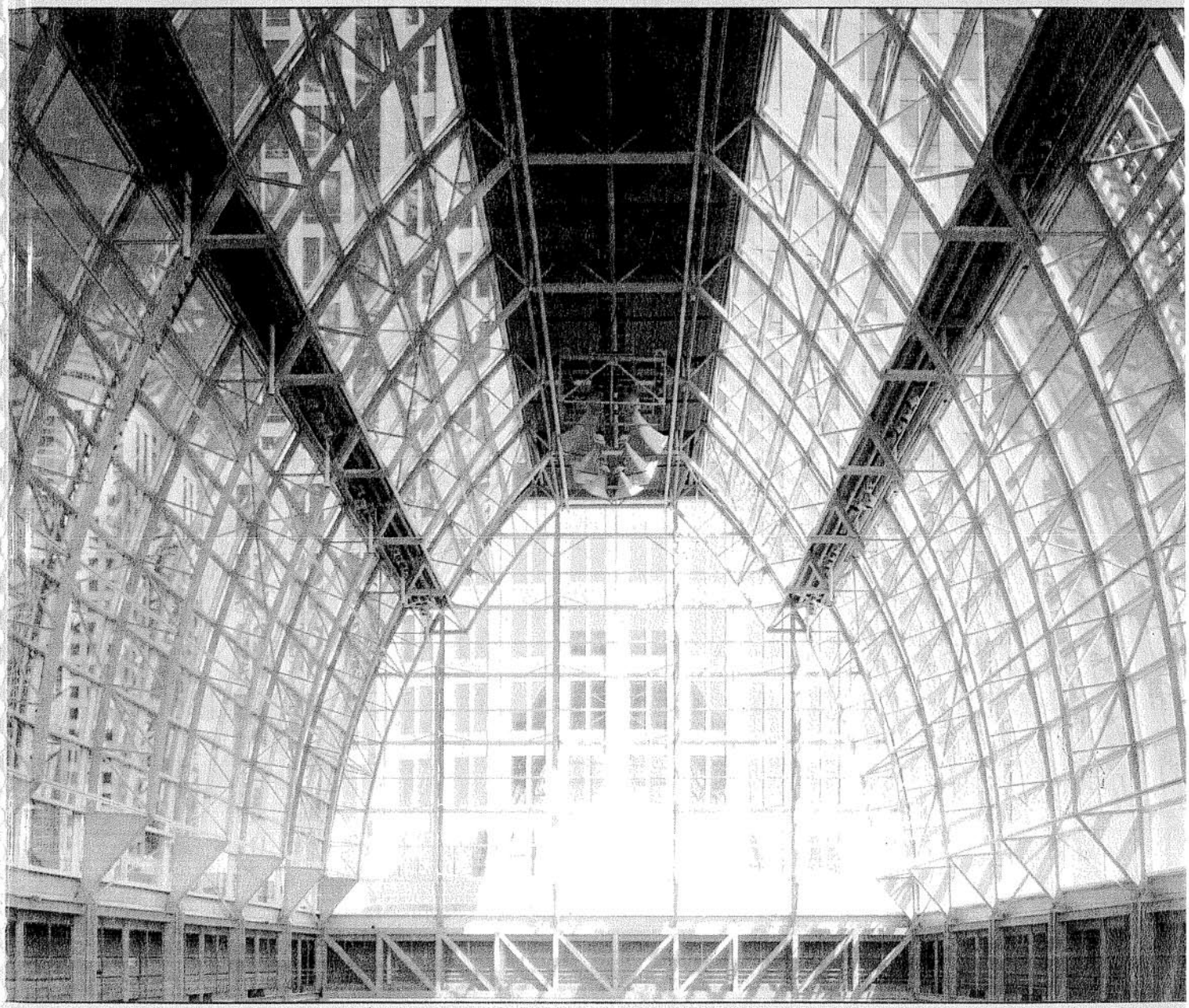
SECOND EDITION

ADVANCED ENGINEERING MATHEMATICS



Michael D. Greenberg

SELECTED FORMULAS

FIRST-ORDER LINEAR: $y' + p(x)y = q(x)$.

General solution $\quad y(x) = e^{-\int p(x)\,dx}\left(\int e^{\int p(x)\,dx} q(x)\,dx + C\right)$

If $y(a) = b$, $\quad y(x) = e^{-\int_a^x p(\xi)\,d\xi}\left(\int_a^x e^{\int_a^\xi p(\zeta)\,d\zeta} q(\xi)\,d\xi + b\right)$

LEGENDRE EQUATION: $(1 - x^2)y'' - 2xy' + \lambda y = 0$

Bounded solutions $P_n(x)$ on $-1 \le x \le 1$ if $\lambda = n(n+1)$, $\quad n = 0, 1, 2, \ldots$

BESSEL EQUATION: $x^2y'' + xy' + \left(x^2 - \nu^2\right) y = 0$

General solution $\quad y(x) = \begin{cases} AJ_\nu(x) + BY_\nu(x) \\ CH_\nu^{(1)}(x) + DH_\nu^{(2)}(x) \end{cases}$

MODIFIED BESSEL EQUATION: $x^2y'' + xy' + \left(-x^2 - \nu^2\right) y = 0$

General solution $\quad y(x) = AI_\nu(x) + BK_\nu(x)$

REDUCIBLE TO A BESSEL EQUATION: $\dfrac{d}{dx}\left(x^a \dfrac{dy}{dx}\right) + bx^c y = 0$

Solution $\quad y(x) = x^{\nu/\alpha} Z_{|\nu|}\left(\alpha\sqrt{|b|}\, x^{1/\alpha}\right), \qquad \alpha = \dfrac{2}{c - a + 2}, \quad \nu = \dfrac{1 - a}{c - a + 2},$

where $Z_{|\nu|}$ denotes $J_{|\nu|}$ and $Y_{|\nu|}$ if $b > 0$, and $I_{|\nu|}$ and $K_{|\nu|}$ if $b < 0$

MATRICES: $\quad \mathbf{A}^{-1} = \dfrac{1}{\det\mathbf{A}}\,\mathrm{adj}\mathbf{A}. \quad (\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}, \quad (\mathbf{A}^{\mathrm{T}})^{-1} = (\mathbf{A}^{-1})^{\mathrm{T}}, \quad (\mathbf{AB})^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}$

CARTESIAN COORDINATES: $\quad u = u(x, y, z), \quad \mathbf{v} = v_x\hat{\mathbf{i}} + v_y\hat{\mathbf{j}} + v_z\hat{\mathbf{k}}$

$$\mathbf{R} = x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}}. \quad d\mathbf{R} = dx\hat{\mathbf{i}} + dy\hat{\mathbf{j}} + dz\hat{\mathbf{k}}.$$

$$dA = \begin{cases} dy\,dz & \text{(constant-}x\text{ surface)} \\ dx\,dz & \text{(constant-}y\text{ surface)} \\ dx\,dy & \text{(constant-}z\text{ surface)} \end{cases}, \qquad dV = dx\,dy\,dz$$

$$\nabla u = \left(\hat{\mathbf{i}}\frac{\partial}{\partial x} + \hat{\mathbf{j}}\frac{\partial}{\partial y} + \hat{\mathbf{k}}\frac{\partial}{\partial z}\right) u = \frac{\partial u}{\partial x}\hat{\mathbf{i}} + \frac{\partial u}{\partial y}\hat{\mathbf{j}} + \frac{\partial u}{\partial z}\hat{\mathbf{k}}$$

$$\nabla^2 u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}$$

$$\nabla \cdot \mathbf{v} = \frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z}$$

$$\nabla \times \mathbf{v} = \left(\frac{\partial v_z}{\partial y} - \frac{\partial v_y}{\partial z}\right)\hat{\mathbf{i}} - \left(\frac{\partial v_z}{\partial x} - \frac{\partial v_x}{\partial z}\right)\hat{\mathbf{j}} + \left(\frac{\partial v_y}{\partial x} - \frac{\partial v_x}{\partial y}\right)\hat{\mathbf{k}} = \begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ v_x & v_y & v_z \end{vmatrix}$$

CYLINDRICAL COORDINATES: $u = u(r, \theta, z)$, $\mathbf{v} = v_r \hat{\mathbf{e}}_r + v_\theta \hat{\mathbf{e}}_\theta + v_z \hat{\mathbf{e}}_z$

$$x = r\cos\theta, \quad y = r\sin\theta, \quad z = z$$

$$\mathbf{R} = r\hat{\mathbf{e}}_r + z\hat{\mathbf{e}}_z, \quad d\mathbf{R} = dr\hat{\mathbf{e}}_r + r\,d\theta\hat{\mathbf{e}}_\theta + dz\hat{\mathbf{e}}_z$$

$$dA = \begin{cases} r\,d\theta\,dz & \text{(constant-}r\text{ surface)} \\ dr\,dz & \text{(constant-}\theta\text{ surface)} \\ r\,dr\,d\theta & \text{(constant-}z\text{ surface)} \end{cases}, \qquad dV = r\,dr\,d\theta\,dz$$

$$\frac{d\hat{\mathbf{e}}_r}{d\theta} = \hat{\mathbf{e}}_\theta, \quad \frac{d\hat{\mathbf{e}}_\theta}{d\theta} = -\hat{\mathbf{e}}_r$$

$$\nabla u = \frac{\partial u}{\partial r}\hat{\mathbf{e}}_r + \frac{1}{r}\frac{\partial u}{\partial \theta}\hat{\mathbf{e}}_\theta + \frac{\partial u}{\partial z}\hat{\mathbf{e}}_z$$

$$\nabla^2 u = \frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2} + \frac{\partial^2 u}{\partial z^2}$$

$$\nabla \cdot \mathbf{v} = \frac{1}{r}\frac{\partial}{\partial r}(r v_r) + \frac{1}{r}\frac{\partial}{\partial \theta} v_\theta + \frac{\partial}{\partial z} v_z$$

$$\nabla \times \mathbf{v} = \left(\frac{1}{r}\frac{\partial v_z}{\partial \theta} - \frac{\partial v_\theta}{\partial z}\right)\hat{\mathbf{e}}_r + \left(\frac{\partial v_r}{\partial z} - \frac{\partial v_z}{\partial r}\right)\hat{\mathbf{e}}_\theta + \frac{1}{r}\left(\frac{\partial (r v_\theta)}{\partial r} - \frac{\partial v_r}{\partial \theta}\right)\hat{\mathbf{e}}_z$$

SPHERICAL COORDINATES: $u = u(\rho, \phi, \theta)$, $\mathbf{v} = v_\rho \hat{\mathbf{e}}_\rho + v_\phi \hat{\mathbf{e}}_\phi + v_\theta \hat{\mathbf{e}}_\theta$

$$x = \rho\sin\phi\cos\theta, \quad y = \rho\sin\phi\sin\theta, \quad z = \rho\cos\phi$$

$$\mathbf{R} = \rho\hat{\mathbf{e}}_\rho, \quad d\mathbf{R} = d\rho\hat{\mathbf{e}}_\rho + \rho\,d\phi\hat{\mathbf{e}}_\phi + \rho\sin\phi\,d\theta\hat{\mathbf{e}}_\theta$$

$$dA = \begin{cases} \rho^2\,|\sin\phi|\,d\phi\,d\theta & \text{(constant-}\rho\text{ surface)} \\ \rho\,|\sin\phi|\,d\rho\,d\theta & \text{(constant-}\phi\text{ surface)} \\ \rho d\rho\,d\phi & \text{(constant-}\theta\text{ surface)} \end{cases}, \qquad dV = \rho^2\,|\sin\phi|\,d\rho\,d\phi\,d\theta$$

$$\frac{\partial \hat{\mathbf{e}}_\rho}{\partial \rho} = 0, \quad \frac{\partial \hat{\mathbf{e}}_\rho}{\partial \phi} = \hat{\mathbf{e}}_\phi, \quad \frac{\partial \hat{\mathbf{e}}_\rho}{\partial \theta} = \sin\phi\,\hat{\mathbf{e}}_\theta,$$

$$\frac{\partial \hat{\mathbf{e}}_\phi}{\partial \rho} = 0, \quad \frac{\partial \hat{\mathbf{e}}_\phi}{\partial \phi} = -\hat{\mathbf{e}}_\rho, \quad \frac{\partial \hat{\mathbf{e}}_\phi}{\partial \theta} = \cos\phi\,\hat{\mathbf{e}}_\theta,$$

$$\frac{\partial \hat{\mathbf{e}}_\theta}{\partial \rho} = 0, \quad \frac{\partial \hat{\mathbf{e}}_\theta}{\partial \phi} = 0, \quad \frac{\partial \hat{\mathbf{e}}_\theta}{\partial \theta} = -\sin\phi\,\hat{\mathbf{e}}_\rho - \cos\phi\,\hat{\mathbf{e}}_\phi$$

$$\nabla u = \frac{\partial u}{\partial \rho}\hat{\mathbf{e}}_\rho + \frac{1}{\rho}\frac{\partial u}{\partial \phi}\hat{\mathbf{e}}_\phi + \frac{1}{\rho\sin\phi}\frac{\partial u}{\partial \theta}\hat{\mathbf{e}}_\theta$$

$$\nabla^2 u = \frac{1}{\rho^2}\left[\frac{\partial}{\partial \rho}\left(\rho^2\frac{\partial u}{\partial \rho}\right) + \frac{1}{\sin\phi}\frac{\partial}{\partial \phi}\left(\sin\phi\frac{\partial u}{\partial \phi}\right) + \frac{1}{\sin^2\phi}\frac{\partial^2 u}{\partial \theta^2}\right]$$

$$\nabla \cdot \mathbf{v} = \frac{1}{\rho^2}\frac{\partial}{\partial \rho}(\rho^2 v_\rho) + \frac{1}{\rho\sin\phi}\frac{\partial}{\partial \phi}(v_\phi\sin\phi) + \frac{1}{\rho\sin\phi}\frac{\partial v_\theta}{\partial \theta}$$

$$\nabla \times \mathbf{v} = \frac{1}{\rho\sin\phi}\left(\frac{\partial}{\partial \phi}(v_\theta\sin\phi) - \frac{\partial v_\phi}{\partial \theta}\right)\hat{\mathbf{e}}_\rho + \frac{1}{\rho}\left(\frac{1}{\sin\phi}\frac{\partial v_\rho}{\partial \theta} - \frac{\partial(\rho v_\theta)}{\partial \rho}\right)\hat{\mathbf{e}}_\phi + \frac{1}{\rho}\left(\frac{\partial(\rho v_\phi)}{\partial \rho} - \frac{\partial v_\rho}{\partial \phi}\right)\hat{\mathbf{e}}_\theta$$

AREA ELEMENT: If $x = x(u,v),\ y = y(u,v),\ z = z(u,v):\quad dA = \sqrt{EG - F^2}\, du\, dv,$

$$E = x_u^2 + y_u^2 + z_u^2, \quad F = x_u x_v + y_u y_v + z_u z_v, \quad G = x_v^2 + y_v^2 + z_v^2$$

If $z = f(x,y):\quad dA = \sqrt{1 + f_x^2 + f_y^2}\, dx\, dy$

VOLUME ELEMENT: $$dV = \left|\frac{\partial(x,y,z)}{\partial(u,v,w)}\right| du\, dv\, dw = \left\| \begin{matrix} x_u & x_v & x_w \\ y_u & y_v & y_w \\ z_u & z_v & z_w \end{matrix} \right\| du\, dv\, dw$$

LEIBNIZ RULE: $$\frac{d}{dt}\int_{a(t)}^{b(t)} f(x,t)\, dx = \int_{a(t)}^{b(t)} \frac{\partial f}{\partial t}\, dx + b'(t) f(b(t),t) - a'(t) f(a(t),t)$$

FOURIER SERIES:

$f(x)$ 2ℓ-periodic: $$f(x) = a_0 + \sum_{n=1}^{\infty}\left(a_n \cos\frac{n\pi x}{\ell} + b_n \sin\frac{n\pi x}{\ell}\right)$$

$$a_0 = \frac{1}{2\ell}\int_{-\ell}^{\ell} f(x)\, dx, \quad a_n = \frac{1}{\ell}\int_{-\ell}^{\ell} f(x)\cos\frac{n\pi x}{\ell}\, dx, \quad b_n = \frac{1}{\ell}\int_{-\ell}^{\ell} f(x)\sin\frac{n\pi x}{\ell}\, dx$$

$f(x)$ defined only on $0 < x < L$:

HRC: $$f(x) = a_0 + \sum_{n=1}^{\infty} a_n \cos\frac{n\pi x}{L}, \qquad a_0 = \frac{1}{L}\int_0^L f(x)\, dx, \quad a_n = \frac{2}{L}\int_0^L f(x)\cos\frac{n\pi x}{L}\, dx$$

HRS: $$f(x) = \sum_{n=1}^{\infty} b_n \sin\frac{n\pi x}{L}, \qquad b_n = \frac{2}{L}\int_0^L f(x)\sin\frac{n\pi x}{L}\, dx$$

QRC: $$f(x) = \sum_{n=1,3,\ldots}^{\infty} a_n \cos\frac{n\pi x}{2L}, \qquad a_n = \frac{2}{L}\int_0^L f(x)\cos\frac{n\pi x}{2L}\, dx$$

QRS: $$f(x) = \sum_{n=1,3,\ldots}^{\infty} b_n \sin\frac{n\pi x}{2L}, \qquad b_n = \frac{2}{L}\int_0^L f(x)\sin\frac{n\pi x}{2L}\, dx$$

FOURIER INTEGRAL: $$f(x) = \int_0^{\infty} [a(\omega)\cos\omega x + b(\omega)\sin\omega x]\, d\omega \qquad (-\infty < x < \infty)$$

$$a(\omega) = \frac{1}{\pi}\int_{-\infty}^{\infty} f(x)\cos\omega x, \quad b(\omega) = \frac{1}{\pi}\int_{-\infty}^{\infty} f(x)\sin\omega x$$

