\ 1/(n^2+1) < 1/n^2$

9. $|z + 1 - 2i| \leqq r < R = 4$

11. $|z| \leqq 2 - \delta\ (\delta > 0)$

13. Nowhere

15. $|z| \leqq \sqrt{5} - \delta\ (\delta > 0)$

Chapter 15 Review Questions and Problems, page 698

11. $\infty,\ e^{3z}$

13. $1,\ \frac{1}{2}\,\mathrm{Ln}[(1+z)/(1-z)]$

15. ∞

17. $1/\sqrt{\pi},\ [1 - \pi(z-2i)^2]^{-1}$

19. 1/3

21. $-1 - (z - \pi i) - (z - \pi i)^2/2! - \cdots, \quad R = \infty$

23. $\frac{1}{2} + \frac{1}{4}(z+1) + \frac{1}{8}(z+1)^2 + \frac{1}{16}(z+1)^3 + \cdots, \quad R = 2$

25. $1 + 3z + 6z^2 + 10z^3 + \cdots,\ R = 1$. Differentiate the geometric series.

27. $i + (z+i) - i(z+i)^2 - (z+i)^3 + \cdots, \quad R = 1$

29. $-(z - \frac{1}{2}\pi) + \dfrac{1}{3!}(z - \frac{1}{2}\pi)^3 - \dfrac{1}{5!}(z - \frac{1}{2}\pi)^5 + - \cdots,\ R = \infty$

Problem Set 16.1, page 707

1. $\dfrac{1}{z^4} + \dfrac{1}{z^3} + \dfrac{1}{z^2} + \dfrac{1}{z} + 1 + z + z^2 + \cdots,\ R = 1$

3. $\dfrac{1}{z^3} - \dfrac{1}{z^2} + \dfrac{1}{2z} - \dfrac{1}{6} + \dfrac{1}{24}z - \dfrac{1}{120}z^2 + - \cdots,\ R = \infty$

5. $\dfrac{1}{z^3} + \dfrac{1}{z^5} + \dfrac{1}{2z^7} + \dfrac{1}{6z^9} + \cdots,\ R = \infty$

7. $\dfrac{e^{z-1}e}{z-1} = e\sum_{n=0}^{\infty}\dfrac{(z-1)^{n-1}}{n!} = e\left[\dfrac{1}{z-1} + 1 + \dfrac{z-1}{2!} + \cdots\right],\ R = \infty$

9. $-\sum_{n=0}^{\infty}\left(\dfrac{i}{2}\right)^{n+1}(z-i)^{n-1} = -\dfrac{i/2}{z-i} + \dfrac{1}{4} + \dfrac{i}{8}(z-i) - \dfrac{1}{16}(z-i)^2 - \cdots,$
$R = 2$

11. $-\sum_{n=0}^{\infty}(z+i)^{n-1} = -\dfrac{1}{z+i} - 1 - (z+i) - \cdots,\ R = 1$

13. $-\dfrac{3}{z-1} + 2 + (z-1)$

15. $\sum_{n=0}^{\infty} z^{3n}$, $|z| < 1$, $\quad -\sum_{n=0}^{\infty} \frac{1}{z^{3n+3}}$, $\quad |z| > 1$

17. $\sum_{n=0}^{\infty} z^{4n+2}$, $|z| < 1$, $\quad -\sum_{n=0}^{\infty} \frac{1}{z^{4n+2}}$, $\quad |z| > 1$

19. $\frac{i}{(z-i)^2} + \frac{1}{z-i} + i + (z-i)$

21. $(1-4z)\sum_{n=0}^{\infty} z^{4n}$, $|z| < 1$, $\left(\frac{4}{z^3} - \frac{1}{z^4}\right)\sum_{n=0}^{\infty} \frac{1}{z^{4n}}$, $|z| > 1$

23. $\sum_{n=0}^{\infty} \frac{(-1)^{n+1}(z+\frac{1}{2}\pi)^{2n-1}}{(2n)!}$, $|z+\frac{1}{2}\pi| > 0$

Problem Set 16.2, page 711

1. $\pm\frac{1}{2}, \pm\frac{3}{2}, \cdots$ (poles of 2nd order), ∞ (essential singularity)
3. $0, \pm\sqrt{\pi}, \pm\sqrt{2\pi}, \cdots$ (simple poles), ∞ (essential singularity)
5. ∞ (essential singularity)
7. $\pm 1, \pm i$ (fourth-order poles), ∞ (essential singularity)
9. $\pm i$ (essential singularities)
13. $-16i$ (fourth order)
15. $\pm 1, \pm 2, \cdots$ (third order)
17. $\pm i/\sqrt{3}$ (simple)
19. $\pm 2i$ (simple), $0, \pm 2\pi i, \pm 4\pi i, \cdots$ (second order)
21. $0, \pm 2\pi, \pm 4\pi, \cdots$ (fourth order)
23. $f(z) = (z-z_0)^n g(z)$, $g(z_0) \neq 0$, hence $f^2(z) = (z-z_0)^{2n} g^2(z)$.

Problem Set 16.3, page 717

1. i, $4i$
3. $-\frac{1}{4}i$ (at $z = 2i$), $\frac{1}{4}i$ (at $-2i$)
5. $1/5!$ (at $z = 0$)
7. 1 (at $\pm n\pi$)
9. $-\frac{1}{4}$ (at $z = 1$), $\frac{1}{4}$ (at $z = -1$)
11. -1 (at $z = \pm\frac{1}{2}\pi, \pm\frac{3}{2}\pi, \cdots$)
15. $e^{1/z} = 1 + 1/z + \cdots$, *Ans.* $2\pi i$
17. Simple poles at $\pm\frac{1}{2}$. *Ans.* $-4i$
19. $-4\pi i \sinh \frac{1}{2}\pi$
21. $-4i \sinh \frac{1}{2}$
23. 0
25. $\frac{1}{2}$ (at $z = \frac{1}{2}$), 2 (at $z = \frac{1}{3}$). *Ans.* $5\pi i$

Problem Set 16.4, page 725

1. $2\pi/\sqrt{13}$
3. $2\pi/35$
5. $2\pi/3$
7. 0
9. π
11. $2\pi/3$
13. $\pi/16$
15. 0
17. $\pi/2$
19. 0
21. 0
23. π
25. 0
27. $-\pi/2$

Chapter 16 Review Questions and Problems, page 726

17. $2\pi i/3$
19. 5π
21. $\frac{1}{2}\pi \cos 10$
23. $\pi i/4$
25. 0 (n even), $(-1)^{(n-1)/2}\, 2\pi i/(n-1)!$ (n odd)
27. $6\pi i$
29. $2\pi/7$
31. $4\pi/\sqrt{3}$
33. 0
35. $\pi/2$

Problem Set 17.1, page 733

3. Only the size
5. $x = c$, $w = -y + ic$, $y = k$, $w = -k + ix$
7. $-3\pi/4 < \text{Arg } w < 3\pi/4$, $|w| < 1/8$
9. $|w| > 3$
11. $|w| \leqq 16$, $v \geqq 0$
13. Annulus $3 < |w| < 5$
15. $\ln 2 \leqq u \leqq \ln 3$, $\pi/4 \leqq v \leqq \pi/2$
17. ± 1, $\pm i$
19. $0, \pm 1, \pm 2, \cdots$
21. $-a/2$
23. a and 0, $\sqrt[3]{a}$
25. $M = e^x = 1$ when $x = 0$, $J = e^{2x}$
27. $M = 1/|z| = 1$ on the unit circle, $J = 1/|z|^2$

Problem Set 17.2, page 737

5. $z = \dfrac{-iw}{-2w + 3}$
7. $z = \dfrac{5i}{4w - 2}$

9. $z = 0$
11. $z = \frac{1}{2} + i \pm \sqrt{\frac{1}{4} + i}$
13. $z = \pm i$
15. $w = 4/z$, etc.
17. $a - d = 0$, $b/c = 1$ by (5)
19. $w = az/d$ $(a \neq 0, d \neq 0)$

Problem Set 17.3, page 741

5. Apply the inverse g of f on both sides of $z_1 = f(z_1)$ to get $g(z_1) = g(f(z_1)) = z_1$.
7. $w = (z + 2i)/(z - 2i)$
9. $w = z - 4$
11. $w = 1/z$
13. $w = (3iz + 1)/z$
15. $w = (z + 1)/(-3z + 1)$
17. $w = (2z - i)/(-iz - 2)$
19. $w = (z^4 - i)/(-iz^4 + 1)$

Problem Set 17.4, page 745

1. Annulus $1 \leqq |w| \leqq e^2$
3. $1/\sqrt{e} < |w| < \sqrt{e}$, $3\pi/4 < \arg w < 5\pi/4$
5. $1 < |w| < e$, $v > 0$
7. w-plane without 0
9. $u^2/\cosh^2 1 + v^2/\sinh^2 1 \leqq 1$, $u \geqq 0$
11. Elliptic annulus bounded by $u^2/\cosh^2 1 + v^2/\sinh^2 1 = 1$ and $u^2/\cosh^2 5 + v^2/\sinh^2 5 = 1$
13. $\pm(2n + 1)\pi/2$, $n = 0, 1, \cdots$
15. $0 < \text{Im } t < \pi$ is the image of R under $t = z^2$. *Ans.* $e^t = e^{z^2}$
17. $0, \pm i, \pm 2i, \cdots$
19. $u^2/\cosh^2 1 + v^2/\sinh^2 1 \leqq 1$, $v < 0$
21. $v < 0$
23. $-1 \leqq u \leqq 1$, $v = 0$ $(c = 0)$, $u^2/\cosh^2 c + v^2/\sinh^2 c = 1$ $(c \neq 0)$
25. $\ln 2 \leqq u \leqq \ln 3$, $\pi/4 \leqq v \leqq \pi/2$

Problem Set 17.5, page 747

1. w moves once around the unit circle.
5. $-5/3$, 2 sheets
7. $-i/2$, 3 sheets
9. 0, 2 sheets

Chapter 17 Review Questions and Problems, page 747

11. $u = \frac{1}{4}v^2 - 1, \frac{1}{4}v^2 - 1$ **13.** $|w| = 20.25, |\arg w| < \pi/2$
15. The domain between $u = \frac{1}{4} - v^2$ and $u = 1 - \frac{1}{4}v^2$
17. $|w + \frac{1}{2}| = \frac{1}{2}$ **19.** $u = 1$ **21.** $|\arg w| < \pi/4$
23. $0, (\pm 1 \pm i)/\sqrt{2}$ **25.** $\pi/8 \pm n\pi/2, n = 0, 1, \cdots$
27. $0, \pm i/\sqrt{2}$ **29.** $w = iz$ **31.** $w = 1/z$
33. $w = z/(z + 2)$ **35.** $\pm\sqrt{2}$ **37.** $2 \pm \sqrt{6}$
39. $1 + i \pm \sqrt{1 + 2i}$ **41.** $w = e^{3z}$ **43.** $z^2/2k$
45. $iz^3 + 1$

Problem Set 18.1, page 753

1. $20x + 200, 20z + 200$ **3.** $110 - 50xy, 110 + 25iz^2$
5. $F = (110/\ln 2)\text{Ln } z$ **7.** $F = 200 - (100/\ln 2)\text{ Ln } z$
13. Use Fig. 388 in Sec. 17.4 with the z- and w-planes interchanged, and $\cos z = \sin(z + \frac{1}{2}\pi)$.
15. $\Phi = 220 - 110xy$

Problem Set 18.2, page 757

1. $u^2 - v^2 = e^{2x}(\cos^2 y - \sin^2 y), \Phi_{xx} = 4e^{2x}(\cos^2 y - \sin^2 y) = -\Phi_{yy}, \nabla^2\Phi = 0$
3. Straightforward calculation, involving the chain rule and the Cauchy-Riemann equations
5. See Fig. 389 in Sec. 17.4. $\Phi = \sin^2 x \cosh^2 y - \cos^2 x \sinh^2 y$.
9. (i) $\Phi = U_1(1 - xy)$. (ii) $w = iz^2$ maps R onto $-2 \leqq u \leqq 0$, thus $\Phi^* = U_1(1 + \frac{1}{2}u) = U_1(1 + \frac{1}{2}(-2xy))$.
11. By Theorem 1 in Sec. 17.2
13. $\Phi = 10[1 - (1/\pi)\text{ Arg }(z - 4)], F = 10[1 + (i/\pi)\text{ Ln }(z - 4)]$
15. Corresponding rays in the w-plane make equal angles, and the mapping is conformal.

Problem Set 18.3, page 760

3. $(100/d)y$. Rotate through $\pi/2$. **5.** $100 - 240\theta/\pi$
7. Re $F(z) = 100 + (200/\pi)$ Re $(\arcsin z)$ **9.** $(240/\pi)$ Arg z
11. $T_0 + (2/\pi)(T_1 - T_0)$ Arg z **13.** $50 + (400/\pi)$ Arg z

Problem Set 18.4, page 766

1. $V = iV_2 = iK, \Psi = -Kx = const, \Phi = Ky = const$
3. $F(z) = Kz$ (K positive real)
5. $V = (1 + 2i)K, F = (1 - 2i)Kz$
7. $F(z) = z^3$
9. Hyperbolas $(x + 1)y = const$. Flow around a corner formed by $x = -1$ and the x-axis.
11. $y/(x^2 + y^2) = c$ or $x^2 + (y - k)^2 = k^2$
13. $F(z) = z/r_0 + r_0/z$

15. Use that $w = \arccos z$ is $w = \cos z$ with the roles of the z- and w-planes interchanged.

Problem Set 18.5, page 771

5. $1 - r^2 \cos 2\theta$

7. $2(r \sin \theta - \frac{1}{2}r^2 \sin 2\theta + \frac{1}{3}r^3 \sin 3\theta - + \cdots)$

9. $\frac{3}{4}r^2 \sin 2\theta - \frac{1}{4}r^6 \sin 6\theta$

11. $\frac{1}{3}\pi^2 - 4(r \cos \theta - \frac{1}{4}r^2 \cos 2\theta + \frac{1}{9}r^3 \cos 3\theta - + \cdots)$

13. $\dfrac{4}{\pi}\left(r \sin \theta - \dfrac{1}{9}\, r^3 \sin 3\theta + \dfrac{1}{25}\, r^5 \sin 5\theta - + \cdots\right)$

Problem Set 18.6, page 774

1. No; $|z|^2$ is not analytic. **3.** Use (2). $F(\frac{1}{3}) = \frac{25}{9}$ **5.** $\Phi(4, -4) = -12$

7. Use (3). $\Phi(1, 1) = -2$ **11.** $|F(e^{i\theta})|^2 = 2 - 2 \cos 2\theta$, $\theta = \pi/2$, Max $= 2$

13. $|F(z)| = [\cos^2 2x + \sinh^2 2y]^{1/2}$, $z = \pm i$, Max $= [1 + \sinh^2 2]^{1/2} = \cosh 2 = 3.7622$

15. No

Chapter 18 Review Questions and Problems, page 775

11. $\Phi = 10(1 - x + y)$, $F = 10 - 10(1 + i)z$

13. $(20/\ln 10)$ Ln z

15. $(10/\ln 10)(\ln 100 - \ln r)$

17. Arg $z = const$

19. $(-i/\pi)$ Ln z

23. $T(x, y) = x(2y + 1) = const$

25. Circles $(x - c)^2 + y^2 = c^2$

27. $F(z) = \dfrac{c}{2\pi}$ Ln $(z - 5)$, Arg $(z - 5) = c$

29. $20 + \dfrac{80}{\pi}\left(r \sin \theta + \dfrac{1}{3}\, r^3 \sin 3\theta + \dfrac{1}{5}\, r^5 \sin 5\theta + \cdots\right)$

Problem Set 19.1, page 786

1. $0.9817 \cdot 10^2$, $-0.1010 \cdot 10^3$, $0.5787 \cdot 10^{-2}$, $-0.1360 \cdot 10^5$

3. $0.36443/(17.862 - 17.798) = 0.36443/0.064 = 5.6942$, $0.3644/(17.86 - 17.80) = 0.3644/0.06 = 6.073$, $0.364/(17.9 - 17.8) = 3.64$, impossible

5. $\dfrac{0.36443(17.862 + 17.798)}{17.862^2 - 17.798^2} = \dfrac{0.36443 \cdot 35.660}{319.05 - 316.77} = \dfrac{12.996}{2.28} = 5.7000,$

$\dfrac{13.00}{2.28} = 5.702,$ $\dfrac{13.0}{2.28} = 5.70,$ $\dfrac{13}{2.3} = 5.7,$ $\dfrac{10}{2} = 5$

7. 19.95, 0.049, 0.05013; 20, 0, 0.05

9. In the present calculation, (b) is more accurate than (a).

11. $-0.126 \cdot 10^{-2}$, $-0.402 \cdot 10^{-3}$; $-0.267 \cdot 10^{-6}$, $-0.847 \cdot 10^{-7}$

13. Add first, then round.

15. $\dfrac{a_1}{a_2} = \dfrac{\tilde{a}_1 + \epsilon_1}{\tilde{a}_2 + \epsilon_2} = \dfrac{\tilde{a}_1 + \epsilon_1}{\tilde{a}_2}\left(1 - \dfrac{\epsilon_2}{\tilde{a}_2} + \dfrac{{\epsilon_2}^2}{{\tilde{a}_2}^2} - + \cdots\right) \approx \dfrac{\tilde{a}_1}{\tilde{a}_2} + \dfrac{\epsilon_1}{\tilde{a}_2} - \dfrac{\epsilon_2}{\tilde{a}_2} \cdot \dfrac{\tilde{a}_1}{\tilde{a}_2},$

hence $\left|\left(\frac{a_1}{a_2} - \frac{\tilde{a}_1}{\tilde{a}_2}\right) \Big/ \left|\frac{a_1}{a_2}\right|\right| \approx \left|\frac{\epsilon_1}{a_1} - \frac{\epsilon_2}{a_2}\right| \leqq |\epsilon_{r1}| + |\epsilon_{r2}| \leqq \beta_{r1} + \beta_{r2}$

19. (a) 19/21 = 0.904761905, $\epsilon_{\text{chop}} = \epsilon_{\text{round}} = 0.1905 \cdot 10^{-5}$, $\epsilon_{r,\text{chop}} = \epsilon_{r,\text{round}} = 0.2106 \cdot 10^{-5}$, etc.

Problem Set 19.2, page 796

1. $g = 1.4 \sin x$, 1.37263 ($= x_5$)
5. $g = x^4 + 0.2$, 0.20165 ($= x_3$)
7. 2.403 ($= x_5$, exact to 3S)
9. 0.904557 ($= x_3$)
11. 1.834243 ($= x_4$)
13. $x_0 = 4.5$, $x_4 = 4.73004$ (6S exact)
15. (a) 0.5, 0.375, 0.377968, 0.377964; (b) $1/\sqrt{7} = 0.377964\,473$
17. $x_{n+1} = (2x_n + 7/x_n^2)/3$, 1.912931 ($= x_3$)
19. (a) **Algorithm Bisect** (f, a_0, b_0, N) **Bisection Method**

This algorithm computes an interval $[a_n, b_n]$ containing a solution of $f(x) = 0$ (f continuous) or it computes a solution c_n, given an initial interval $[a_0, b_0]$ such that $f(a_0)f(b_0) < 0$. Here N is determined by $(b - a)/2^N \leqq \beta$, β the required accuracy.

INPUT: Initial interval $[a_0, b_0]$, maximum number of iterations N.
OUTPUT: Interval $[a_N, b_N]$ containing a solution, or a solution c_n.
For $n = 0, 1, \cdots, N - 1$ do:
  Compute $c_n = \frac{1}{2}(a_n + b_n)$.
  If $f(c_n) = 0$ then OUTPUT c_n. Stop. [*Procedure completed*]
  Else continue.
  If $f(a_n)f(c_n) < 0$ then $a_{n+1} = a_n$ and $b_{n+1} = c_n$.
  Else set $a_{n+1} = c_n$ and $b_{n+1} = b_n$.
End
OUTPUT $[a_N, b_N]$. Stop.
[*Procedure completed*]
End BISECT

Note that $[a_N, b_N]$ gives $(a_N + b_N)/2$ as an approximation of the zero and $(b_N - a_N)/2$ as a corresponding error bound.

(b) 0.739085; (c) 1.30980, 0.429494
21. 1.834243
23. 0.904557

Problem Set 19.3, page 808

1. $L_0(x) = -2x + 19$, $L_1(x) = 2x - 18$, $p_1(x) = 0.1082x + 1.2234$, $p_1(9.4) = 2.2405$
3. 0.9971, 0.9943, 0.9915 (0.9916 4D), 0.9861 (0.9862 4D), 0.9835, 0.9809
5. $p_2(x) = -0.44304x^2 + 1.30906x - 0.02322$, $p_2(0.75) = 0.70929$
7. $p_2(x) = -0.1434x^2 + 1.0895x$, $p_2(0.5) = 0.5089$, $p_2(1.5) = 1.3116$
9. $L_0 = -\frac{1}{6}(x - 1)(x - 2)(x - 3)$, $L_1 = \frac{1}{2}x(x - 2)(x - 3)$, $L_2 = -\frac{1}{2}x(x - 1)(x - 3)$, $L_3 = \frac{1}{6}x(x - 1)(x - 2)$; $p_3(x) = 1 + 0.039740x - 0.335187x^2 + 0.060645x^3$; $p_3(0.5) = 0.943654$ (6S-exact 0.938470), $p_3(1.5) = 0.510116$ (0.511828), $p_3(2.5) = -0.047993$ (-0.048384)

13. $p_2(x) = 0.9461x - 0.2868x(x-1)/2 = -0.1434x^2 + 1.0895x$
15. 0.722, 0.786
17. $\delta f_{1/2} = 0.057\,839$, $\delta f_{3/2} = 0.069\,704$, etc.

Problem Set 19.4, page 815

9. $[-1.39(x-5)^2 + 0.58(x-5)^3]'' = 0.004$ at $x = 5.8$ (due to roundoff; should be 0).
11. $1 - \frac{5}{4}x^2 + \frac{1}{4}x^4$
13. $4 - 12x^2 - 8x^3$, $4 - 12x^2 + 8x^3$. Yes
15. $1 - x^2$, $-2(x-1) - (x-1)^2 + 2(x-1)^3$,
$-1 + 2(x-2) + 5(x-2)^2 - 6(x-2)^3$
17. Curvature $f''/(1 + f'^2)^{3/2} \approx f''$ if $|f'|$ is small
19. Use that the third derivative of a cubic polynomial is constant, so that g''' is piecewise constant, hence constant throughout under the present assumption. Now integrate three times.

Problem Set 19.5, page 828

1. 0.747131
3. 0.69377 (5S-exact 0.69315)
5. 1.566 (4S-exact 1.557)
7. 0.894 (3S-exact 0.908)
9. $J_{h/2} + \epsilon_{h/2} = 1.55963 - 0.00221 = 1.55742$ (6S-exact 1.55741)
11. $J_{h/2} + \epsilon_{h/2} = 0.90491 + 0.00349 = 0.90840$ (5S-exact 0.90842)
13. 0.94508, 0.94583 (5S-exact 0.94608)
15. 0.94614588, 0.94608693 (8S-exact 0.94608307)
17. 0.946083 (6S-exact)
19. 0.9774586 (7S-exact 0.9774377)
21. $x - 2 = t$, 1.098609 (7S-exact 1.098612)
23. $x = \frac{1}{2}(t+1)$, 0.7468241268 (10S-exact 0.7468241330)
25. (a) $M_2 = 2$, $M_2^* = \frac{1}{4}$, $|KM_2| = 2/(12n^2)$, $n = 183$, (b) $f^{(iv)} = 24/x^5$, $2m = 14$
27. 0.08, 0.32, 0.176, 0.256 (exact)
29. $5(0.1040 - \frac{1}{2}\cdot 0.1760 + \frac{1}{3}\cdot 0.1344 - \frac{1}{4}\cdot 0.0384) = 0.256$

Chapter 19 Review Questions and Problems, page 830

17. 4.266, 4.38, 6.0, impossible
19. 49.980, 0.020; 49.980, 0.020008
21. $17.5565 \leqq s \leqq 17.5675$
23. The same as that of $\bar{a}$.
25. -0.2, -0.20032, -0.200323
27. 3, 2.822785, 2.801665, 2.801386, 2.801386
29. 2.95647, 2.96087
31. 0.26, $M_2 = 6$, $M_2^* = 0$, $-0.02 \leqq \epsilon \leqq 0$, $0.24 \leqq a \leqq 0.26$
33. 1.001005, $-0.001476 \leqq \epsilon \leqq 0$

Problem Set 20.1, page 839

1. $x_1 = -2.4$, $x_2 = 5.3$
3. No solution
5. $x_1 = 2$, $x_2 = 1$
7. $x_1 = 6.78$, $x_2 = -11.3$, $x_3 = 15.82$

9. $x_1 = 0$, $x_2 = t_1$ arbitrary, $x_3 = 5t_1 + 10$
11. $x_1 = t_1$, $x_2 = t_2$, both arbitrary, $x_3 = 1.25t_1 - 2.25t_2$
13. $x_1 = 1.5$, $x_2 = -3.5$, $x_3 = 4.5$, $x_4 = -2.5$

Problem Set 20.2, page 844

1. $\begin{bmatrix} 1 & 0 \\ 5 & 1 \end{bmatrix} \begin{bmatrix} 3 & 2 \\ 0 & 1 \end{bmatrix}$, $x_1 = 4.2$, $x_2 = 1.3$ **3.** $\begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 4 & -6 \\ 0 & 5 \end{bmatrix}$, $x_1 = -4$, $x_2 = 3$

5. $\begin{bmatrix} 1 & 0 & 0 \\ \frac{2}{3} & 1 & 0 \\ \frac{1}{2} & 6 & 1 \end{bmatrix} \begin{bmatrix} 6 & 4 & 3 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{2} \end{bmatrix}$, $x_1 = 0.5$, $x_2 = -2.5$, $x_3 = 3.0$

7. $\begin{bmatrix} 1 & 0 & 0 \\ 6 & 1 & 0 \\ 3 & 9 & 1 \end{bmatrix} \begin{bmatrix} 3 & 9 & 6 \\ 0 & -6 & 3 \\ 0 & 0 & -3 \end{bmatrix}$, $x_1 = -1/30$, $x_2 = 2/15$, $x_3 = 1/5$

9. $\begin{bmatrix} 3 & 0 & 0 \\ 2 & 3 & 0 \\ 4 & 1 & 3 \end{bmatrix} \begin{bmatrix} 3 & 2 & 4 \\ 0 & 3 & 1 \\ 0 & 0 & 3 \end{bmatrix}$, $x_1 = 3$, $x_2 = 6$, $x_3 = 2$

11. $\begin{bmatrix} 2 & 0 & 0 \\ 3 & 5 & 0 \\ 4 & 8 & 7 \end{bmatrix} \begin{bmatrix} 2 & 3 & 4 \\ 0 & 5 & 8 \\ 0 & 0 & 7 \end{bmatrix}$, $x_1 = 4$, $x_2 = 0$, $x_3 = -2$

13. $\begin{bmatrix} 2 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 2 & 1 & 1 & 0 \\ 0 & 2 & 1 & 2 \end{bmatrix} \begin{bmatrix} 2 & 1 & 2 & 0 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 2 \end{bmatrix}$, $x_1 = 6$, $x_2 = -2$, $x_3 = 0$, $x_4 = 14$

17. $\begin{bmatrix} -2 & 4 & -1 \\ -2 & 3 & 0 \\ 7 & -12 & 2 \end{bmatrix}$ **19.** $\begin{bmatrix} \frac{341}{6} & -10 & \frac{7}{6} \\ -\frac{15}{2} & \frac{4}{3} & -\frac{1}{6} \\ -17 & 3 & -\frac{1}{3} \end{bmatrix}$

Problem Set 20.3, page 850

3. Exact 21.5, 0, −13.8 **5.** Exact 2, 1, 4 **7.** Exact 0.5, 0.5, 0.5
9. (a) $\mathbf{x}^{(3)T} = [0.49982 \quad 0.50001 \quad 0.50002]$, (b) $\mathbf{x}^{(3)T} = [0.50333 \quad 0.49985 \quad 0.49968]$
11. 6, 15, 46, 96 steps; spectral radius 0.09, 0.35, 0.72, 0.85, approximately
13. $[1.99934 \quad 1.00043 \quad 3.99684]^T$ (Jacobi, step 5); $[2.00004 \quad 0.998059 \quad 4.00072]^T$ (Gauss-Seidel)
17. $\sqrt{306} = 17.49$, 12, 12 **19.** $\sqrt{18k^2} = 4.24|k|$, $4|k|$, $4|k|$

Problem Set 20.4, page 858

1. 12, $\sqrt{62} = 7.87$, 6, $[\frac{1}{6} \quad -1 \quad \frac{5}{6}]$ **3.** 14, $\sqrt{50} = 7.07$, 4, $[-1 \quad 1 \quad \frac{3}{4} \quad -\frac{3}{4}]$
5. 1.9, $\sqrt{1.35} = 1.16$, 1, [0.3 −0.1 0.5 1.0]
7. 6, $\sqrt{6}$, 1, [1 1 1 1 1 1] **11.** $\| \mathbf{A} \|_1 = 17$, $\| \mathbf{A}^{-1} \|_1 = 17$, $\kappa = 289$
13. $\kappa = 100 \cdot 100$ **15.** $\kappa = 1.2 \cdot \frac{120}{98} = 1.469$
17. $46 \leqq 6 \cdot 17$ or $7 \cdot 17$ **19.** $[0 \quad 1]^T$, $[1 \quad -0.4]^T$, 289
21. $[-0.6 \quad 2.8]^T$ **23.** 27, 748, 28375, 943656, 29070279

Problem Set 20.5, page 862

1. $-11.4 + 5.4x$ **3.** $8.95 - 0.388x$
5. $s = -675 + 90t$, $v_{\text{av}} = 90$ km/h **9.** $4 - 0.75x - 0.125x^2$
11. $5.248 + 1.543x$, $3.900 + 0.5321x + 2.021x^2$
13. $-2.448 + 16.23x$, $-9.114 + 13.73x + 2.500x^2$,
$-2.270 + 1.466x$ $-1.778x^2 + 2.852x^3$

Problem Set 20.7, page 871

1. $5 \leqq \lambda \leqq 9$ **3.** 5, 0, 7; radii 4, 6, 6
5. $|\lambda - 4i| \leqq \sqrt{2} + 0.1$, $|\lambda| \leqq 0.1$, $|\lambda - 9i| \leqq \sqrt{2}$
7. $t_{11} = 100$, $t_{22} = t_{33} = 1$
9. They lie in the intervals with endpoints $a_{jj} \pm (n-1)10^{-6}$. (Why?)
11. 0 lies in no Gerschgorin disk, by (3) with $>$; hence $\det \mathbf{A} = \lambda_1 \cdots \lambda_n \neq 0$.
13. $\rho(\mathbf{A}) \leqq$ Row sum norm $\| \mathbf{A} \|_\infty = \max_j \sum_k |a_{jk}| = \max_j$ ($|a_{jj}|$ + Gerschgorin radius)
15. $\sqrt{153} = 12.37$ **17.** $\sqrt{122} = 11.05$ **19.** $6 \leqq \lambda \leqq 10$, $8 \leqq \lambda \leqq 8$

Problem Set 20.8, page 875

1. $q = 4, 4.493, 4.4999$; $|\epsilon| \leqq 1.5, 0.1849, 0.0206$
3. $q = 8, 8.1846, 8.2252$; $|\epsilon| \leqq 1, 0.4769, 0.2200$
5. $q = 4, 4.786, 4.917$; $|\epsilon| \leqq 1.63, 0.619, 0.399$
7. $q = 5.5, 5.5738, 5.6018$; $|\epsilon| \leqq 0.5, 0.3115, 0.1899$; eigenvalues (4S) 1.697, 3.382, 5.303, 5.618
9. $\mathbf{y} = \mathbf{Ax} = \lambda \mathbf{x}$, $\mathbf{y}^T\mathbf{x} = \lambda \mathbf{x}^T\mathbf{x}$, $\mathbf{y}^T\mathbf{y} = \lambda^2 \mathbf{x}^T\mathbf{x}$,
$\epsilon^2 \leqq \mathbf{y}^T\mathbf{y}/\mathbf{x}^T\mathbf{x} - (\mathbf{y}^T\mathbf{x}/\mathbf{x}^T\mathbf{x})^2 = \lambda^2 - \lambda^2 = 0$
11. $q = 1, \cdots, -2.8993$ approximates -3 (0 of the given matrix),
$|\epsilon| \leqq 1.633, \cdots, 0.7024$ (Step 8)

Problem Set 20.9, page 882

1.
$$\begin{bmatrix} 3.500000 & -1.802776 & 0 \\ -1.802776 & 6.730769 & 1.846154 \\ 0 & 1.846154 & 1.769230 \end{bmatrix}$$