FOURIER TRANSFORM: $$F\{f(x)\} = \hat{f}(\omega) = \int_{-\infty}^{\infty} f(x) e^{-i\omega x}\, dx$$

$$F^{-1}\{\hat{f}(\omega)\} = f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{f}(\omega) e^{i\omega x}\, d\omega$$

LAPLACE TRANSFORM: $$L\{f(t)\} = \overline{f}(s) = \int_0^{\infty} f(t) e^{-st}\, dt$$

$$L^{-1}\{\overline{f}(s)\} = f(t) = \frac{1}{2\pi i}\int_{\gamma - i\infty}^{\gamma + i\infty} \overline{f}(s) e^{st}\, ds$$

DIVERGENCE THEOREM: $\int_{\mathcal{V}} \nabla \cdot \mathbf{v}\, dV = \int_{\mathcal{S}} \hat{\mathbf{n}} \cdot \mathbf{v}\, dA$

GREEN'S FIRST IDENTITY: $\int_{\mathcal{V}} (\nabla u \cdot \nabla v + u \nabla^2 v)\, dV = \int_{\mathcal{S}} u \frac{\partial v}{\partial n}\, dA$

GREEN'S SECOND IDENTITY: $\int_{\mathcal{V}} (u \nabla^2 v - v \nabla^2 u)\, dV = \int_{\mathcal{S}} \left(u \frac{\partial v}{\partial n} - v \frac{\partial u}{\partial n} \right) dA$

STOKES'S THEOREM: $\int_{\mathcal{S}} \hat{\mathbf{n}} \cdot \nabla \times \mathbf{v}\, dA = \oint_{\mathcal{C}} \mathbf{v} \cdot d\mathbf{R}$

GREEN'S THEOREM: $\int_{\mathcal{S}} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA = \oint_{\mathcal{C}} P\, dx + Q\, dy$

STURM–LIOUVILLE EQUATION: $[p(x)y']' + q(x)y + \lambda w(x) y = 0$

SINE INTEGRAL FUNCTION: $Si(x) = \int_0^x \frac{\sin t}{t}\, dt, \qquad Si(\infty) = \frac{\pi}{2}$

EXPONENTIAL INTEGRAL FUNCTION: $E_1(x) = \int_x^\infty \frac{e^{-t}}{t}\, dt \quad (x > 0)$

GAMMA FUNCTION: $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t}\, dt \quad (x > 0)$

ERROR FUNCTION: $\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt, \quad \operatorname{erf}(\infty) = 1$

TRIGONOMETRIC FUNCTION IDENTITIES:

$$\cos x = \frac{e^{ix} + e^{-ix}}{2}, \quad \sin x = \frac{e^{ix} - e^{-ix}}{2i}$$
$$\cos(ix) = \cosh x, \quad \sin(ix) = i \sinh x$$
$$\cos^2 x + \sin^2 x = 1$$
$$\cos(A \pm B) = \cos A \cos B \mp \sin A \sin B$$
$$\sin(A \pm B) = \sin A \cos B \pm \sin B \cos A$$
$$\cos A \cos B = [\cos(A+B) + \cos(A-B)]/2$$
$$\sin A \cos B = [\sin(A+B) + \sin(A-B)]/2$$
$$\sin A \sin B = [\cos(A-B) - \cos(A+B)]/2$$

HYPERBOLIC FUNCTION IDENTITIES:

$$\cosh x = \frac{e^{x} + e^{-x}}{2}, \quad \sinh x = \frac{e^{x} - e^{-x}}{2}$$
$$\cosh(ix) = \cos x, \quad \sinh(ix) = i \sin x$$
$$\cosh^2 x - \sinh^2 x = 1$$
$$\cosh(A \pm B) = \cosh A \cosh B \pm \sinh A \sinh B$$
$$\sinh(A \pm B) = \sinh A \cosh B \pm \sinh B \cosh A$$

TAYLOR SERIES: $f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 + \cdots$

$$\frac{1}{1-x} = 1 + x + x^2 + \cdots, \qquad |x| < 1 \quad \text{(Geometric Series)}$$
$$e^x = 1 + x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \cdots, \qquad |x| < \infty$$
$$\cos x = 1 - \frac{1}{2!}x^2 + \frac{1}{4!}x^4 - \cdots, \qquad |x| < \infty$$
$$\sin x = x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 - \cdots, \qquad |x| < \infty$$

Advanced Engineering Mathematics

SECOND EDITION

Michael D. Greenberg
Department of Mechanical Engineering
University of Delaware, Newark, Delaware

PRENTICE HALL
Upper Saddle River, New Jersey 07458

Library of Congress Cataloging-in-Publication Data

Greenberg, Michael D., date–
Advanced engineering mathematics / Michael D. Greenberg. — 2nd ed.
p. cm.
Includes bibliographical references and index.
ISBN 0-13-321431-1
1. Engineering mathematics. I. Title.
TA330.G725 1998
515'.14--dc21 97-43585
CIP

Technical Consultant: Dr. E. Murat Sozer

Acquisition editor: George Lobell
Editorial director: Tim Bozik
Editor-in-chief: Jerome Grant
Editorial assistant: Gale Epps
Executive managing editor: Kathleen Schiaparelli
Managing editor: Linda Mihatov Behrens
Production editor: Nick Romanelli
Director of creative services: Paula Maylahn
Art manager: Gus Vibal
Art director / cover designer: Jayne Conte
Cover photos: Timothy Hursley
Marketing manager: Melody Marcus
Marketing assistant: Jennifer Pan
Assistant vice president of production and manufacturing: David Riccardi
Manufacturing buyer: Alan Fischer

© 1998, 1988 by Prentice-Hall, Inc.
Simon & Schuster / A Viacom Company
Upper Saddle River, New Jersey 07458

All rights reserved. No part of this book may be reproduced, in any form or by any means, without permission in writing from the publisher.

Printed in the United States of America.

ISBN 0-13-321431-1

Prentice-Hall International (UK) Limited, London
Prentice-Hall of Australia Pty. Limited, Sydney
Prentice-Hall Canada, Inc., Toronto
Prentice-Hall Hispanoamericana, S.A., Mexico
Prentice-Hall of India Private Limited, New Delhi
Prentice-Hall of Japan, Inc., Tokyo
Simon & Schuster Asia Pte. Ltd., Singapore
Editora Prentice-Hall do Brasil, Ltda., Rio de Janeiro

Advanced
Engineering Mathematics

Contents

Part I: Ordinary Differential Equations

1 INTRODUCTION TO DIFFERENTIAL EQUATIONS 1

1.1 Introduction 1
1.2 Definitions 2
1.3 Introduction to Modeling 9

2 EQUATIONS OF FIRST ORDER 18

2.1 Introduction 18
2.2 The Linear Equation 19
2.2.1 Homogeneous case 19
2.2.2 Integrating factor method 22
2.2.3 Existence and uniqueness for the linear equation 25
2.2.4 Variation-of-parameter method 27
2.3 Applications of the Linear Equation 34
2.3.1 Electrical circuits 34
2.3.2 Radioactive decay; carbon dating 39
2.3.3 Population dynamics 41
2.3.4 Mixing problems 42
2.4 Separable Equations 46
2.4.1 Separable equations 46
2.4.2 Existence and uniqueness (optional) 48
2.4.3 Applications 53
2.4.4 Nondimensionalization (optional) 56
2.5 Exact Equations and Integrating Factors 62
2.5.1 Exact differential equations 62
2.5.2 Integrating factors 66
Chapter 2 Review 71

3 LINEAR DIFFERENTIAL EQUATIONS OF SECOND ORDER AND HIGHER 73

3.1 Introduction 73
3.2 Linear Dependence and Linear Independence 76

3.3 Homogeneous Equation: General Solution 83
3.3.1 General solution 83
3.3.2 Boundary-value problems 88
3.4 Solution of Homogeneous Equation: Constant Coefficients 91
3.4.1 Euler's formula and review of the circular and hyperbolic functions 91
3.4.2 Exponential solutions 95
3.4.3 Higher-order equations ($n > 2$) 99
3.4.4 Repeated roots 102
3.4.5 Stability 105
3.5 Application to Harmonic Oscillator: Free Oscillation 110
3.6 Solution of Homogeneous Equation: Nonconstant Coefficients 117
3.6.1 Cauchy–Euler equation 118
3.6.2 Reduction of order (optional) 123
3.6.3 Factoring the operator (optional) 126
3.7 Solution of Nonhomogeneous Equation 133
3.7.1 General solution 134
3.7.2 Undetermined coefficients 136
3.7.3 Variation of parameters 141
3.7.4 Variation of parameters for higher-order equations (optional) 144
3.8 Application to Harmonic Oscillator: Forced Oscillation 149
3.8.1 Undamped case 149
3.8.2 Damped case 152
3.9 Systems of Linear Differential Equations 156
3.9.1 Examples 157
3.9.2 Existence and uniqueness 160
3.9.3 Solution by elimination 162
Chapter 3 Review 171

4 POWER SERIES SOLUTIONS 173

4.1 Introduction 173
4.2 Power Series Solutions 176
4.2.1 Review of power series 176
4.2.2 Power series solution of differential equations 182
4.3 The Method of Frobenius 193
4.3.1 Singular points 193
4.3.2 Method of Frobenius 195
4.4 Legendre Functions 212
4.4.1 Legendre polynomials 212
4.4.2 Orthogonality of the P_n's 214
4.4.3 Generating functions and properties 215
4.5 Singular Integrals; Gamma Function 218
4.5.1 Singular integrals 218
4.5.2 Gamma function 223
4.5.3 Order of magnitude 225
4.6 Bessel Functions 230
4.6.1 $\nu \neq$ integer 231

4.6.2 $v =$ integer 233
4.6.3 General solution of Bessel equation 235
4.6.4 Hankel functions (optional) 236
4.6.5 Modified Bessel equation 236
4.6.6 Equations reducible to Bessel equations 238
Chapter 4 Review 245

5 LAPLACE TRANSFORM 247

5.1 Introduction 247
5.2 Calculation of the Transform 248
5.3 Properties of the Transform 254
5.4 Application to the Solution of Differential Equations 261
5.5 Discontinuous Forcing Functions; Heaviside Step Function 269
5.6 Impulsive Forcing Functions; Dirac Impulse Function (Optional) 275
5.7 Additional Properties 281
Chapter 5 Review 290

6 QUANTITATIVE METHODS: NUMERICAL SOLUTION OF DIFFERENTIAL EQUATIONS 292

6.1 Introduction 292
6.2 Euler's Method 293
6.3 Improvements: Midpoint Rule and Runge–Kutta 299
6.3.1 Midpoint rule 299
6.3.2 Second-order Runge–Kutta 302
6.3.3 Fourth-order Runge–Kutta 304
6.3.4 Empirical estimate of the order (optional) 307
6.3.5 Multi-step and predictor-corrector methods (optional) 308
6.4 Application to Systems and Boundary-Value Problems 313
6.4.1 Systems and higher-order equations 313
6.4.2 Linear boundary-value problems 317
6.5 Stability and Difference Equations 323
6.5.1 Introduction 323
6.5.2 Stability 324
6.5.3 Difference equations (optional) 328
Chapter 6 Review 335

7 QUALITATIVE METHODS: PHASE PLANE AND NONLINEAR DIFFERENTIAL EQUATIONS 337

7.1 Introduction 337
7.2 The Phase Plane 338
7.3 Singular Points and Stability 348
7.3.1 Existence and uniqueness 348
7.3.2 Singular points 350
7.3.3 The elementary singularities and their stability 352
7.3.4 Nonelementary singularities 357
7.4 Applications 359

7.4.1 Singularities of nonlinear systems 360
7.4.2 Applications 363
7.4.3 Bifurcations 368
7.5 Limit Cycles, van der Pol Equation, and the Nerve Impulse 372
7.5.1 Limit cycles and the van der Pol equation 372
7.5.2 Application to the nerve impulse and visual perception 375
7.6 The Duffing Equation: Jumps and Chaos 380
7.6.1 Duffing equation and the jump phenomenon 380
7.6.2 Chaos 383
Chapter 7 Review 389

Part II: Linear Algebra

8 SYSTEMS OF LINEAR ALGEBRAIC EQUATIONS; GAUSS ELIMINATION 391

8.1 Introduction 391
8.2 Preliminary Ideas and Geometrical Approach 392
8.3 Solution by Gauss Elimination 396
8.3.1 Motivation 396
8.3.2 Gauss elimination 401
8.3.3 Matrix notation 402
8.3.4 Gauss–Jordan reduction 404
8.3.5 Pivoting 405
Chapter 8 Review 410

9 VECTOR SPACE 412

9.1 Introduction 412
9.2 Vectors; Geometrical Representation 412
9.3 Introduction of Angle and Dot Product 416
9.4 n-Space 418
9.5 Dot Product, Norm, and Angle for n-Space 421
9.5.1 Dot product, norm, and angle 421
9.5.2 Properties of the dot product 423
9.5.3 Properties of the norm 425
9.5.4 Orthogonality 426
9.5.5 Normalization 427
9.6 Generalized Vector Space 430
9.6.1 Vector space 430
9.6.2 Inclusion of inner product and/or norm 433
9.7 Span and Subspace 439
9.8 Linear Dependence 444
9.9 Bases, Expansions, Dimension 448
9.9.1 Bases and expansions 448
9.9.2 Dimension 450
9.9.3 Orthogonal bases 453
9.10 Best Approximation 457

9.10.1 Best approximation and orthogonal projection 458
9.10.2 Kronecker delta 461
Chapter 9 Review 462

10 MATRICES AND LINEAR EQUATIONS 465

10.1 Introduction 465
10.2 Matrices and Matrix Algebra 465
10.3 The Transpose Matrix 481
10.4 Determinants 486
10.5 Rank; Application to Linear Dependence and to Existence and Uniqueness for $\mathbf{Ax} = \mathbf{c}$ 495
10.5.1 Rank 495
10.5.2 Application of rank to the system $\mathbf{Ax} = \mathbf{c}$ 500
10.6 Inverse Matrix, Cramer's Rule, Factorization 508
10.6.1 Inverse matrix 508
10.6.2 Application to a mass-spring system 514
10.6.3 Cramer's rule 517
10.6.4 Evaluation of $\mathbf{A}^{-1}$ by elementary row operations 518
10.6.5 LU-factorization 520
10.7 Change of Basis (Optional) 526
10.8 Vector Transformation (Optional) 530
Chapter 10 Review 539

11 THE EIGENVALUE PROBLEM 541

11.1 Introduction 541
11.2 Solution Procedure and Applications 542
11.2.1 Solution and applications 542
11.2.2 Application to elementary singularities in the phase plane 549
11.3 Symmetric Matrices 554
11.3.1 Eigenvalue problem $\mathbf{Ax} = \lambda\mathbf{x}$ 554
11.3.2 Nonhomogeneous problem $\mathbf{Ax} = \lambda\mathbf{x} + \mathbf{c}$ (optional) 561
11.4 Diagonalization 569
11.5 Application to First-Order Systems with Constant Coefficients (optional) 583
11.6 Quadratic Forms (Optional) 589
Chapter 11 Review 596

12 EXTENSION TO COMPLEX CASE (OPTIONAL) 599

12.1 Introduction 599
12.2 Complex n-Space 599
12.3 Complex Matrices 603
Chapter 12 Review 611

Part III: Scalar and Vector Field Theory

13 DIFFERENTIAL CALCULUS OF FUNCTIONS OF SEVERAL VARIABLES 613

13.1 Introduction 613
13.2 Preliminaries 614
13.2.1 Functions 614
13.2.2 Point set theory definitions 614
13.3 Partial Derivatives 620
13.4 Composite Functions and Chain Differentiation 625
13.5 Taylor's Formula and Mean Value Theorem 629
13.5.1 Taylor's formula and Taylor series for $f(x)$ 630
13.5.2 Extension to functions of more than one variable 636
13.6 Implicit Functions and Jacobians 642
13.6.1 Implicit function theorem 642
13.6.2 Extension to multivariable case 645
13.6.3 Jacobians 649
13.6.4 Applications to change of variables 652
13.7 Maxima and Minima 656
13.7.1 Single variable case 656
13.7.2 Multivariable case 658
13.7.3 Constrained extrema and Lagrange multipliers 665
13.8 Leibniz Rule 675
Chapter 13 Review 681

14 VECTORS IN 3-SPACE 683

14.1 Introduction 683
14.2 Dot and Cross Product 683
14.3 Cartesian Coordinates 687
14.4 Multiple Products 692
14.4.1 Scalar triple product 692
14.4.2 Vector triple product 693
14.5 Differentiation of a Vector Function of a Single Variable 695
14.6 Non-Cartesian Coordinates (Optional) 699
14.6.1 Plane polar coordinates 700
14.6.2 Cylindrical coordinates 704
14.6.3 Spherical coordinates 705
14.6.4 Omega method 707
Chapter 14 Review 712

15 CURVES, SURFACES, AND VOLUMES 714

15.1 Introduction 714
15.2 Curves and Line Integrals 714
15.2.1 Curves 714
15.2.2 Arc length 716
15.2.3 Line integrals 718
15.3 Double and Triple Integrals 723
15.3.1 Double integrals 723
15.3.2 Triple integrals 727
15.4 Surfaces 733

15.4.1 Parametric representation of surfaces 733
15.4.2 Tangent plane and normal 734
15.5 Surface Integrals 739
15.5.1 Area element dA 739
15.5.2 Surface integrals 743
15.6 Volumes and Volume Integrals 748
15.6.1 Volume element dV 749
15.6.2 Volume integrals 752
Chapter 15 Review 755