3.
$$\begin{bmatrix} 0.980000 & -0.441814 & 0 \\ -0.441814 & 0.870164 & 0.371803 \\ 0 & 0.371803 & 0.489836 \end{bmatrix}$$

5. Eigenvalues 8, 3, 1

$$\begin{bmatrix} 5.64516 & -2.50867 & 0 \\ -2.50867 & 5.307219 & 0.374953 \\ 0 & 0.374953 & 1.04762 \end{bmatrix}, \begin{bmatrix} 7.45139 & -1.56325 & 0 \\ -1.56325 & 3.544142 & 0.0983071 \\ 0 & 0.0983071 & 1.00446 \end{bmatrix},$$

$$\begin{bmatrix} 7.91494 & -0.646602 & 0 \\ -0.646602 & 3.08458 & 0.0312469 \\ 0 & 0.0312469 & 1.000482 \end{bmatrix}$$

7.
$$\begin{bmatrix} 18.3171 & 0.881767 & 0 \\ 0.881767 & 8.29042 & 0.360275 \\ 0 & 0.360275 & 1.39250 \end{bmatrix}, \begin{bmatrix} 18.3786 & 0.396511 & 0 \\ 0.396511 & 8.24727 & 0.0600924 \\ 0 & 0.0600924 & 1.37414 \end{bmatrix},$$

$$\begin{bmatrix} 18.3910 & 0.177669 & 0 \\ 0.177669 & 8.23540 & 0.0100214 \\ 0 & 0.0100214 & 1.37363 \end{bmatrix}$$

9.
$$\begin{bmatrix} 7.00224 & 0.0571287 & 0 \\ 0.0571287 & 4.00088 & 0.0249333 \\ 0 & 0.0249333 & 0.996875 \end{bmatrix}, \begin{bmatrix} 7.00298 & 0.0326363 & 0 \\ 0.0326363 & 4.00034 & 0.00621221 \\ 0 & 0.00621221 & 0.996681 \end{bmatrix},$$

$$\begin{bmatrix} 7.00322 & 0.0186419 & 0 \\ 0.0186419 & 4.00011 & 0.00154782 \\ 0 & 0.00154782 & 0.996669 \end{bmatrix}$$

Chapter 20 Review Questions and Problems, page 883

17. $[4 \quad -1 \quad 2]^T$ **19.** $[6 \quad -3 \quad 1]^T$

21. All nonzero entries of the factors are 1.

23.
$$\begin{bmatrix} 2.8193 & -1.5904 & -0.0482 \\ -1.5904 & 1.2048 & -0.0241 \\ -0.0482 & -0.0241 & 0.1205 \end{bmatrix}$$ (4D-values) **25.** Exact $[-2 \quad 1 \quad 2]^T$

27. 15, $\sqrt{89}$, 8 **29.** 7, $\sqrt{21}$, 4 **31.** 14, $\sqrt{78}$, 7

33. 6 **35.** 9 **37.** $11.5 \cdot 4.4578 = 51.2651$

39. $y = 1.98 + 0.98x$

41. Centers 1, 1, 1, radii 2.5, 1, 2.5 ($\lambda = 2.944, 0.028 \pm 0.290i$, 3D)

43. Centers 5, 6, 8; radii 2, 1, 1, ($\lambda = 4.1864, 6.4707, 8.3429$, 5S)

45.
$$\begin{bmatrix} 1.5 & -2.23607 & 0 \\ -2.23607 & 5.8 & -3.1 \\ 0 & -3.1 & 6.7 \end{bmatrix}$$, Step 3: $$\begin{bmatrix} 9.44973 & -1.06216 & 0 \\ -1.06216 & 4.28682 & -0.00308 \\ 0 & -0.00308 & 0.26345 \end{bmatrix}$$

Problem Set 21.1, page 897

1. $y = e^x$, 0.0382, 0.1245 (error of x_5, x_{10})
3. $y = x - \tanh x$ (set $y - x = u$), 0.009292, 0.0188465 (error of x_5, x_{10})
5. $y = e^x$, 0.001275, 0.004200 (error of x_5, x_{10})
7. $y = 1/(1 - x^2/2)$, 0.00029, 0.01187 (error of x_5, x_{10})
9. $y = 1/(1 - x^2/2)$, 0.03547, 0.28715 (error of x_5, x_{10})
11. $y = 1/(1 - x^2/2)$; error -10^{-8}, $-4 \cdot 10^{-8}$, $\cdots$, $-6 \cdot 10^{-7}$, $+10^{-5}$; about $1.3 \cdot 10^{-5}$ by (10)
13. $y = xe^x$; error $\cdot 10^5$ (for $x = 1, \cdots, 3$) 19, 46, 85, 139, 213, 315, 454, 640, 889, 1219
15. $y = 3 \cos x - 2 \cos^2 x$; error $\cdot 10^7$: 0.18, 0.74, 1.73, 3.28, 5.59, 9.04, 14.33, 22.77, 36.80, 61.42
17. $y = 1/(x^5 + 1)$, 0.000307, -0.000259 (error of x_5, x_{10})
19. The errors are for E.-C. 0.02000, 0.06287, 0.05076, for Improved E.-C. -0.000455, 0.012086, 0.009601, for RK 0.0000012, 0.000016, 0.000536.

Problem Set 21.2, page 901

3. $y = e^{-0.1x^2}$; errors 10^{-6} to $6 \cdot 10^{-6}$
5. $y = \tan x$; $y_4, \cdots, y_{10}$ (error $\cdot 10^5$): 0.422798 (−0.48), 0.546315 (−1.2), 0.684161 (−2.4), 0.842332 (−4.4), 1.029714 (−7.5), 1.260288 (−13), 1.557626 (−22)
7. RK-error smaller, error $\cdot 10^5 = 0.4, 0.3, 0.2, 5.6$ (for $x = 0.4, 0.6, 0.8, 1.0$)
9. $y_4 = 4.229\,690$, $y_5 = 4.556\,859$, $y_6 = 5.360\,657$, $y_7 = 8.082\,563$
11. Errors between $-6 \cdot 10^{-7}$ and $+3 \cdot 10^{-7}$. Solution $e^x - x - 1$
13. Errors $\cdot 10^5$ from $x = 0.3$ to 0.7: -5, -11, -19, -31, -47
15. (a) 0, 0.02, 0.0884, 0.215 848, $y_4 = 0.417\,818$, $y_5 = 0.708\,887$ (poor).
(b) By 30–50%

Problem Set 21.3, page 908

3. $y_1 = e^x$, $y_2 = -e^x$, errors range from ± 0.02 to ± 0.23, monotone.
5. $y_1' = y_2$, $y_2' = -4y_1$, $y = y_1 = 1, 0.84, 0.52, 0.0656, -0.4720$; $y = \cos 2x$
7. $y_1 = 4e^{-x} \sin x$, $y_2 = 4e^{-x} \cos x$; errors from 0 to about ± 0.1
9. Errors smaller by about a factor 10^4
11. $y = 0.198669, 0.389494, 0.565220, 0.719632, 0.847790$;
$y' = 0.980132, 0.922062, 0.830020, 0.709991, 0.568572$
13. $y_1 = e^{-3x} - e^{-5x}$, $y_2 = e^{-3x} + e^{-5x}$; $y_1 = 0.1341, 0.1807, 0.1832, 0.1657, 0.1409$; $y_2 = 1.348, 0.9170, 0.6300, 0.4368, 0.3054$
17. You get the exact solution, except for a roundoff error [e.g., $y_1 = 2.761\,608$, $y(0.2) = 2.7616$ (exact), etc.]. Why?
19. $y = 0.198669, 0.389494, 0.565220, 0.719631, 0.847789$;
$y' = 0.980132, 0.922061, 0.830019, 0.709988, 0.568568$

Problem Set 21.4, page 916

3. 105, 155, 105, 115; Step 5: 104.94, 154.97, 104.97, 114.98
5. 0.108253, 0.108253, 0.324760, 0.324760; Step 10: 0.108538, 0.108396, 0.324902, 0.324831

7. 0, 0, 0, 0. All equipotential lines meet at the corners (why?). Step 5: 0.29298, 0.14649, 0.14649, 0.073245
9. $-3u_{11} + u_{12} = -200$, $u_{11} - 3u_{12} = -100$
11. $u_{12} = u_{32} = 31.25$, $u_{21} = u_{23} = 18.75$, $u_{jk} = 25$ at the others
13. $u_{21} = u_{23} = 0.25$, $u_{12} = u_{32} = -0.25$, $u_{jk} = 0$ else
15. (a) $u_{11} = -u_{12} = -66$. (b) Reduce to 4 equations by symmetry.
$u_{11} = u_{31} = -u_{15} = -u_{35} = -92.92$, $u_{21} = -u_{25} = -87.45$,
$u_{12} = u_{32} = -u_{14} = -u_{34} = -64.22$, $u_{22} = -u_{24} = -53.98$,
$u_{13} = u_{23} = u_{33} = 0$
17. $\sqrt{3}$, $u_{11} = u_{21} = 0.0849$, $u_{12} = u_{22} = 0.3170$. (0.1083, 0.3248 are 4S-values of the solution of the linear system of the problem.)

Problem Set 21.5, page 921

5. $u_{11} = 0.766$, $u_{21} = 1.109$, $u_{12} = 1.957$, $u_{22} = 3.293$
7. **A** as in Example 1, right sides -2, -2, -2, -2. Solution $u_{11} = u_{21} = 1.14286$, $u_{12} = u_{22} = 1.42857$
11. $-4u_{11} + u_{21} + u_{12} = -3$, $u_{11} - 4u_{21} + u_{22} = -12$, $u_{11} - 4u_{12} + u_{22} = 0$, $2u_{21} + 2u_{12} - 12u_{22} = -14$, $u_{11} = u_{22} = 2$, $u_{21} = 4$, $u_{12} = 1$. Here $-14/3 = -\frac{4}{3}(1 + 2.5)$ with 4/3 from the stencil.
13. $\mathbf{b} = [-380 \quad -190, \quad -190, \quad 0]^T$; $u_{11} = 140$, $u_{21} = u_{12} = 90$, $u_{22} = 30$

Problem Set 21.6, page 927

5. 0.1636, 0.2545 ($t = 0.04$, $x = 0.2$, 0.4), 0.1074, 0.1752 ($t = 0.08$), 0.0735, 0.1187 ($t = 0.12$), 0.0498, 0.0807 ($t = 0.16$), 0.0339, 0.0548 ($t = 0.2$; exact 0.0331, 0.0535)
7. Substantially less accurate, 0.15, 0.25 ($t = 0.04$), 0.100, 0.163 ($t = 0.08$)
9. Step 5 gives 0, 0.06279, 0.09336, 0.08364, 0.04707, 0.
11. Step 2: 0 (exact 0), 0.0453 (0.0422), 0.0672 (0.0658), 0.0671 (0.0628), 0.0394 (0.0373), 0 (0)
13. 0.1018, 0.1673, 0.1673, 0.1018 ($t = 0.04$), 0.0219, 0.0355, $\cdots$ ($t = 0.20$)
15. 0.3301, 0.5706, 0.4522, 0.2380 ($t = 0.04$), 0.06538, 0.10604, 0.10565, 0.6543 ($t = 0.20$)

Problem Set 21.7, page 930

1. For $x = 0.2$, 0.4 we obtain 0.012, 0.02 ($t = 0.2$), 0.004, 0.008 ($t = 0.4$), -0.004, -0.008 ($t = 0.6$), etc.
3. $u(x, 1) = 0$, -0.05, -0.10, -0.15, -0.075, 0
5. 0.190, 0.308, 0.308, 0.190 (0.178, 0.288, 0.288, 0.178 exact to 3D)
7. 0, 0.354, 0.766, 1.271, 1.679, 1.834, $\cdots$ ($t = 0.1$); 0, 0.575, 0.935, 1.135, 1.296, 1.357, $\cdots$ ($t = 0.2$)

Chapter 21 Review Questions and Problems, page 930

17. $y = \tan x$; 0 (0), 0.10050 (-0.00017), 0.20304 (-0.00033), 0.30981 (-0.00047), 0.42341 (-0.00062), 0.54702 (-0.00072)

19. 0.1003349 ($0.8 \cdot 10^{-7}$) 0.2027099 ($1.6 \cdot 10^{-7}$), 0.3093360 ($2.1 \cdot 10^{-7}$), 0.4227930 ($2.3 \cdot 10^{-7}$), 0.5463023 ($1.8 \cdot 10^{-7}$)
25. $y(0.4) = 1.822798$, $y(0.5) = 2.046315$, $y(0.6) = 2.284161$, $y(0.7) = 2.542332$, $y(0.8) = 2.829714$, $y(0.9) = 3.160288$, $y(1.0) = 3.557626$
27. $y_1 = 3e^{-9x}$, $y_2 = -5e^{-9x}$, [1.23251 −2.05419], [0.506362 −0.843937], · · ·, [0.035113 −0.058522]
29. 1.96, 7.86, 29.46
31. $u(P_{11}) = u(P_{31}) = 270$, $u(P_{21}) = u(P_{13}) = u(P_{23}) = u(P_{33}) = 30$, $u(P_{12}) = u(P_{32}) = 90$, $u(P_{22}) = 60$
35. 0.06279, 0.09336, 0.08364, 0.04707
37. 0, −0.352, −0.153, 0.153, 0.352, 0 if $t = 0.12$ and 0, 0.344, 0.166, −0.166, −0.344, 0 if $t = 0.24$
39. 0.010956, 0.017720, 0.017747, 0.010964 if $t = 0.2$

Problem Set 22.1, page 939

3. $f = 3(x_1 - 2)^2 + 2(x_2 + 4)^2 - 44$. Step 3: $[2.0055 \quad -3.9975]^T$
5. $f = 0.5(x_1 - 1)^2 + 0.7(x_2 + 3)^2 - 5.8$, Step 3: $[0.99406 \quad -3.0015]^T$
7. $f = 0.2(x_1 - 0.2)^2 + x_2^2 - 0.008$. Step 3: $[0.493 \quad -0.011]^T$, Step 6: $[0.203 \quad 0.004]^T$

Problem Set 22.2, page 943

1. x_3, x_4 unused time on M_1, M_2, respectively
3. No
11. $f_{max} = f(0, 5) = 10$
13. $f_{max} = f(9, 6) = 36$
15. $f_{min} = f(3.5, 2.5) = -30$
17. $x_1/3 + x_2/2 \leqq 100$, $x_1/3 + x_2/6 \leqq 80$, $f = 150x_1 + 100x_2$, $f_{max} = f(210, 60) = 37500$
19. $0.5x_1 + 0.75x_2 \leqq 45$ (copper), $0.5x_1 + 0.25x_2 \leqq 30$, $f = 120x_1 + 100x_2$, $f_{max} = f(45, 30) = 8400$

Problem Set 22.3, page 946

1. $f(120/11, 60/11) = 480/11$
3. $f\left(\dfrac{2100}{3}, \dfrac{200}{2/3}\right) = 78000$
5. Matrices with Rows 2 and 3 and Columns 4 and 5 interchanged
7. $f(0, \frac{5}{10}) = -10$
9. $f(5, 4, 6) = 478$

Problem Set 22.4, page 952

1. $f(4, 4) = 72$
3. $f(20, 30) = 50$
5. $f(10, 5) = 5500$
7. $f(1, 1, 0) = 12$
9. $f(\frac{1}{2}, 0, \frac{1}{2}) = 3$

Chapter 22 Review Questions and Problems, page 952

11. Step 5: $[0.353 \quad -0.028]^T$. Slower
13. Of course! Step 5: $[-1.003 \quad 1.897]^T$
21. $f(2, 4) = 100$
23. $f(3, 6) = -54$
25. $f(50, 100) = 150$

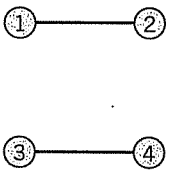
Problem Set 23.1, page 958

9.
$$\begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}$$

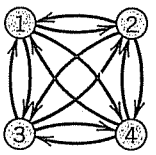
11.
$$\begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

13.
$$\begin{bmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

15.



15.



21.

Vertex	Edge e_1	Edge e_2	Edge e_3
1	1	1	0
2	0	0	1
3	1	0	0
4	0	1	1

23.

Vertex					
1	−1	0	0	1	−1
2	1	−1	1	0	0
3	0	1	−1	−1	1

25.

Vertex	Incident Edges
1	$-e_1, -e_2, e_3, -e_4$
2	e_1
3	$e_2, -e_3$
4	e_4

Problem Set 23.2, page 962

1. 4 **3.** 5 **5.** 4

9. The idea is to go backward. There is a v_{k-1} adjacent to v_k and labeled $k - 1$, etc. Now the only vertex labeled 0 is s. Hence $\lambda(v_0) = 0$ implies $v_0 = s$, so that $v_0 - v_1 - \cdots - v_{k-1} - v_k$ is a path $s \to v_k$ that has length k.

15. No; there is no way of traveling along (3, 4) only once.

21. From m to $100m$, $10m$, $2.5m$, $m + 4.6$

Problem Set 23.3, page 966

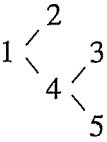
1. (1, 2), (2, 4), (4, 3); $L_2 = 6$, $L_3 = 18$, $L_4 = 14$

3. (1, 2), (1, 4), (2, 3); $L_2 = 2$, $L_3 = 5$, $L_4 = 5$

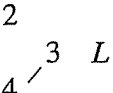
5. (1, 4), (2, 4), (3, 4), (3, 5); $L_2 = 4$, $L_3 = 3$, $L_4 = 2$, $L_5 = 8$

7. (1, 5), (2, 3), (2, 6), (3, 4), (3, 5); $L_2 = 9$, $L_3 = 7$, $L_4 = 8$, $L_5 = 4$, $L_6 = 14$

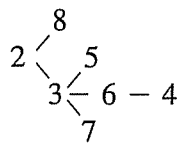
Problem Set 23.4, page 969

1.


$L = 12$

3.


$L = 10$

5.


$L = 28$

9. $1 - 3 - 4 \begin{cases} 2 \\ 5 - 6, \end{cases}$ $L = 38$ **11.** Yes

15. G is connected. If G were not a tree, it would have a cycle, but this cycle would provide two paths between any pair of its vertices, contradicting the uniqueness.

19. If we add an edge (u, v) to T, then since T is connected, there is a path $u \to v$ in T which, together with (u, v), forms a cycle.

Problem Set 23.5, page 972

1. (1, 2), (1, 4), (3, 4), (4, 5), $L = 12$

3. (1, 2), (2, 8), (8, 7), (8, 6), (6, 5), (2, 4), (4, 3), $L = 40$

5. (1, 4), (3, 4), (2, 4), (3, 5), $L = 20$

7. (1, 2), (1, 3), (1, 4), (2, 6), (3, 5), $L = 32$

11. If G is a tree

13. A shortest spanning tree of the largest connected graph that contains vertex 1

Problem Set 23.6, page 978

1. $1 - 2 - 5$, $\Delta f = 2$; $1 - 4 - 2 - 5$, $\Delta f = 2$, etc.

3. $1 - 2 - 4 - 6$, $\Delta f = 2$; $1 - 2 - 3 - 5 - 6$, $\Delta f = 1$, etc.

5. $f_{12} = 4$, $f_{13} = 1$, $f_{14} = 4$, $f_{42} = 4$, $f_{43} = 0$, $f_{25} = 8$, $f_{35} = 1$, $f = 9$

7. $f_{12} = 4$, $f_{13} = 3$, $f_{24} = 4$, $f_{35} = 3$, $f_{54} = 2$, $f_{46} = 6$, $f_{56} = 1$, $f = 7$

9. {4, 5, 6}, 28 **11.** {2, 4, 6}, 50

13. $1 - 2 - 3 - 7$, $\Delta f = 2$; $1 - 4 - 5 - 6 - 7$, $\Delta f = 1$;
$1 - 2 - 3 - 6 - 7$, $\Delta f = 1$; $f_{\max} = 14$

15. {3, 5, 7}, 22 **17.** $S = \{1, 4\}$, cap $(S, T) = 6 + 8 = 14$

19. If $f_{ij} < c_{ij}$ as well as $f_{ij} > 0$

Problem Set 23.7, page 982

3. (2, 3) and (5, 6)

5. $1 - 2 - 5$, $\Delta_t = 2$; $1 - 4 - 2 - 5$, $\Delta_t = 1$; $f = 6 + 2 + 1 = 9$

7. $1 - 2 - 4 - 6$, $\Delta_t = 2$; $1 - 3 - 5 - 6$, $\Delta_t = 1$; $f = 4 + 2 + 1 = 7$

9. By considering only edges with one labeled end and one unlabeled end

17. $S = \{1, 2, 4, 5\}$, $T = \{3, 6\}$, cap $(S, T) = 14$

Problem Set 23.8, page 986

1. No **3.** No **5.** Yes, $S = \{1, 4, 5, 8\}$

7. Yes; a graph is not bipartite if it has a nonbipartite subgraph.

9. $1 - 2 - 3 - 5$

11. (1, 5), (2, 3) by inspection. The augmenting path $1 - 2 - 3 - 5$ gives $1 - 2 - 3 - 5$, that is, (1, 2), (3, 5).

13. (1, 4), (2, 3), (5, 7) by inspection. Or (1, 2), (3, 4), (5, 7) by the use of the path $1 - 2 - 3 - 4$.

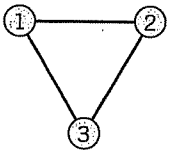
15. 3 **19.** 3 **23.** No; K_5 is not planar.

25. K_3

Chapter 23 Review Questions and Problems, page 987

13. $\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$ **15.** $\begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}$

17. $\begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$ **19.**



21.

Vertex	Incident Edges
1	$e_2, -e_3$
2	$-e_1, e_3$
3	$e_1, -e_2$

23. 4

25. 4

27. $L_2 = 10, L_3 = 15, L_4 = 13$

29. $1 - 4 - 3 - 2, L = 16$

33. $f = 7$

Problem Set 24.1, page 996

1. $q_L = 19, q_M = 20, q_U = 20.5$

3. $q_L = 38, q_M = 44, q_U = 54$

5. $q_L = 69.7, q_M = 70.5, q_U = 71.2$

7. $q_L = 2.3, q_M = 2.4, q_U = 2.45$

9. $q_L = 399, q_M = 401, q_U = 401$

11. $\bar{x} = 19.875, s = 0.835$, IQR $= 1.5$

13. $\bar{x} = 70.49, s = 1.047$, IQR $= 1.5$

15. $\bar{x} = 400.4, s = 1.618$, IQR $= 2$

17. 0 0 300

19. 3.54, 1.29

Problem Set 24.2, page 999

1. 4 outcomes: HH, HT, TH, TT (H = Head, T = Tail)

3. $6^2 = 36$ outcomes (1, 1), (1, 2), $\cdots$, (6, 6)

5. Infinitely many outcomes $S, S^cS, S^cS^cS, \cdots$ (S = "Six")

7. The space of ordered triples of nonnegative numbers

9. The space of ordered pairs of numbers

11. Yes

13. $E = \{S, S^cS, S^cS^cS\}, E^c = \{S^cS^cS^cS, S^cS^cS^cS^cS, \cdots\}$ (S = "Six")

Problem Set 24.3, page 1005

1. (a) $0.9^3 = 72.9\%$, (b) $\frac{90}{100} \cdot \frac{89}{99} \cdot \frac{88}{98} = 72.65\%$

3. $\frac{490}{500} \cdot \frac{489}{499} \cdot \frac{488}{498} \cdot \frac{487}{497} \cdot \frac{486}{496} = 90.35\%$

5. $1 - \frac{1}{25} = 0.96$

7. $1 - 0.75^2 = 0.4375 < 0.5$

9. $P(MMM) + P(MMFM) + P(MFMM) + P(FMMM) = \frac{1}{8} + 3 \cdot \frac{1}{16} = \frac{5}{16}$

11. $\frac{6}{36} + \frac{27}{36} - \frac{3}{36} = \frac{30}{36}$ by Theorem 3, or by counting outcomes
13. $0.08 + 0.04 - 0.08 \cdot 0.04 = 11.68\%$
15. $0.95^4 = 81.5\%$ **17.** $1 - 0.97^4 = 11.5\%$

Problem Set 24.4, page 1010

3. In 40 320 ways
5. $\binom{20}{3} = 1140$
7. 210, 70, 112, 28
9. 9!/(2!3!4!) = 1260. *Ans.* 1/1260
11. $\binom{52}{13} = 635\,013\,559\,600$
13. 1/84, 5/21
15. 676 000

Problem Set 24.5, page 1015

1. $k = 1/55$ by (6)
3. $k = 1/8$ by (10)
5. No because of (6)
7. $1 - P(X \leqq 3) = 0.5$
9. $P(X > 1200) = \int_{1.2}^{2} 6[0.25 - (x - 1.5)^2]\,dx = 0.896$. *Ans.* $0.896^3 = 72\%$
11. $k = 2.5$; 50%
13. $k = 1.1565$; 26.9%
17. $X > b$, $X \geqq b$, $X < c$, $X \leqq c$, etc.

Problem Set 24.6, page 1019

1. 2/3, 1/18
3. 3.5, 2.917
5. 4, 16/3
7. \$643.50
9. $\mu = 1/\theta = 25$; $P = 20.2\%$
11. $\frac{1}{2}$, $\frac{1}{20}$, $(X - \frac{1}{2})\sqrt{20}$
13. 750, 1, 0.002
15. $15c - 500c^3 = 0.97$, $c = 0.0855$

Problem Set 24.7, page 1025

1. 0.0625, 0.25, 0.9375, 0.9375
3. 64%
5. 0.265
7. $f(x) = 0.5^x e^{-0.5}/x!$, $f(0) + f(1) = e^{-0.5}(1.0 + 0.5) = 0.91$. *Ans.* 9%
9. $1 - e^{-0.2} = 18\%$
11. $0.99^{100} = 36.6\%$
13. $\frac{120}{286}$, $\frac{135}{286}$, $\frac{30}{286}$, $\frac{1}{286}$

Problem Set 24.8, page 1031

1. 0.1587, 0.6306, 0.5, 0.4950
3. 17.29, 10.71, 19.152
5. 16%
7. 31.1%, 95.5%
9. About 23
11. About 58%
13. $t = 1084$ hours

Problem Set 24.9, page 1040

1. 1/8, 3/16, 3/8
3. 2/9, 2/9, 1/2
5. $f_2(y) = 1/(\beta_2 - \alpha_2)$ if $\alpha_2 < y < \beta_2$ and 0 elsewhere
7. 27.45 mm, 0.38 mm
9. 25.26 cm, 0.0078 cm

13. Independent, $f_1(x) = 0.1e^{-0.1x}$ if $x > 0$, $f_2(y) = 0.1e^{-0.1y}$ if $y > 0$, 36.8%
15. 50%
17. No

Chapter 24 Review Questions and Problems, page 1041

21. $Q_L = 22.3$, $Q_M = 23.3$, $Q_U = 23.5$
23. $\bar{x} = 22.89$, $s = 1.028$, $s^2 = 1.056$
25. *H*, *TH*, *TTH*, etc.
27. $f(0) = 0.80816$, $f(1) = 0.18367$, $f(2) = 0.00816$
29. Always $B \subseteq A \cup B$. If also $A \subseteq B$, then $B = A \cup B$, etc.
31. 7/3, 8/9
33. 118.019, 1.98, 1.65%
35. 0, 2
37. $\mu = 100/30$
39. 16%, 2.3% (see Fig. 520 in Sec. 24.8)

Problem Set 25.2, page 1048

3. $l = p^k(1-p)^{n-k}$, $\hat{p} = k/n$, k = number of successes in n trials
5. 11/20
7. $l = f(x)$, $\partial(\ln l)/\partial p = 1/p - (x-1)/(1-p) = 0$, $\hat{p} = 1/x$
9. $\hat{\mu} = \bar{x}$
11. $\hat{\theta} = n/\Sigma x_j = 1/\bar{x}$
13. $\hat{\theta} = 1$
15. Variability larger than perhaps expected

Problem Set 25.3, page 1057

1. $\text{CONF}_{0.95}\{37.47 \leqq \mu \leqq 43.87\}$
3. Shorter by a factor $\sqrt{2}$
5. 4, 16
7. Cf. Example 2. $n = 166$
9. $\text{CONF}_{0.99}\{20.07 \leqq \mu \leqq 20.33\}$
11. $\text{CONF}_{0.99}\{63.71 \leqq \mu \leqq 66.29\}$
13. $c = 1.96$, $\bar{x} = 87$, $s^2 = 87 \cdot 413/500 = 71.86$, $k \approx cs/\sqrt{n} = 0.743$, $\text{CONF}_{0.95}\{86 \leqq \mu \leqq 88\}$, $\text{CONF}_{0.95}\{0.17 \leqq p \leqq 0.18\}$
15. $\text{CONF}_{0.95}\{0.00045 \leqq \sigma^2 \leqq 0.00131\}$
17. $\text{CONF}_{0.95}\{0.73 \leqq \sigma^2 \leqq 5.19\}$
19. $\text{CONF}_{0.95}\{23 \leqq \sigma^2 \leqq 548\}$. Hence a larger sample would be desirable.
21. Normal distributions, means -27, 81, 133, variances 16, 144, 400
23. $Z = X + Y$ is normal with mean 105 and variance 1.25.
Ans. $P(104 \leqq Z \leqq 106) = 63\%$

Problem Set 25.4, page 1067

1. $t = \sqrt{7}(0.286 - 0)/4.31 = 0.18 < c = 1.94$; do not reject the hypothesis.
3. $c = 6090 > 6019$; do not reject the hypothesis.
5. $\sigma^2/n = 1$, $c = 28.36$; do not reject the hypothesis.
7. $\mu < 28.76$ or $\mu > 31.24$
9. Alternative $\mu \neq 1000$, $t = \sqrt{20}(996 - 1000)/5 = -3.58 < c = -2.09$ (Table A9, 19 degrees of freedom). Reject the hypothesis $\mu = 1000$ g.
11. Test $\mu = 0$ against $\mu \neq 0$. $t = 2.11 < c = 2.36$ (7 degrees of freedom). Do not reject the hypothesis.
13. $\alpha = 5\%$, $c = 16.92 > 9 \cdot 0.5^2/0.4^2 = 14.06$; do not reject hypothesis.
15. $t_0 = \sqrt{10 \cdot 9 \cdot 17/19}\,(21.8 - 20.2)/\sqrt{9 \cdot 0.6^2 + 8 \cdot 0.5^2} = 6.3 > c = 1.74$ (17 degrees of freedom). Reject the hypothesis and assert that B is better.

17. $v_0 = 50/30 = 1.67 < c = 2.59$ [(9, 15) degrees of freedom]; do not reject the hypothesis.

Problem Set 25.5, page 1071

1. LCL $= 1 - 2.58 \cdot 0.03/\sqrt{6} = 0.968$, UCL $= 1.032$

3. $n = 10$

5. Choose 4 times the original sample size (why?).

7. $2.58\sqrt{0.024}/\sqrt{2} = 0.283$, UCL $= 27.783$, LCL $= 27.217$

11. In 30% (5%) of the cases, approximately

13. UCL $= np + 3\sqrt{np(1-p)}$, CL $= np$, LCL $= np - 3\sqrt{np(1-p)}$

15. CL $= \mu = 2.5$, UCL $= \mu + 3\sqrt{\mu} = 7.2$, LCL $= \mu - 3\sqrt{\mu}$ is negative in (b) and we set LCL $= 0$.

Problem Set 25.6, page 1076

1. 0.9825, 0.9384, 0.4060

3. 0.8187, 0.6703, 0.1353

5. $P(A;\theta) \approx e^{-30\theta}(1 + 30\theta)$

7. $P(A;\theta) \approx e^{-50\theta}$

9. 19.5%, 14.7%

11. $(1-\theta)^5$, $(1-\theta)^5 + 5\theta(1-\theta)^4$

13. Because n is finite

15. $\Phi((9 - 12 + \frac{1}{2})/\sqrt{12(1 - 0.12)}) = 0.22$ (if $c = 9$)

17. $(1 - \frac{1}{2})^3 + 3 \cdot \frac{1}{2}(1 - \frac{1}{2})^2 = \frac{1}{2}$

Problem Set 25.7, page 1079

1. $\chi_0^2 = (30 - 50)^2/50 + (70 - 50)^2/50 = 16 > c = 3.84$; no

3. 41

5. $\chi_0^2 = 2.33 < c = 11.07$. Yes

7. $e_j = np_j = 370/5 = 74$, $\chi_0^2 = 984/74 = 13.3$, $c = 9.49$. Reject the hypothesis.

9. $\chi_0^2 = 1 < 3.84$; yes

13. Combining the results for $x = 10, 11, 12$, we have $K - r - 1 = 9$ ($r = 1$ since we estimated the mean, $\frac{10094}{2608} \approx 3.87$). $\chi_0^2 = 12.98 < c = 16.92$. Do not reject.