16 SCALAR AND VECTOR FIELD THEORY 757

16.1 Introduction 757
16.2 Preliminaries 758
16.2.1 Topological considerations 758
16.2.2 Scalar and vector fields 758
16.3 Divergence 761
16.4 Gradient 766
16.5 Curl 774
16.6 Combinations; Laplacian 778
16.7 Non-Cartesian Systems; Div, Grad, Curl, and Laplacian (Optional) 782
16.7.1 Cylindrical coordinates 783
16.7.2 Spherical coordinates 786
16.8 Divergence Theorem 792
16.8.1 Divergence theorem 792
16.8.2 Two-dimensional case 802
16.8.3 Non-Cartesian coordinates (optional) 803
16.9 Stokes's Theorem 810
16.9.1 Line integrals 814
16.9.2 Stokes's theorem 814
16.9.3 Green's theorem 818
16.9.4 Non-Cartesian coordinates (optional) 820
16.10 Irrotational Fields 826
16.10.1 Irrotational fields 826
16.10.2 Non-Cartesian coordinates 835
Chapter 16 Review 841

Part IV: Fourier Methods and Partial Differential Equations

17 FOURIER SERIES, FOURIER INTEGRAL, FOURIER TRANSFORM 844

17.1 Introduction 844
17.2 Even, Odd, and Periodic Functions 846
17.3 Fourier Series of a Periodic Function 850
17.3.1 Fourier series 850
17.3.2 Euler's formulas 857
17.3.3 Applications 859

17.3.4 Complex exponential form for Fourier series 864
17.4 Half- and Quarter-Range Expansions 869
17.5 Manipulation of Fourier Series (Optional) 873
17.6 Vector Space Approach 881
17.7 The Sturm–Liouville Theory 887
17.7.1 Sturm–Liouville problem 887
17.7.2 Lagrange identity and proofs (optional) 897
17.8 Periodic and Singular Sturm–Liouville Problems 905
17.9 Fourier Integral 913
17.10 Fourier Transform 919
17.10.1 Transition from Fourier integral to Fourier transform 920
17.10.2 Properties and applications 922
17.11 Fourier Cosine and Sine Transforms, and Passage from Fourier Integral to Laplace Transform (Optional) 934
17.11.1 Cosine and sine transforms 934
17.11.2 Passage from Fourier integral to Laplace transform 937
Chapter 17 Review 940

18 DIFFUSION EQUATION 943

18.1 Introduction 943
18.2 Preliminary Concepts 944
18.2.1 Definitions 944
18.2.2 Second-order linear equations and their classification 946
18.2.3 Diffusion equation and modeling 948
18.3 Separation of Variables 954
18.3.1 The method of separation of variables 954
18.3.2 Verification of solution (optional) 964
18.3.3 Use of Sturm–Liouville theory (optional) 965
18.4 Fourier and Laplace Transforms (Optional) 981
18.5 The Method of Images (Optional) 992
18.5.1 Illustration of the method 992
18.5.2 Mathematical basis for the method 994
18.6 Numerical Solution 998
18.6.1 The finite-difference method 998
18.6.2 Implicit methods: Crank–Nicolson, with iterative solution (optional) 1005
Chapter 18 Review 1015

19 WAVE EQUATION 1017

19.1 Introduction 1017
19.2 Separation of Variables; Vibrating String 1023
19.2.1 Solution by separation of variables 1023
19.2.2 Traveling wave interpretation 1027
19.2.3 Using Sturm–Liouville theory (optional) 1029
19.3 Separation of Variables; Vibrating Membrane 1035
19.4 Vibrating String: d'Alembert's Solution 1043
19.4.1 d'Alembert's solution 1043

19.4.2 Use of images 1049
19.4.3 Solution by integral transforms (optional) 1051
Chapter 19 Review 1055

20 LAPLACE EQUATION 1058

20.1 Introduction 1058
20.2 Separation of Variables; Cartesian Coordinates 1059
20.3 Separation of Variables; Non-Cartesian Coordinates 1070
20.3.1 Plane polar coordinates 1070
20.3.2 Cylindrical coordinates (optional) 1077
20.3.3 Spherical coordinates (optional) 1081
20.4 Fourier Transform (Optional) 1088
20.5 Numerical Solution 1092
20.5.1 Rectangular domains 1092
20.5.2 Nonrectangular domains 1097
20.5.3 Iterative algorithms (optional) 1100
Chapter 20 Review 1106

Part V: Complex Variable Theory

21 FUNCTIONS OF A COMPLEX VARIABLE 1108

21.1 Introduction 1108
21.2 Complex Numbers and the Complex Plane 1109
21.3 Elementary Functions 1114
21.3.1 Preliminary ideas 1114
21.3.2 Exponential function 1116
21.3.3 Trigonometric and hyperbolic functions 1118
21.3.4 Application of complex numbers to integration and the solution of differential equations 1120
21.4 Polar Form, Additional Elementary Functions, and Multi-valuedness 1125
21.4.1 Polar form 1125
21.4.2 Integral powers of z and de Moivre's formula 1127
21.4.3 Fractional powers 1128
21.4.4 The logarithm of z 1129
21.4.5 General powers of z 1130
21.4.6 Obtaining single-valued functions by branch cuts 1131
21.4.7 More about branch cuts (optional) 1132
21.5 The Differential Calculus and Analyticity 1136
Chapter 21 Review 1148

22 CONFORMAL MAPPING 1150

22.1 Introduction 1150
22.2 The Idea Behind Conformal Mapping 1150
22.3 The Bilinear Transformation 1158

22.4 Additional Mappings and Applications 1166
22.5 More General Boundary Conditions 1170
22.6 Applications to Fluid Mechanics 1174
Chapter 22 Review 1180

23 THE COMPLEX INTEGRAL CALCULUS 1182

23.1 Introduction 1182
23.2 Complex Integration 1182
23.2.1 Definition and properties 1182
23.2.2 Bounds 1186
23.3 Cauchy's Theorem 1189
23.4 Fundamental Theorem of the Complex Integral Calculus 1195
23.5 Cauchy Integral Formula 1199
Chapter 23 Review 1207

24 TAYLOR SERIES, LAURENT SERIES, AND THE RESIDUE THEOREM 1209

24.1 Introduction 1209
24.2 Complex Series and Taylor Series 1209
24.2.1 Complex series 1209
24.2.2 Taylor series 1214
24.3 Laurent Series 1225
24.4 Classification of Singularities 1234
24.5 Residue Theorem 1240
24.5.1 Residue theorem 1240
24.5.2 Calculating residues 1242
24.5.3 Applications of the residue theorem 1243
Chapter 24 Review 1258

REFERENCES 1260

APPENDICES

A Review of Partial Fraction Expansions 1263
B Existence and Uniqueness of Solutions of Systems of Linear Algebraic Equations 1267
C Table of Laplace Transforms 1271
D Table of Fourier Transforms 1274
E Table of Fourier Cosine and Sine Transforms 1276
F Table of Conformal Maps 1278

ANSWERS TO SELECTED EXERCISES 1282

INDEX 1315

Preface

Purpose and Prerequisites

This book is written primarily for a single- or multi-semester course in applied mathematics for students of engineering or science, but it is also designed for self-study and reference. By self-study we do not necessarily mean outside the context of a formal course. Even within a course setting, if the text can be read independently and understood, then more pedagogical options become available to the instructor.

The prerequisite is a full year sequence in calculus, but the book is written so as to be usable at both the undergraduate level and also for first-year graduate students of engineering and science. The flexibility that permits this broad range of use is described below in the section on Course Use.

Changes from the First Edition

Principal changes from the first edition are as follows:

1. **Part I on ordinary differential equations.** In the first edition we assumed that the reader had previously completed a first course in ordinary differential equations. However, differential equations is traditionally the first major topic in books on advanced engineering mathematics so we begin this edition with a seven chapter sequence on ordinary differential equations. Just as the book becomes increasingly sophisticated from beginning to end, these seven chapters are written that way as well, with the final chapter on nonlinear equations being the most challenging.

2. **Incorporation of a computer-algebra-system.** Several powerful computer environments are available, such as *Maple*, *Mathematica*, and MATLAB. We selected *Maple*, as a representative and user-friendly software. In addition to an Instructor's Manual, a brief student supplement is also available, which presents parallel discussions of *Mathematica* and MATLAB.

3. **Revision of existing material and format.** Pedagogical improvements that evolved through eight years of class use led to a complete rewriting rather than minor modifications of the text. The end-of-section exercises are refined and expanded.

Format

The book is comprised of five parts:

I Ordinary Differential Equations
II Linear Algebra
III Multivariable Calculus and Field Theory
IV Fourier Methods and Partial Differential Equations
V Complex Variable Theory

This breakdown is explicit only in the Contents, to suggest the major groupings of the chapters. Within the text there are no formal divisions between parts, only between chapters.

Each chapter begins with an introduction and (except for the first chapter) ends with a chapter review. Likewise, each section ends with a review called a closure, which is often followed by a section on computer software that discusses the *Maple* commands that are relevant to the material covered in that section; see, for example, pages 29–31. Subsections are used extensively to offer the instructor more options in terms of skipping or including material.

Course Use at Different Levels

To illustrate how the text might serve at different levels, we begin by outlining how we have been using it for courses at the University of Delaware: a sophomore/junior level mathematics course for mechanical engineers, and a first-year graduate level two-semester sequence in applied mathematics for students of mechanical, civil, and chemical engineering, and materials science. We denote these courses as U, G1, and G2, respectively.

Sophomore/junior level course (U). This course follows the calculus/differential equations sequence taught in the mathematics department. We cover three main topics:

Linear Algebra: Chapter 8, Sections 9.1–9.5 (plus a one lecture overview of Secs. 9.7–9.9), 10.1–10.6, and 11.1–11.3. The focus is n-space and applications, such as the mass-spring system in Sec. 10.6.2, Markov population dynamics in Sec. 11.2, and orthogonal modes of vibration in Sec. 11.3.

Field Theory: Chapters 14 and 16. The heart of this material is Chapter 16. Having skipped Chapter 15, we distribute a one page "handout" on the area element formula (18) in Sec. 15.5 since that formula is needed for the surface integrals that occur in Chapter 16. Emphasis is placed on the physical applications in the sections on the divergence theorem and irrotational fields since those applications lead to two of the three chief partial differential equations that will be studied in the third part of the course—the diffusion equation and the Laplace equation.

Fourier Series and PDE's: Sections 17.1–17.4, 18.1, 18.3, 18.6.1, 19.1–19.2.2, 20.1–20.3.1, 20.5.1–20.5.2. Solutions are by separation of variables, using only the half- and quarter-range Fourier series, and by finite differences.

First semester of graduate level course (G1). Text coverage is as follows: Sections 4.4–4.6, 5.1–5.6, Chapter 9, Secs. 11.1–11.4, 11.6, 13.5–13.8, 14.6, 15.4–15.6, Chapter 16, Secs. 17.3, 17.6–17.11, 18.1–18.3.1, 18.3.3–18.4, 19.1–19.2, 20.1–20.4. As in "U" we do cover the important Chapter 16, although quickly. Otherwise, the approach complements that in "U." For instance, in Chapter 9, "U" focuses on n-space, but "G1" focuses on generalized vector space (Sec. 9.6), to get ready for the Sturm–Liouville theory (Section 17.7); in Chapter 11 we emphasize the more advanced sections on diagonalization and quadratic forms, as well as Section 11.3.2 on the eigenvector expansion method in finite-dimensional space, so we can use that method to solve nonhomogeneous partial differential equations in later chapters. Likewise, in covering Chapter 17 we assume that the student has worked with Fourier series before so we move quickly, emphasizing the vector space approach (Sec. 17.6), the Sturm–Liouville theory, and the Fourier integral and transform. When we come

to partial differential equations we use Sturm–Liouville eigenfunction expansions (rather than the half- and quarter-range formulas that suffice in "U"), integral transforms, delta functions, and Bessel and Legendre functions. In solving the diffusion equation in "U" we work only with the homogeneous equation and constant end conditions, but in "G1" we discuss the nonhomogeneous equation and nonconstant end conditions, uniqueness, and so on; these topics are discussed in the exercises.

Second semester of graduate level course (G2). In the second semester we complete the partial differential equation coverage with the methods of images and Green's functions, then turn to complex variable theory, the variational calculus, and an introduction to perturbation methods. For Green's functions we use a "handout," and for the variational calculus and perturbation methods we copy the relevant chapters from M.D. Greenberg, *Foundations of Applied Mathematics* (Englewood Cliffs, NJ: Prentice Hall, 1978). (If you are interested in using any of these materials please contact the College Mathematics Editor office at Prentice-Hall, Inc., One Lake Street, Upper Saddle River, NJ 07458.)

Text coverage is as follows: Chapters 21–24 on complex variable theory; then we return to PDE's, first covering Secs. 18.5–18.6, 19.3–19.4, and 20.3.2–20.4 that were skipped in "G1"; "handouts" on Green's functions, perturbation methods, and the variational calculus.

Shorter courses and optional Sections. A number of sections and subsections are listed as Optional in the Contents, as a guide to instructors in using this text for shorter or more introductory courses. In the chapters on field theory, for example, one could work only with Cartesian coordinates, and avoid the more difficult non-Cartesian case, by omitting those optional sections. We could have labeled the Sturm–Liouville theory section (17.7) as optional but chose not to, because it is such an important topic. Nonetheless, if one wishes to omit it, as we do in "U," that is possible, since subsequent use of the Sturm–Liouville theory in the PDE chapters is confined to optional sections and exercises.

Let us mention Chapter 4, in particular, since its development of series solutions, the method of Frobenius, and Legendre and Bessel functions might seem more detailed than you have time for in your course. One minimal route is to cover only Sections 4.2.2 on power series solutions of ordinary differential equations (ODE's) and 4.4.1 on Legendre polynomials, since the latter does not depend on the more detailed Frobenius material in Section 4.3. Then one can have Legendre functions available when the Laplace equation is studied in spherical coordinates. You might also want to cover Bessel functions but do not want to use class time to go through the Frobenius material. In my own course ("G1") I deal with Bessel functions by using a "handout" that is simpler and shorter, which complements the more thorough treatment in the text.

Exercises

Exercises are of different kinds and arranged, typically, as follows. First, and usually near the beginning of the exercise group, are exercises that follow up on the text or fill in gaps or relate to proofs of theorems stated in that section, thus engaging the student more fully in the reading (e.g., Exercises 1–3 in Section 7.2, Exercise 8 in Section 16.8). Second, there are usually numerous "drill type" exercises that ask the reader to mimic steps or calculations that are essentially like those demonstrated in the text (e.g., there are 19 matrices to invert by hand in Exercise 1 of Section 10.6, and again by computer software in Exercise 3).

Third, there are exercises that require the use of a computer, usually employing software that is explained at the end of the section or in an earlier section; these vary from drill type (e.g., Exercise 1, Section 10.6) to more substantial calculations (e.g., Exercise 15, Section 19.2). Fourth, there are exercises that involve physical applications (e.g., Exercises 8, 9, and 12 of Section 16.10, on the stream function, the entropy of an ideal gas, and integrating the equation of motion of fluid mechanics to obtain the Bernoulli equation). And, fifth, there are exercises intended to extend the text and increase its value as a reference book. In these, we usually guide the student through the steps so that the exercise becomes more usable for subsequent reference or self-study (e.g., see Exercises 17–22 of Section 18.3). Answers to selected exercises (which are denoted in the text by underlining the exercise number) are provided at the end of the book; a more complete set is available for instructors in the Instructor's Manual.

Specific Pedagogical Decisions

In Chapter 2 we consider the linear first-order equation and then the case of separable first-order equations. It is tempting to reverse the order, as some authors do, but we prefer to elevate the linear/nonlinear distinction, which grows increasingly important in engineering mathematics; to do that, it seems best to begin with the linear equation.

It is stated, at the beginning of Chapter 3 on linear differential equations of second order and higher, that the reader is expected to be familiar with the theory of the existence and uniqueness of solutions of linear algebraic equations, especially the role of the determinant of the coefficient matrix, even though this topic occurs later in the text. The instructor is advised to handle this need either by assigning, as prerequisite reading, the brief summary of the needed information given in Appendix B or, if a tighter blending of the differential equation and linear algebra material is desired, by covering Sections 8.1–10.6 before continuing with Chapter 3. Similarly, it is stated at the beginning of Chapter 3 that an elementary knowledge of the complex plane and complex numbers is anticipated. If the class does not meet that prerequisite, then Section 21.2 should be covered before Chapter 3. Alternatively, we could have made that material the first section of Chapter 3, but it seemed better to keep the major topics together—in this case, to keep the complex variable material together.

Some authors prefer to cover second-order equations in one chapter and then higher-order equations in another. My opinion about that choice is that: (i) it is difficult to grasp clearly the second-order case (especially insofar as the case of repeated roots is concerned) without seeing the extension to higher order, and (ii) the higher-order case can be covered readily, so that it becomes more efficient to cover both cases simultaneously.

Finally, let us explain why Chapter 8, on systems of linear algebraic equations and Gauss elimination, is so brief. Just as one discusses the real number axis before discussing functions that map one real axis to another, it seems best to discuss vectors before discussing matrices, which map one vector space into another. But to discuss vectors, span, linear dependence, bases, and expansions, one needs to know the essentials regarding the existence and uniqueness of solutions of systems of linear algebraic equations. Thus, Chapter 8 is intended merely to suffice until, having introduced matrices in Chapter 10, we can provide a more complete discussion.

Appendices

Appendix A reviews partial fraction expansions, needed in the application of Laplace and Fourier transforms. Appendix B summarizes the theory of the existence and uniqueness of solutions of linear algebraic equations, especially the role of the determinant of the coefficient matrix, and is a minimum prerequisite for Chapter 3. Appendices C through F are tables of transforms and conformal maps.

Instructor's Manual

An Instructor's Manual will be available to instructors from the office of the Mathematics Editor, College Department, Prentice-Hall, Inc., 1 Lake Street, Upper Saddle River, NJ 07458. Besides solutions to exercises, this manual contains additional pedagogical ideas for lecture material and some additional coverage, such as the Fast Fourier Transform, that can be used as "handouts."