15. $\chi_0^2 = 49/20 + 49/60 = 3.27 < c = 3.84$ (1 degree of freedom, $\alpha = 5\%$), which supports the claim.

17. 42 even digits, accept.

Problem Set 25.8, page 1082

3. $(\frac{1}{2})^{18}(1 + 18 + 153 + 816) = 0.0038$

5. Hypothesis: A and B are equally good. Then the probability of at least 7 trials favorable to A is $\frac{1}{2}^{18} + 8 \cdot \frac{1}{2}^{18} = 3.5\%$. Reject the hypothesis.

7. Hypothesis $\mu = 0$. Alternative $\mu > 0$, $\bar{x} = 1.58$, $t = \sqrt{10} \cdot 1.58/1.23 = 4.06 > c = 1.83$ ($\alpha = 5\%$). Hypothesis rejected.

9. $\bar{x} = 9.67$, $s = 11.87$, $t_0 = 9.67/(11.87/\sqrt{15}) = 3.15 > c = 1.76$ ($\alpha = 5\%$). Hypothesis rejected.

11. Consider $y_j = x_j - \tilde{\mu}_0$.

13. $P(T \leqq 2) = 0.1\%$ from Table A12. Reject.

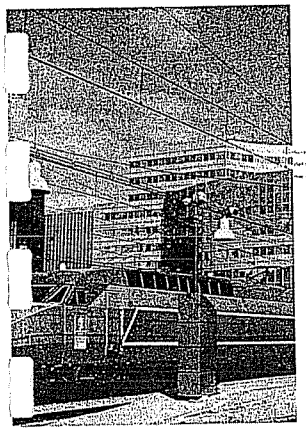
15. $P(T \leqq 15) = 10.8\%$. Do not reject.
17. $P(T \leqq 2) = 2.8\%$. Reject.

Problem Set 25.9, page 1091

1. $y = 1.9 + x$ **3.** $y = 6.7407 + 3.068x$ **5.** $y = 4 + 4.8x$, 172 ft
7. $y = -1146 + 4.32x$ **9.** $y = 0.5932 + 0.1138x$, $R = 1/0.1138$
11. $q_0 = 76$, $K = 2.36\sqrt{76/(7 \cdot 944)} = 0.253$, $\text{CONF}_{0.95}\{-1.58 \leqq \kappa_1 \leqq -1.06\}$
13. $3s_x^2 = 500$, $3s_{xy} = 33.5$, $k_1 = 0.067$, $3s_y^2 = 2.268$, $q_0 = 0.023$, $K = 0.021$
$\text{CONF}_{0.95}\{0.046 \leqq \kappa_1 \leqq 0.088\}$

Chapter 25 Review Questions and Problems, page 1092

21. $\hat{\mu} = 5.33$, $\tilde{\sigma}^2 = 1.722$ **23.** It will double.
25. $\text{CONF}_{0.99}\{19.1 \leqq \mu \leqq 33.7\}$ **27.** $\text{CONF}_{0.95}\{0.726 \leqq \mu \leqq 0.751\}$
29. $\text{CONF}_{0.95}\{1.373 \leqq \mu \leqq 1.451\}$ **31.** $\text{CONF}_{0.99}\{0.05 \leqq \sigma^2 \leqq 10\}$
33. $c = 14.74 > 14.5$; reject μ_0. **35.** $\Phi\left(\dfrac{14.74 - 14.40}{\sqrt{0.025}}\right) = 0.9842$
37. $30.14/3.8 = 7.93 < 8.25$. Reject.
39. $v_0 = 2.5 < 6.0$ [(9, 4) degrees of freedom]; accept the hypothesis.
41. Decrease by a factor $\sqrt{2}$. By a factor $2.58/1.96 = 1.32$.
43. 0.9953, 0.9825, 0.9384, etc. **45.** $y = 1.70 + 0.55x$



APPENDIX 3

Auxiliary Material

A3.1 Formulas for Special Functions

For tables of numeric values, see Appendix 5.

Exponential function e^x (Fig. 544)

$$e = 2.71828\ 18284\ 59045\ 23536\ 02874\ 71353$$

(1) $$e^x e^y = e^{x+y}, \qquad e^x/e^y = e^{x-y}, \qquad (e^x)^y = e^{xy}$$

Natural logarithm (Fig. 545)

(2) $$\ln(xy) = \ln x + \ln y, \qquad \ln(x/y) = \ln x - \ln y, \qquad \ln(x^a) = a \ln x$$

$\ln x$ is the inverse of e^x, and $e^{\ln x} = x$, $e^{-\ln x} = e^{\ln(1/x)} = 1/x$.

Logarithm of base ten $\log_{10} x$ or simply $\log x$

(3) $$\log x = M \ln x, \qquad M = \log e = 0.43429\ 44819\ 03251\ 82765\ 11289\ 18917$$

(4) $$\ln x = \frac{1}{M} \log x, \qquad \frac{1}{M} = \ln 10 = 2.30258\ 50929\ 94045\ 68401\ 79914\ 54684$$

$\log x$ is the inverse of 10^x, and $10^{\log x} = x$, $10^{-\log x} = 1/x$.

Sine and cosine functions (Figs. 546, 547). In calculus, angles are measured in radians, so that $\sin x$ and $\cos x$ have period 2π.

$\sin x$ is odd, $\sin(-x) = -\sin x$, and $\cos x$ is even, $\cos(-x) = \cos x$.

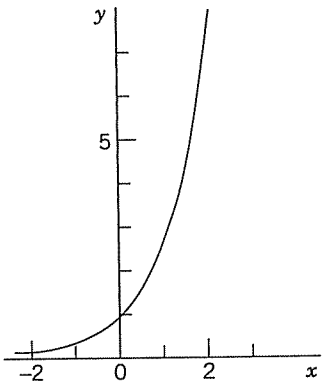


Fig. 544. Exponential function e^x

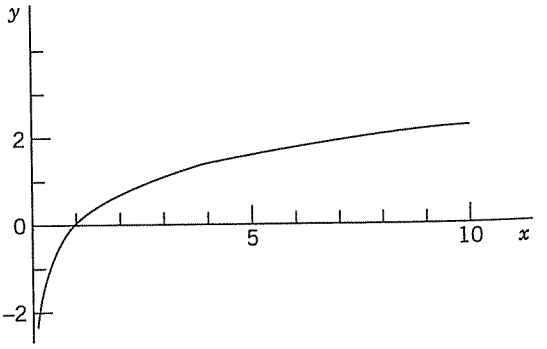


Fig. 545. Natural logarithm $\ln x$

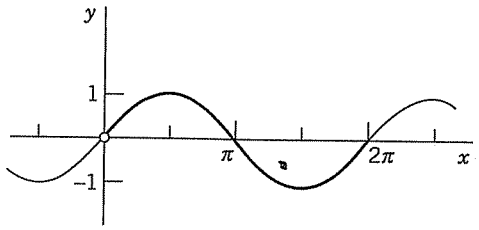


Fig. 546. sin x

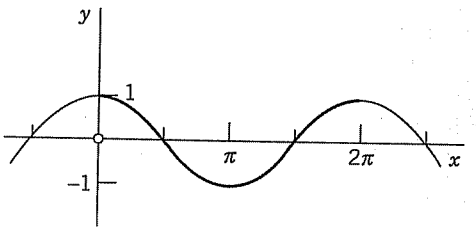


Fig. 547. cos x

$$1^\circ = 0.01745\ 32925\ 19943 \text{ radian}$$

$$1 \text{ radian} = 57^\circ\ 17'\ 44.80625''$$

$$= 57.29577\ 95131^\circ$$

(5) $$\sin^2 x + \cos^2 x = 1$$

(6)
$$\begin{cases} \sin (x + y) = \sin x \cos y + \cos x \sin y \\ \sin (x - y) = \sin x \cos y - \cos x \sin y \\ \cos (x + y) = \cos x \cos y - \sin x \sin y \\ \cos (x - y) = \cos x \cos y + \sin x \sin y \end{cases}$$

(7) $$\sin 2x = 2 \sin x \cos x, \qquad \cos 2x = \cos^2 x - \sin^2 x$$

(8)
$$\begin{cases} \sin x = \cos \left(x - \dfrac{\pi}{2}\right) = \cos \left(\dfrac{\pi}{2} - x\right) \\ \cos x = \sin \left(x + \dfrac{\pi}{2}\right) = \sin \left(\dfrac{\pi}{2} - x\right) \end{cases}$$

(9) $$\sin (\pi - x) = \sin x, \qquad \cos (\pi - x) = -\cos x$$

(10) $$\cos^2 x = \tfrac{1}{2}(1 + \cos 2x), \qquad \sin^2 x = \tfrac{1}{2}(1 - \cos 2x)$$

(11)
$$\begin{cases} \sin x \sin y = \tfrac{1}{2}[-\cos (x + y) + \cos (x - y)] \\ \cos x \cos y = \tfrac{1}{2}[\cos (x + y) + \cos (x - y)] \\ \sin x \cos y = \tfrac{1}{2}[\sin (x + y) + \sin (x - y)] \end{cases}$$

(12)
$$\begin{cases} \sin u + \sin v = 2 \sin \dfrac{u + v}{2} \cos \dfrac{u - v}{2} \\ \cos u + \cos v = 2 \cos \dfrac{u + v}{2} \cos \dfrac{u - v}{2} \\ \cos v - \cos u = 2 \sin \dfrac{u + v}{2} \sin \dfrac{u - v}{2} \end{cases}$$

(13) $$A \cos x + B \sin x = \sqrt{A^2 + B^2} \cos (x \pm \delta), \qquad \tan \delta = \frac{\sin \delta}{\cos \delta} = \mp \frac{B}{A}$$

(14) $$A \cos x + B \sin x = \sqrt{A^2 + B^2} \sin (x \pm \delta), \qquad \tan \delta = \frac{\sin \delta}{\cos \delta} = \pm \frac{A}{B}$$

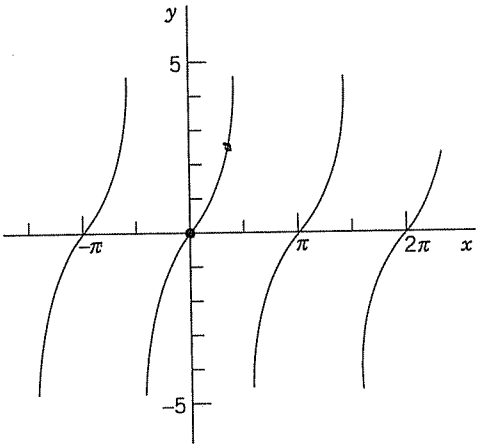


Fig. 548. tan x

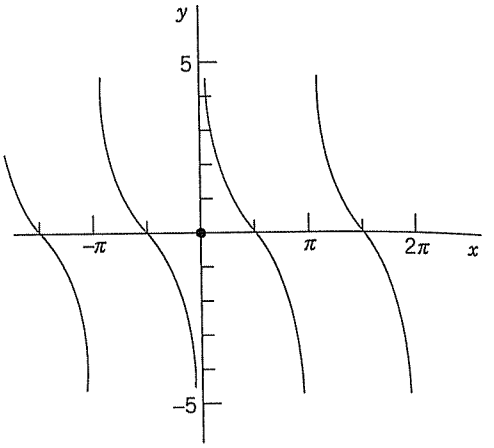


Fig. 549. cot x

Tangent, cotangent, secant, cosecant (Figs. 548, 549)

$$(15)\quad \tan x = \frac{\sin x}{\cos x}, \qquad \cot x = \frac{\cos x}{\sin x}, \qquad \sec x = \frac{1}{\cos x}, \qquad \csc x = \frac{1}{\sin x}$$

$$(16)\quad \tan (x + y) = \frac{\tan x + \tan y}{1 - \tan x \tan y}, \qquad \tan (x - y) = \frac{\tan x - \tan y}{1 + \tan x \tan y}$$

Hyperbolic functions (hyperbolic sine $\sinh x$, etc.; Figs. 550, 551)

$$(17)\quad \sinh x = \tfrac{1}{2}(e^x - e^{-x}), \qquad \cosh x = \tfrac{1}{2}(e^x + e^{-x})$$

$$(18)\quad \tanh x = \frac{\sinh x}{\cosh x}, \qquad \coth x = \frac{\cosh x}{\sinh x}$$

$$(19)\quad \cosh x + \sinh x = e^x, \qquad \cosh x - \sinh x = e^{-x}$$

$$(20)\quad \cosh^2 x - \sinh^2 x = 1$$

$$(21)\quad \sinh^2 x = \tfrac{1}{2}(\cosh 2x - 1), \qquad \cosh^2 x = \tfrac{1}{2}(\cosh 2x + 1)$$

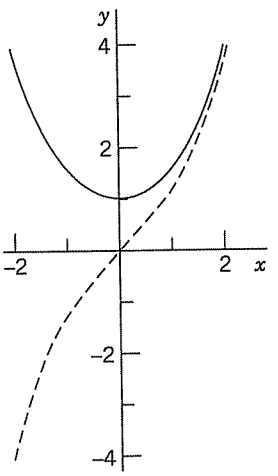


Fig. 550. $\sinh x$ (dashed) and $\cosh x$

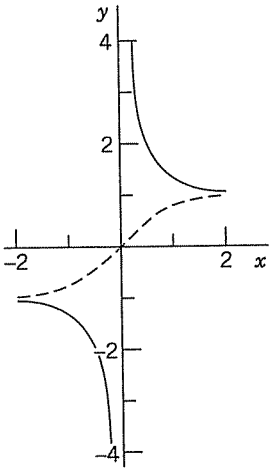


Fig. 551. $\tanh x$ (dashed) and $\coth x$

(22)
$$\begin{cases} \sinh(x \pm y) = \sinh x \cosh y \pm \cosh x \sinh y \\ \cosh(x \pm y) = \cosh x \cosh y \pm \sinh x \sinh y \end{cases}$$

(23)
$$\tanh(x \pm y) = \frac{\tanh x \pm \tanh y}{1 \pm \tanh x \tanh y}$$

Gamma function (Fig. 552 and Table A2 in App. 5). The gamma function $\Gamma(\alpha)$ is defined by the integral

(24)
$$\Gamma(\alpha) = \int_0^\infty e^{-t} t^{\alpha-1}\, dt \qquad (\alpha > 0)$$

which is meaningful only if $\alpha > 0$ (or, if we consider complex α, for those α whose real part is positive). Integration by parts gives the important *functional relation of the gamma function,*

(25)
$$\Gamma(\alpha + 1) = \alpha\Gamma(\alpha).$$

From (24) we readily have $\Gamma(1) = 1$; hence if α is a positive integer, say k, then by repeated application of (25) we obtain

(26)
$$\Gamma(k + 1) = k! \qquad (k = 0, 1, \cdots).$$

This shows that *the gamma function can be regarded as a generalization of the elementary factorial function.* [Sometimes the notation $(\alpha - 1)!$ is used for $\Gamma(\alpha)$, even for noninteger values of α, and the gamma function is also known as the **factorial function.**]

By repeated application of (25) we obtain

$$\Gamma(\alpha) = \frac{\Gamma(\alpha + 1)}{\alpha} = \frac{\Gamma(\alpha + 2)}{\alpha(\alpha + 1)} = \cdots = \frac{\Gamma(\alpha + k + 1)}{\alpha(\alpha + 1)(\alpha + 2) \cdots (\alpha + k)}$$

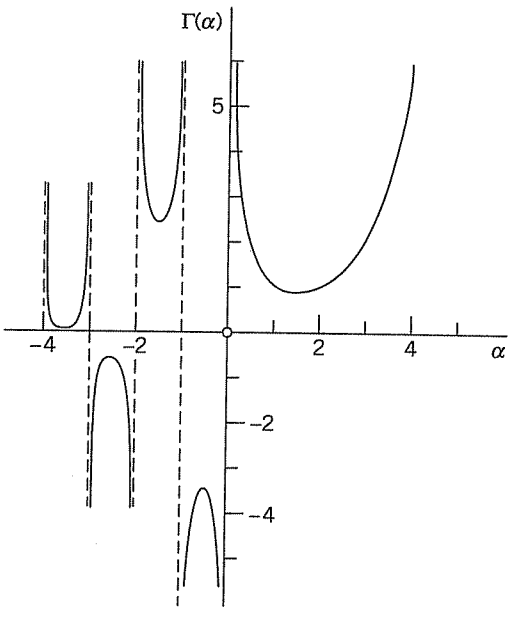


Fig. 552. Gamma function

and we may use this relation

$$\Gamma(\alpha) = \frac{\Gamma(\alpha + k + 1)}{\alpha(\alpha + 1) \cdots (\alpha + k)} \qquad (\alpha \neq 0, -1, -2, \cdots) \tag{27}$$

for defining the gamma function for negative α ($\neq -1, -2, \cdots$), choosing for k the smallest integer such that $\alpha + k + 1 > 0$. *Together with* (24), *this then gives a definition of* $\Gamma(\alpha)$ *for all* α *not equal to zero or a negative integer* (Fig. 552).

It can be shown that the gamma function may also be represented as the limit of a product, namely, by the formula

$$\Gamma(\alpha) = \lim_{n \to \infty} \frac{n!\, n^\alpha}{\alpha(\alpha + 1)(\alpha + 2) \cdots (\alpha + n)} \qquad (\alpha \neq 0, -1, \cdots). \tag{28}$$

From (27) or (28) we see that, for complex α, the gamma function $\Gamma(\alpha)$ is a meromorphic function with simple poles at $\alpha = 0, -1, -2, \cdots$.

An approximation of the gamma function for large positive α is given by the **Stirling formula**

$$\Gamma(\alpha + 1) \approx \sqrt{2\pi\alpha} \left(\frac{\alpha}{e}\right)^\alpha \tag{29}$$

where e is the base of the natural logarithm. We finally mention the special value

$$\Gamma(\tfrac{1}{2}) = \sqrt{\pi}. \tag{30}$$

Incomplete gamma functions

$$P(\alpha, x) = \int_0^x e^{-t} t^{\alpha-1}\, dt, \qquad Q(\alpha, x) = \int_x^\infty e^{-t} t^{\alpha-1}\, dt \qquad (\alpha > 0) \tag{31}$$

$$\Gamma(\alpha) = P(\alpha, x) + Q(\alpha, x) \tag{32}$$

Beta function

$$B(x, y) = \int_0^1 t^{x-1}(1 - t)^{y-1}\, dt \qquad (x > 0, y > 0) \tag{33}$$

Representation in terms of gamma functions:

$$B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x + y)} \tag{34}$$

Error function (Fig. 553 and Table A4 in App. 5)

$$\operatorname{erf} x = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt \tag{35}$$

$$\operatorname{erf} x = \frac{2}{\sqrt{\pi}} \left(x - \frac{x^3}{1!3} + \frac{x^5}{2!5} - \frac{x^7}{3!7} + - \cdots\right) \tag{36}$$

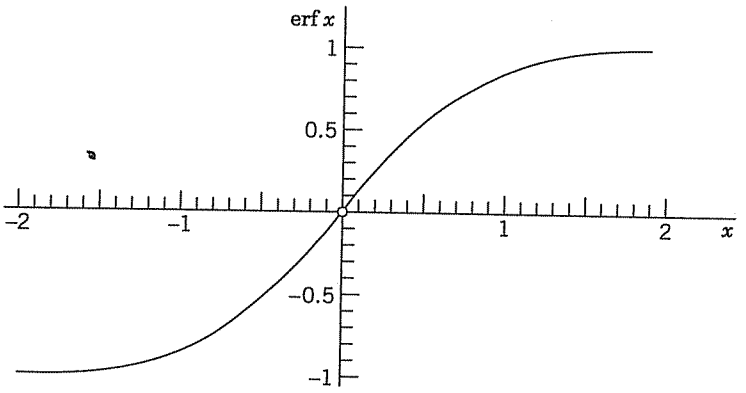


Fig. 553. Error function

$\operatorname{erf}(\infty) = 1$, *complementary error function*

$$\operatorname{erfc} x = 1 - \operatorname{erf} x = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-t^2}\, dt \tag{37}$$

Fresnel integrals[1] (Fig. 554)

$$C(x) = \int_0^x \cos(t^2)\, dt, \qquad S(x) = \int_0^x \sin(t^2)\, dt \tag{38}$$

$C(\infty) = \sqrt{\pi/8}$, $S(\infty) = \sqrt{\pi/8}$, *complementary functions*

$$\begin{aligned} c(x) &= \sqrt{\frac{\pi}{8}} - C(x) = \int_x^\infty \cos(t^2)\, dt \\ s(x) &= \sqrt{\frac{\pi}{8}} - S(x) = \int_x^\infty \sin(t^2)\, dt \end{aligned} \tag{39}$$

Sine integral (Fig. 555 and Table A4 in App. 5)

$$\operatorname{Si}(x) = \int_0^x \frac{\sin t}{t}\, dt \tag{40}$$

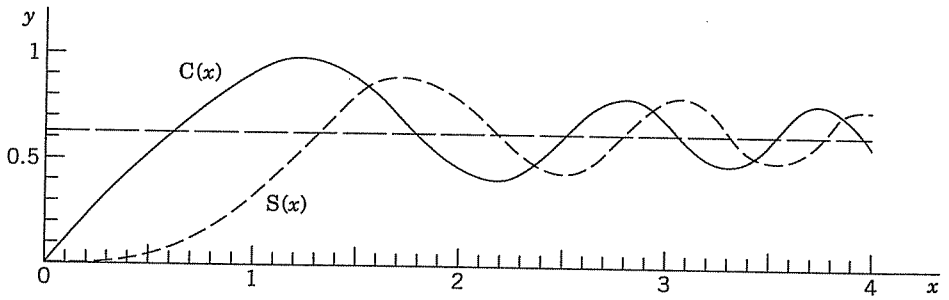


Fig. 554. Fresnel integrals

[1] AUGUSTIN FRESNEL (1788–1827), French physicist and mathematician. For tables see Ref. [GR1].

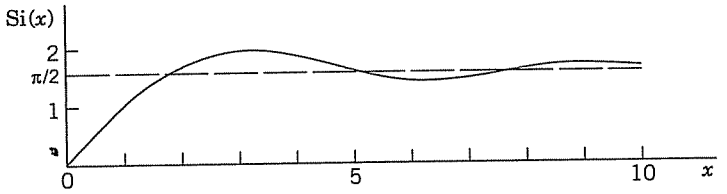


Fig. 555. Sine integral

$\mathrm{Si}(\infty) = \pi/2$, *complementary function*

(41) $$\mathrm{si}(x) = \frac{\pi}{2} - \mathrm{Si}(x) = \int_x^\infty \frac{\sin t}{t}\, dt$$

Cosine integral (Table A4 in App. 5)

(42) $$\mathrm{ci}(x) = \int_x^\infty \frac{\cos t}{t}\, dt \qquad (x > 0)$$

Exponential integral

(43) $$\mathrm{Ei}(x) = \int_x^\infty \frac{e^{-t}}{t}\, dt \qquad (x > 0)$$

Logarithmic integral

(44) $$\mathrm{li}(x) = \int_0^x \frac{dt}{\ln t}$$

A3.2 Partial Derivatives

For differentiation formulas, see inside of front cover.

Let $z = f(x, y)$ be a real function of two independent real variables, x and y. If we keep y constant, say, $y = y_1$, and think of x as a variable, then $f(x, y_1)$ depends on x alone. If the derivative of $f(x, y_1)$ with respect to x for a value $x = x_1$ exists, then the value of this derivative is called the **partial derivative** of $f(x, y)$ *with respect to* x *at the point* (x_1, y_1) and is denoted by

$$\left.\frac{\partial f}{\partial x}\right|_{(x_1,y_1)} \qquad \text{or by} \qquad \left.\frac{\partial z}{\partial x}\right|_{(x_1,y_1)}$$

Other notations are

$$f_x(x_1, y_1) \qquad \text{and} \qquad z_x(x_1, y_1);$$

these may be used when subscripts are not used for another purpose and there is no danger of confusion.

We thus have, by the definition of the derivative,

$$(1) \qquad \left.\frac{\partial f}{\partial x}\right|_{(x_1,y_1)} = \lim_{\Delta x \to 0} \frac{f(x_1 + \Delta x, y_1) - f(x_1, y_1)}{\Delta x}.$$

The partial derivative of $z = f(x, y)$ with respect to y is defined similarly; we now keep x constant, say, equal to x_1, and differentiate $f(x_1, y)$ with respect to y. Thus

$$(2) \qquad \left.\frac{\partial f}{\partial y}\right|_{(x_1,y_1)} = \left.\frac{\partial z}{\partial y}\right|_{(x_1,y_1)} = \lim_{\Delta y \to 0} \frac{f(x_1, y_1 + \Delta y) - f(x_1, y_1)}{\Delta y}.$$

Other notations are $f_y(x_1, y_1)$ and $z_y(x_1, y_1)$.

It is clear that the values of those two partial derivatives will in general depend on the point (x_1, y_1). Hence the partial derivatives $\partial z/\partial x$ and $\partial z/\partial y$ at a variable point (x, y) are functions of x and y. The function $\partial z/\partial x$ is obtained as in ordinary calculus by differentiating $z = f(x, y)$ with respect to x, ***treating y as a constant,*** and $\partial z/\partial y$ is obtained by differentiating z with respect to y, ***treating x as a constant.***

EXAMPLE 1 Let $z = f(x, y) = x^2y + x \sin y$. Then

$$\frac{\partial f}{\partial x} = 2xy + \sin y, \qquad \frac{\partial f}{\partial y} = x^2 + x \cos y. \qquad \blacksquare$$

The partial derivatives $\partial z/\partial x$ and $\partial z/\partial y$ of a function $z = f(x, y)$ have a very simple **geometric interpretation.** The function $z = f(x, y)$ can be represented by a surface in space. The equation $y = y_1$ then represents a vertical plane intersecting the surface in a curve, and the partial derivative $\partial z/\partial x$ at a point (x_1, y_1) is the slope of the tangent (that is, $\tan \alpha$ where α is the angle shown in Fig. 556) to the curve. Similarly, the partial derivative $\partial z/\partial y$ at (x_1, y_1) is the slope of the tangent to the curve $x = x_1$ on the surface $z = f(x, y)$ at (x_1, y_1).

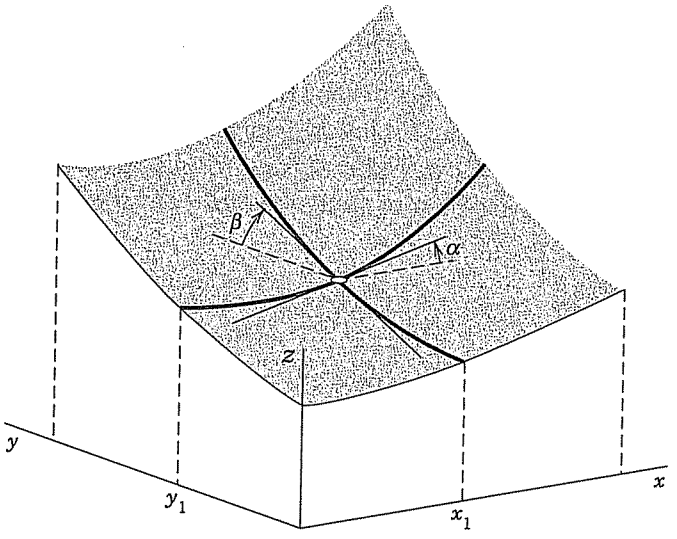


Fig. 556. Geometrical interpretation of first partial derivatives

The partial derivatives $\partial z/\partial x$ and $\partial z/\partial y$ are called *first partial derivatives* or *partial derivatives of first order.* By differentiating these derivatives once more, we obtain the four *second partial derivatives* (or *partial derivatives of second order*)[2]

$$(3) \qquad \begin{aligned} \frac{\partial^2 f}{\partial x^2} &= \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right) = f_{xx} \\ \frac{\partial^2 f}{\partial x\, \partial y} &= \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right) = f_{yx} \\ \frac{\partial^2 f}{\partial y\, \partial x} &= \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right) = f_{xy} \\ \frac{\partial^2 f}{\partial y^2} &= \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial y}\right) = f_{yy}. \end{aligned}$$

It can be shown that if all the derivatives concerned are continuous, then the two mixed partial derivatives are equal, so that the order of differentiation does not matter (see Ref. [GR4] in App. 1), that is,

$$(4) \qquad \frac{\partial^2 z}{\partial x\, \partial y} = \frac{\partial^2 z}{\partial y\, \partial x}.$$

EXAMPLE 2 For the function in Example 1.

$$f_{xx} = 2y, \qquad f_{xy} = 2x + \cos y = f_{yx}, \qquad f_{yy} = -x \sin y.$$

By differentiating the second partial derivatives again with respect to x and y, respectively, we obtain the *third partial derivatives* or *partial derivatives of the third order* of f, etc.

If we consider a function $f(x, y, z)$ of ***three independent variables,*** then we have the three first partial derivatives $f_x(x, y, z)$, $f_y(x, y, z)$, and $f_z(x, y, z)$. Here f_x is obtained by differentiating f with respect to x, ***treating both y and z as constants.*** Thus, analogous to (1), we now have

$$\left.\frac{\partial f}{\partial x}\right|_{(x_1, y_1, z_1)} = \lim_{\Delta x \to 0} \frac{f(x_1 + \Delta x, y_1, z_1) - f(x_1, y_1, z_1)}{\Delta x},$$

etc. By differentiating f_x, f_y, f_z again in this fashion we obtain the second partial derivatives of f, etc.

EXAMPLE 3 Let $f(x, y, z) = x^2 + y^2 + z^2 + xy\, e^z$. Then

$$\begin{aligned} &f_x = 2x + y\, e^z, && f_y = 2y + x\, e^z, && f_z = 2z + xy\, e^z, \\ &f_{xx} = 2, && f_{xy} = f_{yx} = e^z, && f_{xz} = f_{zx} = y\, e^z, \\ &f_{yy} = 2, && f_{yz} = f_{zy} = x\, e^z, && f_{zz} = 2 + xy\, e^z. \end{aligned}$$

[2] **CAUTION!** In the subscript notation the subscripts are written in the order in which we differentiate, whereas in the "∂" notation the order is opposite.

A3.3 Sequences and Series

See also Chap. 15.

Monotone Real Sequences

We call a real sequence $x_1, x_2, \cdots, x_n, \cdots$ a **monotone sequence** if it is either **monotone increasing,** that is,

$$x_1 \leqq x_2 \leqq x_3 \leqq \cdots$$

or **monotone decreasing,** that is,

$$x_1 \geqq x_2 \geqq x_3 \geqq \cdots.$$

We call $x_1, x_2, \cdots$ a **bounded sequence** if there is a positive constant K such that $|x_n| < K$ for all n.

THEOREM 1 *If a real sequence is bounded and monotone, it converges.*

PROOF Let $x_1, x_2, \cdots$ be a bounded monotone increasing sequence. Then its terms are smaller than some number B and, since $x_1 \leqq x_n$ for all n, they lie in the interval $x_1 \leqq x_n \leqq B$, which will be denoted by I_0. We bisect I_0; that is, we subdivide it into two parts of equal length. If the right half (together with its endpoints) contains terms of the sequence, we denote it by I_1. If it does not contain terms of the sequence, then the left half of I_0 (together with its endpoints) is called I_1. This is the first step.

In the second step we bisect I_1, select one half by the same rule, and call it I_2, and so on (see Fig. 557 on p. A70).

In this way we obtain shorter and shorter intervals $I_0, I_1, I_2, \cdots$ with the following properties. Each I_m contains all I_n for $n > m$. No term of the sequence lies to the right of I_m, and, since the sequence is monotone increasing, all x_n with n greater than some number N lie in I_m; of course, N will depend on m, in general. The lengths of the I_m approach zero as m approaches infinity. Hence there is precisely one number, call it L, that lies in all those intervals,[3] and we may now easily prove that the sequence is convergent with the limit L.

In fact, given an $\epsilon > 0$, we choose an m such that the length of I_m is less than ϵ. Then L and all the x_n with $n > N(m)$ lie in I_m, and, therefore, $|x_n - L| < \epsilon$ for all those n. This completes the proof for an increasing sequence. For a decreasing sequence the proof is the same, except for a suitable interchange of "left" and "right" in the construction of those intervals. ▩

[3]This statement seems to be obvious, but actually it is not; it may be regarded as an axiom of the real number system in the following form. Let $J_1, J_2, \cdots$ be closed intervals such that each J_m contains all J_n with $n > m$, and the lengths of the J_m approach zero as m approaches infinity. Then there is precisely one real number that is contained in all those intervals. This is the so-called **Cantor–Dedekind axiom,** named after the German mathematicians GEORG CANTOR (1845–1918), the creator of set theory, and RICHARD DEDEKIND (1831–1916), known for his fundamental work in number theory. For further details see Ref. [GR2] in App. 1. (An interval I is said to be **closed** if its two endpoints are regarded as points belonging to I. It is said to be **open** if the endpoints are not regarded as points of I.)