Acknowledgements

I am grateful for a great deal of support in writing this second edition, but especially to Dr. E. Murat Sozer, who acted as a Technical Consultant. Dr. Sozer's work on the latex preparation of text and the figures went far beyond his original desire to learn the latex system and whose generous help was always extraordinary in quantity and quality. Sincere thanks also to my mathematics editor, George Lobell, for his insight and support, to Nick Romanelli in charge of production, to Anita Z. Hoover at the University of Delaware for extensive help with the latex computer system, and to these outside reviewers of the developing manuscript: Gregory Adams (Bucknell University), James M. W. Brownjohn (Nanyang Technical University), Melvin G. Davidson (Western Washington University), John H. Ellison (Grove City College), Mark Farris (Midwestern State University), Amitabha Ghosh (Rochester Institute of Technology), Evans M. Harrell, II (Georgia Tech.), Allen Hesse (Rochester Institute of Technology), Chung-yau Lam (Nanyang Technical University), Mohan Manoharan (Nanyang Technical University), James G. McDaniel (Boston University), Carruth McGehee (Lousiana State University), William Moss (Clemson University), Jean-Paul Nicol (Auburn University), John A. Pfaltzgraff (University of North Carolina, Chapel Hill), Mohammad Tavakoli (Chaffey College), David E. Weidner (University of Delaware), and Jingyi Zhu (University of Utah). I also thank these graduate students in this department for their help with working and debugging exercises: Rishikesh Bhalerao, Nadeem Faiz, Santosh Prabhu, and Rahul Rao and Zhaohui Chen.

I'm grateful to my wife, Yisraela, for her deep support and love when this task looked like more than I could handle, and for assuming many of my responsibilities, to give me the needed time. I dedicate this book to her.

Most of all, I am grateful to the Lord for bringing this book back to life and watching over all aspects of its writing and production: " From whence cometh my help? My help cometh from the Lord, who made heaven and earth." (Psalm 121)

Michael D. Greenberg

Chapter 1

Introduction to Differential Equations

1.1 Introduction

The mathematical formulation of problems in engineering and science usually leads to equations involving derivatives of one or more unknown functions. Such equations are called differential equations.

Consider, for instance, the motion of a body of mass m along a straight line, which we designate as an x axis. Let the mass be subjected to a force $F(t)$ along that axis, where t is the time. Then according to Newton's second law of motion

$$m\frac{d^2x}{dt^2} = F(t), \tag{1}$$

where $x(t)$ is the mass's displacement measured from the origin. If we prescribe the displacement $x(t)$ and wish to determine the force $F(t)$ required to produce that displacement, then the solution is simple: according to (1), we merely differentiate the given $x(t)$ twice and multiply by m.

However, if we know the applied force $F(t)$ and wish to determine the displacement $x(t)$ that results, then (1) is a "differential equation" on $x(t)$ since it involves the derivative, more precisely the second derivative, of the unknown function $x(t)$ with respect to t. To solve for x we need to "undo" the differentiations. That is, we need to integrate (1), twice in fact. For definiteness and simplicity, suppose that $F(t) = F_0$ is a constant. Then, integrating (1) once with respect to t gives

$$m\frac{dx}{dt} = F_0 t + A, \tag{2}$$

where A is an arbitrary constant of integration, and integrating again gives

$$mx = \frac{F_0}{2}t^2 + At + B,$$

or,

$$x(t) = \frac{1}{m}\left(\frac{F_0}{2}t^2 + At + B\right). \tag{3}$$

The constants of integration, A and B, can be found from (2) and (3) if the displacement x and velocity dx/dt are prescribed at the initial time ($t = 0$). If both $x(0)$ and $\frac{dx}{dt}(0)$ are zero, for instance, then (by setting $t = 0$) we find from (2) that $A = 0$, and then from (3) that $B = 0$. Thus, (3) gives the solution as $x(t) = F_0t^2/2m$, and this solution holds for all $t \geq 0$.

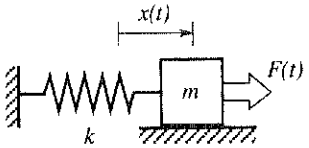


Figure 1. Mass/spring system.

Unfortunately, most differential equations cannot be solved this easily, that is, by merely undoing the derivatives. For instance, suppose that the mass is restrained by a coil spring that supplies a restoring force proportional to the displacement x, with constant of proportionality k (Fig. 1). Then in place of (1), the differential equation governing the motion is

$$m\frac{d^2x}{dt^2} = -kx + F(t)$$

or,

$$m\frac{d^2x}{dt^2} + kx = F(t). \tag{4}$$

After one integration, (4) becomes

$$m\frac{dx}{dt} + k\int x(t)\,dt = \int F(t)\,dt + A, \tag{5}$$

where A is the constant of integration. Since $F(t)$ is a prescribed function, the integral of $F(t)$ can be evaluated, but since $x(t)$ is the unknown, the integral of $x(t)$ cannot be evaluated, and we cannot proceed with our solution–by–integration.

Thus, we see that solving differential equations is not merely a matter of undoing the derivatives by direct integration. Indeed, the theory and technique involved is considerable, and will occupy us for these first seven chapters.

1.2 Definitions

In this section we introduce some of the basic terminology.

Differential equation. By a **differential equation** we mean an equation containing one or more derivatives of the function under consideration. Here are some examples of differential equations that we will study in later chapters:

$$m\frac{d^2x}{dt^2} + kx = F(t), \tag{1}$$

$$L\frac{d^2i}{dt^2} + \frac{1}{C}i = \frac{dE}{dt}, \tag{2}$$

$$\frac{d^2\theta}{dt^2} + \frac{g}{l}\sin\theta = 0, \tag{3}$$

$$\frac{dx}{dt} = cx, \tag{4}$$

$$\frac{d^2y}{dx^2} = C\sqrt{1 + \left(\frac{dy}{dx}\right)^2}, \tag{5}$$

$$EI\frac{d^4y}{dx^4} = -w(x). \tag{6}$$

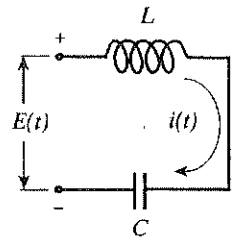


Figure 1. Electrical circuit, equation (2).

Equation (1) is the differential equation governing the linear displacement $x(t)$ of a body of mass m, subjected to an applied force $F(t)$ and a restraining spring of stiffness k, as mentioned in the preceding section.

Equation (2) governs the current $i(t)$ in an electrical circuit containing an inductor with inductance L, a capacitor with capacitance C, and an applied voltage source of strength $E(t)$ (Fig. 1), where t is the time.

Equation (3) governs the angular motion $\theta(t)$ of a pendulum of length l, under the action of gravity, where g is the acceleration of gravity and t is the time (Fig. 2).

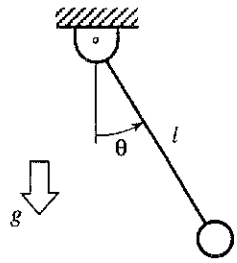


Figure 2. Pendulum, equation (3).

Equation (4) governs the population $x(t)$ of a single species, where t is the time and c is a net birth/death rate constant.

Equation (5) governs the shape of a flexible cable or string, hanging under the action of gravity, where $y(x)$ is the deflection and C is a constant that depends upon the mass density of the cable and the tension at the midpoint $x = 0$ (Fig. 3).

Finally, equation (6) governs the deflection $y(x)$ of a beam subjected to a loading $w(x)$ (Fig. 4), where E and I are physical constants of the beam material and cross section, respectively.

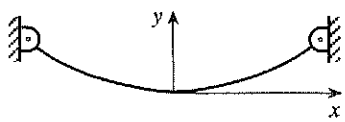


Figure 3. Hanging cable, equation (5).

Ordinary and partial differential equations. We classify a differential equation as an **ordinary differential equation** if it contains ordinary derivatives with respect to a single independent variable, and as a **partial differential equation** if it contains partial derivatives with respect to two or more independent variables. Thus, equations (1)–(6) are ordinary differential equations (often abbreviated as **ODE**'s). The independent variable is t in (1)–(4) and x in (5) and (6).

Some representative and important partial differential equations (**PDE**'s) are as follows:

$$\alpha^2\frac{\partial^2 u}{\partial x^2} = \frac{\partial u}{\partial t}, \tag{7}$$

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = 0, \tag{8}$$

$$c^2\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) = \frac{\partial^2 u}{\partial t^2}, \tag{9}$$

$$\frac{\partial^4 u}{\partial x^4} + 2\frac{\partial^4 u}{\partial x^2 \partial y^2} + \frac{\partial^4 u}{\partial y^4} = 0. \tag{10}$$

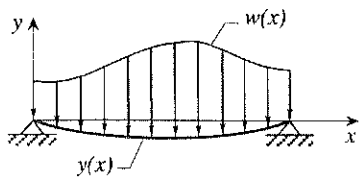


Figure 4. Loaded beam, equation (6).

Equation (7) is the *heat equation*, governing the time-varying temperature distribution $u(x,t)$ in a one-dimensional rod or slab; x locates the point under consideration within the material, t is the time, and α^2 is a material property called the diffusivity.

Equation (8) is the *Laplace equation*, governing the steady-state temperature distribution $u(x,y,z)$ within a three-dimensional body; x, y, z are the coordinates of the point within the material.

Equation (9) is the *wave equation*, governing the deflection $u(x,y,t)$ of a vibrating membrane such as a drum head.

Equation (10) is the *biharmonic equation*, governing the stream function $u(x,y)$ in the case of the slow (creeping) motion of a viscous fluid such as a film of wet paint.

Besides the possibility of having more than one independent variable, there could be more than one dependent variable. For instance,

$$\begin{aligned}
\frac{dx_1}{dt} &= -(k_{21}+k_{31})x_1 + k_{12}x_2 + k_{13}x_3 \\
\frac{dx_2}{dt} &= k_{21}x_1 - (k_{12}+k_{32})x_2 + k_{32}x_3 \\
\frac{dx_3}{dt} &= k_{31}x_1 + k_{32}x_2 - (k_{13}+k_{23})x_3
\end{aligned} \tag{11}$$

is a set, or system, of three ODE's governing the three unknowns $x_1(t), x_2(t), x_3(t)$; (11) arises in chemical kinetics, where x_1, x_2, x_3 are concentrations of three reacting chemical species, such as in a combustion chamber, where the k_{ij}'s are reaction rate constants, and where the reactions are, in chemical jargon, first-order reactions. Similarly,

$$\begin{aligned}
\frac{\partial E_2}{\partial x} - \frac{\partial E_1}{\partial y} &= 0 \\
\frac{\partial E_1}{\partial x} + \frac{\partial E_2}{\partial y} &= \frac{\sigma(x,y)}{\epsilon}
\end{aligned} \tag{12}$$

is a system of two PDE's governing the two unknowns $E_1(x,y)$ and $E_2(x,y)$, which are the x and y components of the electric field intensity, respectively, $\sigma(x,y)$ is the charge distribution density, and ϵ is the permittivity; these are the Maxwell's equations governing two-dimensional electrostatics.

At this point, we limit our subsequent attention to *ordinary* differential equations. We will not return to partial differential equations until much later on in this book. Thus, we will generally omit the adjective "ordinary," for brevity, and will speak only of "differential equations" over the next several chapters.

Order. We define the **order** of a differential equation as the order of the highest derivative therein. Thus, (4) is of first order, (1), (2), (3), and (5) are of second order, (6) is of fourth order, and (11) is a system of first-order ODE's.

More generally,

$$F\left(x, u(x), u'(x), u''(x), \ldots, u^{(n)}(x)\right) = 0 \tag{13}$$

is said to be an nth-order differential equation on the unknown $u(x)$, where n is the order of the highest derivative present in (13). Here, we use the standard prime notation for derivatives: $u'(x)$ denotes du/dx, $u''(x)$ denotes the second derivative, ..., and $u^{(n)}(x)$ denotes the nth derivative. In the fourth-order differential equation (6), for instance, in which the dependent variable is y rather than u, $F(x, y, y', y'', y''', y'''') = EIy'''' + w(x)$, which happens not to contain y, y', y'', or y'''.

Solution. A function is said to be a **solution** of a differential equation, over a particular domain of the independent variable, if its substitution into the equation reduces that equation to an identity everywhere within that domain.

EXAMPLE 1. The function $y(x) = 4\sin x - x\cos x$ is a solution of the differential equation

$$y'' + y = 2\sin x \tag{14}$$

on the entire x axis because its substitution into (14) yields

$$(-4\sin x + 2\sin x + x\cos x) + (4\sin x - x\cos x) = 2\sin x,$$

which is an identity for all x. Note that we said "a" solution rather than "the" solution since there are many solutions of (14):

$$y(x) = A\sin x + B\cos x - x\cos x \tag{15}$$

is a solution for *any* values of the constants A and B, as is verified by substitution of (15) into (14). [In a later chapter, we will be in a position to derive the solution (15), and to show that it is the most general possible solution, that is, that every solution of (14) can be expressed in the form (15).] ∎

EXAMPLE 2. The function $y(x) = 1/x$ is a solution of the differential equation

$$y' + y^2 = 0 \tag{16}$$

over any interval that does not contain the origin since its substitution into (16) gives $-1/x^2 + 1/x^2 = 0$, which relation is an identity, provided that $x \neq 0$. ∎

EXAMPLE 3. Whereas (14) admits an *infinity* of solutions [one for each choice of A and B in (15)], the equation

$$|y'| + |y| + 3 = 0 \tag{17}$$

evidently admits *none* since the two nonnegative terms and one positive term cannot possibly sum to zero for any choice of $y(x)$. ∎

In applications, however, one normally expects that if a problem is formulated carefully then it should indeed have a solution, and that the solution should be

unique, that is, there should be one and only one. Thus, the issues of *existence* (Does the equation have any solution?) and *uniqueness* (If it does have a solution, is that solution unique?) are of important interest.

Initial-value problems and boundary-value problems. Besides the differential equation to be satisfied, the unknown function is often subjected to conditions at one or more points on the interval under consideration. Conditions specified at a single point (often the left end point of the interval), are called **initial conditions**, and the differential equation together with those initial conditions is called an **initial-value problem**. Conditions specified at both ends are called **boundary conditions**, and the differential equation together with the boundary conditions is called a **boundary-value problem**. For initial-value problems the independent variable is often the time, though not necessarily, and for boundary-value problems the independent variable is often a space variable.

EXAMPLE 4. *Straight-Line Motion of a Mass.* Consider once again the problem of predicting the straight-line motion of a body of mass m subjected to a force $F(t)$. According to Newton's second law of motion, the governing differential equation on the displacement $x(t)$ is $mx'' = F(t)$. Besides the differential equation, suppose that we wish to impose the conditions $x(0) = 0$ and $x'(0) = V$; that is, the initial displacement and velocity are 0 and V, respectively. Then the complete problem statement is the initial-value problem

$$\begin{aligned} mx''(t) &= F(t), \qquad (0 \le t < \infty) \\ x(0) &= 0, \quad x'(0) = V. \end{aligned} \tag{18}$$

That is, $x(t)$ is to satisfy the differential equation $mx'' = F(t)$ on the interval $0 \le t < \infty$ *and* the initial conditions $x(0) = 0$ and $x'(0) = V$. ∎

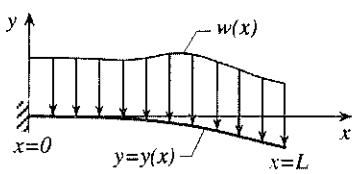


Figure 5. Loaded cantilever beam.

EXAMPLE 5. *Deflection of a Loaded Cantilever Beam.* Consider the deflection $y(x)$ of a cantilever beam of length L, under a loading $w(x)$ newtons per meter (Fig. 5). Using the so-called Euler beam theory, one finds that the governing problem is as follows:

$$\begin{aligned} EIy'''' &= -w(x) \qquad (0 \le x \le L) \\ y(0) = 0, \quad y'(0) &= 0, \quad y''(L) = 0, \quad y'''(L) = 0, \end{aligned} \tag{19}$$

where E and I are known physical constants. The appended conditions are boundary conditions because some are specified at one end, and some at the other end, and (19) is therefore a boundary-value problem. The physical significance of the boundary conditions is as follows: $y(0) = 0$ is true simply by virtue of our chosen placement of the origin of the x, y coordinate system; $y'(0) = 0$ follows since the beam is cantilevered out of the wall, so that its slope at $x = 0$ is zero; $y''(L) = 0$ and $y'''(L) = 0$ because the "moment" and "shear force," respectively, are zero at the end of the beam. ∎

Linear and nonlinear differential equations. An nth-order differential equation is said to be **linear** if it is expressible in the form

$$\boxed{a_0(x)y^{(n)}(x) + a_1(x)y^{(n-1)}(x) + \cdots + a_n(x)y(x) = f(x),} \tag{20}$$

where $a_0(x), \ldots, a_n(x)$ are functions of the independent variable x alone, and **nonlinear** otherwise. Thus, equations (1), (2), (4), and (6) are linear, and (3) and (5) are nonlinear. If $f(x) = 0$, we say that (20) is **homogeneous**; if not, it is **nonhomogeneous**. If $a_0(x)$ does not vanish on the x interval of interest, then we may divide (20) by $a_0(x)$ (to normalize the leading coefficient) and re-express it as

$$y^{(n)}(x) + p_1(x)y^{(n-1)}(x) + \cdots + p_n(x)y(x) = q(x). \tag{21}$$

We will find that the theory of linear differential equations is quite comprehensive insofar as all of our major concerns – the existence and uniqueness of solutions, and how to *find* them, especially if the coefficients $a_0(x), \ldots, a_n(x)$ are constants. Even in the nonconstant coefficient case the theory provides substantial guidance.

Nonlinear equations are, in general, far more difficult, and the available theory is not nearly as comprehensive as for linear equations. Whereas for linear equations solutions can generally be found either in closed form or as infinite series, for nonlinear equations one might focus instead upon obtaining qualitative information about the solution, rather than the solution itself, or upon pursuing numerical solutions by computer simulation, or both.