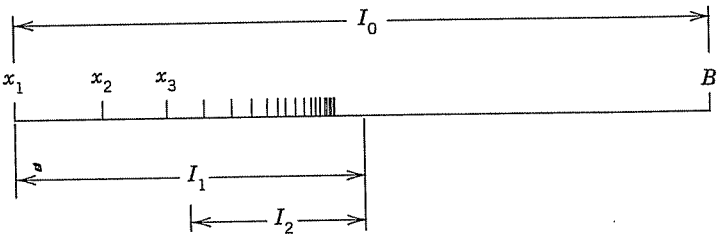


Fig. 557. Proof of Theorem 1

Real Series

THEOREM 2

Leibniz Test for Real Series

Let $x_1, x_2, \cdots$ be real and monotone decreasing to zero, that is,

(1) (a) $x_1 \geqq x_2 \geqq x_3 \geqq \cdots$, (b) $\lim_{m\to\infty} x_m = 0.$

Then the series with terms of alternating signs

$$x_1 - x_2 + x_3 - x_4 + - \cdots$$

converges, and for the remainder R_n after the nth term we have the estimate

(2) $$|R_n| \leqq x_{n+1}.$$

PROOF Let s_n be the nth partial sum of the series. Then, because of (1a),

$$s_1 = x_1, \qquad s_2 = x_1 - x_2 \leqq s_1,$$

$$s_3 = s_2 + x_3 \geqq s_2, \qquad s_3 = s_1 - (x_2 - x_3) \leqq s_1,$$

so that $s_2 \leqq s_3 \leqq s_1$. Proceeding in this fashion, we conclude that (Fig. 558)

(3) $$s_1 \geqq s_3 \geqq s_5 \geqq \cdots \geqq s_6 \geqq s_4 \geqq s_2$$

which shows that the odd partial sums form a bounded monotone sequence, and so do the even partial sums. Hence, by Theorem 1, both sequences converge, say,

$$\lim_{n\to\infty} s_{2n+1} = s, \qquad \lim_{n\to\infty} s_{2n} = s^*.$$

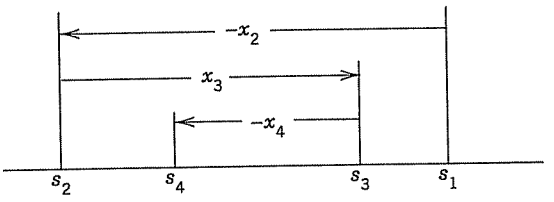


Fig. 558. Proof of the Leibniz test

Now, since $s_{2n+1} - s_{2n} = x_{2n+1}$, we readily see that (lb) implies

$$s - s^* = \lim_{n\to\infty} s_{2n+1} - \lim_{n\to\infty} s_{2n} = \lim_{n\to\infty} (s_{2n+1} - s_{2n}) = \lim_{n\to\infty} x_{2n+1} = 0.$$

Hence $s^* = s$, and the series converges with the sum s.

We prove the estimate (2) for the remainder. Since $s_n \to s$, it follows from (3) that

$$s_{2n+1} \geqq s \geqq s_{2n} \qquad \text{and also} \qquad s_{2n-1} \geqq s \geqq s_{2n}.$$

By subtracting s_{2n} and s_{2n-1}, respectively, we obtain

$$s_{2n+1} - s_{2n} \geqq s - s_{2n} \geqq 0, \qquad 0 \geqq s - s_{2n-1} \geqq s_{2n} - s_{2n-1}.$$

In these inequalities, the first expression is equal to x_{2n+1}, the last is equal to $-x_{2n}$, and the expressions between the inequality signs are the remainders R_{2n} and R_{2n-1}. Thus the inequalities may be written

$$x_{2n+1} \geqq R_{2n} \geqq 0, \qquad 0 \geqq R_{2n-1} \geqq -x_{2n}$$

and we see that they imply (2). This completes the proof. ∎

A3.4 Grad, Div, Curl, ∇^2 in Curvilinear Coordinates

To simplify formulas we write Cartesian coordinates $x = x_1$, $y = x_2$, $z = x_3$. We denote curvilinear coordinates by q_1, q_2, q_3. Through each point P there pass three coordinate surfaces $q_1 = const$, $q_2 = const$, $q_3 = const$. They intersect along coordinate curves. We assume the three coordinate curves through P to be **orthogonal** (perpendicular to each other). We write coordinate transformations as

(1) $$x_1 = x_1(q_1, q_2, q_3), \qquad x_2 = x_2(q_1, q_2, q_3), \qquad x_3 = x_3(q_1, q_2, q_3).$$

Corresponding transformations of grad, div, curl, and ∇^2 can all be written by using

(2) $$h_j^2 = \sum_{k=1}^{3} \left(\frac{\partial x_k}{\partial q_j}\right)^2.$$

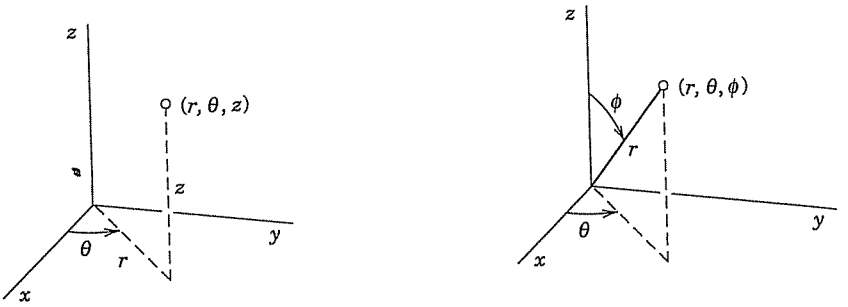
Next to Cartesian coordinates, most important are **cylindrical coordinates** $q_1 = r$, $q_2 = \theta$, $q_3 = z$ (Fig. 559a on p. A72) defined by

(3) $$x_1 = q_1 \cos q_2 = r \cos \theta, \qquad x_2 = q_1 \sin q_2 = r \sin \theta, \qquad x_3 = q_3 = z$$

and **spherical coordinates** $q_1 = r$, $q_2 = \theta$, $q_3 = \phi$ (Fig. 559b) defined by[4]

(4) $$x_1 = q_1 \cos q_2 \sin q_3 = r \cos \theta \sin \phi, \qquad x_2 = q_1 \sin q_2 \sin q_3 = r \sin \theta \sin \phi$$
$$x_3 = q_1 \cos q_3 = r \cos \phi.$$

[4]This is the notation used in calculus and in many other books. It is logical since in it, θ plays the same role as in polar coordinates. **CAUTION!** Some books interchange the roles of θ and ϕ.



(a) Cylindrical coordinates (b) Spherical coordinates

Fig. 559. Special curvilinear coordinates

In addition to the general formulas for any orthogonal coordinates q_1, q_2, q_3, we shall give additional formulas for these important special cases.

Linear Element *ds*. In Cartesian coordinates,

$$ds^2 = dx_1^2 + dx_2^2 + dx_3^2 \qquad \text{(Sec. 9.5).}$$

For the q-coordinates,

(5) $$ds^2 = h_1^2\,dq_1^2 + h_2^2\,dq_2^2 + h_3^2\,dq_3^2.$$

(5′) $$ds^2 = dr^2 + r^2\,d\theta^2 + dz^2 \qquad \text{(Cylindrical coordinates).}$$

For polar coordinates set $dz^2 = 0$.

(5″) $$ds^2 = dr^2 + r^2\sin^2\phi\,d\theta^2 + r^2\,d\phi^2 \qquad \text{(Spherical coordinates).}$$

Gradient. grad $f = \nabla f = [f_{x_1},\ f_{x_2},\ f_{x_3}]$ (partial derivatives; Sec. 9.7). In the q-system, with **u, v, w** denoting unit vectors in the positive directions of the q_1, q_2, q_3 coordinate curves, respectively,

(6) $$\text{grad } f = \nabla f = \frac{1}{h_1}\frac{\partial f}{\partial q_1}\mathbf{u} + \frac{1}{h_2}\frac{\partial f}{\partial q_2}\mathbf{v} + \frac{1}{h_3}\frac{\partial f}{\partial q_3}\mathbf{w}$$

(6′) $$\text{grad } f = \nabla f = \frac{\partial f}{\partial r}\mathbf{u} + \frac{1}{r}\frac{\partial f}{\partial \theta}\mathbf{v} + \frac{\partial f}{\partial z}\mathbf{w} \qquad \text{(Cylindrical coordinates)}$$

(6″) $$\text{grad } f = \nabla f = \frac{\partial f}{\partial r}\mathbf{u} + \frac{1}{r\sin\phi}\frac{\partial f}{\partial \theta}\mathbf{v} + \frac{1}{r}\frac{\partial f}{\partial \phi}\mathbf{w} \qquad \text{(Spherical coordinates).}$$

Divergence div $\mathbf{F} = \nabla\bullet\mathbf{F} = (F_1)_{x_1} + (F_2)_{x_2} + (F_3)_{x_3}$ ($\mathbf{F} = [F_1, F_2, F_3]$, Sec. 9.8);

(7) $$\text{div } \mathbf{F} = \nabla\bullet\mathbf{F} = \frac{1}{h_1h_2h_3}\left[\frac{\partial}{\partial q_1}(h_2h_3F_1) + \frac{\partial}{\partial q_2}(h_3h_1F_2) + \frac{\partial}{\partial q_3}(h_1h_2F_3)\right]$$

(7′) $$\text{div } \mathbf{F} = \nabla\bullet\mathbf{F} = \frac{1}{r}\frac{\partial}{\partial r}(rF_1) + \frac{1}{r}\frac{\partial F_2}{\partial \theta} + \frac{\partial F_3}{\partial z} \qquad \text{(Cylindrical coordinates)}$$

$$(7'') \quad \operatorname{div} \mathbf{F} = \nabla \cdot \mathbf{F} = \frac{1}{r^2}\frac{\partial}{\partial r}(r^2 F_1) + \frac{1}{r\sin\phi}\frac{\partial F_2}{\partial \theta} + \frac{1}{r\sin\phi}\frac{\partial}{\partial \phi}(\sin\phi\, F_3) \qquad \text{(Spherical coordinates).}$$

Laplacian $\nabla^2 f = \nabla \cdot \nabla f = \operatorname{div}(\operatorname{grad} f) = f_{x_1x_1} + f_{x_2x_2} + f_{x_3x_3}$ (Sec. 9.8):

$$(8) \quad \nabla^2 f = \frac{1}{h_1h_2h_3}\left[\frac{\partial}{\partial q_1}\left(\frac{h_2h_3}{h_1}\frac{\partial f}{\partial q_1}\right) + \frac{\partial}{\partial q_2}\left(\frac{h_3h_1}{h_2}\frac{\partial f}{\partial q_2}\right) + \frac{\partial}{\partial q_3}\left(\frac{h_1h_2}{h_3}\frac{\partial f}{\partial q_3}\right)\right]$$

$$(8') \quad \nabla^2 f = \frac{\partial^2 f}{\partial r^2} + \frac{1}{r}\frac{\partial f}{\partial r} + \frac{1}{r^2}\frac{\partial^2 f}{\partial \theta^2} + \frac{\partial^2 f}{\partial z^2} \qquad \text{(Cylindrical coordinates)}$$

$$(8'') \quad \nabla^2 f = \frac{\partial^2 f}{\partial r^2} + \frac{2}{r}\frac{\partial f}{\partial r} + \frac{1}{r^2\sin^2\phi}\frac{\partial^2 f}{\partial \theta^2} + \frac{1}{r^2}\frac{\partial^2 f}{\partial \phi^2} + \frac{\cot\phi}{r^2}\frac{\partial f}{\partial \phi} \qquad \text{(Spherical coordinates).}$$

Curl (Sec. 9.9):

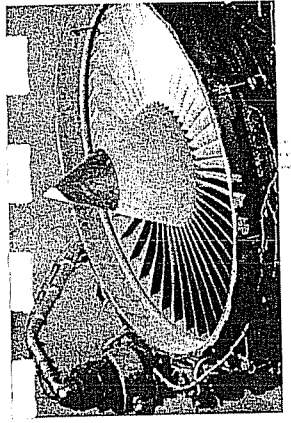
$$(9) \quad \operatorname{curl} \mathbf{F} = \nabla \times \mathbf{F} = \frac{1}{h_1h_2h_3}\begin{vmatrix} h_1\mathbf{u} & h_2\mathbf{v} & h_3\mathbf{w} \\ \dfrac{\partial}{\partial q_1} & \dfrac{\partial}{\partial q_2} & \dfrac{\partial}{\partial q_3} \\ h_1F_1 & h_2F_2 & h_3F_3 \end{vmatrix}.$$

For cylindrical coordinates we have in (9) (as in the previous formulas)

$$h_1 = h_r = 1, \qquad h_2 = h_\theta = q_1 = r, \qquad h_3 = h_z = 1$$

and for spherical coordinates we have

$$h_1 = h_r = 1, \qquad h_2 = h_\theta = q_1 \sin q_3 = r\sin\phi, \qquad h_3 = h_\phi = q_1 = r.$$



APPENDIX 4

Additional Proofs

Section 2.6, page 73

PROOF OF THEOREM 1 Uniqueness[1]

Assuming that the problem consisting of the ODE

$$y'' + p(x)y' + q(x)y = 0 \tag{1}$$

and the two initial conditions

$$y(x_0) = K_0, \qquad y'(x_0) = K_1 \tag{3}$$

has two solutions $y_1(x)$ and $y_2(x)$ on the interval I in the theorem, we show that their difference

$$y(x) = y_1(x) - y_2(x)$$

is identically zero on I; then $y_1 \equiv y_2$ on I, which implies uniqueness.

Since (1) is homogeneous and linear, y is a solution of that ODE on I, and since y_1 and y_2 satisfy the same initial conditions, y satisfies the conditions

$$y(x_0) = 0, \qquad y'(x_0) = 0. \tag{10}$$

We consider the function

$$z(x) = y(x)^2 + y'(x)^2$$

and its derivative

$$z' = 2yy' + 2y'y''.$$

From the ODE we have

$$y'' = -py' - qy.$$

By substituting this in the expression for z' we obtain

$$z' = 2yy' - 2py'^2 - 2qyy'. \tag{11}$$

Now, since y and y' are real,

$$(y \pm y')^2 = y^2 \pm 2yy' + y'^2 \geqq 0.$$

[1]This proof was suggested by my colleague, Prof. A. D. Ziebur. In this proof we use formula numbers that have not yet been used in Sec. 2.6.

From this and the definition of z we obtain the two inequalities

(12) (a) $2yy' \leqq y^2 + y'^2 = z$, (b) $-2yy' \leqq y^2 + y'^2 = z$.

From (12b) we have $2yy' \geqq -z$. Together, $|2yy'] \leqq z$. For the last term in (11) we now obtain

$$-2qyy' \leqq |-2qyy'| = |q||2yy'| \leqq |q|z.$$

Using this result as well as $-p \leqq |p|$ and applying (12a) to the term $2yy'$ in (11), we find

$$z' \leqq z + 2|p|y'^2 + |q|z.$$

Since $y'^2 \leqq y^2 + y'^2 = z$, from this we obtain

$$z' \leqq (1 + 2|p| + |q|)z$$

or, denoting the function in parentheses by h,

(13a) $$z' \leqq hz \qquad \text{for all } x \text{ on } I.$$

Similarly, from (11) and (12) it follows that

(13b)
$$\begin{aligned} -z' &= -2yy' + 2py'^2 + 2qyy' \\ &\leqq z + 2|p|z + |q|z = hz. \end{aligned}$$

The inequalities (13a) and (13b) are equivalent to the inequalities

(14) $$z' - hz \leqq 0, \qquad z' + hz \geqq 0.$$

Integrating factors for the two expressions on the left are

$$F_1 = e^{-\int h(x)\,dx} \qquad \text{and} \qquad F_2 = e^{\int h(x)\,dx}.$$

The integrals in the exponents exist because h is continuous. Since F_1 and F_2 are positive, we thus have from (14)

$$F_1(z' - hz) = (F_1 z)' \leqq 0 \qquad \text{and} \qquad F_2(z' + hz) = (F_2 z)' \geqq 0.$$

This means that $F_1 z$ is nonincreasing and $F_2 z$ is nondecreasing on I. Since $z(x_0) = 0$ by (10), when $x \leqq x_0$ we thus obtain

$$F_1 z \geqq (F_1 z)_{x_0} = 0, \qquad F_2 z \leqq (F_2 z)_{x_0} = 0$$

and similarly, when $x \geqq x_0$,

$$F_1 z \leqq 0, \qquad F_2 z \geqq 0.$$

Dividing by F_1 and F_2 and noting that these functions are positive, we altogether have

$$z \leqq 0, \qquad z \geqq 0 \qquad \text{for all } x \text{ on } I.$$

This implies that $z = y^2 + y'^2 \equiv 0$ on I. Hence $y \equiv 0$ or $y_1 \equiv y_2$ on I. ■

Section 5.4, pages 184

PROOF OF THEOREM 2 Frobenius Method. Basis of Solutions. Three Cases

The formula numbers in this proof are the same as in the text of Sec. 5.4. An additional formula not appearing in Sec. 5.4 will be called (A) (see below).

The ODE in Theorem 2 is

$$y'' + \frac{b(x)}{x}\,y' + \frac{c(x)}{x^2}\,y = 0, \tag{1}$$

where $b(x)$ and $c(x)$ are analytic functions. We can write it

$$x^2y'' + xb(x)y' + c(x)y = 0. \tag{1'}$$

The indicial equation of (1) is

$$r(r-1) + b_0r + c_0 = 0. \tag{4}$$

The roots r_1, r_2 of this quadratic equation determine the general form of a basis of solutions of (1), and there are three possible cases as follows.

Case 1. Distinct Roots not Differing by an Integer. A first solution of (1) is of the form

$$y_1(x) = x^{r_1}(a_0 + a_1x + a_2x^2 + \cdots) \tag{5}$$

and can be determined as in the power series method. For a proof that in this case, the ODE (1) has a second independent solution of the form

$$y_2(x) = x^{r_2}(A_0 + A_1x + A_2x^2 + \cdots), \tag{6}$$

see Ref. [A11] listed in App. 1.

Case 2. Double Root. The indicial equation (4) has a double root r if and only if $(b_0 - 1)^2 - 4c_0 = 0$, and then $r = \frac{1}{2}(1 - b_0)$. A first solution

$$y_1(x) = x^r(a_0 + a_1x + a_2x^2 + \cdots), \qquad r = \tfrac{1}{2}(1 - b_0), \tag{7}$$

can be determined as in Case 1. We show that a second independent solution is of the form

$$y_2(x) = y_1(x)\ln x + x^r(A_1x + A_2x^2 + \cdots) \qquad (x > 0). \tag{8}$$

We use the method of reduction of order (see Sec. 2.1), that is, we determine $u(x)$ such that $y_2(x) = u(x)y_1(x)$ is a solution of (1). By inserting this and the derivatives

$$y_2' = u'y_1 + uy_1', \qquad y_2'' = u''y_1 + 2u'y_1' + uy_1''$$

into the ODE (1′) we obtain

$$x^2(u''y_1 + 2u'y_1' + uy_1'') + xb(u'y_1 + uy_1') + cuy_1 = 0.$$

Since y_1 is a solution of (1′), the sum of the terms involving u is zero, and this equation reduces to

$$x^2y_1u'' + 2x^2y_1'u' + xby_1u' = 0.$$

By dividing by x^2y_1 and inserting the power series for b we obtain

$$u'' + \left(2\frac{y_1'}{y_1} + \frac{b_0}{x} + \cdots\right)u' = 0.$$

Here and in the following the dots designate terms that are constant or involve positive powers of x. Now from (7) if follows that

$$\frac{y_1'}{y_1} = \frac{x^{r-1}[ra_0 + (r+1)a_1x + \cdots]}{x^r[a_0 + a_1x + \cdots]}$$

$$= \frac{1}{x}\left(\frac{ra_0 + (r+1)a_1x + \cdots}{a_0 + a_1x + \cdots}\right) = \frac{r}{x} + \cdots.$$

Hence the previous equation can be written

(A) $$u'' + \left(\frac{2r + b_0}{x} + \cdots\right)u' = 0.$$

Since $r = (1 - b_0)/2$, the term $(2r + b_0)/x$ equals $1/x$, and by dividing by u' we thus have

$$\frac{u''}{u'} = -\frac{1}{x} + \cdots.$$

By integration we obtain $\ln u' = -\ln x + \cdots$, hence $u' = (1/x)e^{(\cdots)}$. Expanding the exponential function in powers of x and integrating once more, we see that u is of the form

$$u = \ln x + k_1x + k_2x^2 + \cdots.$$

Inserting this into $y_2 = uy_1$, we obtain for y_2 a representation of the form (8).

Case 3. Roots Differing by an Integer. We write $r_1 = r$ and $r_2 = r - p$ where p is a ***positive*** integer. A first solution

(9) $$y_1(x) = x^{r_1}(a_0 + a_1x + a_2x^2 + \cdots)$$

can be determined as in Cases 1 and 2. We show that a second independent solution is of the form

(10) $$y_2(x) = ky_1(x)\ln x + x^{r_2}(A_0 + A_1x + A_2x^2 + \cdots)$$

where we may have $k \neq 0$ or $k = 0$. As in Case 2 we set $y_2 = uy_1$. The first steps are literally as in Case 2 and give Eq. (A),

$$u'' + \left(\frac{2r + b_0}{x} + \cdots\right)u' = 0.$$

Now by elementary algebra, the coefficient $b_0 - 1$ of r in (4) equals minus the sum of the roots,

$$b_0 - 1 = -(r_1 + r_2) = -(r + r - p) = -2r + p.$$

Hence $2r + b_0 \triangleq p + 1$, and division by u' gives

$$\frac{u''}{u'} = -\left(\frac{p+1}{x} + \cdots\right).$$

The further steps are as in Case 2. Integrating, we find

$$\ln u' = -(p+1)\ln x + \cdots, \qquad \text{thus} \qquad u' = x^{-(p+1)}e^{(\cdots)}$$

where dots stand for some series of nonnegative integer powers of x. By expanding the exponential function as before we obtain a series of the form

$$u' = \frac{1}{x^{p+1}} + \frac{k_1}{x^p} + \cdots + \frac{k_{p-1}}{x^2} + \frac{k_p}{x} + k_{p+1} + k_{p+2}x + \cdots.$$

We integrate once more. Writing the resulting logarithmic term first, we get

$$u = k_p \ln x + \left(-\frac{1}{px^p} - \cdots - \frac{k_{p-1}}{x} + k_{p+1}x + \cdots\right).$$

Hence, by (9) we get for $y_2 = uy_1$ the formula

$$y_2 = k_p y_1 \ln x + x^{r_1-p}\left(-\frac{1}{p} - \cdots - k_{p-1}x^{p-1} + \cdots\right)(a_0 + a_1x + \cdots).$$

But this is of the form (10) with $k = k_p$ since $r_1 - p = r_2$ and the product of the two series involves nonnegative integer powers of x only. ∎

Section 5.7, page 205

THEOREM

Reality of Eigenvalues

If p, q, r, and p' in the Sturm–Liouville equation (1) *of Sec. 5.7 are real-valued and continuous on the interval $a \leqq x \leqq b$ and $r(x) > 0$ throughout that interval (or $r(x) < 0$ throughout that interval), then all the eigenvalues of the Sturm–Liouville problem* (1), (2), *Sec. 5.7, are real.*

PROOF Let $\lambda = \alpha + i\beta$ be an eigenvalue of the problem and let

$$y(x) = u(x) + iv(x)$$

be a corresponding eigenfunction; here α, β, u, and v are real. Substituting this into (1), Sec. 5.7, we have

$$(pu' + ipv')' + (q + \alpha r + i\beta r)(u + iv) = 0.$$

This complex equation is equivalent to the following pair of equations for the real and the imaginary parts:

$$(pu')' + (q + \alpha r)u - \beta r v = 0$$
$$(pv')' + (q + \alpha r)v + \beta r u = 0.$$

Multiplying the first equation by v, the second by $-u$ and adding, we get

$$-\beta(u^2 + v^2)r = u(pv')' - v(pu')'$$
$$= [(pv')u - (pu')v]'.$$

The expression in brackets is continuous on $a \leqq x \leqq b$, for reasons similar to those in the proof of Theorem 1, Sec. 5.7. Integrating over x from a to b, we thus obtain

$$-\beta \int_a^b (u^2 + v^2)r \, dx = \Big[p(uv' - u'v) \Big]_a^b.$$

Because of the boundary conditions the right side is zero; this is as in that proof. Since y is an eigenfunction, $u^2 + v^2 \not\equiv 0$. Since y and r are continuous and $r > 0$ (or $r < 0$) on the interval $a \leqq x \leqq b$, the integral on the left is not zero. Hence, $\beta = 0$, which means that $\lambda = \alpha$ is real. This completes the proof. ∎

Section 7.7, page 308

THEOREM

Determinants

The definition of a determinant

(7) $$D = \det \mathbf{A} = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix}$$

as given in Sec. 7.7 is unambiguous, that is, it yields the same value of D no matter which rows or columns we choose in developings.

PROOF In this proof we shall use formula numbers not yet used in Sec. 7.7.

We shall prove first that *the same value is obtained no matter which* ***row*** *is chosen.*

The proof is by induction. The statement is true for a second-order determinant, for which the developments by the first row $a_{11}a_{22} + a_{12}(-a_{21})$ and by the second row $a_{21}(-a_{12}) + a_{22}a_{11}$ give the same value $a_{11}a_{22} - a_{12}a_{21}$. Assuming the statement to be true for an $(n - 1)$st-order determinant, we prove that it is true for an nth-order determinant.

For this purpose we expand D in terms of each of two arbitrary rows, say, the ith and the jth, and compare the results. Without loss of generality let us assume $i < j$.

First Expansion. We expand D by the ith row. A typical term in this expansion is

$$a_{ik}C_{ik} = a_{ik} \cdot (-1)^{i+k}M_{ik}. \tag{19}$$

The minor M_{ik} of a_{ik} in D is an $(n-1)$st-order determinant. By the induction hypothesis we may expand it by any row. We expand it by the row corresponding to the jth row of D. This row contains the entries a_{jl} $(l \neq k)$. It is the $(j-1)$st row of M_{ik}, because M_{ik} does not contain entries of the ith row of D, and $i < j$. We have to distinguish between two cases as follows.

Case I. If $l < k$, then the entry a_{jl} belongs to the lth column of M_{ik} (see Fig. 560). Hence the term involving a_{jl} in this expansion is

$$a_{jl} \cdot (\text{cofactor of } a_{jl} \text{ in } M_{ik}) = a_{jl} \cdot (-1)^{(j-1)+l}M_{ikjl} \tag{20}$$

where M_{ikjl} is the minor of a_{jl} in M_{ik}. Since this minor is obtained from M_{ik} by deleting the row and column of a_{jl}, it is obtained from D by deleting the ith and jth rows and the kth and lth columns of D. We insert the expansions of the M_{ik} into that of D. Then it follows from (19) and (20) that the terms of the resulting representation of D are of the form

$$a_{ik}a_{jl} \cdot (-1)^b M_{ikjl} \qquad (l < k) \tag{21a}$$

where

$$b = i + k + j + l - 1.$$

Case II. If $l > k$, the only difference is that then a_{jl} belongs to the $(l-1)$st column of M_{ik}, because M_{ik} does not contain entries of the kth column of D, and $k < l$. This causes an additional minus sign in (20), and, instead of (21a), we therefore obtain

$$-a_{ik}a_{jl} \cdot (-1)^b M_{ikjl} \qquad (l > k) \tag{21b}$$

where b is the same as before.

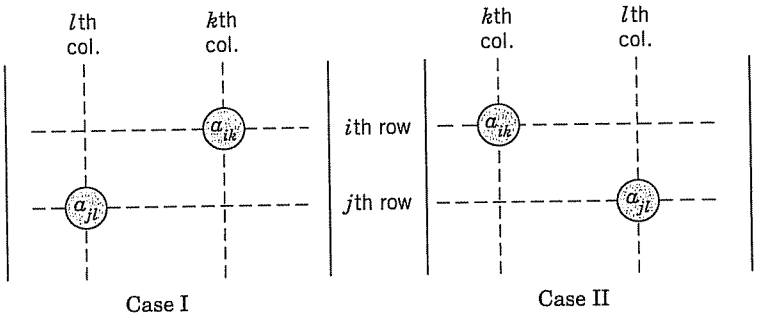


Fig. 560. Cases I and II of the two expansions of D

Second Expansion. We now expand D at first by the jth row. A typical term in this expansion is

$$a_{jl}C_{jl} = a_{jl} \cdot (-1)^{j+l}M_{jl}. \tag{22}$$

By the induction hypothesis we may expand the minor M_{jl} of a_{jl} in D by its ith row, which corresponds to the ith row of D, since $j > i$.

Case I. If $k > l$, the entry a_{ik} in that row belongs to the $(k-1)$st column of M_{jl}, because M_{jl} does not contain entries of the lth column of D, and $l < k$ (see Fig. 560). Hence the term involving a_{ik} in this expansion is

$$a_{ik} \cdot (\text{cofactor of } a_{ik} \text{ in } M_{jl}) = a_{ik} \cdot (-1)^{i+(k-1)}M_{ikjl}, \tag{23}$$

where the minor M_{ikjl} of a_{ik} in M_{jl} is obtained by deleting the ith and jth rows and the kth and lth columns of D [and is, therefore, identical with M_{ikjl} in (20), so that our notation is consistent]. We insert the expansions of the M_{jl} into that of D. It follows from (22) and (23) that this yields a representation whose terms are identical with those given by (21a) when $l < k$.

Case II. If $k < l$, then a_{ik} belongs to the kth column of M_{jl}, we obtain an additional minus sign, and the result agrees with that characterized by (21b).

We have shown that the two expansions of D consist of the same terms, and this proves our statement concerning rows.

The proof of the statement concerning ***columns*** is quite similar; if we expand D in terms of two arbitrary columns, say, the kth and the lth, we find that the general term involving $a_{jl}a_{ik}$ is exactly the same as before. This proves that not only all column expansions of D yield the same value, but also that their common value is equal to the common value of the row expansions of D.

This completes the proof and shows that *our definition of an nth-order determinant is unambiguous.* ▩

Section 9.3, page 377

PROOF OF FORMULA (2)

We prove that in right-handed Cartesian coordinates, the vector product

$$\mathbf{v} = \mathbf{a} \times \mathbf{b} = [a_1, \quad a_2, \quad a_3] \times [b_1, \quad b_2, \quad b_3]$$

has the components

$$v_1 = a_2b_3 - a_3b_2, \qquad v_2 = a_3b_1 - a_1b_3, \qquad v_3 = a_1b_2 - a_2b_1. \tag{2}$$

We need only consider the case $\mathbf{v} \neq \mathbf{0}$. Since $\mathbf{v}$ is perpendicular to both $\mathbf{a}$ and $\mathbf{b}$, Theorem 1 in Sec. 9.2 gives $\mathbf{a} \cdot \mathbf{v} = 0$ and $\mathbf{b} \cdot \mathbf{v} = 0$; in components [see (2), Sec. 9.2],

$$\begin{aligned} a_1v_1 + a_2v_2 + a_3v_3 &= 0 \\ b_1v_1 + b_2v_2 + b_3v_3 &= 0. \end{aligned} \tag{3}$$

Multiplying the first equation by b_3, the last by a_3, and subtracting, we obtain

$$(a_3b_1 - a_1b_3)v_1 = (a_2b_3 - a_3b_2)v_2.$$

Multiplying the first equation by b_1, the last by a_1, and subtracting, we obtain

$$(a_1b_2 - a_2b_1)v_2 = (a_3b_1 - a_1b_3)v_3.$$

We can easily verify that these two equations are satisfied by

$$(4) \qquad v_1 = c(a_2b_3 - a_3b_2), \qquad v_2 = c(a_3b_1 - a_1b_3), \qquad v_3 = c(a_1b_2 - a_2b_1)$$

where c is a constant. The reader may verify by inserting that (4) also satisfies (3). Now each of the equations in (3) represents a plane through the origin in $v_1v_2v_3$-space. The vectors **a** and **b** are normal vectors of these planes (see Example 6 in Sec. 9.2). Since $\mathbf{v} \neq \mathbf{0}$, these vectors are not parallel and the two planes do not coincide. Hence their intersection is a straight line L through the origin. Since (4) is a solution of (3) and, for varying c, represents a straight line, we conclude that (4) represents L, and every solution of (3) must be of the form (4). In particular, the components of **v** must be of this form, where c is to be determined. From (4) we obtain

$$|\mathbf{v}|^2 = {v_1}^2 + {v_2}^2 + {v_3}^2 = c^2[(a_2b_3 - a_3b_2)^2 + (a_3b_1 - a_1b_3)^2 + (a_1b_2 - a_2b_1)^2].$$

This can be written

$$|\mathbf{v}|^2 = c^2[({a_1}^2 + {a_2}^2 + {a_3}^2)({b_1}^2 + {b_2}^2 + {b_3}^2) - (a_1b_1 + a_2b_2 + a_3b_3)^2],$$

as can be verified by performing the indicated multiplications in both formulas and comparing. Using (2) in Sec. 9.2, we thus have

$$|\mathbf{v}|^2 = c^2[(\mathbf{a} \cdot \mathbf{a})(\mathbf{b} \cdot \mathbf{b}) - (\mathbf{a} \cdot \mathbf{b})^2].$$

By comparing this with formula (12) in Team Project 24 of Problem Set 9.3 we conclude that $c = \pm 1$.