The tendency in science and engineering, until around 1960, when high-speed digital computers became widely available, was to try to get by almost exclusively with linear theory. For instance, consider the nonlinear equation (3), namely,

$$\theta'' + \frac{g}{l}\sin\theta = 0, \tag{22}$$

governing the motion of a pendulum, where $\theta(t)$ is the angular displacement from the vertical and t is the time. If one is willing to limit one's attention to small motions, that is, where θ is small compared to unity (i.e., 1 radian), then one can use the approximation

$$\sin\theta = \theta - \frac{1}{3!}\theta^3 + \frac{1}{5!}\theta^5 - \cdots \approx \theta$$

to replace the nonlinear equation (2) by the approximate "linearized" equation

$$\theta'' + \frac{g}{l}\theta = 0, \tag{23}$$

which (as we shall see in Chapter 3) is readily solved.

Unfortunately, the linearized version (23) is not only less and less accurate as larger motions are considered, it may even be incorrect in a qualitative sense as well. That is, from a phenomenological standpoint, replacing a nonlinear differential equation by an approximate linear one may amount to "throwing out the baby with the bathwater."

Thus, it is extremely important for us to keep the distinction between linear and nonlinear clearly in mind as we proceed with our study of differential equations. Aside from Sections 2.4 and 2.5, most of our study of nonlinear equations takes place in Chapters 6 and 7.

Closure. Notice that we have begun, in this section, to classify differential equations, that is, to categorize them by types. Thus far we have distinguished ODE's (ordinary differential equations) from PDE's (partial differential equations), established the order of a differential equation, distinguished initial-value problems from boundary-value problems, linear equations from nonlinear ones, and homogeneous equations from nonhomogeneous ones.

Why do we classify so extensively? Because the most general differential equation is far too difficult for us to deal with. The most reasonable program, then, is to break the set of all possible differential equations into various categories and to try to develop theory and solution strategies that are tailored to the specific nature of a given category. Historically, however, the early work on differential equations – by such mathematicians as *Leonhard Euler* (1707–1783), *Jakob* (James) *Bernoulli* (1654–1705) and his brother *Johann* (John) (1667–1748), *Joseph-Louis Lagrange* (1736–1813), *Alexis-Claude Clairaut* (1713–1765), and *Jean le Rond d'Alembert* (1717–1783) – generally involved attempts at solving specific equations rather than the development of a general theory.

From an applications point of view, we shall find that in many cases diverse physical phenomena are governed by the same differential equation. For example, consider equations (1) and (2) and observe that they are actually the same equation, to within a change in the names of the various quantities: $m \to L$, $k \to 1/C$, $F(t) \to dE(t)/dt$, and $x(t) \to i(t)$. Thus, to within these correspondences, their solutions are identical. We speak of the mechanical system and the electrical circuit as *analogs* of each other. This idea is deeper and more general than can be seen from this one example, and the result is that if one knows a lot about mechanical systems, for example, then one thereby knows a lot about electrical, biological, and social systems, for example, to whatever extent they are governed by differential equations of the same form.

Or, returning to PDE's for the moment, consider equation (7), which we introduced as the one-dimensional heat equation. Actually, (7) governs *any* one-dimensional diffusion process, be it the diffusion of heat by conduction, or the diffusion of material such as a pollutant in a river. Thus, when one is studying heat conduction one is also learning about all diffusion processes because the governing differential equation is the same. The significance of this fact can hardly be overstated as a justification for a careful study of the mathematical field of differential equations, or as a cause for marvel at the underlying design of the physical universe.

EXERCISES 1.2

1. Determine the order of each differential equation, and whether or not the given functions are solutions of that equation.

(a) $(y')^2 = 4y$; $y_1(x) = x^2$, $y_2(x) = 2x^2$, $y_3(x) = e^{-x}$

(b) $2yy' = 9\sin 2x$; $y_1(x) = \sin x$, $y_2(x) = 3\sin x$, $y_3(x) = e^x$

(c) $y'' - 9y = 0$; $y_1(x) = e^{3x} - e^x$, $y_2(x) = 3\sinh 3x$, $y_3(x) = 2e^{3x} - e^{-3x}$

(d) $(y')^2 - 4xy' + 4y = 0$; $y_1(x) = x^2 - x$, $y_2(x) = 2x - 1$

(e) $y'' + 9y = 0$; $y_1(x) = 4\sin 3x + 3\cos 3x$, $y_2(x) = 6\sin(3x + 2)$

(f) $y'' - y' - 2y = 6$; $y_1(x) = 5e^{2x} - 3$, $y_2(x) = -2e^{-x} - 3$

(g) $y''' - 6y'' + 12y' - 8y = 32 - 16x$; $y_1(x) = 2x - 1 + (A + Bx + Cx^2)e^{2x}$ for any constants A, B, C

(h) $y' + 2xy = 1$; $y_1(x) = Ae^{-x^2}\int_0^x e^{t^2}\,dt$, $y_2(x) = e^{-x^2}\int_a^x e^{t^2}\,dt$ for any constants A and a.

2. Verify that $u(x,t) = Ax + B + (C\sin\kappa x + D\cos\kappa x)\exp(-\kappa^2\alpha^2 t)$ is a solution of (7) for any constants A, B, C, D, κ. NOTE: We will sometimes use the notation exp( ) in place of $e^{(\ )}$ because it takes up less vertical space.

3. Verify that $u(x,y,z) = A\sin ax\sin by\sinh cz$ is a solution of (8) for any constants A, a, b, c, provided that $a^2 + b^2 = c^2$.

4. (a) Verify that $u(x,t) = (Ax + B)(Ct + D) + (E\sin\kappa x + F\cos\kappa x)(G\sin\kappa ct + H\cos\kappa ct)$ is a solution of the one-dimensional wave equation

$$c^2\frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 u}{\partial t^2},$$

for any constants $A, B, \ldots, H, \kappa$.

(b) Verify that $u(x,t) = f(x - ct) + g(x + ct)$ is a solution of that equation for any twice-differentiable functions f and g.

(c) For what value(s) of the constant m is $u(x,t) = \sin(x + mt)$ a solution of that equation ?

5. For what value(s) of the constant λ will $y = \exp(\lambda x)$ be a solution of the given differential equation? If there are no such λ's, state that.

(a) $y' + 3y = 0$ (b) $y' + 3y^2 = 0$
(c) $y'' - 3y' + 2y = 0$ (d) $y'' - 2y' + y = 0$
(e) $y''' - y' = 0$ (f) $y''' - 2y'' - y' + 2y = 0$
(g) $y'''' - 6y'' + 5y = 0$ (h) $y'' + 5yy' + y = 0$

6. First, verify that the given function is a solution of the given differential equation, for any constants A, B. Then, solve for A, B so that y satisfies the given initial or boundary conditions.

(a) $y'' + 4y = 8x^2$; $y(x) = 2x^2 - 1 + A\sin 2x + B\cos 2x$; $y(0) = 1$, $y'(0) = 0$

(b) $y'' - y = x^2$; $y(x) = -x^2 - 2 + A\sinh x + B\cosh x$; $y(0) = -2$, $y'(0) = 0$

(c) $y'' + 2y' + y = 0$; $y(x) = (A + Bx)e^x$; $y(0) = 1$, $y(2) = 0$

(d) $y'' - y' = 0$; $y(x) = A + Be^x$; $y'(0) = 1$, $y(3) = 0$

7. Classify each equation as linear or nonlinear:

(a) $y' + e^x y = 4$ (b) $yy' = x + y$
(c) $e^x y' = x - 2y$ (d) $y' - \exp y = \sin x$
(e) $y'' + (\sin x)y = x^2$ (f) $y'' - y = \exp x$
(g) $yy''' + 4y = 3x$ (h) $y''' = y$

8. Recall that the nonlinear equation (5) governs the deflection $y(x)$ of the flexible cable shown in Fig. 3. Supposing that the sag is small compared to the span, suggest a linearized version of (5) that can be expected to give good accuracy in predicting the shape $y(x)$.

1.3 Introduction to Modeling

Emphasis in this book is on the mathematical analysis that begins once the problem has been formulated – that is, once the modeling phase has been completed. Detailed discussion of the modeling is handled best within applications courses, such

as Heat Transfer, Fluid Mechanics, and Circuit Theory. However, we wish to emphasize the close relationship between the mathematics and the underlying physics, and to motivate the mathematics more fully. Thus, besides the purely mathematical examples in the text, we will include physical applications and some of the underlying modeling as well.

Our intention in this section is only to illustrate the nature of the modeling process, and we will do so through two examples. We suggest that you pay special attention to Example 1 because we will come back to it at numerous points later on in the text.

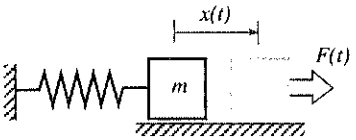


Figure 1. Mechanical oscillator.

EXAMPLE 1. *Mechanical Oscillator.* Consider a block of mass m lying on a table and restrained laterally by an ordinary coil spring (Fig. 1), and denote by x the displacement of the mass (measured as positive to the right) from its "equilibrium position;" that is, when $x = 0$ the spring is neither stretched nor compressed. We imagine the mass to be disturbed from its equilibrium position by an initial disturbance and/or an applied force $F(t)$, where t is the time, and we seek the differential equation governing the resulting displacement history $x(t)$.

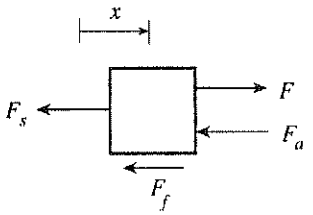


Figure 2. The forces, if $x > 0$ and $x' > 0$.

Our first step is to identify the relevant physics which, in this case, is Newton's second law of motion. Since the motion is constrained to be along a straight line, we need consider only the forces in the x direction, and these are shown in Fig. 2: F_s is the force exerted by the spring on the mass (the spring force, for brevity), F_a is the aerodynamic drag, F_f is the force exerted on the bottom of the mass due to its sliding friction, and F is the applied force, the driving force. How do we know if F_s, F_f, and F_a act to the left or to the right? The idea is to make assumptions on the signs of the displacement $x(t)$ and the velocity $x'(t)$ at the instant under consideration. For definiteness, suppose that $x > 0$ and $x' > 0$. Then it follows that each of the forces F_s, F_f, and F_a is directed to the left, as shown in Fig. 2. (The equation of motion that we obtain will be insensitive to those assumptions, as we shall see.) Newton's second law then gives

$$(\text{mass})(x \text{ acceleration}) = \text{sum of } x \text{ forces}$$

or,

$$mx'' = F - F_s - F_f - F_a, \tag{1}$$

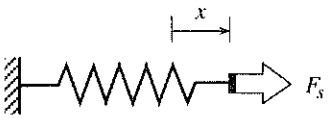


Figure 3. Spring force and displacement.

and we now need to express each of the forces F_s, F_f, and F_a in terms of the dependent and independent variables x and t.

Consider F_s first. If one knows enough about the geometry of the spring and the material of which it is made, one can derive an expression for F_s as a function of the extension x, as might be discussed in a course in Advanced Strength of Materials. In practice, however, one can proceed empirically, and more simply, by actually applying various positive (i.e., to the right, in the positive x direction) and negative forces (to the left) to the spring and measuring the resulting displacement x (Fig. 3). For a typical coil spring, the resulting graph will be somewhat as sketched in Fig. 4, where its steepening at A and B is due to the coils becoming completely compressed and completely extended, respectively. Thus, F_s in (1) is the function the graph of which is shown as the curve AB. (Ignore the dashed line L for the moment.)

Next, consider the friction force F_f. The modeling of F_f will depend upon the nature of the contact between the mass and the table – in particular, upon whether it is dry or

lubricated. Let us suppose it is lubricated, so that the mass rides on a thin film of lubricant such as oil. To model F_f, then, we must consider the fluid mechanics of the lubricating film. The essential idea is that the stress τ (force per unit area) on the bottom of the mass is proportional to the gradient du/dy of the fluid velocity u (Fig. 5), where the constant of proportionality is the coefficient of viscosity μ: $\tau = \mu du/dy$. But $u(y)$ is found, in a course in Fluid Mechanics, to be a linear function, namely,

$$u(y) = \frac{u(h) - u(0)}{h} y = \frac{x'(t) - 0}{h} y = \frac{x'(t)}{h} y,$$

so

$$\tau = \mu \frac{du}{dy} = \frac{\mu}{h} x'(t).$$

Thus,

$$\begin{aligned} F_f &= (\text{stress } \tau)\ (\text{area } A \text{ of bottom of block}) \\ &= \left(\frac{\mu x'(t)}{h} \right) (A). \end{aligned}$$

That is, it is of the form

$$F_f = cx'(t), \tag{2}$$

for some constant c that we may consider as known. Thus, the upshot is that the friction force is proportional to the velocity. We will call c the *damping coefficient* because, as we will see in Chapter 3, the effect of the cx' term in the governing differential equation is to cause the motion to "damp out."

Likewise, one can model the aerodynamic drag force F_a, but let us neglect F_a on the tentative assumption that it can be shown to be small compared to the other two forces. Then (1) becomes

$$mx''(t) + cx'(t) + F_s(x) = F(t). \tag{3}$$

Equation (3) is nonlinear because $F_s(x)$ is not a linear function of x, as seen from its graph AB in Fig. 4. As a final simplifying approximation, let us suppose that the x motion is small enough, say between a and b in Fig. 4, so that we can linearize F_s, and hence the governing differential equation, by approximating the graph of F_s by its tangent line L. Since L is a straight line through the origin, it follows that we can express

$$F_s(x) \approx kx. \tag{4}$$

We call k the *spring stiffness*.

Thus, the final form of our governing differential equation, or equation of motion, is the linearized approximation

$$\boxed{mx'' + cx' + kx = F(t),} \tag{5}$$

on $0 \le t < \infty$, where the constants m, c, k and the applied force $F(t)$ are known. Equation (5) is important, and we will return to it repeatedly.

To this equation we wish to append suitable initial or boundary conditions. This particular problem is most naturally of initial-value type since we envision initiating the motion

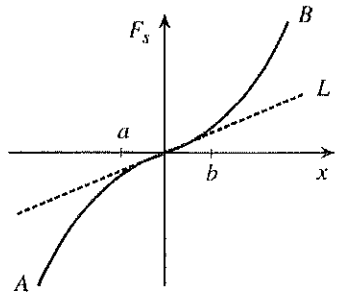


Figure 4. Force-displacement graph.

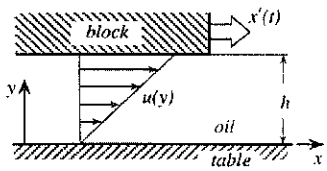


Figure 5. Lubricating film.

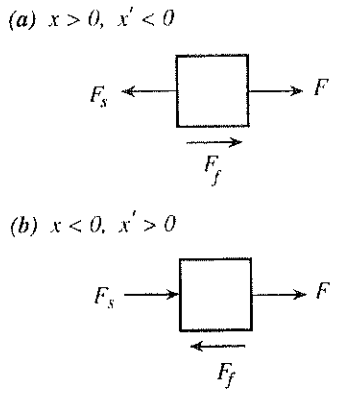


Figure 6. Other assumptions on the signs of x and x'.

in some manner at the initial time, say $t = 0$, and then considering the motion that results. Thus, to (5) we append initial conditions

$$x(0) = x_0 \quad \text{and} \quad x'(0) = x_0', \tag{6}$$

for some (positive, negative, or zero) specified constants x_0 and x_0'. It should be plausible intuitively that we do need to specify both the initial displacement $x(0)$ and the initial velocity $x'(0)$ if we are to ensure a unique resulting motion. In any case, the theoretical appropriateness of the conditions (6) are covered in Chapter 3.

The differential equation (5) and initial conditions (6) comprise our resulting mathematical model of the physical system. By no means is there an exact correspondence between the model and the system since approximations were made in modeling the forces F_s and F_f, and in neglecting F_a entirely. Indeed, even our use of Newtonian mechanics, rather than relativistic mechanics, was an approximation.

This completes the modeling phase. The next step would be to solve the differential equation (5) subject to the initial conditions (6), for the motion $x(t)$.

COMMENT 1. Let us examine our claim that the resulting differential equation is insensitive to the assumptions made as to the signs of x and x'. In place of our assumption that $x > 0$ and $x' > 0$ at the instant under consideration, suppose we assume that $x > 0$ and $x' < 0$. Since $x > 0$, it follows that F_s acts to the left, and since $x' < 0$, it follows that F_f acts to the right. Then (Fig. 6a)

$$mx'' = F - F_s + F_f, \tag{7}$$

where we continue to neglect F_a. The sign of the F_f term is different in (7), compared with (1), because F_f now acts to the right, but notice that F_f now needs to be written as $F_f = c(-x'(t))$, rather than $cx'(t)$ since x' is negative. Further, F_s is still kx, so (7) becomes

$$mx'' = F(t) - kx + (-cx'), \tag{8}$$

which is indeed equivalent to (5), as claimed.

Next, what if $x < 0$ and $x' > 0$? This time (Fig. 6b)

$$mx'' = F + F_s - F_f, \tag{9}$$

which differs from (1) in the sign of the F_s term. But now F_s needs to be written as $F_s = k(-x(t))$ since x is negative. Further, F_f is cx', so (9) becomes

$$mx'' = F + k(-x) - cx',$$

which, again, is equivalent to (5). The remaining case, where $x < 0$ and $x' < 0$, is left for the exercises.

COMMENT 2. The approximation (4) was introduced from consideration of the graph shown in Fig. 4, but it amounts to expanding $F_s(x)$ in a Taylor series about the equilibrium point $x = 0$, as

$$F_s(x) = F_s(0) + F_s'(0)x + \frac{F_s''(0)}{2!}x^2 + \cdots$$

and linearizing – that is, cutting off after the first-order term:

$$\begin{aligned} F_s(x) &\approx F_s(0) + F_s'(0)x \\ &= 0 + kx = kx. \end{aligned}$$

This idea, the simplification of a differential equation by such tangent-line approximation, is of great importance in modeling.