We show that $c = +1$. This can be done as follows.

If we change the lengths and directions of **a** and **b** continuously and so that at the end $\mathbf{a} = \mathbf{i}$ and $\mathbf{b} = \mathbf{j}$ (Fig. 186a in Sec. 9.3), then **v** will change its length and direction continuously, and at the end, $\mathbf{v} = \mathbf{i} \times \mathbf{j} = \mathbf{k}$. Obviously we may effect the change so that both **a** and **b** remain different from the zero vector and are not parallel at any instant. Then **v** is never equal to the zero vector, and since the change is continuous and c can only assume the values $+1$ or -1, it follows that at the end c must have the same value as before. Now at the end $\mathbf{a} = \mathbf{i}$, $\mathbf{b} = \mathbf{j}$, $\mathbf{v} = \mathbf{k}$ and, therefore, $a_1 = 1$, $b_2 = 1$, $v_3 = 1$, and the other components in (4) are zero. Hence from (4) we see that $v_3 = c = +1$. This proves Theorem 1.

For a left-handed coordinate system, $\mathbf{i} \times \mathbf{j} = -\mathbf{k}$ (see Fig. 186b in Sec. 9.3), resulting in $c = -1$. This proves the statement right after formula (2). ∎

Section 9.9, page 416

PROOF OF THE INVARIANCE OF THE CURL

This proof will follow from two theorems (A and B), which we prove first.

THEOREM A

Transformation Law for Vector Components

For any vector $\mathbf{v}$ *the components* v_1, v_2, v_3 *and* v_1^*, v_2^*, v_3^* *in any two systems of Cartesian coordinates* x_1, x_3, x_3 *and* x_1^*, x_2^*, x_3^*, *respectively, are related by*

$$
\begin{aligned}
v_1^* &= c_{11}v_1 + c_{12}v_2 + c_{13}v_3 \\
v_2^* &= c_{21}v_1 + c_{22}v_2 + c_{23}v_3 \\
v_3^* &= c_{31}v_1 + c_{32}v_2 + c_{33}v_3,
\end{aligned} \tag{1}
$$

and conversely

$$
\begin{aligned}
v_1 &= c_{11}v_1^* + c_{21}v_2^* + c_{31}v_3^* \\
v_2 &= c_{12}v_1^* + c_{22}v_2^* + c_{32}v_3^* \\
v_3 &= c_{13}v_1^* + c_{23}v_2^* + c_{33}v_3^*
\end{aligned} \tag{2}
$$

with coefficients

$$
\begin{aligned}
c_{11} &= \mathbf{i}^* \cdot \mathbf{i} & c_{12} &= \mathbf{i}^* \cdot \mathbf{j} & c_{13} &= \mathbf{i}^* \cdot \mathbf{k} \\
c_{21} &= \mathbf{j}^* \cdot \mathbf{i} & c_{22} &= \mathbf{j}^* \cdot \mathbf{j} & c_{23} &= \mathbf{j}^* \cdot \mathbf{k} \\
c_{31} &= \mathbf{k}^* \cdot \mathbf{i} & c_{32} &= \mathbf{k}^* \cdot \mathbf{j} & c_{33} &= \mathbf{k}^* \cdot \mathbf{k}
\end{aligned} \tag{3}
$$

satisfying

$$
\sum_{j=1}^{3} c_{kj}c_{mj} = \delta_{km} \qquad (k, m = 1, 2, 3), \tag{4}
$$

where the **Kronecker delta**[2] *is given by*

$$
\delta_{km} = \begin{cases} 0 & (k \neq m) \\ 1 & (k = m) \end{cases}
$$

and $\mathbf{i}, \mathbf{j}, \mathbf{k}$ *and* $\mathbf{i}^*, \mathbf{j}^*, \mathbf{k}^*$ *denote the unit vectors in the positive* x_1-, x_2-, x_3- *and* x_1^*-, x_2^*-, x_3^*-*directions, respectively.*

---

[2]LEOPOLD KRONECKER (1823–1891), German mathematician at Berlin, who made important contributions to algebra, group theory, and number theory.

We shall keep our discussion completely independent of Chap. 7, but readers familiar with matrices should recognize that we are dealing with **orthogonal transformations and matrices** and that our present theorem follows from Theorem 2 in Sec. 8.3.

PROOF The representation of $\mathbf{v}$ in the two systems are

$$(5) \qquad (a) \quad \mathbf{v} = v_1\mathbf{i} + v_2\mathbf{j} + v_3\mathbf{k} \qquad (b) \quad \mathbf{v} = v_1^*\mathbf{i}^* + v_2^*\mathbf{j}^* + v_3^*\mathbf{k}^*.$$

Since $\mathbf{i}^* \cdot \mathbf{i}^* = 1$, $\mathbf{i}^* \cdot \mathbf{j}^* = 0$, $\mathbf{i}^* \cdot \mathbf{k}^* = 0$, we get from (5b) simply $\mathbf{i}^* \cdot \mathbf{v} = v_1^*$ and from this and (5a)

$$v_1^* = \mathbf{i}^* \cdot \mathbf{v} = \mathbf{i}^* \cdot v_1\mathbf{i} + \mathbf{i}^* \cdot v_2\mathbf{j} + \mathbf{i}^* \cdot v_3\mathbf{k} = v_1\mathbf{i}^* \cdot \mathbf{i} + v_2\mathbf{i}^* \cdot \mathbf{j} + v_3\mathbf{i}^* \cdot \mathbf{k}.$$

Because of (3), this is the first formula in (1), and the other two formulas are obtained similarly, by considering $\mathbf{j}^* \cdot \mathbf{v}$ and then $\mathbf{k}^* \cdot \mathbf{v}$. Formula (2) follows by the same idea, taking $\mathbf{i} \cdot \mathbf{v} = v_1$ from (5a) and then from (5b) and (3)

$$v_1 = \mathbf{i} \cdot \mathbf{v} = v_1^*\mathbf{i} \cdot \mathbf{i}^* + v_2^*\mathbf{i} \cdot \mathbf{j}^* + v_3^*\mathbf{i} \cdot \mathbf{k}^* = c_{11}v_1^* + c_{21}v_2^* + c_{31}v_3^*,$$

and similarly for the other two components.

We prove (4). We can write (1) and (2) briefly as

$$(6) \qquad (a) \quad v_j = \sum_{m=1}^{3} c_{mj}v_m^*, \qquad (b) \quad v_k^* = \sum_{j=1}^{3} c_{kj}v_j.$$

Substituting v_j into v_k^*, we get

$$v_k^* = \sum_{j=1}^{3} c_{kj} \sum_{m=1}^{3} c_{mj}v_m^* = \sum_{m=1}^{3} v_m^* \left(\sum_{j=1}^{3} c_{kj}c_{mj} \right),$$

where $k = 1, 2, 3$. Taking $k = 1$, we have

$$v_1^* = v_1^* \left(\sum_{j=1}^{3} c_{1j}c_{1j} \right) + v_2^* \left(\sum_{j=1}^{3} c_{1j}c_{2j} \right) + v_3^* \left(\sum_{j=1}^{3} c_{1j}c_{3j} \right).$$

For this to hold for *every* vector $\mathbf{v}$, the first sum must be 1 and the other two sums 0. This proves (4) with $k = 1$ for $m = 1, 2, 3$. Taking $k = 2$ and then $k = 3$, we obtain (4) with $k = 2$ and 3, for $m = 1, 2, 3$. ∎

THEOREM B

Transformation Law for Cartesian Coordinates

The transformation of any Cartesian $x_1x_2x_3$-coordinate system into any other Cartesian $x_1^*x_2^*x_3^*$-coordinate system is of the form

$$(7) \qquad x_m^* = \sum_{j=1}^{3} c_{mj}x_j + b_m, \quad m = 1, 2, 3,$$

with coefficients (3) *and constants b_1, b_2, b_3; conversely,*

$$(8) \qquad x_k = \sum_{n=1}^{3} c_{nk}x_n^* + \tilde{b}_k, \qquad k = 1, 2, 3.$$

Theorem B follows from Theorem A by noting that the most general transformation of a Cartesian coordinate system into another such system may be decomposed into a transformation of the type just considered and a translation; and under a translation, corresponding coordinates differ merely by a constant.

PROOF OF THE INVARIANCE OF THE CURL

We write again x_1, x_2, x_3 instead of x, y, z, and similarly x_1^*, x_2^*, x_3^* for other Cartesian coordinates, assuming that both systems are right-handed. Let a_1, a_2, a_3 denote the components of curl $\mathbf{v}$ in the $x_1x_2x_3$-coordinates, as given by (1), Sec. 9.9, with

$$x = x_1, \qquad y = x_2, \qquad z = x_3.$$

Similarly, let a_1^*, a_2^*, a_3^* denote the components of curl $\mathbf{v}$ in the $x_1^*x_2^*x_3^*$-coordinate system. We prove that the length and direction of curl $\mathbf{v}$ are independent of the particular choice of Cartesian coordinates, as asserted. We do this by showing that the components of curl $\mathbf{v}$ satisfy the transformation law (2), which is characteristic of vector components. We consider a_1. We use (6a), and then the chain rule for functions of several variables (Sec. 9.6). This gives

$$a_1 = \frac{\partial v_3}{\partial x_2} - \frac{\partial v_2}{\partial x_3} = \sum_{m=1}^{3} \left(c_{m3} \frac{\partial v_m^*}{\partial x_2} - c_{m2} \frac{\partial v_m^*}{\partial x_3} \right)$$

$$= \sum_{m=1}^{3} \sum_{j=1}^{3} \left(c_{m3} \frac{\partial v_m^*}{\partial x_j^*} \frac{\partial x_j^*}{\partial x_2} - c_{m2} \frac{\partial v_m^*}{\partial x_j^*} \frac{\partial x_j^*}{\partial x_3} \right).$$

From this and (7) we obtain

$$a_1 = \sum_{m=1}^{3} \sum_{j=1}^{3} (c_{m3}c_{j2} - c_{m2}c_{j3}) \frac{\partial v_m^*}{\partial x_j^*}$$

$$= (c_{33}c_{22} - c_{32}c_{23}) \left(\frac{\partial v_3^*}{\partial x_2^*} - \frac{\partial v_2^*}{\partial x_3^*} \right) + \cdots$$

$$= (c_{33}c_{22} - c_{32}c_{23})a_1^* + (c_{13}c_{32} - c_{12}c_{33})a_2^* + (c_{23}c_{12} - c_{22}c_{13})a_3^*.$$

Note what we did. The double sum had $3 \times 3 = 9$ terms, 3 of which were zero (when $m = j$), and the remaining 6 terms we combined in pairs as we needed them in getting a_1^*, a_2^*, a_3^*.

We now use (3), Lagrange's identity (see Team Project 24 in Problem Set 9.3) and $\mathbf{k}^* \times \mathbf{j}^* = -\mathbf{i}^*$ and $\mathbf{k} \times \mathbf{j} = -\mathbf{i}$. Then

$$c_{33}c_{22} - c_{32}c_{23} = (\mathbf{k}^* \cdot \mathbf{k})(\mathbf{j}^* \cdot \mathbf{j}) - (\mathbf{k}^* \cdot \mathbf{j})(\mathbf{j}^* \cdot \mathbf{k})$$

$$= (\mathbf{k}^* \times \mathbf{j}^*) \cdot (\mathbf{k} \times \mathbf{j}) = \mathbf{i}^* \cdot \mathbf{i} = c_{11}, \qquad \text{etc.}$$

Hence $a_1 = c_{11}a_1^* + c_{21}a_2^* + c_{31}a_3^*$. This is of the form of the first formula in (2) in Theorem A, and the other two formulas of the form (2) are obtained similarly. This proves the theorem for right-handed systems. If the $x_1x_2x_3$-coordinates are left-handed, then $\mathbf{k} \times \mathbf{j} = +\mathbf{i}$, but then there is a minus sign in front of the determinant in (1), Sec. 9.9. ■

Section 10.2, pages 426–427

PROOF OF THEOREM 1, PART (b) We prove that if

$$(1) \qquad \int_C \mathbf{F}(\mathbf{r}) \cdot d\mathbf{r} = \int_C (F_1\,dx + F_2\,dy + F_3\,dz)$$

with continuous F_1, F_2, F_3 in a domain D is independent of path in D, then $F = \operatorname{grad} f$ in D for some f; in components

$$(2') \qquad F_1 = \frac{\partial f}{\partial x}, \qquad F_2 = \frac{\partial f}{\partial y}, \qquad F_3 = \frac{\partial f}{\partial z}.$$

We choose any fixed A: (x_0, y_0, z_0) in D and any B: (x, y, z) in D and define f by

$$(3) \qquad f(x, y, z) = f_0 + \int_A^B (F_1\,dx^* + F_2\,dy^* + F_3\,dz^*)$$

with any constant f_0 and any path from A to B in D. Since A is fixed and we have independence of path, the integral depends only on the coordinates x, y, z, so that (3) defines a function $f(x, y, z)$ in D. We show that $\mathbf{F} = \operatorname{grad} f$ with this f, beginning with the first of the three relations (2′). Because of independence of path we may integrate from A to B_1: (x_1, y, z) and then parallel to the x-axis along the segment B_1B in Fig. 561 with B_1 chosen so that the whole segment lies in D. Then

$$f(x, y, z) = f_0 + \int_A^{B_1} (F_1\,dx^* + F_2\,dy^* + F_3\,dz^*) + \int_{B_1}^B (F_1\,dx^* + F_2\,dy^* + F_3\,dz^*).$$

We now take the partial derivative with respect to x on both sides. On the left we get $\partial f/\partial x$. We show that on the right we get F_1. The derivative of the first integral is zero because A: (x_0, y_0, z_0) and B_1: (x_1, y, z) do not depend on x. We consider the second integral. Since on the segment B_1B, both y and z are constant, the terms $F_2\,dy^*$ and

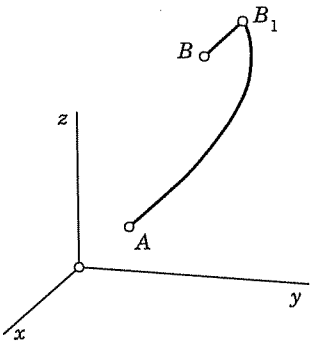


Fig. 561. Proof of Theorem 1

$F_3\,dz^*$ do not contribute to the derivative of the integral. The remaining part can be written as a definite integral,

$$\int_{B_1}^{B} F_1\,dx^* = \int_{x_1}^{x} F_1(x^*, y, z)\,dx^*.$$

Hence its partial derivative with respect to x is $F_1(x, y, z)$, and the first of the relations $(2')$ is proved. The other two formulas in $(2')$ follow by the same argument. ▩

Section 13.4, page 620

PROOF OF THEOREM 2 Cauchy–Riemann Equations

We prove that Cauchy–Riemann equations

$$(1) \qquad u_x = v_y, \qquad u_y = -v_x$$

are sufficient for a complex function $f(z) = u(x, y) + iv(x, y)$ to be analytic; precisely, *if the real part u and the imaginary part v of $f(z)$ satisfy* (1) *in a domain D in the complex plane and if the partial derivatives in* (1) *are* ***continuous*** *in D, then $f(z)$ is analytic in D.*

In this proof we write $\Delta z = \Delta x + i\Delta y$ and $\Delta f = f(z + \Delta z) - f(z)$. The idea of proof is as follows.

(a) We express Δf in terms of first partial derivatives of u and v, by applying the mean value theorem of Sec. 9.6.

(b) We get rid of partial derivatives with respect to y by applying the Cauchy–Riemann equations.

(c) We let Δz approach zero and show that then $\Delta f/\Delta z$ as obtained approaches a limit, which is equal to $u_x + iv_x$, the right side of (4) in Sec. 13.4, regardless of the way of approach to zero.

(a) Let P: (x, y) be any fixed point in D. Since D is a domain, it contains a neighborhood of P. We can choose a point Q: $(x + \Delta x, y + \Delta y)$ in this neighborhood such that the straight-line segment PQ is in D. Because of our continuity assumptions we may apply the mean value theorem in Sec. 9.6. This yields

$$u(x + \Delta x, y + \Delta y) - u(x, y) = (\Delta x)u_x(M_1) + (\Delta y)u_y(M_1)$$
$$v(x + \Delta x, y + \Delta y) - v(x, y) = (\Delta x)v_x(M_2) + (\Delta y)v_y(M_2)$$

where M_1 and M_2 ($\neq M_1$ in general!) are suitable points on that segment. The first line is Re Δf and the second is Im Δf, so that

$$\Delta f = (\Delta x)u_x(M_1) + (\Delta y)u_y(M_1) + i\big[(\Delta x)v_x(M_2) + (\Delta y)v_y(M_2)\big].$$

(b) $u_y = -v_x$ and $v_y = u_x$ by the Cauchy–Riemann equations, so that

$$\Delta f = (\Delta x)u_x(M_1) - (\Delta y)v_x(M_1) + i\big[(\Delta x)v_x(M_2) + (\Delta y)u_x(M_2)\big].$$

Also $\Delta z = \Delta x + i\Delta y$, so that we can write $\Delta x = \Delta z - i\Delta y$ in the first term and $\Delta y = (\Delta z - \Delta x)/i = -i(\Delta z - \Delta x)$ in the second term. This gives

$$\Delta f = (\Delta z - i\Delta y)u_x(M_1) + i(\Delta z - \Delta x)v_x(M_1) + i[(\Delta x)v_x(M_2) + (\Delta y)u_x(M_2)].$$

By performing the multiplications and reordering we obtain

$$\Delta f = (\Delta z)u_x(M_1) - i\Delta y\{u_x(M_1) - u_x(M_2)\} + i[(\Delta z)v_x(M_1) - \Delta x\{v_x(M_1) - v_x(M_2)\}].$$

Division by Δz now yields

$$\text{(A)} \quad \frac{\Delta f}{\Delta z} = u_x(M_1) + iv_x(M_1) - \frac{i\Delta y}{\Delta z}\{u_x(M_1) - u_x(M_2)\} - \frac{i\Delta x}{\Delta z}\{v_x(M_1) - v_x(M_2)\}.$$

(c) We finally let Δz approach zero and note that $|\Delta y/\Delta z| \leqq 1$ and $|\Delta x/\Delta z| \leqq 1$ in (A). Then Q: $(x + \Delta x, y + \Delta y)$ approaches P: (x, y), so that M_1 and M_2 must approach P. Also, since the partial derivatives in (A) are assumed to be continuous, they approach their value at P. In particular, the differences in the braces $\{\cdots\}$ in (A) approach zero. Hence the limit of the right side of (A) exists and is independent of the path along which $\Delta z \to 0$. We see that this limit equals the right side of (4) in Sec. 13.4. This means that $f(z)$ is analytic at every point z in D, and the proof is complete. ▩

Section 14.2, pages 647–648

GOURSAT'S PROOF OF CAUCHY'S INTEGRAL THEOREM Goursat proved Cauchy's integral theorem without assuming that $f'(z)$ is continuous, as follows.

We start with the case when C is the boundary of a triangle. We orient C counterclockwise. By joining the midpoints of the sides we subdivide the triangle into four congruent triangles (Fig. 562). Let C_{I}, C_{II}, C_{III}, C_{IV} denote their boundaries. We claim that (see Fig. 562).

$$\text{(1)} \qquad \oint_C f\,dz = \oint_{C_{\mathrm{I}}} f\,dz + \oint_{C_{\mathrm{II}}} f\,dz + \oint_{C_{\mathrm{III}}} f\,dz + \oint_{C_{\mathrm{IV}}} f\,dz.$$

Indeed, on the right we integrate along each of the three segments of subdivision in both possible directions (Fig. 562), so that the corresponding integrals cancel out in pairs, and the sum of the integrals on the right equals the integral on the left. We now pick an integral on the right that is biggest in absolute value and call its path C_1. Then, by the triangle inequality (Sec. 13.2),

$$\left|\oint_C f\,dz\right| \leqq \left|\oint_{C_{\mathrm{I}}} f\,dz\right| + \left|\oint_{C_{\mathrm{II}}} f\,dz\right| + \left|\oint_{C_{\mathrm{III}}} f\,dz\right| + \left|\oint_{C_{\mathrm{IV}}} f\,dz\right| \leqq 4\left|\oint_{C_1} f\,dz\right|.$$

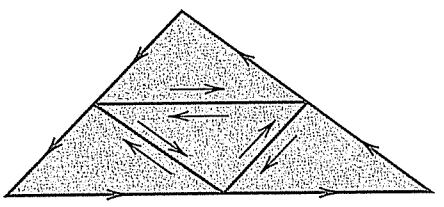


Fig. 562. Proof of Cauchy's integral theorem

We now subdivide the triangle bounded by C_1 as before and select a triangle of subdivision with boundary C_2 for which

$$\left|\oint_{C_1} f\,dz\right| \leqq 4\left|\oint_{C_2} f\,dz\right|. \qquad \text{Then} \qquad \left|\oint_{C} f\,dz\right| \leqq 4^2\left|\oint_{C_2} f\,dz\right|.$$

Continuing in this fashion, we obtain a sequence of triangles $T_1, T_2, \cdots$ with boundaries $C_1, C_2, \cdots$ that are similar and such that T_n lies in T_m when $n > m$, and

$$\left|\oint_{C} f\,dz\right| \leqq 4^n\left|\oint_{C_n} f\,dz\right|, \qquad n = 1, 2, \cdots. \tag{2}$$

Let z_0 be the point that belongs to all these triangles. Since f is differentiable at $z = z_0$, the derivative $f'(z_0)$ exists. Let

$$h(z) = \frac{f(z) - f(z_0)}{z - z_0} - f'(z_0). \tag{3}$$

Solving this algebraically for $f(z)$ we have

$$f(z) = f(z_0) + (z - z_0)f'(z_0) + h(z)(z - z_0).$$

Integrating this over the boundary C_n of the triangle T_n gives

$$\oint_{C_n} f(z)\,dz = \oint_{C_n} f(z_0)\,dz + \oint_{C_n} (z - z_0)f'(z_0)\,dz + \oint_{C_n} h(z)(z - z_0)\,dz.$$

Since $f(z_0)$ and $f'(z_0)$ are constants and C_n is a closed path, the first two integrals on the right are zero, as follows from Cauchy's proof, which is applicable because the integrands do have continuous derivatives (0 and *const,* respectively). We thus have

$$\oint_{C_n} f(z)\,dz = \oint_{C_n} h(z)(z - z_0)\,dz.$$

Since $f'(z_0)$ is the limit of the difference quotient in (3), for given $\epsilon > 0$ we can find a $\delta > 0$ such that

$$|h(z)| < \epsilon \qquad \text{when} \qquad |z - z_0| < \delta. \tag{4}$$

We may now take n so large that the triangle T_n lies in the disk $|z - z_0| < \delta$. Let L_n be the length of C_n. Then $|z - z_0| < L_n$ for all z on C_n and z_0 in T_n. From this and (4) we have $|h(z)(z - z_0)| < \epsilon L_n$. The *ML*-inequality in Sec. 14.1 now gives

$$\left|\oint_{C_n} f(z)\,dz\right| = \left|\oint_{C_n} h(z)(z - z_0)\,dz\right| \leqq \epsilon L_n \cdot L_n = \epsilon {L_n}^2. \tag{5}$$

Now denote the length of C by L. Then the path C_1 has the length $L_1 = L/2$, the path C_2 has the length $L_2 = L_1/2 = L/4$, etc., and C_n has the length $L_n = L/2^n$. Hence ${L_n}^2 = L^2/4^n$. From (2) and (5) we thus obtain

$$\left|\oint_{C} f\,dz\right| \leqq 4^n\left|\oint_{C_n} f\,dz\right| \leqq 4^n\epsilon {L_n}^2 = 4^n\epsilon\,\frac{L^2}{4^n} = \epsilon L^2.$$

By choosing ϵ (> 0) sufficiently small we can make the expression on the right as small as we please, while the expression on the left is the definite value of an integral. Consequently, this value must be zero, and the proof is complete.

The proof for *the case in which C is the boundary of a polygon* follows from the previous proof by subdividing the polygon into triangles (Fig. 563). The integral corresponding to each such triangle is zero. The sum of these integrals is equal to the integral over C, because we integrate along each segment of subdivision in both directions, the corresponding integrals cancel out in pairs, and we are left with the integral over C.

The case of a general simple closed path C can be reduced to the preceding one by inscribing in C a closed polygon P of chords, which approximates C "sufficiently accurately," and it can be shown that there is a polygon P such that the integral over P differs from that over C by less than any preassigned positive real number $\tilde{\epsilon}$, no matter how small. The details of this proof are somewhat involved and can be found in Ref. [D6] listed in App. 1. ▩

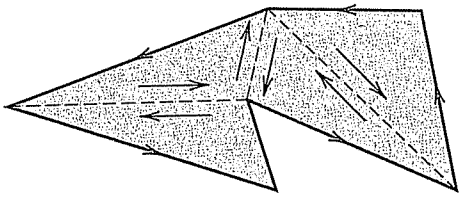


Fig. 563. Proof of Cauchy's integral theorem for a polygon

Section 15.1, page 667

PROOF OF THEOREM 4 Cauchy's Convergence Principle for Series

(a) In this proof we need two concepts and a theorem, which we list first.

1. A **bounded sequence** $s_1, s_2, \cdots$ is a sequence whose terms all lie in a disk of (sufficiently large, finite) radius K with center at the origin; thus $|s_n| < K$ for all n.

2. A **limit point** a of a sequence $s_1, s_2, \cdots$ is a point such that, given an $\epsilon > 0$, there are infinitely many terms satisfying $|s_n - a| < \epsilon$. (Note that this does *not* imply convergence, since there may still be infinitely many terms that do not lie within that circle of radius ϵ and center a.)

Example: $\frac{1}{4}, \frac{3}{4}, \frac{1}{8}, \frac{7}{8}, \frac{1}{16}, \frac{15}{16}, \cdots$ has the limit points 0 and 1 and diverges.

3. A bounded sequence in the complex plane has at least one limit point. (Bolzano–Weierstrass theorem; proof below. Recall that "sequence" always mean *infinite* sequence.)

(b) We now turn to the actual proof that $z_1 + z_2 + \cdots$ converges if and only if for every $\epsilon > 0$ we can find an N such that

$$(1) \qquad |z_{n+1} + \cdots + z_{n+p}| < \epsilon \qquad \text{for every } n > N \text{ and } p = 1, 2, \cdots.$$

Here, by the definition of partial sums,

$$s_{n+p} - s_n = z_{n+1} + \cdots + z_{n+p}.$$

Writing $n + p = r$, we see from this that (1) is equivalent to

$$(1^*) \qquad |s_r - s_n| < \epsilon \qquad \text{for all } r > N \text{ and } n > N.$$

Suppose that $s_1, s_2, \cdots$ converges. Denote its limit by s. Then for a given $\epsilon > 0$ we can find an N such that

$$|s_n - s| < \frac{\epsilon}{2} \qquad \text{for every } n > N.$$

Hence, if $r > N$ and $n > N$, then by the triangle inequality (Sec. 13.2),

$$|s_r - s_n| = |(s_r - s) - (s_n - s)| \leqq |s_r - s| + |s_n - s| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,$$

that is, (1*) holds.

(c) Conversely, assume that $s_1, s_2, \cdots$ satisfies (1*). We first prove that then the sequence must be bounded. Indeed, choose a fixed ϵ and a fixed $n = n_0 > N$ in (1*). Then (1*) implies that all s_r with $r > N$ lie in the disk of radius ϵ and center s_{n_0} and only *finitely many terms* $s_1, \cdots, s_N$ may not lie in this disk. Clearly, we can now find a circle so large that this disk and these finitely many terms all lie within this new circle. Hence the sequence is bounded. By the Bolzano–Weierstrass theorem, it has at least one limit point, call it s.

We now show that the sequence is convergent with the limit s. Let $\epsilon > 0$ be given. Then there is an N^* such that $|s_r - s_n| < \epsilon/2$ for all $r > N^*$ and $n > N^*$, by (1*). Also, by the definition of a limit point, $|s_n - s| < \epsilon/2$ for *infinitely many* n, so that we can find and fix an $n > N^*$ such that $|s_n - s| < \epsilon/2$. Together, for *every* $r > N^*$,

$$|s_r - s| = |(s_r - s_n) + (s_n - s)| \leqq |s_r - s_n| + |s_n - s| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon;$$

that is, the sequence $s_1, s_2, \cdots$ is convergent with the limit s. ▩

THEOREM

Bolzano–Weierstrass Theorem[3]

A bounded infinite sequence $z_1, z_2, z_3, \cdots$ in the complex plane has at least one limit point.

PROOF It is obvious that we need both conditions: a finite sequence cannot have a limit point, and the sequence 1, 2, 3, $\cdots$, which is infinite but not bounded, has no limit point. To prove the theorem, consider a bounded infinite sequence $z_1, z_2, \cdots$ and let K be such that $|z_n| < K$ for all n. If only finitely many values of the z_n are different, then, since the sequence is infinite, some number z must occur infinitely many times in the sequence, and, by definition, this number is a limit point of the sequence.

We may now turn to the case when the sequence contains infinitely many *different* terms. We draw a large square Q_0 that contains all z_n. We subdivide Q_0 into four congruent squares, which we number 1, 2, 3, 4. Clearly, at least one of these squares (each taken

[3]BERNARD BOLZANO (1781–1848), Austrian mathematician and professor of religious studies, was a pioneer in the study of point sets, the foundation of analysis, and mathematical logic.

For Weierstrass, see Sec. 15.5.

with its complete boundary) must contain infinitely many terms of the sequence. The square of this type with the lowest number (1, 2, 3, or 4) will be denoted by Q_1. This is the first step. In the next step we subdivide Q_1 into four congruent squares and select a square Q_2 by the same rule, and so on. This yields an infinite sequence of squares Q_0, $Q_1, Q_2, \cdots, Q_n, \cdots$ with the property that the side of Q_n approaches zero as n approaches infinity, and Q_m contains all Q_n with $n > m$. It is not difficult to see that the number which belongs to all these squares,[4] call it $z = a$, is a limit point of the sequence. In fact, given an $\epsilon > 0$, we can choose an N so large that the side of the square Q_N is less than ϵ and, since Q_N contains infinitely many z_n, we have $|z_n - a| < \epsilon$ for infinitely many n. This completes the proof. ∎

Section 15.3, pages 681–682

PART (b) OF THE PROOF OF THEOREM 5

We have to show that

$$\sum_{n=2}^{\infty} a_n \left[\frac{(z + \Delta z)^n - z^n}{\Delta z} - nz^{n-1} \right]$$

$$= \sum_{n=2}^{\infty} a_n \, \Delta z \left[(z + \Delta z)^{n-2} + 2z(z + \Delta z)^{n-3} + \cdots + (n-1)z^{n-2} \right],$$

thus,

$$\frac{(z + \Delta z)^n - z^n}{\Delta z} - nz^{n-1}$$

$$= \Delta z \left[(z + \Delta z)^{n-2} + 2z(z + \Delta z)^{n-3} + \cdots + (n-1)z^{n-2} \right].$$

If we set $z + \Delta z = b$ and $z = a$, thus $\Delta z = b - a$, this becomes simply

$$\text{(7a)} \qquad \frac{b^n - a^n}{b - a} - na^{n-1} = (b - a)A_n \qquad (n = 2, 3, \cdots),$$

where A_n is the expression in the brackets on the right,

$$\text{(7b)} \qquad A_n = b^{n-2} + 2ab^{n-3} + 3a^2b^{n-4} + \cdots + (n-1)a^{n-2};$$

thus, $A_2 = 1$, $A_3 = b + 2a$, etc. We prove (7) by induction. When $n = 2$, then (7) holds, since then

$$\frac{b^2 - a^2}{b - a} - 2a = \frac{(b + a)(b - a)}{b - a} - 2a = b - a = (b - a)A_2.$$

Assuming that (7) holds for $n = k$, we show that it holds for $n = k + 1$. By adding and subtracting a term in the numerator and then dividing we first obtain

$$\frac{b^{k+1} - a^{k+1}}{b - a} = \frac{b^{k+1} - ba^k + ba^k - a^{k+1}}{b - a} = b \, \frac{b^k - a^k}{b - a} + a^k.$$

---

[4] The fact that such a unique number $z = a$ exists seems to be obvious, but it actually follows from an axiom of the real number system, the so-called *Cantor–Dedekind axiom*: see footnote 3 in App. A3.3.

By the induction hypothesis, the right side equals $b[(b-a)A_k + ka^{k-1}] + a^k$. Direct calculation shows that this is equal to

$$(b-a)\{bA_k + ka^{k-1}\} + aka^{k-1} + a^k.$$

From (7b) with $n = k$ we see that the expression in the braces $\{\cdots\}$ equals

$$b^{k-1} + 2ab^{k-2} + \cdots + (k-1)ba^{k-2} + ka^{k-1} = A_{k+1}.$$

Hence our result is

$$\frac{b^{k+1} - a^{k+1}}{b-a} = (b-a)A_{k+1} + (k+1)a^k.$$

Taking the last term to the left, we obtain (7) with $n = k+1$. This proves (7) for any integer $n \geqq 2$ and completes the proof. ∎

Section 18.2, page 754

ANOTHER PROOF OF THEOREM 1 ***without the use of a harmonic conjugate***

We show that if $w = u + iv = f(z)$ is analytic and maps a domain D conformally onto a domain D^* and $\Phi^*(u, v)$ is harmonic in D^*, then

(1) $$\Phi(x, y) = \Phi^*(u(x, y), v(x, y))$$

is harmonic in D, that is, $\nabla^2\Phi = 0$ in D. We make no use of a harmonic conjugate of Φ^*, but use straightforward differentiation. By the chain rule,

$$\Phi_x = \Phi_u^* u_x + \Phi_v^* v_x.$$

We apply the chain rule again, underscoring the terms that will drop out when we form $\nabla^2\Phi$:

$$\Phi_{xx} = \underline{\Phi_u^* u_{xx}} + (\Phi_{uu}^* u_x + \underline{\Phi_{uv}^* v_x})u_x$$
$$+ \underline{\Phi_v^* v_{xx}} + (\underline{\Phi_{vu}^* u_x} + \Phi_{vv}^* v_x)v_x.$$

Φ_{yy} is the same with each x replaced by y. We form the sum $\nabla^2\Phi$. In it, $\Phi_{vu}^* = \Phi_{uv}^*$ is multiplied by

$$u_x v_x + u_y v_y$$

which is 0 by the Cauchy–Riemann equations. Also $\nabla^2 u = 0$ and $\nabla^2 v = 0$. There remains

$$\nabla^2\Phi = \Phi_{uu}^*(u_x^2 + u_y^2) + \Phi_{vv}^*(v_x^2 + v_y^2).$$

By the Cauchy–Riemann equations this becomes

$$\nabla^2\Phi = (\Phi_{uu}^* + \Phi_{vv}^*)(u_x^2 + v_x^2)$$

and is 0 since Φ^* is harmonic. ∎

APPENDIX 5

Tables

For Tables of Laplace transforms see Secs. 6.8 and 6.9.
For Tables of Fourier transforms see Sec. 11.10.

If you have a Computer Algebra System (CAS), you may not need the present tables, but you may still find them convenient from time to time.

Table A1 Bessel Functions

For more extensive tables see Ref. [GR1] in App. 1.

x	$J_0(x)$	$J_1(x)$	x	$J_0(x)$	$J_1(x)$	x	$J_0(x)$	$J_1(x)$
0.0	1.0000	0.0000	3.0	−0.2601	0.3391	6.0	0.1506	−0.2767
0.1	0.9975	0.0499	3.1	−0.2921	0.3009	6.1	0.1773	−0.2559
0.2	0.9900	0.0995	3.2	−0.3202	0.2613	6.2	0.2017	−0.2329
0.3	0.9776	0.1483	3.3	−0.3443	0.2207	6.3	0.2238	−0.2081
0.4	0.9604	0.1960	3.4	−0.3643	0.1792	6.4	0.2433	−0.1816
0.5	0.9385	0.2423	3.5	−0.3801	0.1374	6.5	0.2601	−0.1538
0.6	0.9120	0.2867	3.6	−0.3918	0.0955	6.6	0.2740	−0.1250
0.7	0.8812	0.3290	3.7	−0.3992	0.0538	6.7	0.2851	−0.0953
0.8	0.8463	0.3688	3.8	−0.4026	0.0128	6.8	0.2931	−0.0652
0.9	0.8075	0.4059	3.9	−0.4018	−0.0272	6.9	0.2981	−0.0349
1.0	0.7652	0.4401	4.0	−0.3971	−0.0660	7.0	0.3001	−0.0047
1.1	0.7196	0.4709	4.1	−0.3887	−0.1033	7.1	0.2991	0.0252
1.2	0.6711	0.4983	4.2	−0.3766	−0.1386	7.2	0.2951	0.0543
1.3	0.6201	0.5220	4.3	−0.3610	−0.1719	7.3	0.2882	0.0826
1.4	0.5669	0.5419	4.4	−0.3423	−0.2028	7.4	0.2786	0.1096
1.5	0.5118	0.5579	4.5	−0.3205	−0.2311	7.5	0.2663	0.1352
1.6	0.4554	0.5699	4.6	−0.2961	−0.2566	7.6	0.2516	0.1592
1.7	0.3980	0.5778	4.7	−0.2693	−0.2791	7.7	0.2346	0.1813
1.8	0.3400	0.5815	4.8	−0.2404	−0.2985	7.8	0.2154	0.2014
1.9	0.2818	0.5812	4.9	−0.2097	−0.3147	7.9	0.1944	0.2192
2.0	0.2239	0.5767	5.0	−0.1776	−0.3276	8.0	0.1717	0.2346
2.1	0.1666	0.5683	5.1	−0.1443	−0.3371	8.1	0.1475	0.2476
2.2	0.1104	0.5560	5.2	−0.1103	−0.3432	8.2	0.1222	0.2580
2.3	0.0555	0.5399	5.3	−0.0758	−0.3460	8.3	0.0960	0.2657
2.4	0.0025	0.5202	5.4	−0.0412	−0.3453	8.4	0.0692	0.2708
2.5	−0.0484	0.4971	5.5	−0.0068	−0.3414	8.5	0.0419	0.2731
2.6	−0.0968	0.4708	5.6	0.0270	−0.3343	8.6	0.0146	0.2728
2.7	−0.1424	0.4416	5.7	0.0599	−0.3241	8.7	−0.0125	0.2697
2.8	−0.1850	0.4097	5.8	0.0917	−0.3110	8.8	−0.0392	0.2641
2.9	−0.2243	0.3754	5.9	0.1220	−0.2951	8.9	−0.0653	0.2559

$J_0(x) = 0$ for $x = 2.40483, 5.52008, 8.65373, 11.7915, 14.9309, 18.0711, 21.2116, 24.3525, 27.4935, 30.6346$

$J_1(x) = 0$ for $x = 3.83171, 7.01559, 10.1735, 13.3237, 16.4706, 19.6159, 22.7601, 25.9037, 29.0468, 32.1897$

Table A1 ***(continued)***

x	$Y_0(x)$	$Y_1(x)$	x	$Y_0(x)$	$Y_1(x)$	x	$Y_0(x)$	$Y_1(x)$
0.0	$(-\infty)$	$(-\infty)$	2.5	0.498	0.146	5.0	−0.309	0.148
0.5	−0.445	−1.471	3.0	0.377	0.325	5.5	−0.339	−0.024
1.0	0.088	−0.781	3.5	0.189	0.410	6.0	−0.288	−0.175
1.5	0.382	−0.412	4.0	−0.017	0.398	6.5	−0.173	−0.274
2.0	0.510	−0.107	4.5	−0.195	0.301	7.0	−0.026	−0.303

Table A2 Gamma Function [see (24) in App. A3.1]

α	$\Gamma(\alpha)$	α	$\Gamma(\alpha)$	α	$\Gamma(\alpha)$	α	$\Gamma(\alpha)$	α	$\Gamma(\alpha)$
1.00	1.000 000	1.20	0.918 169	1.40	0.887 264	1.60	0.893 515	1.80	0.931 384
1.02	0.988 844	1.22	0.913 106	1.42	0.886 356	1.62	0.895 924	1.82	0.936 845
1.04	0.978 438	1.24	0.908 521	1.44	0.885 805	1.64	0.898 642	1.84	0.942 612
1.06	0.968 744	1.26	0.904 397	1.46	0.885 604	1.66	0.901 668	1.86	0.948 687
1.08	0.959 725	1.28	0.900 718	1.48	0.885 747	1.68	0.905 001	1.88	0.955 071
1.10	0.951 351	1.30	0.897 471	1.50	0.886 227	1.70	0.908 639	1.90	0.961 766
1.12	0.943 590	1.32	0.894 640	1.52	0.887 039	1.72	0.912 581	1.92	0.968 774
1.14	0.936 416	1.34	0.892 216	1.54	0.888 178	1.74	0.916 826	1.94	0.976 099
1.16	0.929 803	1.36	0.890 185	1.56	0.889 639	1.76	0.921 375	1.96	0.983 743
1.18	0.923 728	1.38	0.888 537	1.58	0.891 420	1.78	0.926 227	1.98	0.991 708
1.20	0.918 169	1.40	0.887 264	1.60	0.893 515	1.80	0.931 384	2.00	1.000 000

Table A3 Factorial Function and Its Logarithm with Base 10

n	$n!$	$\log(n!)$	n	$n!$	$\log(n!)$	n	$n!$	$\log(n!)$
1	1	0.000 000	6	720	2.857 332	11	39 916 800	7.601 156
2	2	0.301 030	7	5 040	3.702 431	12	479 001 600	8.680 337
3	6	0.778 151	8	40 320	4.605 521	13	6 227 020 800	9.794 280
4	24	1.380 211	9	362 880	5.559 763	14	87 178 291 200	10.940 408
5	120	2.079 181	10	3 628 800	6.559 763	15	1 307 674 368 000	12.116 500

Table A4 Error Function, Sine and Cosine Integrals [see (35), (40), (42) in App. A3.1]

x	erf x	Si(x)	ci(x)	x	erf x	Si(x)	ci(x)
0.0	0.0000	0.0000	∞	2.0	0.9953	1.6054	−0.4230
0.2	0.2227	0.1996	1.0422	2.2	0.9981	1.6876	−0.3751
0.4	0.4284	0.3965	0.3788	2.4	0.9993	1.7525	−0.3173
0.6	0.6039	0.5881	0.0223	2.6	0.9998	1.8004	−0.2533
0.8	0.7421	0.7721	−0.1983	2.8	0.9999	1.8321	−0.1865
1.0	0.8427	0.9461	−0.3374	3.0	1.0000	1.8487	−0.1196
1.2	0.9103	1.1080	−0.4205	3.2	1.0000	1.8514	−0.0553
1.4	0.9523	1.2562	−0.4620	3.4	1.0000	1.8419	0.0045
1.6	0.9763	1.3892	−0.4717	3.6	1.0000	1.8219	0.0580
1.8	0.9891	1.5058	−0.4568	3.8	1.0000	1.7934	0.1038
2.0	0.9953	1.6054	−0.4230	4.0	1.0000	1.7582	0.1410

Table A5 Binomial Distribution

Probability function $f(x)$ [see (2), Sec. 24.7] and distribution function $F(x)$

n	x	$p = 0.1$		$p = 0.2$		$p = 0.3$		$p = 0.4$		$p = 0.5$	
		$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$
		0.		**0.**		**0.**		**0.**		**0.**	
1	0	9000	0.9000	8000	0.8000	7000	0.7000	6000	0.6000	5000	0.5000
	1	1000	1.0000	2000	1.0000	3000	1.0000	4000	1.0000	5000	1.0000
2	0	8100	0.8100	6400	0.6400	4900	0.4900	3600	0.3600	2500	0.2500
	1	1800	0.9900	3200	0.9600	4200	0.9100	4800	0.8400	5000	0.7500
	2	0100	1.0000	0400	1.0000	0900	1.0000	1600	1.0000	2500	1.0000
3	0	7290	0.7290	5120	0.5120	3430	0.3430	2160	0.2160	1250	0.1250
	1	2430	0.9720	3840	0.8960	4410	0.7840	4320	0.6480	3750	0.5000
	2	0270	0.9990	0960	0.9920	1890	0.9730	2880	0.9360	3750	0.8750
	3	0010	1.0000	0080	1.0000	0270	1.0000	0640	1.0000	1250	1.0000
4	0	6561	0.6561	4096	0.4096	2401	0.2401	1296	0.1296	0625	0.0625
	1	2916	0.9477	4096	0.8192	4116	0.6517	3456	0.4752	2500	0.3125
	2	0486	0.9963	1536	0.9728	2646	0.9163	3456	0.8208	3750	0.6875
	3	0036	0.9999	0256	0.9984	0756	0.9919	1536	0.9744	2500	0.9375
	4	0001	1.0000	0016	1.0000	0081	1.0000	0256	1.0000	0625	1.0000
5	0	5905	0.5905	3277	0.3277	1681	0.1681	0778	0.0778	0313	0.0313
	1	3281	0.9185	4096	0.7373	3602	0.5282	2592	0.3370	1563	0.1875
	2	0729	0.9914	2048	0.9421	3087	0.8369	3456	0.6826	3125	0.5000
	3	0081	0.9995	0512	0.9933	1323	0.9692	2304	0.9130	3125	0.8125
	4	0005	1.0000	0064	0.9997	0284	0.9976	0768	0.9898	1563	0.9688
	5	0000	1.0000	0003	1.0000	0024	1.0000	0102	1.0000	0313	1.0000
6	0	5314	0.5314	2621	0.2621	1176	0.1176	0467	0.0467	0156	0.0156
	1	3543	0.8857	3932	0.6554	3025	0.4202	1866	0.2333	0938	0.1094
	2	0984	0.9841	2458	0.9011	3241	0.7443	3110	0.5443	2344	0.3438
	3	0146	0.9987	0819	0.9830	1852	0.9295	2765	0.8208	3125	0.6563
	4	0012	0.9999	0154	0.9984	0595	0.9891	1382	0.9590	2344	0.8906
	5	0001	1.0000	0015	0.9999	0102	0.9993	0369	0.9959	0938	0.9844
	6	0000	1.0000	0001	1.0000	0007	1.0000	0041	1.0000	0156	1.0000
7	0	4783	0.4783	2097	0.2097	0824	0.0824	0280	0.0280	0078	0.0078
	1	3720	0.8503	3670	0.5767	2471	0.3294	1306	0.1586	0547	0.0625
	2	1240	0.9743	2753	0.8520	3177	0.6471	2613	0.4199	1641	0.2266
	3	0230	0.9973	1147	0.9667	2269	0.8740	2903	0.7102	2734	0.5000
	4	0026	0.9998	0287	0.9953	0972	0.9712	1935	0.9037	2734	0.7734
	5	0002	1.0000	0043	0.9996	0250	0.9962	0774	0.9812	1641	0.9375
	6	0000	1.0000	0004	1.0000	0036	0.9998	0172	0.9984	0547	0.9922
	7	0000	1.0000	0000	1.0000	0002	1.0000	0016	1.0000	0078	1.0000
8	0	4305	0.4305	1678	0.1678	0576	0.0576	0168	0.0168	0039	0.0039
	1	3826	0.8131	3355	0.5033	1977	0.2553	0896	0.1064	0313	0.0352
	2	1488	0.9619	2936	0.7969	2965	0.5518	2090	0.3154	1094	0.1445
	3	0331	0.9950	1468	0.9437	2541	0.8059	2787	0.5941	2188	0.3633
	4	0046	0.9996	0459	0.9896	1361	0.9420	2322	0.8263	2734	0.6367
	5	0004	1.0000	0092	0.9988	0467	0.9887	1239	0.9502	2188	0.8555
	6	0000	1.0000	0011	0.9999	0100	0.9987	0413	0.9915	1094	0.9648
	7	0000	1.0000	0001	1.0000	0012	0.9999	0079	0.9993	0313	0.9961
	8	0000	1.0000	0000	1.0000	0001	1.0000	0007	1.0000	0039	1.0000

Table A6 Poisson Distribution

Probability function $f(x)$ [see (5), Sec. 24.7] and distribution function $F(x)$

x	$\mu = 0.1$		$\mu = 0.2$		$\mu = 0.3$		$\mu = 0.4$		$\mu = 0.5$	
	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$
0	**0.**9048	0.9048	**0.**8187	0.8187	**0.**7408	0.7408	**0.**6703	0.6703	**0.**6065	0.6065
1	0905	0.9953	1637	0.9825	2222	0.9631	2681	0.9384	3033	0.9098
2	0045	0.9998	0164	0.9989	0333	0.9964	0536	0.9921	0758	0.9856
3	0002	1.0000	0011	0.9999	0033	0.9997	0072	0.9992	0126	0.9982
4	0000	1.0000	0001	1.0000	0003	1.0000	0007	0.9999	0016	0.9998
5							0001	1.0000	0002	1.0000

x	$\mu = 0.6$		$\mu = 0.7$		$\mu = 0.8$		$\mu = 0.9$		$\mu = 1$	
	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$
0	**0.**5488	0.5488	**0.**4966	0.4966	**0.**4493	0.4493	**0.**4066	0.4066	**0.**3679	0.3679
1	3293	0.8781	3476	0.8442	3595	0.8088	3659	0.7725	3679	0.7358
2	0988	0.9769	1217	0.9659	1438	0.9526	1647	0.9371	1839	0.9197
3	0198	0.9966	0284	0.9942	0383	0.9909	0494	0.9865	0613	0.9810
4	0030	0.9996	0050	0.9992	0077	0.9986	0111	0.9977	0153	0.9963
5	0004	1.0000	0007	0.9999	0012	0.9998	0020	0.9997	0031	0.9994
6			0001	1.0000	0002	1.0000	0003	1.0000	0005	0.9999
7									0001	1.0000

x	$\mu = 1.5$		$\mu = 2$		$\mu = 3$		$\mu = 4$		$\mu = 5$	
	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$
0	**0.**2231	0.2231	**0.**1353	0.1353	**0.**0498	0.0498	**0.**0183	0.0183	**0.**0067	0.0067
1	3347	0.5578	2707	0.4060	1494	0.1991	0733	0.0916	0337	0.0404
2	2510	0.8088	2707	0.6767	2240	0.4232	1465	0.2381	0842	0.1247
3	1255	0.9344	1804	0.8571	2240	0.6472	1954	0.4335	1404	0.2650
4	0471	0.9814	0902	0.9473	1680	0.8153	1954	0.6288	1755	0.4405
5	0141	0.9955	0361	0.9834	1008	0.9161	1563	0.7851	1755	0.6160
6	0035	0.9991	0120	0.9955	0504	0.9665	1042	0.8893	1462	0.7622
7	0008	0.9998	0034	0.9989	0216	0.9881	0595	0.9489	1044	0.8666
8	0001	1.0000	0009	0.9998	0081	0.9962	0298	0.9786	0653	0.9319
9			0002	1.0000	0027	0.9989	0132	0.9919	0363	0.9682
10					0008	0.9997	0053	0.9972	0181	0.9863
11					0002	0.9999	0019	0.9991	0082	0.9945
12					0001	1.0000	0006	0.9997	0034	0.9980
13							0002	0.9999	0013	0.9993
14							0001	1.0000	0005	0.9998
15									0002	0.9999
16									0000	1.0000

Table A7 Normal Distribution

Values of the distribution function $\Phi(z)$ [see (3), Sec. 24.8]. $\Phi(-z) = 1 - \Phi(z)$

z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$
	0.		**0.**		**0.**		**0.**		**0.**		**0.**
0.01	5040	0.51	6950	1.01	8438	1.51	9345	2.01	9778	2.51	9940
0.02	5080	0.52	6985	1.02	8461	1.52	9357	2.02	9783	2.52	9941
0.03	5120	0.53	7019	1.03	8485	1.53	9370	2.03	9788	2.53	9943
0.04	5160	0.54	7054	1.04	8508	1.54	9382	2.04	9793	2.54	9945
0.05	5199	0.55	7088	1.05	8531	1.55	9394	2.05	9798	2.55	9946
0.06	5239	0.56	7123	1.06	8554	1.56	9406	2.06	9803	2.56	9948
0.07	5279	0.57	7157	1.07	8577	1.57	9418	2.07	9808	2.57	9949
0.08	5319	0.58	7190	1.08	8599	1.58	9429	2.08	9812	2.58	9951
0.09	5359	0.59	7224	1.09	8621	1.59	9441	2.09	9817	2.59	9952
0.10	5398	0.60	7257	1.10	8643	1.60	9452	2.10	9821	2.60	9953
0.11	5438	0.61	7291	1.11	8665	1.61	9463	2.11	9826	2.61	9955
0.12	5478	0.62	7324	1.12	8686	1.62	9474	2.12	9830	2.62	9956
0.13	5517	0.63	7357	1.13	8708	1.63	9484	2.13	9834	2.63	9957
0.14	5557	0.64	7389	1.14	8729	1.64	9495	2.14	9838	2.64	9959
0.15	5596	0.65	7422	1.15	8749	1.65	9505	2.15	9842	2.65	9960
0.16	5636	0.66	7454	1.16	8770	1.66	9515	2.16	9846	2.66	9961
0.17	5675	0.67	7486	1.17	8790	1.67	9525	2.17	9850	2.67	9962
0.18	5714	0.68	7517	1.18	8810	1.68	9535	2.18	9854	2.68	9963
0.19	5753	0.69	7549	1.19	8830	1.69	9545	2.19	9857	2.69	9964
0.20	5793	0.70	7580	1.20	8849	1.70	9554	2.20	9861	2.70	9965
0.21	5832	0.71	7611	1.21	8869	1.71	9564	2.21	9864	2.71	9966
0.22	5871	0.72	7642	1.22	8888	1.72	9573	2.22	9868	2.72	9967
0.23	5910	0.73	7673	1.23	8907	1.73	9582	2.23	9871	2.73	9968
0.24	5948	0.74	7704	1.24	8925	1.74	9591	2.24	9875	2.74	9969
0.25	5987	0.75	7734	1.25	8944	1.75	9599	2.25	9878	2.75	9970
0.26	6026	0.76	7764	1.26	8962	1.76	9608	2.26	9881	2.76	9971
0.27	6064	0.77	7794	1.27	8980	1.77	9616	2.27	9884	2.77	9972
0.28	6103	0.78	7823	1.28	8997	1.78	9625	2.28	9887	2.78	9973
0.29	6141	0.79	7852	1.29	9015	1.79	9633	2.29	9890	2.79	9974
0.30	6179	0.80	7881	1.30	9032	1.80	9641	2.30	9893	2.80	9974
0.31	6217	0.81	7910	1.31	9049	1.81	9649	2.31	9896	2.81	9975
0.32	6255	0.82	7939	1.32	9066	1.82	9656	2.32	9898	2.82	9976
0.33	6293	0.83	7967	1.33	9082	1.83	9664	2.33	9901	2.83	9977
0.34	6331	0.84	7995	1.34	9099	1.84	9671	2.34	9904	2.84	9977
0.35	6368	0.85	8023	1.35	9115	1.85	9678	2.35	9906	2.85	9978
0.36	6406	0.86	8051	1.36	9131	1.86	9686	2.36	9909	2.86	9979
0.37	6443	0.87	8078	1.37	9147	1.87	9693	2.37	9911	2.87	9979
0.38	6480	0.88	8106	1.38	9162	1.88	9699	2.38	9913	2.88	9980
0.39	6517	0.89	8133	1.39	9177	1.89	9706	2.39	9916	2.89	9981
0.40	6554	0.90	8159	1.40	9192	1.90	9713	2.40	9918	2.90	9981
0.41	6591	0.91	8186	1.41	9207	1.91	9719	2.41	9920	2.91	9982
0.42	6628	0.92	8212	1.42	9222	1.92	9726	2.42	9922	2.92	9982
0.43	6664	0.93	8238	1.43	9236	1.93	9732	2.43	9925	2.93	9983
0.44	6700	0.94	8264	1.44	9251	1.94	9738	2.44	9927	2.94	9984
0.45	6736	0.95	8289	1.45	9265	1.95	9744	2.45	9929	2.95	9984
0.46	6772	0.96	8315	1.46	9279	1.96	9750	2.46	9931	2.96	9985
0.47	6808	0.97	8340	1.47	9292	1.97	9756	2.47	9932	2.97	9985
0.48	6844	0.98	8365	1.48	9306	1.98	9761	2.48	9934	2.98	9986
0.49	6879	0.99	8389	1.49	9319	1.99	9767	2.49	9936	2.99	9986
0.50	6915	1.00	8413	1.50	9332	2.00	9772	2.50	9938	3.00	9987

Table A8 Normal Distribution

Values of z for given values of $\Phi(z)$ [see (3), Sec. 24.8] and $D(z) = \Phi(z) - \Phi(-z)$
Example: $z = 0.279$ if $\Phi(z) = 61\%$; $z = 0.860$ if $D(z) = 61\%$.

%	$z(\Phi)$	$z(D)$	%	$z(\Phi)$	$z(D)$	%	$z(\Phi)$	$z(D)$
1	−2.326	0.013	41	−0.228	0.539	81	0.878	1.311
2	−2.054	0.025	42	−0.202	0.553	82	0.915	1.341
3	−1.881	0.038	43	−0.176	0.568	83	0.954	1.372
4	−1.751	0.050	44	−0.151	0.583	84	0.994	1.405
5	−1.645	0.063	45	−0.126	0.598	85	1.036	1.440
6	−1.555	0.075	46	−0.100	0.613	86	1.080	1.476
7	−1.476	0.088	47	−0.075	0.628	87	1.126	1.514
8	−1.405	0.100	48	−0.050	0.643	88	1.175	1.555
9	−1.341	0.113	49	−0.025	0.659	89	1.227	1.598
10	−1.282	0.126	50	0.000	0.674	90	1.282	1.645
11	−1.227	0.138	51	0.025	0.690	91	1.341	1.695
12	−1.175	0.151	52	0.050	0.706	92	1.405	1.751
13	−1.126	0.164	53	0.075	0.722	93	1.476	1.812
14	−1.080	0.176	54	0.100	0.739	94	1.555	1.881
15	−1.036	0.189	55	0.126	0.755	95	1.645	1.960
16	−0.994	0.202	56	0.151	0.772	96	1.751	2.054
17	−0.954	0.215	57	0.176	0.789	97	1.881	2.170
18	−0.915	0.228	58	0.202	0.806	97.5	1.960	2.241
19	−0.878	0.240	59	0.228	0.824	98	2.054	2.326
20	−0.842	0.253	60	0.253	0.842	99	2.326	2.576
21	−0.806	0.266	61	0.279	0.860	99.1	2.366	2.612
22	−0.772	0.279	62	0.305	0.878	99.2	2.409	2.652
23	−0.739	0.292	63	0.332	0.896	99.3	2.457	2.697
24	−0.706	0.305	64	0.358	0.915	99.4	2.512	2.748
25	−0.674	0.319	65	0.385	0.935	99.5	2.576	2.807
26	−0.643	0.332	66	0.412	0.954	99.6	2.652	2.878
27	−0.613	0.345	67	0.440	0.974	99.7	2.748	2.968
28	−0.583	0.358	68	0.468	0.994	99.8	2.878	3.090
29	−0.553	0.372	69	0.496	1.015	99.9	3.090	3.291
30	−0.524	0.385	70	0.524	1.036			
31	−0.496	0.399	71	0.553	1.058	99.91	3.121	3.320
32	−0.468	0.412	72	0.583	1.080	99.92	3.156	3.353
33	−0.440	0.426	73	0.613	1.103	99.93	3.195	3.390
34	−0.412	0.440	74	0.643	1.126	99.94	3.239	3.432
35	−0.385	0.454	75	0.674	1.150	99.95	3.291	3.481
36	−0.358	0.468	76	0.706	1.175	99.96	3.353	3.540
37	−0.332	0.482	77	0.739	1.200	99.97	3.432	3.615
38	−0.305	0.496	78	0.772	1.227	99.98	3.540	3.719
39	−0.279	0.510	79	0.806	1.254	99.99	3.719	3.891
40	−0.253	0.524	80	0.842	1.282			

Table A9 *t*-Distribution

Values of z for given values of the distribution function $F(z)$ (see (8) in Sec. 25.3). Example: For 9 degrees of freedom, $z = 1.83$ when $F(z) = 0.95$.

$F(z)$	Number of Degrees of Freedom									
	1	2	3	4	5	6	7	8	9	10
0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.6	0.32	0.29	0.28	0.27	0.27	0.26	0.26	0.26	0.26	0.26
0.7	0.73	0.62	0.58	0.57	0.56	0.55	0.55	0.55	0.54	0.54
0.8	1.38	1.06	0.98	0.94	0.92	0.91	0.90	0.89	0.88	0.88
0.9	3.08	1.89	1.64	1.53	1.48	1.44	1.41	1.40	1.38	1.37
0.95	6.31	2.92	2.35	2.13	2.02	1.94	1.89	1.86	1.83	1.81
0.975	12.7	4.30	3.18	2.78	2.57	2.45	2.36	2.31	2.26	2.23
0.99	31.8	6.96	4.54	3.75	3.36	3.14	3.00	2.90	2.82	2.76
0.995	63.7	9.92	5.84	4.60	4.03	3.71	3.50	3.36	3.25	3.17
0.999	318.3	22.3	10.2	7.17	5.89	5.21	4.79	4.50	4.30	4.14

$F(z)$	Number of Degrees of Freedom									
	11	12	13	14	15	16	17	18	19	20
0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.6	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26
0.7	0.54	0.54	0.54	0.54	0.54	0.54	0.53	0.53	0.53	0.53
0.8	0.88	0.87	0.87	0.87	0.87	0.86	0.86	0.86	0.86	0.86
0.9	1.36	1.36	1.35	1.35	1.34	1.34	1.33	1.33	1.33	1.33
0.95	1.80	1.78	1.77	1.76	1.75	1.75	1.74	1.73	1.73	1.72
0.975	2.20	2.18	2.16	2.14	2.13	2.12	2.11	2.10	2.09	2.09
0.99	2.72	2.68	2.65	2.62	2.60	2.58	2.57	2.55	2.54	2.53
0.995	3.11	3.05	3.01	2.98	2.95	2.92	2.90	2.88	2.86	2.85
0.999	4.02	3.93	3.85	3.79	3.73	3.69	3.65	3.61	3.58	3.55

$F(z)$	Number of Degrees of Freedom									
	22	24	26	28	30	40	50	100	200	∞
0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.6	0.26	0.26	0.26	0.26	0.26	0.26	0.25	0.25	0.25	0.25
0.7	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.52
0.8	0.86	0.86	0.86	0.85	0.85	0.85	0.85	0.85	0.84	0.84
0.9	1.32	1.32	1.31	1.31	1.31	1.30	1.30	1.29	1.29	1.28
0.95	1.72	1.71	1.71	1.70	1.70	1.68	1.68	1.66	1.65	1.65
0.975	2.07	2.06	2.06	2.05	2.04	2.02	2.01	1.98	1.97	1.96
0.99	2.51	2.49	2.48	2.47	2.46	2.42	2.40	2.36	2.35	2.33
0.995	2.82	2.80	2.78	2.76	2.75	2.70	2.68	2.63	2.60	2.58
0.999	3.50	3.47	3.43	3.41	3.39	3.31	3.26	3.17	3.13	3.09

Table A10 Chi-square Distribution

Values of x for given values of the distribution function $F(z)$ (see Sec. 25.3 before (17)). Example: For 3 degrees of freedom, $z = 11.34$ when $F(z) = 0.99$.