COMMENT 3. The final equation for F_s, $F_s = kx$ is well known as **Hooke's law**, after *Robert Hooke* (1635–1703). Hooke published his law of elastic behavior in 1676 as the anagram *ceiiinossttuv* and, two years later, the solution *ut tensio sic vis*: roughly, "as the force, so is the displacement." In view of the complexity with which we can now deal, Hooke's law must look quite modest, but one must appreciate it within its historical context. In spirit, it followed *Galileo Galilei* (1564–1642) who, in breaking lines established by the ancient Greeks, sought to establish a quantitative science, expressed in formulas and mathematical terms. For example, where Aristotle explained the increasing speed of a falling body in terms of the body moving more and more jubilantly as it approached its natural place (the center of the earth, which was believed to coincide with the center of the universe), Galileo sidestepped the question of cause entirely, and instead put forward the formula $v = 9.8t$, where v is the speed (in meters per second) and t is the time (in seconds). It may be argued that such departure from the less productive Greek tradition marked the beginning of modern science.

COMMENT 4. In science and engineering it is useful to think in terms of inputs and outputs. Here, there are three inputs, the two initial conditions and the applied force $F(t)$, and the output is the resulting solution, or response, $x(t)$. ∎

The foregoing introductory example illustrates several general truths about modeling. First, we see that it is not necessarily an easy task and generally requires a sound understanding of the underlying physics. Even in this little example one senses that obtaining suitable expressions for F_f and F_a (if one does include F_a) requires skillful handling of the fluid mechanics of the lubrication film and the aerodynamics of the moving block.

Second, we see that approximations will no doubt be necessary, and the idea is to make them judiciously. In this example we made several approximations. The expression $u(y) = x'(t)y/h$, for instance, is probably quite accurate but is not exact, especially near the ends of the block. Further, one can imagine that as the motion continues, the lubricant will heat up so that the viscosity μ will decrease. This effect is probably negligible, but we mention it in order to suggest that there is virtually no end to the levels of complexity that one may address, or suppress, in the modeling process. The key is to seek a level of complexity that will provide sufficient accuracy for the purpose at hand, and to seek a *uniform* level of approximation. For instance, it would hardly make sense to model F_f with great sophistication and accuracy if F_s is of comparable magnitude and is modeled only crudely.

To stay on this point a bit longer, note several more approximations that were implicit to our discussion. First, we implicitly assumed that the block is rigid, whereas the applied forces will cause some slight distortion of its shape and dimensions; again, neglect of this effect is surely an excellent approximation when considering the motion $x(t)$. Second, and more serious, notice that our empirical determination of $F_s(x)$ was based on a static test whereas, like the block, the spring

is itself in motion. Thus, there is an inertial effect for the spring, analogous to the mx'' term for the mass, that we have neglected. If the mass of the spring is not negligible compared to that of the block, then that approximation may be insufficiently accurate.

Finally, notice carefully that we neglect a particular effect, in modeling, not on the grounds that it is small in an absolute sense, but because it is small relative to other effects. For instance, an aerodynamic force $F_a = 0.001$ newton may seem small numerically, but would not be negligible if F, F_s, and F_f were of comparable size.

Let us consider one more example.

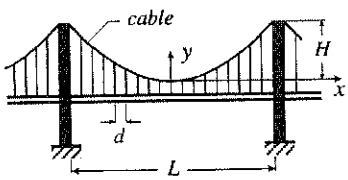


Figure 7. Suspension bridge cable.

EXAMPLE 2. *Suspension Bridge Cable.* To design a suspension bridge cable, one needs to know the relationships among the deflected shape of the cable, the tension in it, and the weight distribution that it needs to support.

In the case of a typical suspension bridge, the roadbed supported by the cables is much heavier than the cables themselves, so let us neglect the weight of the cables, and assume that the loading is due entirely to the roadbed. Consider the configuration shown schematically in Fig. 7.

A cable is a distributed system, rather than one or more point masses, and for such systems a useful modeling approach is to isolate a typical element of the system and apply to it the relevant physical laws. In the present case a typical element is an incremental portion of the cable lying between x and $x + \Delta x$, for any x between $-L/2$ and $L/2$, as shown in Fig. 8: Δs is the arc length, T the tension in the cable, θ the inclination with respect to the horizontal, and ΔW the load supported by the element. If the roadbed is uniform and weighs $2w$ newtons per meter, then each of the two cables supports w newtons per meter, so $\Delta W = w\Delta x$.

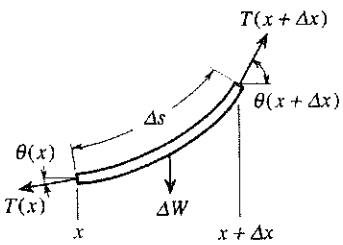


Figure 8. Typical cable element.

Besides neglecting the weight of the cable itself, as mentioned above, there are three additional approximations that are implicit in the foregoing. First, in assuming a uniform load w per unit length, we have really assumed that the vertical support spacing d is very small compared to the span L, so that the intermittent support forces can be distributed as a uniform load. Second, in assuming that the tension is in the direction of the cable we have really assumed that the cable is flexible, a term that we now need to explain. The general state of affairs at any section, such as at the $x + \Delta x$ end of the element, is as shown in Fig. 9, namely, there can be a shear force V, a tensile force T through the centerline, and a moment or "couple" M. (V is the net effect of shearing stresses distributed over the face, and T and M are the net effect of normal stresses distributed over the face.) By a flexible string or cable, we mean one that is unable to support any shear V or moment M; that is, $V = M = 0$. For instance, if one takes a typical household string between the thumb and forefinger of each hand one finds that it offers virtually no resistance to shearing or bending, but considerable resistance to stretching. Thus, when we include only tensile forces in Fig. 8 we are assuming that the cable is flexible. Of course, if we imagine taking the suspension cables on the Golden Gate Bridge "between our fingers" we can imagine quite a bit of resistance to shearing and bending! But the point to keep in mind is that even those heavy cables would offer little resistance to shearing and bending by the enormous loads to which they are actually subjected by the weight of the roadbed.

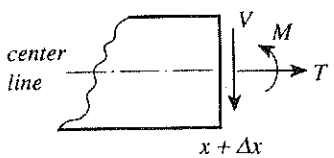


Figure 9. Forces and moments at an end.

Finally, we assume that the cable is inextensible, even under the large tensions that are

anticipated.

If we can accept these assumptions we can now apply the relevant physics which, again, is Newton's second law of motion. But this time there is no acceleration, so the element is in static equilibrium. Thus, the x and y forces must each sum to zero:

$$x: \quad T(x+\Delta x)\cos\theta(x+\Delta x) - T(x)\cos\theta(x) = 0, \tag{10a}$$

$$y: \quad T(x+\Delta x)\sin\theta(x+\Delta x) - T(x)\sin\theta(x) - w\Delta x = 0. \tag{10b}$$

Dividing each of these equations by Δx and letting $\Delta x \to 0$, we obtain

$$\frac{d}{dx}(T\cos\theta) = 0, \tag{11a}$$

$$\frac{d}{dx}(T\sin\theta) = w, \tag{11b}$$

or, upon integration,

$$T\cos\theta = A, \tag{12a}$$

$$T\sin\theta = wx + B, \tag{12b}$$

where A, B are arbitrary constants. Dividing (12b) by (12a), to eliminate the unknown tension $T(x)$, and noting that $\tan\theta = dy/dx$, we obtain the differential equation

$$\frac{dy}{dx} = \frac{w}{A}x + \frac{B}{A} \tag{13}$$

governing the cable shape $y(x)$, which we are trying to determine.

In this case, the solution is simple enough so that we might as well complete the solution. To solve, we merely integrate (13), obtaining

$$y(x) = \frac{w}{2A}x^2 + \frac{B}{A}x + C,$$

where A, B, C are arbitrary constants. To evaluate them, we invoke the associated boundary conditions:

$$y(0) = 0 \qquad \text{(by choice of location of origin)}, \tag{14a}$$

$$y'(0) = 0 \qquad \text{(by symmetry about x = 0)}, \tag{14b}$$

$$y\left(\frac{L}{2}\right) = H \qquad \text{(from Fig. 7)}. \tag{14c}$$

Equation (14a) gives $C = 0$, and (14b) gives $B/A = 0$, and hence $B = 0$. Thus far,

$$y(x) = \frac{w}{2A}x^2,$$

and (14c) then gives $A = wL^2/8H$. Thus, the cable's shape is given by the parabola

$$y(x) = \frac{4H}{L^2}x^2. \tag{15}$$

Finally, the distribution of tension $T(x)$ may be found by squaring and adding (12a) and (12b):

$$T(x) = \sqrt{A^2 + (wx + B)^2} = \sqrt{\left(\frac{wL^2}{8H}\right)^2 + w^2x^2}$$
$$= w\sqrt{x^2 + \frac{L^4}{64H^2}}. \tag{16}$$

In a sense, obtaining $y(x)$ and $T(x)$ marks the end of the analysis, and if we are content that the expressions (15) and (16) are sufficiently accurate, then the next step would be to use them to help in the actual bridge design. Before doing that, however, we should check those results, for there may be errors in our analysis. Also, the approximations that we have made may prove to be insufficiently accurate.

One of the standard ways of checking results is by means of special cases and limiting cases, for which the correct results are known or obvious. Considering (15), we observe first that the parabolic shape looks *reasonable*. Furthermore, (15) implies that $y(x) \to 0$, over $-L/2 \le x \le L/2$, as $H \to 0$ with L fixed, and also that $y(x) \to 0$ at each x, as $L \to \infty$ with H fixed. These results look reasonable too. Turning to (16), observe that the tension becomes infinite throughout the cable as $H \to 0$, as expected. (Try straightening out a loaded washline by pulling on one end!) Finally, consider the limiting case $H \to \infty$, with L fixed. In that case, (16) gives $T(L/2) \to wL/2$, which agrees with the result obtained from a simple consideration of the physics (Exercise 2). ∎

Closure. The purpose of this section is to illustrate the modeling process, whereby one begins with the physical problem at hand and ends up with an equivalent mathematical statement, or model. Actually, we should say approximately equivalent since the modeling process normally involves a number of assumptions and approximations. By no means do we claim to show how to model in general, but only to illustrate the modeling process and the intimate connection between the physics and the mathematics. As we proceed through this text we will attempt to keep that connection in view, even though our emphasis will be on the mathematics.

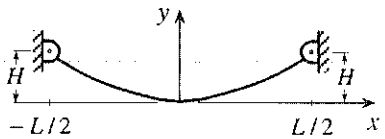
Finally, let us note that when we speak of the physical problem and the physics we intend those words to cover a much broader range of possibilities. For instance, the problem might be in the realm of economics, such as predicting the behavior of the Dow Jones Stock Index as a function of time. In that case the relevant "physical laws" would be economic laws such as the law of supply and demand. Or, the problem might fall in the realm of sociology, ecology, biology, chemistry, and so on. In any case, the general modeling approach is essentially the same, independent of the field of application.

EXERCISES 1.3

1. In Example 1 we showed that the same differential equation, (5), results, independent of whether $x > 0$ and $x' > 0$, or $x > 0$ and $x' < 0$ or $x < 0$ and $x' > 0$. Consider the last remaining case, $x < 0$ and $x' < 0$, and show that, once again, one obtains equation (5).

2. At the end of Example 2, we stated that the result $T(L/2) \to wL/2$, obtained from (16), for the limiting case where $H \to \infty$ with L fixed, agrees with the result obtained from a simple consideration of the physics. Explain that statement.

3. (*Catenary*) In our Suspension Bridge Cable example we neglected the weight of the cable itself relative to the weight of the roadbed. At the other extreme, suppose that the weight of the roadbed (or other loading) is negligible compared to the weight of the cable. Indeed, consider a uniform flexible cable, or catenary, hanging under the action of its own weight only, as sketched in the figure. Then Fig. 8 still holds, but with $\Delta W = \mu \Delta s$, where μ is the weight per unit arc length of the cable.



(a) Proceeding somewhat as in (10)–(12), derive the governing differential equation

$$y'' = C\sqrt{1 + y'^2}, \tag{3.1}$$

where C is an unknown constant.

(b) Since $y(x)$ is symmetric about $x = 0$, it suffices to consider the interval $0 \le x \le L/2$. Then we have the boundary conditions $y(0) = 0$, $y'(0) = 0$, and $y(L/2) = H$. Verify (you need not derive it) that

$$y(x) = \frac{1}{C}(\cosh Cx - 1) \tag{3.2}$$

satisfies (3.1) and the boundary conditions $y(0) = 0$ and $y'(0) = 0$. But it remains to determine C. Invoking the remaining boundary condition, $y(L/2) = H$, show that C satisfies the equation

$$H = \frac{1}{C}\left(\cosh \frac{CL}{2} - 1\right). \tag{3.3}$$

Unfortunately, (3.3) is a transcendental equation for C, so that we cannot solve it explicitly. We can solve it numerically, for given values of H and L, but you need not do that.

(c) As a partial check on these results, notice that they should reduce to the parabolic cable solution in the limiting case where the sag-to-span ratio H/L tends to zero, for then the load per unit x length, due to the weight of the cable, approaches a constant, as it is in Example 2, where the load is due entirely to the uniform roadbed. The problem that we pose for you is to carry out that check. HINT: Think of L as fixed and H tending to zero. For H to approach zero, in (3.3), we need $CL/2$ to approach zero – that is, $C \to 0$. Thus, we can expand the $\cosh Cx - 1$ in (3.2) in a Maclaurin series in C and retain the leading term. Show that that step gives $y(x) \approx Cx^2/2$, and the boundary condition $y(L/2) = H$ enables us to determine C. The result should be identical to (15).

(d) Actually, for small sag-to-span ratio we should be able to neglect the y'^2 term in (3.1), relative to unity, so that (3.1) can be linearized as

$$y'' = C. \tag{3.4}$$

Integrating (3.4) and using the boundary conditions $y(0) = 0$, $y'(0) = 0$, and $y(L/2) = H$, show that one obtains (15) once again.

Chapter 2

Differential Equations of First Order

2.1 Introduction

In studying algebraic equations, one considers the very simple first-degree polynomial equation $ax = b$ first, then the second-degree polynomial equation (quadratic equation), and so on. Likewise, in the theory of differential equations it is reasonable and traditional to begin with first-order equations, and that is the subject of this chapter. In Chapter 3 we turn to equations of second order and higher.

Recall that the general first-order equation is given by

$$F(x, y, y') = 0, \tag{1}$$

where x and y are independent and dependent variables, respectively. In spite of our analogy with algebraic equations, first-order differential equations can fall anywhere in the spectrum of complexity, from extremely simple to hopelessly difficult. Thus, we identify several different subclasses of (1), each of which is susceptible to a particular solution method, and develop them in turn. Specifically, we consider these subclasses: the linear equation $a_0(x)y' + a_1(x)y = f(x)$ in Section 2.2, "separable" equations in Section 2.4, equations that are "exact" (or can be made exact) in Section 2.5, and various other more specialized cases within the exercises.

These subclasses are not mutually exclusive. For instance, a given equation could be both linear and separable, in which case we could solve it by one method or the other. Given such a choice, choose whichever you prefer. In other cases the given equation might not fit into *any* of these subclasses and might be hopelessly difficult from an analytical point of view. Thus, it will be important to complement our analytical methods with numerical solution techniques. But that is a long way off, in Chapter 6. Analytical methods and the general theory of differential equations will occupy us in Chapters 2 through 5.

It should be stressed that the equation types that are susceptible to the analytical solution techniques described in these chapters can also be solved analytically

by computer algebra software that is currently available, such as *Maple*, *Mathematica*, and *MATLAB*, and this approach is discussed here as well. One needs to know both: the underlying theory and solution methodology on one hand, and the available computer software on the other.

2.2 The Linear Equation

The first case that we consider is the general first-order linear differential equation

$$a_0(x)y' + a_1(x)y = f(x). \tag{1}$$

Dividing through by $a_0(x)$ [which is permissible if $a_0(x) \neq 0$ over the x interval of interest], we can re-express (1) in the more concise form

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

We assume that $p(x)$ and $q(x)$ are continuous over the x interval of interest.

2.2.1. Homogeneous case. It is tempting to think that to solve (2) for $y(x)$ we need to get rid of the derivative, and that we can accomplish that merely by integration. It's true that if we integrate (2) with respect to x,

$$\int y'\,dx + \int p\,y\,dx = \int q\,dx,$$

then the first term reduces nicely to y (plus a constant of integration), but the catch is that the $\int p\,y\,dx$ term becomes a stumbling block because we cannot evaluate it since $y(x)$ is unknown! Essentially, then, we have merely converted the differential equation (2) to an "integral equation" – that is, one involving the integral of the unknown function. Thus, we are no better off.

To solve (2), we begin with the simpler special case where $q(x)$ is zero,

$$y' + p(x)y = 0, \tag{3}$$

which is called the **homogeneous** version of (2). To solve (3), divide by y (assuming that y is nonzero on the x interval of interest. This assumption is tentative since y is not yet known) and integrate on x. Using the fact that $y'dx = dy$, from the calculus, we thus obtain

$$\int \frac{dy}{y} + \int p(x)\,dx = 0, \tag{4}$$

and recalling that

$$\int \frac{dx}{x} = \ln|x| + \text{constant}, \tag{5}$$

(4) gives

$$\ln|y| = -\int p(x)\,dx + C,$$

where the arbitrary constant C can include the arbitrary constant of integration from the p integral as well. Thus,

$$|y(x)| = e^{-\int p(x)\,dx + C} = e^C e^{-\int p(x)\,dx} = Be^{-\int p(x)\,dx}, \tag{6}$$

where we have set $e^C \equiv B$ for simplicity. Since C is real, e^C is nonnegative, so B is an arbitrary nonnegative constant: $B \geq 0$. The integral $\int p(x)\,dx$ does indeed exist since we have assumed that $p(x)$ is continuous. Finally, it follows from (6) that $y(x) = \pm B \exp\left(-\int p\,dx\right)$ or,

$$\boxed{y(x) = Ae^{-\int p(x)\,dx}} \tag{7}$$

if we now allow the arbitrary constant A to be positive, zero, or negative.

Observe that our tentative assumption, just below (3), that $y(x) \neq 0$, is now seen to be justified because the exponential function in (7) is strictly positive. [Of course $y(x) = 0$ if $A = 0$, but in that simple case $y(x) = 0$ is seen to satisfy (3) without further ado.] Summarizing, the solution of the homogeneous equation (3) is given by (7), where A is an arbitrary constant.

The presence of the arbitrary constant A, in (7), permits the solution (7) to satisfy an initial condition, if one is specified. Thus, suppose that we seek a solution $y(x)$ of our differential equation $y' + p(x)y = 0$ that satisfies an initial condition $y(a) = b$, for specified values of a and b. For this purpose, it is convenient to re-express (7) as

$$y(x) = Ae^{-\int_a^x p(\xi)\,d\xi}, \tag{8}$$

which is equivalent to (7) since $\int p(x)\,dx$ and $\int_a^x p(\xi)\,d\xi$ differ at most by an additive constant, say D, and the resulting e^D factor in (8) can be absorbed into the arbitrary constant A.

A point of notation: why do we change the integration variable from x to ξ in (8)? Because $\int_a^x p(\xi)\,d\xi$ means that we integrate along some axis, say a ξ axis, from a to x. Thus, x is a fixed endpoint, and is different from the integration variable that runs from a to x. To write $\int_a^x p(x)\,dx$ runs the risk of confusing the x's inside the integral with the fixed endpoint x. The *name* of the integration variable is immaterial, so one calls it a "dummy variable." For instance, $\int_0^x \xi\,d\xi$, $\int_0^x \eta\,d\eta$, and $\int_0^x \rho\,d\rho$ are all the same, namely, $x^2/2$. One often sees the letter ξ used as a dummy x variable because it is the Greek version of x. In Roman letters it is written as xi and is pronounced as ksē. Occasionally, we may be guilty of bad notation and write an integral as $\int_a^x f(x)\,dx$, simply to minimize the letters used, but even then we need to remember the distinction between the x in the upper limit and the x's inside the integral. In fact, this notation is typical in books on engineering and science, where there is less focus on such points of rigor.

Now imposing the condition $y(a) = b$ on (8) gives $y(a) = b = Ae^0 = A$, so $A = b$ and hence

$$\boxed{y(x) = be^{-\int_a^x p(\xi)\,d\xi}.} \tag{9}$$

Thus, (9) satisfies the initial condition $y(a) = b$. To verify that it also satisfies the differential equation (3), we use the **fundamental theorem of the calculus**: If $f(x)$ is continuous on an interval I, namely $x_1 \le x \le x_2$, and

$$F(x) = \int_{x_1}^{x} f(\xi)\, d\xi, \tag{10a}$$

on I, then

$$F'(x) = f(x) \tag{10b}$$

on I. Using this theorem, and chain differentiation [let the $-\int_a^x p(\xi)\, d\xi$ in the exponent be u, say, and write $de^u/dx = (de^u/du)(du/dx)$], it is shown that (9) does indeed satisfy the differential equation $y' + p(x)y = 0$ on an interval I if $p(x)$ is continuous on I.

EXAMPLE 1. Consider the differential equation

$$y' + 2xy = 0 \tag{11}$$

on $-\infty < x < \infty$, over which interval $p(x) = 2x$ is continuous, as we have assumed. Then (7) gives

$$y(x) = Ae^{-\int 2x\, dx} = Ae^{-x^2} \tag{12}$$

on $-\infty < x < \infty$. The graphs of the solutions, (12), are called the solution curves or integral curves corresponding to the given differential equation (11), and these are displayed for several values of A in Fig. 1. Those above and below the x axis correspond to $A > 0$ ($A = 1, 2, 3$) and $A < 0$ ($A = -1, -2$), respectively, and $A = 0$ gives the solution curve $y(x) = 0$. ∎

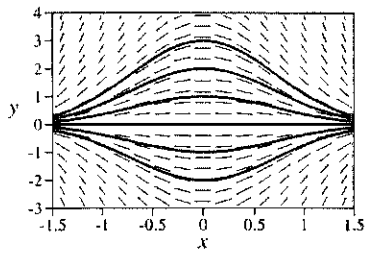


Figure 1. The solution curves and direction field for $y' + 2xy = 0$.

In Example 1 we used the term "solution curve." A **solution curve**, or **integral curve**, corresponding to a differential equation $y'(x) = f(x, y)$, is simply the graph of a solution to that equation.

Besides several solution curves, Fig. 1 also contains a field of lineal elements through a discrete set of points, or grid. By a **lineal element** at a point (x_0, y_0), corresponding to a differential equation $y'(x) = f(x, y)$, we mean a short straight-line segment through the point (x_0, y_0), centered at that point and having the slope $f(x_0, y_0)$. That is, each lineal element has the same slope as the solution curve passing through that point and is therefore a small part of the tangent line to that solution curve. The set of all of the lineal elements is called the **direction field** corresponding to the given differential equation.

In intuitive language, the direction field shows the "flow" of solution curves. Given a sufficiently dense (computer) plot of the direction field, one can visualize the various solution curves, or sketch the solution curve through a given initial point.

EXAMPLE 2. Consider the problem

$$(x + 2)y' - xy = 0, \tag{13a}$$

$$y(0) = 3. \tag{13b}$$

Since an initial value is prescribed, let us use (9) rather than (7), with $p(x) = -x/(x+2)$:

$$\begin{aligned} y(x) &= 3e^{\int_0^x \xi\, d\xi/(\xi+2)} = 3e^{[\xi+2-2\ln|\xi+2|]}\Big|_0^x \\ &= 3e^{[x+2-2\ln|x+2|-(2-2\ln 2)]} = 3e^x e^{\ln|x+2|^{-2}} e^{\ln 2^2} \\ &= 12\frac{e^x}{|x+2|^2} = 12\frac{e^x}{x^2+4x+4}, \end{aligned} \tag{14}$$

where we have used the identity $e^{\ln f} = f$. Of course, we could have used (7) instead, and then imposed the initial condition on that result:

$$y(x) = Ae^{\int x\,dx/(x+2)} = Ae^{x+2-2\ln|x+2|} = A\frac{e^2e^x}{|x+2|^2},$$

where $y(0) = 3 = Ae^2/4$ gives $A = 12e^{-2}$ and, once again, we obtain the solution

$$y(x) = 12\frac{e^x}{x^2+4x+4}. \tag{15}$$

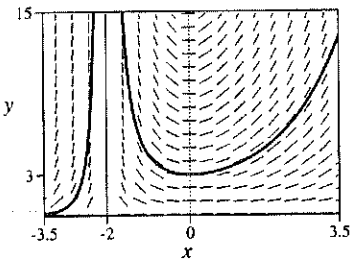


Figure 2. Solution to $(x+2)y' - xy = 0$; $y(0) = 3$, together with the direction field.

The graph of that solution is given in Fig. 2, in which we show the direction field of the equation (13a), as well.

On what x interval is the solution (15) valid? Recall that the solution (9) was guaranteed over the broadest interval, containing the initial point $x = a$, over which $p(x)$ is continuous. In this case $a = 0$ and $p(x) = -x/(x+2)$, which is undefined (infinite) at $x = -2$, so (15) is valid at least on $-2 < x < \infty$. In fact, the interval of validity cannot be extended any further to the left in this example, because whereas we need both y and y' to have uniquely defined finite values satisfying $(x+2)y' - xy = 0$, both y and y' given by (15) are undefined at $x = -2$. Thus, the interval of validity of (15) is $-2 < x < \infty$. ∎

With the homogeneous problem (3) solved, we now turn to the more difficult case, the nonhomogeneous equation (2). We will show how to solve (2) by two different methods: first, by using an integrating factor and, second, by the method of variation of parameters.

2.2.2. Integrating factor method. To solve (2) by the integrating factor method, we begin by multiplying both sides by a yet to be determined function $\sigma(x)$, so that (2) becomes

$$\sigma y' + \sigma p y = \sigma q. \tag{16}$$

[For (2) and (16) to be equivalent, we require $\sigma(x)$ to be nonzero on the x interval of interest, for if $\sigma(x) = 0$ at one or more points, then it does not follow from (16) that $y' + py$ equals q at those points.] The idea is to seek $\sigma(x)$ so that the left-hand side of (16) is a derivative:

$$\sigma y' + \sigma p y = \frac{d}{dx}(\sigma y), \tag{17}$$

because then (16) becomes

$$\frac{d}{dx}(\sigma y) = \sigma q, \tag{18}$$

which can be solved by integration. For instance, to solve the equation $y' + \frac{1}{x}y = 4$, observe that *if* we multiply through by x, then we have $xy' + y = 4x$, or $(xy)' = 4x$, which can be integrated to give $xy = 2x^2 + C$ and hence the solution $y = 2x + C/x$. In this case $\sigma(x)$ was simply x. Such a function $\sigma(x)$ is called an **integrating factor**, and its use is called the integrating factor method. The idea was invented by *Leonhard Euler* (1707–1783), one of the greatest and most prolific mathematicians of all time. He contributed to virtually every branch of mathematics and to the application of mathematics to the science of mechanics.

One might say that the integrating factor method is similar to the familiar method of solving a quadratic equation by completing the square. In completing the square, we add a suitable quantity to both sides so that the left-hand side becomes a "perfect square," and the equation can then be solved by taking square roots. In the integrating factor method we multiply both sides by a suitable quantity so that the left-hand side becomes a "perfect derivative," and the equation can then be solved by integration.

How do we find $\sigma(x)$, if indeed such a function exists? Writing out the right-hand side of (17) gives

$$\sigma y' + \sigma p y = \sigma y' + \sigma' y,$$

which is satisfied identically if we choose $\sigma(x)$ so that

$$\sigma'(x) = p(x)\sigma(x). \tag{19}$$

But (19) is of the same form as (3), with y changed to σ and p to $-p$, so its solution follows from (7) as $\sigma(x) = Ae^{+\int p(x)\,dx}$. From (16), we see that any constant scale factor in σ is inconsequential, so we can take $A = 1$ without loss. Thus, the desired integrating factor is

$$\sigma(x) = e^{\int p(x)\,dx}. \tag{20}$$

Putting this $\sigma(x)$ into (18) gives

$$\frac{d}{dx}\left(e^{\int p(x)\,dx} y\right) = e^{\int p(x)\,dx} q(x),$$

so

$$e^{\int p(x)\,dx} y = \int e^{\int p(x)\,dx} q(x)\,dx + C,$$

or

$$\boxed{y(x) = e^{-\int p(x)\,dx}\left(\int e^{\int p(x)\,dx} q(x)\,dx + C\right),} \tag{21}$$

where C is an arbitrary constant of integration.

Not only does (21) satisfy (2), but it can be seen from the steps leading to (21) that *every* solution to (2) must be of the form (21). Thus, we call (21) the **general solution** of (2).

EXAMPLE 3. Solve

$$y' + 3y = x. \tag{22}$$

With $p(x) = 3$ and $q(x) = x$, we have

$$e^{\int p(x)\,dx} = e^{\int 3\,dx} = e^{3x},$$

so (21) gives

$$y(x) = e^{-3x}\left(\int e^{3x} x\,dx + C\right) = \frac{x}{3} - \frac{1}{9} + Ce^{-3x}. \tag{23}$$

as the general solution to (22). ∎

If we wish to solve for C, in (21), so as to satisfy an initial condition $y(a) = b$, then it is convenient to return to (20) and use the definite integral $\int_a^x p(\xi)\,d\xi$ in place of the indefinite integral $\int p(x)\,dx$. These integrals differ by at most an arbitrary additive constant, say B, and the scale factor e^B that results, in the right-hand side of (20), can be discarded without loss. Thus, let us change both p integrals in (21) to definite integrals, with a as lower limit. Likewise, the q integral in (21) can be changed to a definite integral with a as lower limit since this step changes the integral by an arbitrary additive constant at most, which constant can be absorbed by the arbitrary constant C. Thus, equivalent to (21), we have

$$y(x) = e^{-\int_a^x p(\xi)\,d\xi}\left(\int_a^x e^{\int_a^\xi p(\zeta)\,d\zeta} q(\xi)\,d\xi + C\right),$$

where ζ is zeta. If we impose on this result the initial condition $y(a) = b$, we obtain $y(a) = b = e^{-0}(0 + C)$, where each ξ integral is zero because its lower and upper integration limits are the same. Thus, $C = b$, and

$$\boxed{y(x) = e^{-\int_a^x p(\xi)\,d\xi}\left(\int_a^x e^{\int_a^\xi p(\zeta)\,d\zeta} q(\xi)\,d\xi + b\right).} \tag{24}$$

As a partial check, notice that (24) does reduce to (9) in the event that $q(x) = 0$.

Whereas (21) was the general solution to (2), we call (24) a **particular solution** since it corresponds to one particular solution curve, the solution curve through the point (a, b).

EXAMPLE 4. Solve

$$y' - 2xy = \sin x, \tag{25a}$$

$$y(0) = 3. \tag{25b}$$

This time an initial condition is prescribed, so it is more convenient to use (24) than (21). With $p(x) = -2x$, $q(x) = \sin x$, $a = 0$, and $b = 3$, we have

$$e^{\int_0^x p(\xi)\,d\xi} = e^{\int_0^x (-2\xi)\,d\xi} = e^{-x^2},$$

so that (24) gives the desired solution as

$$y(x) = e^{x^2}\left(\int_0^x e^{-\xi^2}\sin\xi\,d\xi + 3\right). \tag{26}$$

The integral in (26) is said to be **nonelementary** in that it cannot be evaluated in closed form in terms of the so-called **elementary functions**: powers of x, trigonometric and inverse trigonometric functions, exponentials, and logarithms. Thus, we will leave it as it is. It can be evaluated in terms of nonelementary functions, or integrated numerically, but such discussion is beyond the scope of this example. ∎

2.2.3. Existence and uniqueness for the linear equation. A fundamental issue, in the theory of differential equations, is whether a given differential equation $F(x, y, y') = 0$ *has* a solution through a given initial point $y(a) = b$ in the x, y plane. That is the question of existence. If a solution does exist, then our next question is: is that solution unique, or is there more than one such solution? That is the question of uniqueness. Finally, if we do indeed have a unique solution, then over what x interval does it apply?

For the linear problem

$$y' + p(x)y = q(x), \qquad y(a) = b, \tag{27}$$

all of these questions are readily answered. Our proof of existence is said to be constructive because we actually *found* a solution, (24). We are also assured that that solution is unique because our derivation of (24) did not offer any alternatives: each step was uniquely implied by the preceding one, as you can verify by reviewing those steps. Nonetheless, let us show how to prove uniqueness, for (27), by a different line of approach.

Suppose that we have two solutions of (27), $y_1(x)$ and $y_2(x)$, on any x interval I containing the initial point a. That is,

$$y_1' + p(x)y_1 = q(x), \qquad y_1(a) = b, \tag{28a}$$

and

$$y_2' + p(x)y_2 = q(x), \qquad y_2(a) = b. \tag{28b}$$

Next, denote the difference $y_1(x) - y_2(x)$ as $u(x)$, say. If we subtract (28b) from (28a), and use the fact that $(f - g)' = f' - g'$, known from the calculus, we obtain the "homogenized" problem

$$u' + p(x)u = 0, \qquad u(a) = 0, \tag{29}$$

on $u(x)$. But $u' + pu = 0$ implies that $\int du/u + \int p\,dx = 0$, which implies that $\ln|u| = -\int p\,dx + C$, which implies that $|u| = \exp\left(-\int p\,dx + C\right)$ and,

finally, that $u(x) = A\exp(-\int p\,dx)$, where A is an arbitrary constant. Since $u(a) = 0$ and the exponential is nonzero, it follows that A must be zero, so $u(x) = y_1(x) - y_2(x)$ must be identically zero. Thus, $y_1(x)$ and $y_2(x)$ must be identical on I. Since $y_1(x)$ and $y_2(x)$ are *any* two solutions of (27), the solution must be unique.

That approach, in proving uniqueness, is somewhat standard. Namely, we suppose that we have two solutions to the given problem, we let their difference be u, say, obtain a homogenized problem on u, and show that u must be identically zero.

Finally, what is the interval of existence of our solution (24)? The only possible breakdown of (24) is that one or more of the integrals might not exist (be convergent). But since p is continuous, by assumption, it follows that the ξ and ζ integrals of p both exist and, indeed, are continuous functions of x and ξ, respectively. Thus $\exp\left(\int_a^\xi p(\zeta)\,d\zeta\right)$ is continuous too, and since q is also continuous by assumption, then the integral of the exponential times q must exist.

In summary, we have the following result.

THEOREM 2.2.1 *Existence and Uniqueness, for the Linear Equation.*
The linear equation $y' + p(x)y = q(x)$ does admit a solution through an initial point $y(a) = b$ if $p(x)$ and $q(x)$ are continuous at a. That solution is unique, and it exists at least on the largest x interval containing $x = a$, over which $p(x)$ and $q(x)$ are continuous.

In Example 1, for instance, $p(x) = 2x$ and $q(x) = 0$ were continuous for all x, so every solution was valid over $-\infty < x < \infty$. In Example 2, $p(x) = -x/(x+2)$ and $q(x) = 0$, so the broadest interval containing the initial point $x = 0$, over which p and q were continuous, was $-2 < x < \infty$ and, sure enough, we found that the solution (15) was valid over that interval but not beyond it because of the singularity in the solution at $x = -2$. We might think of the solution as "inheriting" that singularity from the singular behavior of $p(x) = -x/(x+2)$ at that point.