$F(z)$	Number of Degrees of Freedom									
	1	2	3	4	5	6	7	8	9	10
0.005	0.00	0.01	0.07	0.21	0.41	0.68	0.99	1.34	1.73	2.16
0.01	0.00	0.02	0.11	0.30	0.55	0.87	1.24	1.65	2.09	2.56
0.025	0.00	0.05	0.22	0.48	0.83	1.24	1.69	2.18	2.70	3.25
0.05	0.00	0.10	0.35	0.71	1.15	1.64	2.17	2.73	3.33	3.94
0.95	3.84	5.99	7.81	9.49	11.07	12.59	14.07	15.51	16.92	18.31
0.975	5.02	7.38	9.35	11.14	12.83	14.45	16.01	17.53	19.02	20.48
0.99	6.63	9.21	11.34	13.28	15.09	16.81	18.48	20.09	21.67	23.21
0.995	7.88	10.60	12.84	14.86	16.75	18.55	20.28	21.95	23.59	25.19

$F(z)$	Number of Degrees of Freedom									
	11	12	13	14	15	16	17	18	19	20
0.005	2.60	3.07	3.57	4.07	4.60	5.14	5.70	6.26	6.84	7.43
0.01	3.05	3.57	4.11	4.66	5.23	5.81	6.41	7.01	7.63	8.26
0.025	3.82	4.40	5.01	5.63	6.26	6.91	7.56	8.23	8.91	9.59
0.05	4.57	5.23	5.89	6.57	7.26	7.96	8.67	9.39	10.12	10.85
0.95	19.68	21.03	22.36	23.68	25.00	26.30	27.59	28.87	30.14	31.41
0.975	21.92	23.34	24.74	26.12	27.49	28.85	30.19	31.53	32.85	34.17
0.99	24.72	26.22	27.69	29.14	30.58	32.00	33.41	34.81	36.19	37.57
0.995	26.76	28.30	29.82	31.32	32.80	34.27	35.72	37.16	38.58	40.00

$F(z)$	Number of Degrees of Freedom									
	21	22	23	24	25	26	27	28	29	30
0.005	8.0	8.6	9.3	9.9	10.5	11.2	11.8	12.5	13.1	13.8
0.01	8.9	9.5	10.2	10.9	11.5	12.2	12.9	13.6	14.3	15.0
0.025	10.3	11.0	11.7	12.4	13.1	13.8	14.6	15.3	16.0	16.8
0.05	11.6	12.3	13.1	13.8	14.6	15.4	16.2	16.9	17.7	18.5
0.95	32.7	33.9	35.2	36.4	37.7	38.9	40.1	41.3	42.6	43.8
0.975	35.5	36.8	38.1	39.4	40.6	41.9	43.2	44.5	45.7	47.0
0.99	38.9	40.3	41.6	43.0	44.3	45.6	47.0	48.3	49.6	50.9
0.995	41.4	42.8	44.2	45.6	46.9	48.3	49.6	51.0	52.3	53.7

$F(z)$	Number of Degrees of Freedom							
	40	50	60	70	80	90	100	> 100 (Approximation)
0.005	20.7	28.0	35.5	43.3	51.2	59.2	67.3	$\frac{1}{2}(h - 2.58)^2$
0.01	22.2	29.7	37.5	45.4	53.5	61.8	70.1	$\frac{1}{2}(h - 2.33)^2$
0.025	24.4	32.4	40.5	48.8	57.2	65.6	74.2	$\frac{1}{2}(h - 1.96)^2$
0.05	26.5	34.8	43.2	51.7	60.4	69.1	77.9	$\frac{1}{2}(h - 1.64)^2$
0.95	55.8	67.5	79.1	90.5	101.9	113.1	124.3	$\frac{1}{2}(h + 1.64)^2$
0.975	59.3	71.4	83.3	95.0	106.6	118.1	129.6	$\frac{1}{2}(h + 1.96)^2$
0.99	63.7	76.2	88.4	100.4	112.3	124.1	135.8	$\frac{1}{2}(h + 2.33)^2$
0.995	66.8	79.5	92.0	104.2	116.3	128.3	140.2	$\frac{1}{2}(h + 2.58)^2$

In the last column, $h = \sqrt{2m - 1}$, where m is the number of degrees of freedom.

Table A11 F-Distribution with (m, n) Degrees of Freedom

Values of z for which the distribution function $F(z)$ [see (13), Sec. 25.4] has the value **0.95**

Example: For (7, 4) d.f., $z = 6.09$ if $F(z) = 0.95$.

n	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m = 6$	$m = 7$	$m = 8$	$m = 9$
1	161	200	216	225	230	234	237	239	241
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
32	4.15	3.29	2.90	2.67	2.51	2.40	2.31	2.24	2.19
34	4.13	3.28	2.88	2.65	2.49	2.38	2.29	2.23	2.17
36	4.11	3.26	2.87	2.63	2.48	2.36	2.28	2.21	2.15
38	4.10	3.24	2.85	2.62	2.46	2.35	2.26	2.19	2.14
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97
150	3.90	3.06	2.66	2.43	2.27	2.16	2.07	2.00	1.94
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93
1000	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88

Table A11 F-Distribution with (m, n) Degrees of Freedom *(continued)*

Values of z for which the distribution function $F(z)$ [see (13), Sec. 25.4] has the value **0.95**

n	$m = 10$	$m = 15$	$m = 20$	$m = 30$	$m = 40$	$m = 50$	$m = 100$	∞
1	242	246	248	250	251	252	253	254
2	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5
3	8.79	8.70	8.66	8.62	8.59	8.58	8.55	8.53
4	5.96	5.86	5.80	5.75	5.72	5.70	5.66	5.63
5	4.74	4.62	4.56	4.50	4.46	4.44	4.41	4.37
6	4.06	3.94	3.87	3.81	3.77	3.75	3.71	3.67
7	3.64	3.51	3.44	3.38	3.34	3.32	3.27	3.23
8	3.35	3.22	3.15	3.08	3.04	3.02	2.97	2.93
9	3.14	3.01	2.94	2.86	2.83	2.80	2.76	2.71
10	2.98	2.85	2.77	2.70	2.66	2.64	2.59	2.54
11	2.85	2.72	2.65	2.57	2.53	2.51	2.46	2.40
12	2.75	2.62	2.54	2.47	2.43	2.40	2.35	2.30
13	2.67	2.53	2.46	2.38	2.34	2.31	2.26	2.21
14	2.60	2.46	2.39	2.31	2.27	2.24	2.19	2.13
15	2.54	2.40	2.33	2.25	2.20	2.18	2.12	2.07
16	2.49	2.35	2.28	2.19	2.15	2.12	2.07	2.01
17	2.45	2.31	2.23	2.15	2.10	2.08	2.02	1.96
18	2.41	2.27	2.19	2.11	2.06	2.04	1.98	1.92
19	2.38	2.23	2.16	2.07	2.03	2.00	1.94	1.88
20	2.35	2.20	2.12	2.04	1.99	1.97	1.91	1.84
22	2.30	2.15	2.07	1.98	1.94	1.91	1.85	1.78
24	2.25	2.11	2.03	1.94	1.89	1.86	1.80	1.73
26	2.22	2.07	1.99	1.90	1.85	1.82	1.76	1.69
28	2.19	2.04	1.96	1.87	1.82	1.79	1.73	1.65
30	2.16	2.01	1.93	1.84	1.79	1.76	1.70	1.62
32	2.14	1.99	1.91	1.82	1.77	1.74	1.67	1.59
34	2.12	1.97	1.89	1.80	1.75	1.71	1.65	1.57
36	2.11	1.95	1.87	1.78	1.73	1.69	1.62	1.55
38	2.09	1.94	1.85	1.76	1.71	1.68	1.61	1.53
40	2.08	1.92	1.84	1.74	1.69	1.66	1.59	1.51
50	2.03	1.87	1.78	1.69	1.63	1.60	1.52	1.44
60	1.99	1.84	1.75	1.65	1.59	1.56	1.48	1.39
70	1.97	1.81	1.72	1.62	1.57	1.53	1.45	1.35
80	1.95	1.79	1.70	1.60	1.54	1.51	1.43	1.32
90	1.94	1.78	1.69	1.59	1.53	1.49	1.41	1.30
100	1.93	1.77	1.68	1.57	1.52	1.48	1.39	1.28
150	1.89	1.73	1.64	1.54	1.48	1.44	1.34	1.22
200	1.88	1.72	1.62	1.52	1.46	1.41	1.32	1.19
1000	1.84	1.68	1.58	1.47	1.41	1.36	1.26	1.08
∞	1.83	1.67	1.57	1.46	1.39	1.35	1.24	1.00

Table A11 F-Distribution with (m, n) Degrees of Freedom *(continued)*

Values of z for which the distribution function $F(z)$ [see (13), Sec. 25.4] has the value **0.99**

n	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m = 6$	$m = 7$	$m = 8$	$m = 9$
1	4052	4999	5403	5625	5764	5859	5928	5981	6022
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07
32	7.50	5.34	4.46	3.97	3.65	3.43	3.26	3.13	3.02
34	7.44	5.29	4.42	3.93	3.61	3.39	3.22	3.09	2.98
36	7.40	5.25	4.38	3.89	3.57	3.35	3.18	3.05	2.95
38	7.35	5.21	4.34	3.86	3.54	3.32	3.15	3.02	2.92
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72
70	7.01	4.92	4.07	3.60	3.29	3.07	2.91	2.78	2.67
80	6.96	4.88	4.04	3.56	3.26	3.04	2.87	2.74	2.64
90	6.93	4.85	4.01	3.54	3.23	3.01	2.84	2.72	2.61
100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59
150	6.81	4.75	3.91	3.45	3.14	2.92	2.76	2.63	2.53
200	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50
1000	6.66	4.63	3.80	3.34	3.04	2.82	2.66	2.53	2.43
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41

Table A11 F-Distribution with (m, n) Degrees of Freedom *(continued)*

Values of z for which the distribution function $F(z)$ [see (13), Sec. 25.4] has the value **0.99**

n	$m = 10$	$m = 15$	$m = 20$	$m = 30$	$m = 40$	$m = 50$	$m = 100$	∞
1	6056	6157	6209	6261	6287	6303	6334	6366
2	99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5
3	27.2	26.9	26.7	26.5	26.4	26.4	26.2	26.1
4	14.5	14.2	14.0	13.8	13.7	13.7	13.6	13.5
5	10.1	9.72	9.55	9.38	9.29	9.24	9.13	9.02
6	7.87	7.56	7.40	7.23	7.14	7.09	6.99	6.88
7	6.62	6.31	6.16	5.99	5.91	5.86	5.75	5.65
8	5.81	5.52	5.36	5.20	5.12	5.07	4.96	4.86
9	5.26	4.96	4.81	4.65	4.57	4.52	4.42	4.31
10	4.85	4.56	4.41	4.25	4.17	4.12	4.01	3.91
11	4.54	4.25	4.10	3.94	3.86	3.81	3.71	3.60
12	4.30	4.01	3.86	3.70	3.62	3.57	3.47	3.36
13	4.10	3.82	3.66	3.51	3.43	3.38	3.27	3.17
14	3.94	3.66	3.51	3.35	3.27	3.22	3.11	3.00
15	3.80	3.52	3.37	3.21	3.13	3.08	2.98	2.87
16	3.69	3.41	3.26	3.10	3.02	2.97	2.86	2.75
17	3.59	3.31	3.16	3.00	2.92	2.87	2.76	2.65
18	3.51	3.23	3.08	2.92	2.84	2.78	2.68	2.57
19	3.43	3.15	3.00	2.84	2.76	2.71	2.60	2.49
20	3.37	3.09	2.94	2.78	2.69	2.64	2.54	2.42
22	3.26	2.98	2.83	2.67	2.58	2.53	2.42	2.31
24	3.17	2.89	2.74	2.58	2.49	2.44	2.33	2.21
26	3.09	2.81	2.66	2.50	2.42	2.36	2.25	2.13
28	3.03	2.75	2.60	2.44	2.35	2.30	2.19	2.06
30	2.98	2.70	2.55	2.39	2.30	2.25	2.13	2.01
32	2.93	2.65	2.50	2.34	2.25	2.20	2.08	1.96
34	2.89	2.61	2.46	2.30	2.21	2.16	2.04	1.91
36	2.86	2.58	2.43	2.26	2.18	2.12	2.00	1.87
38	2.83	2.55	2.40	2.23	2.14	2.09	1.97	1.84
40	2.80	2.52	2.37	2.20	2.11	2.06	1.94	1.80
50	2.70	2.42	2.27	2.10	2.01	1.95	1.82	1.68
60	2.63	2.35	2.20	2.03	1.94	1.88	1.75	1.60
70	2.59	2.31	2.15	1.98	1.89	1.83	1.70	1.54
80	2.55	2.27	2.12	1.94	1.85	1.79	1.65	1.49
90	2.52	2.24	2.09	1.92	1.82	1.76	1.62	1.46
100	2.50	2.22	2.07	1.89	1.80	1.74	1.60	1.43
150	2.44	2.16	2.00	1.83	1.73	1.66	1.52	1.33
200	2.41	2.13	1.97	1.79	1.69	1.63	1.48	1.28
1000	2.34	2.06	1.90	1.72	1.61	1.54	1.38	1.11
∞	2.32	2.04	1.88	1.70	1.59	1.52	1.36	1.00

Table A12 Distribution Function $F(x) = P(T \leqq x)$ of the Random Variable T in Section 25.8

x	$n=3$
	0.
0	167
1	500

x	$n=4$
	0.
0	042
1	167
2	375

x	$n=5$
	0.
0	008
1	042
2	117
3	242
4	408

x	$n=6$
	0.
0	001
1	008
2	028
3	068
4	136
5	235
6	360
7	500

x	$n=7$
	0.
1	001
2	005
3	015
4	035
5	068
6	119
7	191
8	281
9	386
10	500

x	$n=8$
	0.
2	001
3	003
4	007
5	016
6	031
7	054
8	089
9	138
10	199
11	274
12	360
13	452

x	$n=9$
	0.
4	001
5	003
6	006
7	012
8	022
9	038
10	060
11	090
12	130
13	179
14	238
15	306
16	381
17	460

x	$n=10$
	0.
6	001
7	002
8	005
9	008
10	014
11	023
12	036
13	054
14	078
15	108
16	146
17	190
18	242
19	300
20	364
21	431
22	500

x	$n=11$
	0.
8	001
9	002
10	003
11	005
12	008
13	013
14	020
15	030
16	043
17	060
18	082
19	109
20	141
21	179
22	223
23	271
24	324
25	381
26	440
27	500

x	$n=20$
	0.
50	001
51	002
52	002
53	003
54	004
55	005
56	006
57	007
58	008
59	010
60	012
61	014
62	017
63	020
64	023
65	027
66	032
67	037
68	043
69	049
70	056
71	064
72	073
73	082
74	093
75	104
76	117
77	130
78	144
79	159
80	176
81	193
82	211
83	230
84	250
85	271
86	293
87	315
88	339
89	362
90	387
91	411
92	436
93	462
94	487

x	$n=19$
	0.
43	001
44	002
45	002
46	003
47	003
48	004
49	005
50	006
51	008
52	010
53	012
54	014
55	017
56	021
57	025
58	029
59	034
60	040
61	047
62	054
63	062
64	072
65	082
66	093
67	105
68	119
69	133
70	149
71	166
72	184
73	203
74	223
75	245
76	267
77	290
78	314
79	339
80	365
81	391
82	418
83	445
84	473
85	500

x	$n=18$
	0.
38	001
39	002
40	003
41	003
42	004
43	005
44	007
45	009
46	011
47	013
48	016
49	020
50	024
51	029
52	034
53	041
54	048
55	056
56	066
57	076
58	088
59	100
60	115
61	130
62	147
63	165
64	184
65	205
66	227
67	250
68	275
69	300
70	327
71	354
72	383
73	411
74	441
75	470
76	500

x	$n=17$
	0.
32	001
33	002
34	002
35	003
36	004
37	005
38	007
39	009
40	011
41	014
42	017
43	021
44	026
45	032
46	038
47	046
48	054
49	064
50	076
51	088
52	102
53	118
54	135
55	154
56	174
57	196
58	220
59	245
60	271
61	299
62	328
63	358
64	388
65	420
66	452
67	484

x	$n=16$
	0.
27	001
28	002
29	002
30	003
31	004
32	006
33	008
34	010
35	013
36	016
37	021
38	026
39	032
40	039
41	048
42	058
43	070
44	083
45	097
46	114
47	133
48	153
49	175
50	199
51	225
52	253
53	282
54	313
55	345
56	378
57	412
58	447
59	482

x	$n=15$
	0.
23	001
24	002
25	003
26	004
27	006
28	008
29	010
30	014
31	018
32	023
33	029
34	037
35	046
36	057
37	070
38	084
39	101
40	120
41	141
42	164
43	190
44	218
45	248
46	279
47	313
48	349
49	385
50	423
51	461
52	500

x	$n=14$
	0.
18	001
19	002
20	002
21	003
22	005
23	007
24	010
25	013
26	018
27	024
28	031
29	040
30	051
31	063
32	079
33	096
34	117
35	140
36	165
37	194
38	225
39	259
40	295
41	334
42	374
43	415
44	457
45	500

x	$n=13$
	0.
14	001
15	001
16	002
17	003
18	005
19	007
20	011
21	015
22	021
23	029
24	038
25	050
26	064
27	082
28	102
29	126
30	153
31	184
32	218
33	255
34	295
35	338
36	383
37	429
38	476

x	$n=12$
	0.
11	001
12	002
13	003
14	004
15	007
16	010
17	016
18	022
19	031
20	043
21	058
22	076
23	098
24	125
25	155
26	190
27	230
28	273
29	319
30	369
31	420
32	473

INDEX

Page numbers A1, A2, A3, · · · refer to App. 1 to App. 5 at the end of the book.

A

Abel 78
Absolute
  convergence 667, 697
  frequency 994, 1000
  value 607
Absolutely integrable 508
Acceleration 395, 995
Acceptance sampling 1073
Adams–Bashforth methods 899
Adams–Moulton methods 900
Adaptive 824
Addition of
  complex numbers 603
  matrices 275
  means 1038
  normal random variables 1050
  power series 174, 680
  variances 1039
  vectors 276, 324, 367
Addition rule 1002
Adiabatic 561
ADI method 915
Adjacency matrix 956
Adjacent vertices 955
Airfoil 732
Airy's equation 552, 904
Algebraic multiplicity 337, 865
Algorithm 777, 783
  Dijkstra 964
  efficient 962
  Ford–Fulkerson 979
  Gauss 837
  Gauss-Seidel 848
  Greedy 967
  Kruskal 967
  Moore 960
  polynomially bounded 962
  Prim 971
  Runge-Kutta 892, 904
  stable, unstable 783
Aliasing 526
Allowable number of defectives 1073
Alternating
  direction implicit method 915
  path 983
Alternative hypothesis 1058
Ampère 92
Amplification 89
Amplitude spectrum 506
Analytic function 175, 617, 681
Analytic at infinity 711
Angle
  between curves 35
  between vectors 372
Angular speed 381, 765
Annulus 613
Anticommutative 379
AOQ, AOQL 1075–1076
Approximate solution of
  differential equations 9, 886–934
  eigenvalue problems 863–882
  equations 787–796
  systems of equations 833–858
Approximation
  least squares 860
  polynomial 797
  trigonometric 502
A priori estimate 794
AQL 1074
Arc of a curve 391
Archimedian principle 68
Arc length 393
Arctan 634
Area 435, 442, 454
Argand diagram 605
Argument 607
Artificial variable 949
Assignment problem 982
Associated Legendre functions 182
Asymptotically
  equal 191, 1009
  normal 1057
  stable 148

Attractive 148
Augmented matrix 288, 833
Augmenting path 975, 983
  theorem 977, 984
Autonomous 31, 151
Average (*see* Mean value)
Average outgoing quality 1075
Axioms of probability 1001

B

Back substitution 289, 834
Backward
  differences 807
  edge 974, 976
  Euler method 896, 907
Band matrix 914
Bashforth method 899
Basic
  feasible solution 942, 944
  variables 945
Basis 49, 106, 113, 138, 300, 325, 360
Beam 120, 547
Beats 87
Bellman optimality principle 963
Bell-shaped curve 1026
Bernoulli 30
  distribution 1020
  equation 30
  law of large numbers 1032
  numbers 690
Bessel 189
  equation 189, 204
  functions 191, 198, 202, 207, A94
  functions, tables A94
  inequality 215, 504
Beta function A64
Bezier curves 816
BFS 960
Bijective mapping 729
Binary 782
Binomial
  coefficients 1009
  distribution 1020, A96
  series 689
  theorem 1010
Binormal (Fig. 210) 397
Bipartite graph 982, 985
Birthday problem 1010
Bisection method 796
Bolzano–Weierstrass theorem A91
Bonnet 181
Boundary 433
  conditions 203, 540, 558, 571, 587
  point 433, 613
  value problem 203, 558
Bounded
  domain 646
  function 38
  region 433
  sequence A69
Boxplot 995
Branch
  cut 632
  point 746
Breadth first search 960
Buoyance force 68

C

Cable 52, 198, 593
CAD (Computer aided design) 810
Cancellation law 321
Cantor–Dedekind axiom A69
Capacitance 92
Capacitor 92
Capacity
  of a cut set 976
  of an edge 973
Cardano 602
Cardioid 443
Cartesian coordinates 366, 604
CAS (Computer algebra system) vii, 777
Catenary 399
Cauchy 69
  convergence principle 667, A90
  determinant 112
  –Goursat theorem 647
  –Hadamard formula 676
  inequality 660
  integral formula 654
  integral theorem 647, 652
  method of steepest descent 938
  principal value 719, 722
  product 680
  –Riemann equations 37, 618, 621
  –Schwarz inequality 326, 859
Cayley transformation 739

Center 143
of a graph 973
of gravity 436, 457
of a power series 171
Central
differences 808
limit theorem 1057
moments 1019
Centrifugal, centripetal 396
Cgs system: Front cover
Chain rule 401
Characteristic
determinant 336, 864
equation 59, 111, 336, 551, 864
function 542, 574
of a partial differential equation 551
polynomial 336, 864
value 324, 864
vector 324, 864
Chebyshev polynomials 209
Chinese postman problem 963
Chi-square 1055, 1077, A101
Cholesky's method 843
Chopping 782
Chromatic number 987
Circle 391
of convergence 675
Circuit 95
Circular
disk 613
helix 391, 394
membrane 580
Circulation 764
Cissoid 399
Clairaut equation 34
Class intervals 994
Closed
disk 613
integration formula 822, 827
interval A69
path 959
point set 613
region 433
Coefficient matrix 288, 833
Coefficients of a
differential equation 46
power series 171
system of equations 287, 833
Cofactor 309
Collatz's theorem 870
Column 275
space 300
sum norm 849
vector 275
Combination 1007
Combinatorial optimization 954–986
Comparison test 668
Complement 613, 988
Complementary
error function A64
Fresnel integrals A65
sine integral A66
Complementation rule 1002
Complete
bipartite graph 987
graph 958
matching 983
orthonormal set 214
Complex
conjugate numbers 605
exponential function 57, 623
Fourier integral 519
Fourier series 497
function 614
hyperbolic functions 628, 743
impedance 98
indefinite integral 637
integration 637–663, 712–725
line integral 633
logarithm 630, 688
matrices 356
number 602
plane 605
plane, extended 710
potential 761
sequence 664
series 666
sphere 710
trigonometric functions 626, 688
trigonometric polynomial 524
variable 614
vector space 324, 359
Complexity 961
Component 366, 374
Composite transformation 281
Compound interest 9
Compressible fluid 412
Computer
aided design (CAD) 810
algebra system (CAS) vii, 777

Computer (*Cont.*)
graphics 287
software (*see* Software)
Conchoid 399
Condition number 855
Conditionally convergent 667
Conditional probability 1003
Conduction of heat 465, 552, 757
Cone 406, 448
CONF 1049
Confidence
intervals 1049–1058
level 1049
limits 1049
Conformal mapping 730, 754
Conic sections 355
Conjugate
complex numbers 605
harmonic function 622
Connected
graph 960
set 613
Conservative 415, 428
Consistent equations 292, 303
Constraints 937
Consumer's risk 1075
Continuity
of a complex function 615
equation 413
of a vector function 387
Continuous
distribution 1011
random variable 1011, 1034
Contour integral 647
Contraction 789
Control
chart 1068
limit 1068
variables 936
Convergence
absolute 667
circle of 675
conditional 667
interval 172, 676
of an iterative process 793, 848
mean 214
mean-square 214
in norm 214
principle 667
radius 172
Convergence (*Cont.*)
of a sequence 386, 664
of a series 171, 666
superlinear 795
tests 667–672
uniform 691
Conversion of an ODE to a system 134
Convolution 248, 523
Cooling 14
Coordinate transformations A71, A84
Coordinates
Cartesian 366, 604
curvilinear A71
cylindrical 587, A71
polar 137, 443, 580, 607
spherical 588, A71
Coriolis acceleration 396
Corrector 890, 900
Correlation analysis 1089
Cosecant 627, A62
Cosine
of a complex variable 627, 688, 743
hyperbolic 688
integral A66, A95
of a real variable A60
Cotangent 627, A62
Coulomb 92
law 409
Covariance 1039, 1085
Cramer's rule 306–307, 312
Crank–Nicolson method 924
Critical
damping 65
point 31, 142, 730
region 1060
Cross product 377
Crout's method 841
Cubic spline 811
Cumulative
distribution function 1011
frequency 994
Curl 414, 430, 472, A71
Curvature 397
Curve 389
arc length of 393
fitting 859
orientation of 390
piecewise smooth 421
rectifiable 393
simple 391

Curve (*Cont.*)
smooth 421, 638
twisted 391
Curvilinear coordinates A71
Cut set 976
Cycle 959
Cycloid 399
Cylinder 446
flow around 763, 767
Cylindrical coordinates 587, A71

D

D'Alembert's solution 549, 551
Damping 64, 88
Dantzig 944
DATA DESK 991
Decay 5
Decreasing sequence A69
Decrement 69
Dedekind A69
Defect 337
Defective item 1073
Definite complex integral 639
Definiteness 356
Deformation of path 649
Degenerate feasible solution 947
Degree of
precision 822
a vertex 955
Degrees of freedom 1052, 1055, 1066
Deleted neighborhood 712
Delta
Dirac 242
Kronecker 210, A83
De Moivre 610
formula 610
limit theorem 1031
De Morgan's laws 999
Density 1014
Dependent
linearly 49, 74, 106, 297, 300, 325
random variables 1036
Depth first search 960
Derivative
of a complex function 616, 658
directional 404
left-hand 484
right-hand 484
of a vector function 387
DERIVE 778
Descartes 366
Determinant 306, 308
Cauchy 112
characteristic 336, 864
of a matrix 305
of a matrix product 321
Vandermonde 112
DFS 960
Diagonalization 351
Diagonal matrix 284
Diagonally dominant matrix 868
Diameter of a graph 973
Differences 802, 804, 807–808
Difference table 803
Differentiable complex function 616
Differential 19, 429
form 20, 429
geometry 389
operator 59
Differential equation (ODE and PDE)
Airy 552, 904
Bernoulli 30
Bessel 189, 204
Cauchy–Riemann 37, 618, 621
with constant coefficients 53, 111
elliptic 551, 909
Euler–Cauchy 69, 185
exact 20
homogeneous 27, 46, 105, 535
hyperbolic 551, 928
hypergeometric 188
Laguerre 257
Laplace 407, 465, 536, 579, 587, 910
Legendre 177, 204, 590
linear 26, 45, 105, 535
nonhomogeneous 27, 46, 105, 535
nonlinear 45, 151, 535
numeric methods for 886–934
ordinary 4
parabolic 551, 909, 922
partial 535
Poisson 910, 918
separable 12
Sturm–Liouville 203
of vibrating beam 547, 552
of vibrating mass 61, 86, 135, 150, 243, 261, 342, 499
of vibrating membrane 569–586
of vibrating string 538

Differentiation
analytic functions 691
complex functions 616
Laplace transforms 254
numeric 827
power series 174, 680
series 696
vector functions 387
Diffusion equation 464, 552
Diffusivity 552
Digraph 955
Dijkstra's algorithm 964
Dimension of vector space 300, 325, 369
Diocles 399
Dirac's delta 242
Directed
graph 955
line segment 364
path 982
Directional derivative 404
Direction field 10
Direct method 845
Dirichlet 467
discontinuous factor 509
problem 467, 558, 587, 915
Discharge of a source 767
Discrete
Fourier transform 525
random variable 1011, 1033
spectrum 507, 524
Disjoint events 998
Disk 613
Dissipative 429
Distribution 1010
Bernoulli 1020
binomial 1020, A96
chi-square 1055, A101
continuous 1011
discrete 1011, 1033
Fisher's *F*- 1066, A102
-free test 1080
function 1011, 1032
Gauss 1026
hypergeometric 1024
marginal 1035
multinomial 1025
normal 1026, 1047–1057, 1062–1067, A98
Poisson 1022, 1073, A97
Student's *t*- 1053, A100
two-dimensional 1032
Distribution (*Cont.*)
uniform 1015, 1017, 1034
Divergence
theorem of Gauss 459
of vector fields 410, A72
Divergent
sequence 665
series 171, 667
Divided differences 802
Division of complex numbers 604, 609
Domain 401, 613, 646
Doolittle's method 841
Dot product 325, 371
Double
Fourier series 576
integral 433
labeling 968
precision 782
Driving force 84
Drumhead 569
Duffing equation 159
Duhamel's formula 597

E

Eccentricity of a vertex 973
Echelon form 294
Edge 955
coloring 987
incidence list 957
Efficient algorithm 962
Eigenbasis 349
Eigenfunction 204, 542, 559, 574
expansion 210
Eigenspace 336, 865
Eigenvalue problems for
matrices (Chap. 8) 333–363
matrices, numerics 863–882
ODEs (Sturm-Liouville problems) 203–216
PDEs 540–593
systems of ODEs 130–165
Eigenvector 334, 864
EISPACK 778
Elastic membrane 340
Electrical network (*see* Networks)
Electric circuit (*see* Circuit)
Electromechanical analogies 96
Electromotive force 91
Electrostatic
field 750

Electrostatic (*Cont.*)
potential 588–592, 750
Element of matrix 273
Elementary
matrix 296
operations 292
Elimination of first derivative 197
Ellipse 390
Ellipsoid 449
Elliptic
cylinder 448
paraboloid 448
partial differential equation 551, 909
Empty set 998
Engineering system: Front cover
Entire function 624, 661, 711
Entry 273, 309
Equality of
complex numbers 602
matrices 275
vectors 365
Equally likely 1000
Equipotential
lines 750, 762
surfaces 750
Equivalence relation 296
Equivalent linear systems 292
Erf 568, 690, A64, A95
Error 783
bound 784
estimation 785
function 568, 690, A64, A95
propagation 784
Type I, Type II 1060
Essential singularity 708
Estimation of parameters 1046–1057
Euclidean
norm 327
space 327
Euler 69
backward methods for ODEs 896, 907
beta function A64
–Cauchy equation 69, 108, 116, 185, 589
–Cauchy method 887, 890, 903
constant 200
formula 58, 496, 624, 627, 687
formulas for Fourier coefficients 480, 487
graph 963
numbers 690
method for systems 903
Euler (*Cont.*)
trail 963
Evaporation 18
Even function 490
Event 997
Everett interpolation formula 809
Exact
differential equation 20
differential form 20, 429
Existence theorem
differential equations 37, 73, 107, 109, 137, 175
Fourier integral 508
Fourier series 484
Laplace transforms 226
Linear equations 302
Expectation 1016
Experiment 997
Explicit solution 4
Exponential
decay 5
function, complex 57, 623
function, real A60
growth 5, 31
integral A66
Exposed vertex 983
Extended complex plane 710, 736
Extension, periodic 494
Extrapolation 797
Extremum 937

F

Factorial function 192, 1008, A95
Failure 1021
Fair die 1000
Falling body 8
False position 796
Family of curves 5, 35
Faraday 92
Fast Fourier transform 526
F-distribution 1066, A102
Feasible solution 942
Fehlberg 894
Fibonacci numbers 683
Field
conservative 415, 428
of force 385
gravitational 385, 407, 411, 587

Field (*Cont.*)
irrotational 415, 765
scalar 384
vector 384
velocity 385
Finite complex plane 710
First
fundamental form 457
Green's formula 466
shifting theorem 224
Fisher, R. A. 1047, 1066, 1077
F-distribution 1066, A102
Fixed
decimal point 781
point 736, 781, 787
Flat spring 68
Floating point 781
Flow augmenting path 975
Flows in networks 973–981
Fluid flow 412, 463, 761
Flux 412, 450
integral 450
Folium of Descartes 399
Forced oscillations 84, 499
Ford–Fulkerson algorithm 979
Forest 970
Form
Hermitian 361
quadratic 353
skew-Hermitian 361
Forward
differences 804
edge 974, 976
Four-color theorem 987
Fourier 477
–Bessel series 213, 583
coefficients 480, 487
coefficients, complex 497
constants 210
cosine integral 511
cosine series 491
cosine transform 514, 529
double series 576
half-range expansions 494
integral 508, 563
integral, complex 519
-Legendre series 212, 590
matrix 525
series 211, 480, 487
series, complex 497
Fourier (*Cont.*)
series, generalized 210
sine integral 511
sine series 491, 543
sine transform 514, 530
transform 519, 531, 565
transform, discrete 525
transform, fast 526
Fractional linear transformation 734
Fraction defective 1073
Fredholm 201
Free
fall 18
oscillations 61
Frenet formulas 400
Frequency 63
of values in samples 994
Fresnel integrals 690, A65
Friction 18–19
Frobenius 182
method 182
norm 849
theorem 869
Fulkerson 979
Full-wave rectifier 248
Function
analytic 175, 617
Bessel 191, 198, 202, 207, A94
beta A64
bounded 38
characteristic 542, 574
complex 614
conjugate harmonic 622
entire 624, 661, 711
error 568, 690, A64, A95
even 490
exponential 57, 623, A60
factorial 192, 1008, A95
gamma 192, A95
Hankel 202
harmonic 465, 622, 772
holomorphic 617
hyperbolic 628, 743, A62
hypergeometric 188
inverse hyperbolic 634
inverse trigonometric 634
Legendre 177
logarithmic 630, A60
meromorphic 711
Neumann 201

Function (*Cont.*)
odd 490
orthogonal 205, 482
orthonormal 205, 210
periodic 478
probability 1012, 1033
rational 617
scalar 384
space 382
staircase 248
step 234
trigonometric 626, 688, A60
unit step 234
vector 384
Function space 327
Fundamental
form 457
matrix 139
mode 542
period 485
system 49, 106, 113, 138
theorem of algebra 662

G

Galilei 15
Gamma function 192, A95
GAMS 778
Gauss 188
distribution 1026
divergence theorem 459
elimination method 289, 834
hypergeometric equation 188
integration formula 826
–Jordan elimination 317, 844
least squares 860, 1084
quadrature 826
–Seidel iteration 846, 913
General
powers 632
solution 64, 106, 138, 159
Generalized
Fourier series 210
function 242
solution 545
triangle inequality 608
Generating function 181, 216, 258
Geometric
multiplicity 337
series 167, 668, 673, 687, 692
Gerschgorin's theorem 866
Gibbs phenomenon 490, 510
Global error 887
Golden Rule 15, 23
Goodness of fit 1076
Gosset 1066
Goursat 648, A88
Gradient 403, 415, 426, A72
method 938
Graph 955
bipartite 982
complete 958
Euler 963
planar 987
sparse 957
weighted 959
Gravitation 385, 407, 411, 587
Greedy algorithm 967
Greek alphabet: Back cover
Green 439
formulas 466
theorem 439, 466
Gregory–Newton formulas 805
Growth restriction 225
Guldin's theorem 458

H

Hadamard's formula 676
Half-life time 9
Half-plane 613
Half-range Fourier series 494
Half-wave rectifier 248, 489
Halving 819, 824
Hamiltonian cycle 960
Hanging cable 198
Hankel functions 202
Hard spring 159
Harmonic
conjugate 622
function 465, 622, 772
oscillation 63
series 670
Heat
equation 464, 536, 553, 757, 923
potential 758
Heaviside 221
expansions 245
formulas 247

Heaviside (*Cont.*)
function 234
Helicoid 449
Helix 391, 394, 399
Helmholtz equation 572
Henry 92
Hermite
interpolation 816
polynomials 216
Hermitian 357, 361
Hertz 63
Hesse's normal form 375
Heun's method 890
High-frequency line equations 594
Hilbert 201, 326
matrix 858
space 326
Histogram 994
Holomorphic 617
Homogeneous
differential equation 27, 46, 105, 535
system of equations 288, 304, 833
Hooke's law 62
Householder's tridiagonalization 875
Hyperbolic
differential equation 551, 909, 928
functions, complex 628, 743
functions, real A62
paraboloid 448
partial differential equations 551, 928
spiral 399
Hypergeometric
differential equation 188
distribution 1024
functions 188
series 188
Hypocycloid 399
Hypothesis 1058

I

Idempotent matrix 286
Identity
of Lagrange 383
matrix (*see* Unit matrix)
theorem for power series 679
transformation 736
trick 351
Ill-conditioned 851
Image 327, 729
Imaginary
axis 604
part 602
unit 602
Impedance 94, 98
Implicit solution 20
Improper integral 222, 719, 722
Impulse 242
IMSL 778
Incidence
list 957
matrix 958
Inclusion theorem 868
Incomplete gamma function A64
Incompressible 765
Inconsistent equations 292, 303
Increasing sequence A69
Indefinite
integral 637, 650
integration 640
Independence of path 426, 648
Independent
events 1004
random variables 1036
Indicial equation 184
Indirect method 845
Inductance 92
Inductor 92
Inequality
Bessel 215, 504
Cauchy 660
ML- 644
Schur 869
triangle 326, 372, 608
Infinite
dimensional 325
population 1025, 1045
sequence 664
series (*see* Series)
Infinity 710, 736
Initial
condition 6, 48, 137, 540
value problem 6, 38, 48, 107, 886, 902
Injective mapping 729
Inner product 325, 359, 371
space 326
Input 26, 84, 230
Instability (*see* Stability)
Integral
contour 647

Integral (*Cont.*)
definite 639
double 433
equation 252
Fourier 508, 563
improper 719, 722
indefinite 637, 650
line 421, 633
surface 449
theorems, complex 647, 654
theorems, real 439, 453, 469
transform 221, 513
triple 458
Integrating factor 23
Integration
complex functions 637–663, 701–727
Laplace transforms255
numeric 817–827
power series 680
series 695
Integro-differential equation 92
Interest 9, 33
Interlacing of zeros 197
Intermediate value theorem 796
Interpolation 797–815
Hermite 816
Lagrange 798
Newton 802, 805, 807
spline 811
Interquartile range 995
Intersection of events 998
Interval
closed A69
of convergence 172, 676
estimate 1046
open 4, A69
Invariant subspace 865
Inverse
hyperbolic functions 634
mapping principle 733
of a matrix 315, 844
trigonometric functions 634
Inversion 735
Investment 9, 33
Irreducible 869
Irregular boundary 919
Irrotational 415, 765
Isocline 10
Isolated singularity 707
Isotherms 758
Iteration
for eigenvalues 872
for equations 787–794
Gauss–Seidel 846, 913
Jacobi 850
Picard 41

J

Jacobian 436, 733
Jacobi iteration 850
Jerusalem, Shrine of the Book 814
Jordan 316
Joukowski airfoil 732

K

Kirchhoff's laws 92, 973
Kronecker delta 210, A83
Kruskal's algorithm 967
Kutta 892

L

l_1, l_2, l_∞ 853
L_2 863
Labeling 968
Lagrange 50
identity of 383
interpolation 798
Laguerre polynomials 209, 257
Lambert's law 43
LAPACK 778
Laplace 221
equation 407, 465, 536, 579, 587, 910
integrals 512
limit theorem 1031
operator 408
transform 221, 594
Laplacian 443, A73
Latent root 324
Laurent series 701, 712
Law of
absorption 43
cooling 14
gravitation 385
large numbers 1032
mass action 43
the mean (*see* Mean value theorem)
LC-circuit 97

LCL 1068
Least squares 860, 1084
Lebesgue 863
Left-hand
  derivative 484
  limit 484
Left-handed 378
Legendre 177
  differential equation 177, 204, 590
  functions 177
  polynomials 179, 207, 590, 826
Leibniz 14
  convergence test A70
Length
  of a curve 393
  of a vector 365
Leonardo of Pisa 638
Leontief 344
Leslie model 341
Libby 13
Liebmann's method 913
Likelihood function 1047
Limit
  of a complex function 615
  cycle 157
  left-hand 484
  point A90
  right-hand 484
  of a sequence 664
  vector 386
  of a vector function 387
Lineal element 9
Linear
  algebra 271–363
  combination 106, 325
  dependence 49, 74, 106, 108, 297, 325
  differential equation 26, 45, 105, 535
  element 394, A72
  fractional transformation 734
  independence 49, 74, 106, 108, 297, 325
  interpolation 798
  operator 60
  optimization 939
  programming 939
  space (*see* Vector space)
  system of equations 287, 833
  transformation 281, 327
Linearization of systems of ODEs 151
Line integral 421, 633
Lines of force 751
LINPACK 779
Liouville 203
  theorem 661
Lipschitz condition 40
List 957
Ljapunov 148
Local
  error 887
  minimum 937
Logarithm 630, 688, A60
Logarithmic
  decrement 69
  integral A66
  spiral 399
Logistic population law 30
Longest path 959
Loss of significant digits 785
Lotka-Volterra population model 154
Lot tolerance per cent defective 1075
Lower
  control limit 1068
  triangular matrix 283
LTPD 1075
LU-factorization 841

M

Maclaurin 683
  series 683
  trisectrix 399
Magnitude of a vector (*see* Length)
Main diagonal 274, 309
Malthus's law 5, 31
MAPLE 779
Mapping 327, 729
Marconi 63
Marginal distributions 1035
Markov process 285, 341
Mass–spring system 61, 86, 135, 150, 243, 252, 261, 342, 499
Matching 985
MATHCAD 779
MATHEMATICA 779
Mathematical expectation 1019, 1038
MATLAB 779
Matrix
  addition 275
  augmented 288, 833
  band 914

Matrix (*Cont.*)
diagonal 284
eigenvalue problem 333–363, 863–882
Hermitian 357
identity (*see* Unit matrix)
inverse 315
inversion 315, 844
multiplication 278, 321
nonsingular 315
norm 849, 854
normal 362, 869
null (*see* Zero matrix)
orthogonal 345
polynomial 865
scalar 284
singular 315
skew-Hermitian 357
skew-symmetric 283, 345
sparse 812, 912
square 274
stochastic 285
symmetric 283, 345
transpose 282
triangular 283
tridiagonal 812, 875, 914
unit 284
unitary 357
zero 276
Max-flow min-cut theorem 978
Maximum 937
flow 979
likelihood method 1047
matching 983
modulus theorem 772
principle 773
Mean convergence 214
Mean-square convergence 214
Mean value of a (an)
analytic function 771
distribution 1016
function 764
harmonic function 772
sample 996
Mean value theorem 402, 434, 454
Median 994, 1081
Membrane 569–586
Meromorphic function 711
Mesh incidence matrix 278
Method of
false position 796
Method of (*Cont.*)
least squares 860, 1084
moments 1046
steepest descent 938
undetermined coefficients 78, 117, 160
variation of parameters 98, 118, 160
Middle quartile 994
Minimum 937, 942, 946
MINITAB 991
Minor 309
Mixed
boundary value problem 558, 587, 759, 917
triple product 381
Mixing problem 13, 130, 146, 163, 259
Mks system: Front cover
ML-inequality 644
Möbius 453
strip 453, 456
transformation 734
Mode 542, 582
Modeling 2, 6, 13, 61, 84, 130, 159, 222, 340, 499, 538, 569, 750–767
Modified Bessel functions 203
Modulus 607
Molecule 912
Moivre's formula 610
Moment
central 1019
of a distribution 1019
of a force 380
generating function 1026
of inertia 436, 455, 457
of a sample 1046
vector 380
Monotone sequence A69
Moore's shortest path algorithm 960
Morera's theorem 661
Moulton 900
Moving trihedron (*see* Trihedron)
M-test for convergence 969
Multinomial distribution 1025
Multiple point 391
Multiplication of
complex numbers 603, 609
determinants 322
matrices 278, 321
means 1038
power series 174, 680
vectors 219, 371, 377
Multiplication rule for events 1003

Multiplicity 337, 865
Multiply connected domain 646
Multistep method 898
"Multivalued function" 615
Mutually exclusive events 998

N

Nabla 403
NAG 779
Natural
frequency 63
logarithm 630, A60
spline 812
Neighborhood 387, 613
Nested form 786
NETLIB 779
Networks 132, 146, 162, 244, 260, 263, 277, 331
in graph theory 973
Neumann, C. 201
functions 201
problem 558, 587, 917
Newton 14
–Cotes formulas 822
interpolation formulas 802, 805, 807
law of cooling 14
law of gravitation 385
method 800
–Raphson method 800
second law 62
Neyman 1049, 1058
Nicolson 924
Nicomedes 399
Nilpotent matrix 286
NIST 779
Nodal
incidence matrix 277
line 574
Node 142, 797
Nonbasic variables 945
Nonconservative 428
Nonhomogeneous
differential equation 27, 78, 116, 159, 305, 535
system of equations 288, 304
Nonlinear differential equations 45, 151
Nonorientable surface 453
Nonparametric test 1080
Nonsingular matrix 315
Norm 205, 326, 346, 359, 365, 849
Normal
acceleration 395
asymptotically 1057
to a curve 398
derivative 444, 465
distribution 1026, 1047–1057, 1062–1067, A98
two-dimensional 1090
equations 860, 1086
form of a PDE 551
matrix 362, 869
mode 542, 582
plane 398
to a plane 375
random variable 1026
to a surface 447
vector 375, 447
Null
hypothesis 1058
matrix (*see* Zero matrix)
space 301
vector (*see* Zero vector)
Nullity 301
Numeric methods 777–934
differentiation 827
eigenvalues 863–882
equations 787–796
integration 817–827
interpolation 797–816
linear equations 833–858
matrix inversion 315, 844
optimization 936–953
ordinary differential equations (ODEs)
886–908
partial differential equations (PDEs) 909–930
Nyström method 906

O

O 962
Objective function 936
OC curve 1062
Odd function 490
ODE 4 (*see also* Differential equations)
Ohm's law 92
One-dimensional
heat equation 553
wave equation 539
One-sided
derivative 484
test 1060

One-step method 898
One-to-one mapping 729
Open
  disk 613
  integration formula 827
  interval A69
  point set 613
Operating characteristic 1062
Operational calculus 59, 220
Operation count 838
Operator 59, 327
Optimality principle, Bellman's 963
Optimal solution 942
Optimization 936–953, 959–990
Orbit 141
Order 887, 962
  of a determinant 308
  of a differential equation 4, 535
  of an iteration process 793
Ordering 969
Ordinary differential equations 2–269, 886–908
  (*see also* Differential equation)
Orientable surface 452
Orientation of a
  curve 638
  surface 452
Orthogonal
  coordinates A71
  curves 35
  eigenvectors 350
  expansion 210
  functions 205, 482
  matrix 345
  polynomials 209
  series 210
  trajectories 35
  transformation 346
  vectors 326, 371
Orthonormal functions 205, 210
Oscillations
  of a beam 547, 552
  of a cable 198
  in circuits 91
  damped 64, 88
  forced 84
  free 61, 500, 547
  harmonic 63
  of a mass on a spring 61, 86, 135, 150, 243, 252, 261, 342, 499
  of a membrane 569–586
Oscillations (*Cont.*)
  self-sustained 157
  of a string 204, 538, 929
  undamped 62,
Osculating plane 398
Outcome 997
Outer product 377
Outlier 995
Output 26, 230
Overdamping 65
Overdetermined system 292
Overflow 782
Overrelaxation 851
Overtone 542

P

Paired comparison 1065
Pappus's theorem 458
Parabolic differential equation 551, 922
Paraboloid 448
Parachutist 12
Parallelepiped 382
Parallel flow 766
Parallelogram
  equality 326, 372, 612
  law 367
Parameter of a distribution 1016
Parametric representation 389, 446
Parking problem 1023
Parseval's equality 215, 504
Partial
  derivative 388, A66
  differential equation 535, 909
  fractions 231, 245
  pivoting 291, 834
  sum 171, 480, 666
Particular solution 6, 48, 106, 159
Pascal 399
Path
  in a digraph 974
  in a graph 959
  of integration 421, 637
PDE 535, 909 (*see also* Differential equation)
Peaceman-Rachford method 915
Pearson, E. S. 1058
Pearson, K. 1066
Pendulum 68, 152, 156
Period 478

Periodic
extension 494
function 478
Permutation 1006
Perron–Frobenius theorem 344, 869
Pfaffian form 429
Phase
angle 88
of complex number (*see* Argument)
lag 88
plane 141, 147
portrait 141, 147
Picard
iteration method 41
theorem 709
Piecewise
continuous 226
smooth 421, 448, 639
Pivoting 291, 834
Planar graph 987
Plane 315, 375
Plane curve 391
Poincaré 216
Point
estimate 1046
at infinity 710, 736
set 613
source 765
spectrum 507, 524
Poisson 769
distribution 1022, 1073, A97
equation 910, 918
integral formula 769
Polar
coordinates 437, 443, 580, 607–608
form of complex numbers 607
moment on inertia 436
Pole 708
Polynomial
approximation 797
matrix 865
Polynomially bounded 962
Polynomials 617
Chebyshev 209
Hermite 216
Laguerre 207, 257
Legendre 179, 207, 590, 826
trigonometric 502
Population in statistics 1044
Population models 31, 154, 341
Position vector 366
Positive definite 326, 372
Possible values 1012
Postman problem 963
Potential 407, 427, 590, 750, 762
complex 763
theory 465, 749
Power
method for eigenvalues 872
of a test 1061
series 167, 673
series method 167
Precision 782
Predator–prey 154
Predictor–corrector 890, 900
Pre-Hilbert space 326
Prim's algorithm 971
Principal
axes theorem 354
branch 632
diagonal (*see* main diagonal)
directions 340
normal (Fig. 210) 397
part 708
value 607, 630, 632, 719, 722
Prior estimate 794
Probability 1000, 1001
conditional 1003
density 1014, 1034
distribution 1010, 1032
function 1012, 1033
Producer's risk 1075
Product (*see* Multiplication)
Projection of a vector 374
Pseudocode 783
Pure imaginary number 603

Q

QR-factorization method 879
Quadratic
equation 785
form 353
interpolation 799
Qualitative methods 124, 139–165
Quality control 1068
Quartile 995
Quasilinear 551, 909
Quotient of complex numbers 604

R

Rachford method 915
Radiation 7, 561
Radiocarbon dating 13
Radius
  of convergence 172, 675, 686
  of a graph 973
Random
  experiment 997
  numbers 1045
  variable 1010, 1032
Range of a
  function 614
  sample 994
Rank of a matrix 297, 299, 311
Raphson 790
Rational function 617
Ratio test 669
Rayleigh 159
  equation 159
  quotient 872
RC-circuit 97, 237, 240
Reactance 93
Real
  axis 604
  part 602
  sequence 664
  vector space 324, 369
Rectangular
  membrane 571
  pulse 238, 243
  rule 817
  wave 211, 480, 488, 492
Rectifiable curve 393
Rectification of a lot 1075
Rectifier 248, 489, 492
Rectifying plane 398
Reduction of order 50, 116
Region 433, 614
Regression 1083
  coefficient 1085, 1088
  line 1084
Regula falsi 796
Regular
  point of an ODE 183
  Sturm-Liouville problem 206
Rejectable quality level 1075
Rejection region 1060
Relative
  class frequency 994
Relative (*Cont.*)
  error 784
  frequency 1000
Relaxation 850
Remainder 171, 666
Removable singularity 709
Representation 328
Residual 849, 852
Residue 713
Residue theorem 715
Resistance 91
Resonance 86
Response 28, 84
Restoring force 62
Resultant of forces 367
Riccati equation 34
Riemann 618
  sphere 710
  surface 746
Right-hand
  derivative 484
  limit 484
Right-handed 377
Risk 1095
RL-circuit 97, 240
RLC-circuit 95, 241, 244, 499
Robin problem 558, 587
Rodrigues's formula 181
Romberg integration 829
Root 610
Root test 671
Rotation 381, 385, 414, 734, 764
Rounding 782
Row
  echelon form 294
  -equivalent 292, 298
  operations 292
  scaling 838
  space 300
  sum norm 849
  vector 274
Runge–Kutta–Fehlberg 893
Runge–Kutta methods 892, 904
Runge–Kutta–Nyström 906

S

Saddle point 143
Sample 997, 1045
  covariance 1085

Sample (*Cont.*)
distribution function 1076
mean 996, 1045
moments 1046
point 997
range 994
size 997, 1045
space 997
standard deviation 996, 1046
variance 996, 1045
Sampling 1004, 1023, 1073
SAS 991
Sawtooth wave 248, 493, 505
Scalar 276, 364
field 384
function 384
matrix 284
multiplication 276, 368
triple product 381
Scaling 838
Scheduling 987
Schoenberg 810
Schrödinger 242
Schur's inequality 869
Schwartz 242
Schwarz inequality 326
Secant 627, A62
method 794
Second
Green's formula 466
shifting theorem 235
Sectionally continuous
(*see* Piecewise continuous)
Seidel 846
Self-starting 898
Self-sustained oscillations 157
Separable differential equation 12
Separation of variables 12, 540
Sequence 664, A69
Series 666, A69
addition of 680
of Bessel functions 213, 583
binomial 689
convergence of 171, 666
differentiation of 174, 680
double Fourier 576
of eigenfunctions 210
Fourier 211, 480, 487
geometric 167, 668, 673, 687, 692
harmonic 670
Series (*Cont.*)
hypergeometric 188
infinite 666, A70
integration of 680
Laurent 701, 712
Maclaurin 683
multiplication of 174, 680
of orthogonal functions 210
partial sums of 171, 666
power 167, 673
real A69
remainder of 171, 666, 684
sum of 171, 666
Taylor 683
trigonometric 479
value of 171, 666
Serret–Frenet formulas
(*see* Frenet formulas)
Set of points 613
Shifted data problem 232
Shifting theorems 224, 235, 528
Shortest
path 959
spanning tree 967
Shrine of the Book 814
Sifting 242
Significance level 1059
Significant
digit 781
in statistics 1059
Sign test 1081
Similar matrices 350
Simple
curve 391
event 997
graph 955
pole 708
zero 709
Simplex
method 944
table 945
Simply connected 640, 646
Simpson's rule 821
Simultaneous
corrections 850
differential equations 124
linear equations (*see* Linear systems)
Sine
of a complex variable 627, 688, 742
hyperbolic 688

Sine (*Cont.*)
integral 509, 690, A65, A95
of a real variable A60
Single precision 782
Single-valued relation 615
Singular
at infinity 711
matrix 315
point 183, 686, 707
solution 8, 50
Sturm–Liouville problem 206
Singularity 686, 707
Sink 464, 765, 973
SI system: Front cover
Size of a sample 997, 1045
Skew-Hermitian 357, 361
Skewness 1020
Skew-symmetric matrix 283, 345
Skydiver 12
Slack variable 941
Slope field 9
Smooth
curve 421, 638
piecewise 421, 448, 639
surface 448
Sobolev 242
Soft spring 159
Software 778, 991
Solution
of a differential equation 4, 46, 105, 536
general 6, 48, 106, 138
particular 6, 48, 106
singular 8, 50
space 304
steady-state 88
of a system of differential equations 136
of a system of equations 288
vector 288
SOR 851
Sorting 969, 993
Source 464, 765, 973
Span 300
Spanning tree 967
Sparse
graph 957
matrix 812, 912
system of equations 846
Spectral
density 520
mapping theorem 344, 865
Spectral (*Cont.*)
radius 848
representation 520
shift 344, 865, 874
Spectrum 324, 542, 864
Speed 394
Sphere 446
Spherical coordinates 588, A71
Spiral 399
point 144
Spline 811
S-PLUS, SPSS 992
Spring 62
Square
error 503
matrix 274
membrane 575
root 792
wave 211, 480, 488, 492
Stability 31, 148, 783, 822, 922
chart 148
Stagnation point 763
Staircase function 248
Standard
basis 328, 369
deviation 1016
form of a linear ODE 26, 45, 105
Standardized random variable 1018
Stationary point 937
Statistical
inference 1044
tables A96–A106
Steady 413, 463
state 88
Steepest descent 938
Steiner 399, 457
Stem-and-leaf plot 994
Stencil 912
Step-by-step method 886
Step function 234
Step size control 889
Stereographic projection 711
Stiff
ODE 896
system of ODEs 907
Stirling formula 1008, A64
Stochastic
matrix 285
variable 1011
Stokes's theorem 469

Straight line 375, 391
Stream function 762
Streamline 761
Strength of a source 767
Strictly diagonally dominant 868
String 204, 538, 594
Student's t-distribution 1053, A100
Sturm–Liouville problem 203
Subgraph 956
Submarine cable equations 594
Submatrix 302
Subsidiary equation 220, 230
Subspace 300
  invariant 865
Success 1021
Successive
  corrections 850
  overrelaxation 851
Sum (*see* Addition)
Sum of a series 171, 666
Superlinear convergence 795
Superposition principle 106, 138
Surface 445
  area 435, 442, 454
  integral 449
  normal 406
Surjective mapping 729
Symmetric matrix 283, 345
System of
  differential equations 124–165, 258–263, 902
  linear equations (*see* Linear system)
  units: Front cover

T

Tables
  on differentiation: Front cover
  of Fourier transforms 529–531
  of functions A94–A106
  of integrals: Front cover
  of Laplace transforms 265–267
  statistical A96–A106
Tangent 627, A62
  to a curve 392, 397
  hyperbolic 629, A62
  plane 406, 447
  vector 392
Tangential acceleration 395
Target 973
Tarjan 971
Taylor 683
  formula 684
  series 683
Tchebichef (*see* Chebyshev)
t-distribution 1053, A100
Telegraph equations 594
Termination criterion 791
Termwise
  differentiation 696
  integration 695
  multiplication 680
Test
  chi-square 1077
  for convergence 667–672
  of hypothesis 1058–1068
  nonparametric 1080
Tetrahedron 382
Thermal
  conductivity 465, 552
  diffusivity 465, 552
Three
  -eights rule 830
  -sigma limits 1028
Timetabling 987
Torricelli's law 15
Torsion of a curve 397
Torsional vibrations 68
Torus 454
Total differential 19
Trace of a matrix 344, 355, 864
Trail 959
Trajectories 35, 133, 141
Transfer function 230
Transformation
  of Cartesian coordinates A84
  by a complex function 729
  of integrals 437, 439, 459, 469
  linear 281
  linear fractional 734
  orthogonal 346
  similarity 350
  unitary 359
  of vector components A83
Transient state 88
Translation 365, 734
Transmission line equations 593
Transpose of a matrix 282
Transpositions 1081, A106
Trapezoidal rule 817
Traveling salesman problem 960

Tree 966
Trend 1081
Trial 997
Triangle inequality 326, 372, 608
Triangular matrix 283
Tricomi equation 551, 552
Tridiagonalization 875
Tridiagonal matrix 812, 875, 914
Trigonometric
  approximation 502
  form of complex numbers 607, 624
  functions, complex 626, 688
  functions, real A60
  polynomial 502
  series 479
  system 479, 482
Trihedron 398
Triple
  integral 458
  product 381
Trivial solution 27, 304
Truncation error 783
Tuning 543
Twisted curve 391
Two-dimensional
  distribution 1032
  normal distribution 1090
  random variable 1032
  wave equation 571
Two-sided test 1060
Type of a differential equation 551
Type I and II errors 1060

U

UCL 1068
Unconstrained optimization 937
Uncorrelated 1090
Undamped system 62
Underdamping 66
Underdetermined system 292
Underflow 782
Undetermined coefficients 79, 117, 160
Uniform
  convergence 691
  distribution 1015, 1017, 1034
Union of events 998
Uniqueness
  differential equations 37, 73, 107, 137
  Dirichlet problem 774
Uniqueness (*Cont.*)
  Laurent series 705
  linear equations 303
  power series 678
Unit
  binormal vector 398
  circle 611
  impulse 242
  matrix 284
  normal vector 447
  principal normal vector 398
  step function 234
  tangent vector 392, 398
  vector 326
Unitary
  matrix 357
  system of vectors 359
  transformation 359
Unstable (*see* Stability)
Upper control limit 1068

V

Value of a series 171, 666
Vandermonde determinant 112
Van der Pol equation 157
Variable
  complex 614
  random 1010, 1032
  standardized random 1018
  stochastic 1011
Variance of a
  distribution 1016
  sample 996, 1045
Variation of parameters 98, 118, 160
Vector 274, 364
  addition 276, 327, 367
  field 384
  function 384
  moment 380
  norm 853
  product 377
  space 300, 323, 369
  subspace 300
Velocity 394
  field 385
  potential 762
  vector 394
Venn diagram 998
Verhulst 31

Vertex 955
coloring 987
exposed 983
incidence list 957
Vibrations (*see* Oscillations)
Violin string 538
Vizing's theorem 987
Volta 92
Voltage drop 92
Volterra 154, 201, 253
Volume 435
Vortex 767
Vorticity 764

W

Waiting time problem 1013
Walk 959
Wave equation 536, 539, 569, 929
Weber 217
equation 217
Weber (*Cont.*)
functions 201
Website *see* Preface
Weierstrass 618, 696
approximation theorem 797
M-test 696
Weighted graph 959
Weight function 205
Well-conditioned 852
Wessel 605
Wheatstone bridge 296
Work 373, 423
integral 422
Wronskian 75, 108

Z

Zero
of analytic function 709
matrix 276
vector 367

PHOTO CREDITS

Part A Opener: John Sohm/Chromoshohm/Photo Researchers.
Part B Opener: Lester Lefkowitz/Corbis Images.
Part C Opener: Science Photo Library/Photo Researchers.
Part D Opener: Lester Lefkowitz/Corbis Images.
Part E Opener: Walter Hodges/Stone/Getty Images.
Chapter 19, Figure 434: From Dennis Sharp, ARCHITECKTUR. ©1973 der deutschsprachigen Ausgabe Edition Praeger GmbH, Munchen.
Part F Opener: Charles O'Rear/Corbis Images.
Part G Opener: Greg Pease/Stone/Getty Images.
Appendix 1 Opener: Patrick Bennett/Corbis Images.
Appendix 2 Opener: Richard T. Nowitz/Corbis Images.
Appendix 3 Opener: Chris Kapolka/Stone/Getty Images.
Appendix 4 Opener: John Olson/Corbis Images.

Some Constants

$e = 2.71828\ 18284\ 59045\ 23536$
$\sqrt{e} = 1.64872\ 12707\ 00128\ 14685$
$e^2 = 7.38905\ 60989\ 30650\ 22723$

$\pi = 3.14159\ 26535\ 89793\ 23846$
$\pi^2 = 9.86960\ 44010\ 89358\ 61883$
$\sqrt{\pi} = 1.77245\ 38509\ 05516\ 02730$

$\log_{10} \pi = 0.49714\ 98726\ 94133\ 85435$
$\ln \pi = 1.14472\ 98858\ 49400\ 17414$
$\log_{10} e = 0.43429\ 44819\ 03251\ 82765$
$\ln 10 = 2.30258\ 50929\ 94045\ 68402$

$\sqrt{2} = 1.41421\ 35623\ 73095\ 04880$
$\sqrt[3]{2} = 1.25992\ 10498\ 94873\ 16477$
$\sqrt{3} = 1.73205\ 08075\ 68877\ 29353$
$\sqrt[3]{3} = 1.44224\ 95703\ 07408\ 38232$
$\ln 2 = 0.69314\ 71805\ 59945\ 30942$
$\ln 3 = 1.09861\ 22886\ 68109\ 69140$

$\gamma = 0.57721\ 56649\ 01532\ 86061$
$\ln \gamma = -0.54953\ 93129\ 81644\ 82234$
(see Sec. 5.6)
$1^\circ = 0.01745\ 32925\ 19943\ 29577$ rad
1 rad $= 57.29577\ 95130\ 82320\ 87680^\circ$
$= 57^\circ 17' 44.806''$

Polar Coordinates

$$x = r\cos\theta \qquad y = r\sin\theta$$

$$r = \sqrt{x^2 + y^2} \qquad \tan\theta = \frac{y}{x}$$

$$dx\,dy = r\,dr\,d\theta$$

Series

$$\frac{1}{1-x} = \sum_{m=0}^{\infty} x^m \quad (|x| < 1)$$

$$e^x = \sum_{m=0}^{\infty} \frac{x^m}{m!}$$

$$\sin x = \sum_{m=0}^{\infty} \frac{(-1)^m x^{2m+1}}{(2m+1)!}$$

$$\cos x = \sum_{m=0}^{\infty} \frac{(-1)^m x^{2m}}{(2m)!}$$

$$\ln(1-x) = -\sum_{m=1}^{\infty} \frac{x^m}{m} \quad (|x| < 1)$$

$$\arctan x = \sum_{m=0}^{\infty} \frac{(-1)^m x^{2m+1}}{2m+1} \quad (|x| < 1)$$

Greek Alphabet

α	Alpha	ν	Nu
β	Beta	ξ	Xi
γ, Γ	Gamma	o	Omicron
δ, Δ	Delta	π	Pi
ϵ, ε	Epsilon	ρ	Rho
ζ	Zeta	σ, Σ	Sigma
η	Eta	τ	Tau
$\theta, \vartheta, \Theta$	Theta	υ, Υ	Upsilon
ι	Iota	ϕ, φ, Φ	Phi
κ	Kappa	χ	Chi
λ, Λ	Lambda	ψ, Ψ	Psi
μ	Mu	ω, Ω	Omega

Vectors

$$\mathbf{a}\cdot\mathbf{b} = a_1b_1 + a_2b_2 + a_3b_3$$

$$\mathbf{a}\times\mathbf{b} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix}$$

$$\text{grad } f = \nabla f = \frac{\partial f}{\partial x}\mathbf{i} + \frac{\partial f}{\partial y}\mathbf{j} + \frac{\partial f}{\partial z}\mathbf{k}$$

$$\text{div } \mathbf{v} = \nabla\cdot\mathbf{v} = \frac{\partial v_1}{\partial x} + \frac{\partial v_2}{\partial y} + \frac{\partial v_3}{\partial z}$$

$$\text{curl } \mathbf{v} = \nabla\times\mathbf{v} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ v_1 & v_2 & v_3 \end{vmatrix}$$