EXAMPLE 5. *A Narrow Escape.* The condition of continuity of $p(x)$ and $q(x)$ is sufficient to imply the conclusions stated in the theorem, but is not necessary, as illustrated by the problem

$$xy' + 3y = 6x^3. \tag{30}$$

The general solution of (30) is readily found, from (21), to be

$$y(x) = x^3 + \frac{C}{x^3}. \tag{31}$$

The graphs of the solution for several values of C (i.e., the solution curves) are shown in Fig. 3. Now, $p(x) = 3/x$ is continuous for all x except $x = 0$, and $q(x) = 6x^2$ is continuous for all x, so if we append to (30) an initial condition $y(a) = b$, for some $a > 0$, then Theorem 2.2.1 tells us that the solution (31) that passes through that initial point will

be valid over the interval $0 < x < \infty$ at least. For instance, if $a = 1$ and $b = 2.5$ (the point P_1), then $C = 1$, and the solution (31) is valid only over $0 < x < \infty$ since it is undefined at $x = 0$ because of the $1/x^3$ term, as can also be seen from the figure. However, if $a = 1$ and $b = 1$ (the point P_2) then $C = 0$, and the solution (31) is valid over the broader interval $-\infty < x < \infty$ because $C = 0$ removes the singular $1/x^3$ term in (31). That is, if the initial point happens to lie on the solution curve $y(x) = x^3$ through the origin, then the solution $y(x) = x^3$ is valid on $-\infty < x < \infty$; if not, then the solution (30) is valid on $0 < x < \infty$ if $a > 0$ and on $-\infty < x < 0$ if $a < 0$. ∎

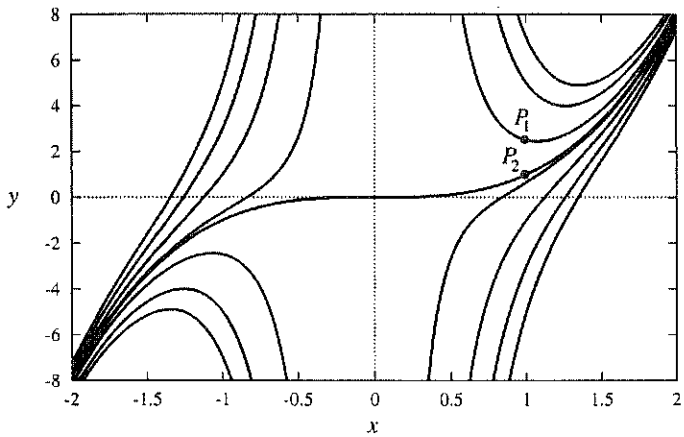


Figure 3. Representative integral curves (31) for the equation (30).

2.2.4. Variation of parameter method. A second method for the solution of the general first-order linear equation

$$y' + p(x)y = q(x) \tag{32}$$

is the method of **variation of parameters**, due to the great French mathematician *Joseph Louis Lagrange* (1736–1813) who, like Euler, also worked on the applications of mathematics to mechanics, especially celestial mechanics.

Lagrange's method is as follows. We begin by considering the homogeneous version of (32),

$$y' + p(x)y = 0, \tag{33}$$

which is more readily solved. Recall that we solved it by integrating

$$\int \frac{dy}{y} + \int p(x)\,dx = 0$$

and obtaining the general solution

$$y_h(x) = Ae^{-\int p(x)\,dx}. \tag{34}$$

We use the subscript h because $y_h(x)$ is called the **homogeneous solution** of (32). That is, it is the solution of the homogeneous version (33) of the original nonhomogeneous equation (32). [In place of $y_h(x)$, some authors write $y_c(x)$ and call it the **complementary solution**.] Lagrange's idea is to try varying the "parameter" A, the arbitrary constant in (34). Thus, we seek a solution $y(x)$ of the *non*homogeneous equation in the form

$$y(x) = A(x)e^{-\int p(x)\,dx}. \tag{35}$$

(The general idea of seeking a solution of a differential equation in a particular form is important and is developed further in subsequent chapters.)

Putting (35) into (32) gives

$$\left(A'e^{-\int p\,dx} + A(-p)e^{-\int p\,dx}\right) + pAe^{-\int p\,dx} = q. \tag{36}$$

Cancelling the two A terms and solving for A' gives

$$A'(x) = q(x)e^{\int p(x)\,dx}, \tag{37}$$

which can be integrated to give

$$A(x) = \int e^{\int p(x)\,dx} q(x)\,dx + C$$

and hence the general solution

$$y(x) = A(x)e^{-\int p(x)\,dx} = e^{-\int p(x)\,dx}\left(\int e^{\int p(x)\,dx} q(x)\,dx + C\right), \tag{38}$$

which is identical to our previous result (21).

It is easy to miss how remarkable is the idea behind Lagrange's method because it starts out looking like a foolish idea and ends up working beautifully. Why do we say that it looks foolish? Because it is completely nonspecific. To explain what we mean by that, let us put Lagrange's idea aside, for a moment, and consider the second-order linear equation $y''+y'-2y = 0$, for instance. In Chapter 3 we will learn to seek solutions of such equations in the exponential form $y = e^{\lambda x}$, where λ is a constant that needs to be determined. Putting that form into $y'' + y' - 2y = 0$ gives the equation $(\lambda^2 + \lambda - 2)e^{\lambda x} = 0$, which implies that λ needs to satisfy the quadratic equation $\lambda^2 + \lambda - 2 = 0$, with roots $\lambda = 1$ and $\lambda = -2$. Thus, we are successful, in this example, in finding two solutions of the assumed form, $y(x) = e^x$ and $y(x) = e^{-2x}$. Notice how easily this idea works. It is easily implemented because most of the work has been done by deciding to look for solutions in the correct form, exponentials, rather than looking within the set of all possible functions. Similarly, if we lose our eyeglasses, the task of finding them is much easier if we know that they are somewhere on our desk, than if we know that they are somewhere in the universe.

Returning to Lagrange's idea, observe that the form (35) is completely nonspecific. That is, *every* function can be expressed in that form by a suitable choice

of $A(x)$. Thus, (35) seems useless in that it does not narrow down the search in the least. That's why we say that at first glance Lagrange's idea looks like a foolish idea.

Next, why do we say that, nevertheless, it works beautifully? Notice that the equation (36), governing $A(x)$, is itself a first-order nonhomogeneous equation of the same form as the original equation (32), and looks even harder than (32)–except for the fact that the two A terms cancel, so that we obtain the simple equation (37) that can be solved by direct integration. The cancellation of the two A terms was not serendipitous. For suppose that $A(x)$ is a constant. Then the A' term in (36) drops out, and the two A terms must cancel to zero because if A is a constant then (35) is a solution of the homogeneous equation!

In Chapter 3 we generalize Lagrange's method to higher-order differential equations.

Closure. In this chapter we begin to solve differential equations. In particular, we consider the general first-order linear equation

$$y' + p(x)y = q(x), \tag{39}$$

where $p(x)$ and $q(x)$ are given. We begin with the homogeneous case,

$$y' + p(x)y = 0 \tag{40}$$

because it is simpler, and find its general solution

$$y(x) = Ae^{-\int p(x)\,dx}. \tag{41}$$

If an initial condition $y(a) = b$ is appended to (40), then (41) gives the particular solution of (40) through the initial point (a, b) as

$$y(x) = be^{-\int_a^x p(\xi)\,d\xi}. \tag{42}$$

Turning next to the full nonhomogeneous equation (39), we derive the general solution

$$y(x) = e^{-\int p(x)\,dx}\left(\int e^{\int p(x)\,dx} q(x)\,dx + C\right), \tag{43}$$

first by the integrating factor method, and then again by the method of variation of parameters. Both of these methods will come up again in subsequent sections and chapters.

If an initial condition $y(a) = b$ is appended to (39), then (43) gives the particular solution of (39) through the initial point (a, b) as

$$y(x) = e^{-\int_a^x p(\xi)\,d\xi}\left(\int_a^x e^{\int_a^\xi p(\zeta)\,d\zeta} q(\xi)\,d\xi + b\right), \tag{44}$$

which solution is unique, and which is valid on the broadest x interval containing $x = a$, on which both $p(x)$ and $q(x)$ are continuous, and possibly even on a broader interval than that.

Finally, we introduce the idea of lineal elements and the direction field of a differential equation $y' = f(x, y)$.

It is noteworthy that we are successful in finding the general solution of the general first-order linear equation $y' + p(x)y = q(x)$ explicitly and in closed form. For other equation types we may not be so successful, as we shall see.

In closing, we call attention to the exercises, to follow, that introduce additional important special cases, the **Bernoulli, Riccati, d'Alembert-Lagrange**, and **Clairaut equations**. In subsequent sections we occasionally refer to those equations and to those exercises.

Computer software. There are now several powerful computer-algebra systems, such as *Mathematica*, *MATLAB*, and *Maple*, that can be used to implement much of the mathematics presented in this text – numerically, symbolically, and graphically. Consider the application of *Maple*, as a representative software, to the material in this section.

There are two types of applications involved in this section. One entails finding the general solution of a given first-order differential equation, or a particular solution satisfying a given initial condition. These can be carried out on *Maple* using the dsolve command ("function," in *Maple* terminology).

For example, to solve the equation $(x+2)y' - xy = 0$ of Example 2, for $y(x)$, enter

$$\text{dsolve}((x+2) * \text{diff}(y(x), x) - x * y(x) = 0, y(x));$$

(including the semicolon) and return; **dsolve** is the differential equation solver, and **diff** is the derivative command for y'. [The command for y'' would be diff$(y(x), x, x)$, and so on for higher derivatives.] The output is the general solution

$$y(x) = \frac{_C1 \ \exp(x)}{x^2 + 4x + 4}$$

where $_C1$ is *Maple* notation for an arbitrary constant.

To solve the same equation, but this time with the initial condition $y(0) = 3$, enter

$$\text{dsolve}(\{(x+2) * \text{diff}(y(x), x) - x * y(x) = 0, \ y(0) = 3\}, \ y(x));$$

and return. The output is the particular solution

$$y(x) = 12\frac{\exp(x)}{x^2 + 4x + 4}$$

which agrees with our result in Example 2.

The dsolve command can cope with differential equations that contain unspecified parameters or functions. For example, to solve $y' + p(x)y = 0$, where $p(x)$ is not specified, enter

$$\text{dsolve}(\text{diff}(y(x), x) - p(x) * y(x) = 0, \ y(x));$$

and return. The output is the general solution

$$y(x) = \exp\left(\int p(x)\,dx\right) _C1$$

The second type of application entails generating the graphical display of various solution curves and/or the direction field for a given differential equation. Both of these tasks can be carried out on *Maple* using the **phaseportrait** command. For example, to obtain the plot shown as Fig. 1, enter

with(DEtools):

to access the phaseportrait command; then return and enter

phaseportrait $(-2 * x * y,\ [x, y],\ x = -1.5..1.5,\ \{[0, -2], [0, -1],$
$[0, 0], [0, 1], [0, 2], [0, 3]\},$ arrows = LINE);

and return. The items within the outer parentheses are as follows:

phaseportrait(right-hand side of $y' = f(x, y)$, [variables], xrange,
{initial points}, optional specification to include direction
field lineal elements and choice of their line thickness);

The yrange is set automatically, but it can be specified as an additional optional item if you wish. All items following the { initial points } are optional, so if you want the yrange to be $-1 \leq y \leq 3$, say, then modify the phaseportrait command as follows:

phaseportrait $(-2 * x * y,\ [x, y],\ x = -1.5..1.5,\ \{[0, -2], [0, -1],$
$[0, 0], [0, 1], [0, 2], [0, 3]\},\ y = -1..3,$ arrows = LINE);

To run phaseportrait over and over, one needs to enter "with(DEtools):" only at the beginning of the session.

To obtain a listing of the mathematical functions and operators (or commands) available in *Maple*, enter **?lib** and return. Within that list one would find such commands as dsolve and phaseportrait. To learn how to use a command enter a question mark, then the command name, then return. For example, type **?dsolve** and return.

In the exercises that follow, and those in subsequent sections, problems are included that require the use of a computer-algebra system such as one of the systems mentioned above. These are important, and we strongly urge you to develop skill in the use of at least one such system in parallel with, but not in place of, developing understanding of the underlying mathematics presented in this text.

EXERCISES 2.2

1. Assuming that $p(x)$ and $q(x)$ are continuous, verify by direct substitution

(a) that (9) satisfies (3) (b) that (21) satisfies (2)

2. In each case find the general solution, both using the "off-the-shelf" formula (21) and then again by actually carrying out the steps of the integrating factor method. That is, find the integrating factor $\sigma(x)$ and then carry out solution steps analogous to those in our derivation of (21). Understand the x interval on which the equation is defined to be the broadest interval on which both $p(x)$ and $q(x)$ are continuous. For example, in part (a) the x interval is $-\infty < x < \infty$, in part (e) it is any interval on which $\tan x$ is continuous (such as $\pi/2 < x < 3\pi/2$), and in part (f) it is either $-\infty < x < 0$ or $0 < x < \infty$ [to ensure the continuity of $p(x) = 2/x$].

(a) $y' - y = 3e^x$ (b) $y' + 4y = 8$
(c) $y' + y = x^2$ (d) $y' = y - \sin 2x$
(e) $y' - (\tan x)y = 6$ (f) $xy' + 2y = x^3$
(g) $xy' - 2y = x^3$ (h) $y' + (\cot x)y = 2\cos x$
(i) $(x-5)(xy' + 3y) = 2$ (j) $\dfrac{dx}{dy} - 6x = e^y$
(k) $y\dfrac{dx}{dy} - y^5 + 3x = 0$ (l) $y^2\dfrac{dx}{dy} + xy - 4y^2 = 1$
(m) $t\dfrac{dx}{dt} - 4t^5 = x$ (n) $\dfrac{dr}{d\theta} + 2(\cot 2\theta)r = 1$

3.(a)–(n) For the equation given in Exercise 2, solve by the method of variation of parameters. That is, first find the homogeneous solution, then vary the parameter, and so on – as we did in (34)–(37) for the general equation (31).

4.(a)–(n) For the equation given in Exercise 2, find the general solution using computer software (such as *Mathematica*, *MATLAB*, or *Maple*). Verify your result by showing that it does satisfy the given differential equation.

5. Solve $xy' + y = 6x^2$ subject to the given initial condition using any method of this section, and state the (broadest) interval of validity of the solution. Also, sketch the graph of the solution, by hand, and label any key values.

(a) $y(1) = 0$ (b) $y(1) = 2$ (c) $y(2) = 2$
(d) $y(-3) = 18$ (e) $y(-3) = -5$ (f) $y(-2) = 8$

6. Solve $xy' + 2y = x + 2$ subject to the given initial condition using any method of this section, and state the (broadest) interval of validity of the solution. Also, sketch the graph of the solution, by hand, and label any key values.

(a) $y(2) = 0$ (b) $y(0) = 1$ (c) $y(-1) = 1$
(d) $y(1) = 1$ (e) $y(-2) = 0$ (f) $y(-3) = 0$

7. Find the general solution using any method of this section. The answer may be left in implicit form, rather than explicit form, if necessary. HINT: Remember that which variable is the independent variable and which is the dependent variable is a matter of viewpoint, and one can change one's viewpoint. In these problems, consider whether it might be better to regard x as a function of y, and recall from the calculus that $dy/dx = 1/(dx/dy)$.

(a) $\dfrac{dy}{dx} = \dfrac{1}{x + 3e^y}$ (b) $\dfrac{dy}{dx} = \dfrac{1}{6x + y^2}$
(c) $(6y^2 - x)\dfrac{dy}{dx} - y = 0$ (d) $(y^2 \sin y + x)\dfrac{dy}{dx} = y$

8. (*Direction fields*) The direction field concept was discussed within Example 1. For the differential equation given, use computer software to plot the direction field over the specified rectangular region in the x, y plane, as well as the integral curve through the specified point P. Also, if you can identify any integral curves exactly, from an inspection of the direction field, then give the equations of those curves, and verify that they do satisfy the given differential equation.

(a) $y' = 2 + (2x - y)^3$ on $|x| \le 4$, $|y| \le 4$; $P = (2, 1)$
(b) $y' = y(y^2 - 4)$ on $|x| \le 4$, $|y| \le 4$; $P = (1, 1)$
(c) $y' = (3 - y^2)^2$ on $|x| \le 2$, $|y| \le 3$; $P = (0, 0)$
(d) $y' + 2y = e^{-x}$ on $|x| \le 3$, $|y| \le 2$; $P = (0, 0.5)$
(e) $y' = x^2/(y^2 - 1)$ on $|x| \le 3$, $|y| \le 3$; $P = (-3, -3)$
(f) $y' + x = y$ on $|x| \le 20$, $0 \le y \le 20$; $P = (0, 1)$
(g) $y' = e^x y$ on $0 \le x \le 50$, $0 \le y \le 50$; $P = (0, 10)$
(h) $y' = x \sin y$ on $0 \le x \le 10$, $0 \le y \le 10$; $P = (2, 2)$

9. (*Bernoulli equation*) The equation

$$\boxed{y' + p(x)y = q(x)y^n,} \tag{9.1}$$

where n is a constant (not necessarily an integer), is called **Bernoulli's equation**, after the Swiss mathematician *Jakob Bernoulli*. Jakob (1654–1705), his brother Johann (1667–1748), and Johann's son Daniel (1700–1782), are the best known of the eight members of the Bernoulli family who were prominent mathematicians and scientists.

(a) Give the general solution of (9.1) for the special cases $n = 0$ and $n = 1$.
(b) If n is neither 0 nor 1, then (9.1) is nonlinear. Nevertheless, show that by transforming the dependent variable from

$y(x)$ to $v(x)$ according to $v = y^{1-n}$ (for $n \neq 0, 1$), (9.1) can be converted to the equation

$$v' + (1-n)p(x)v = (1-n)q(x), \tag{9.2}$$

which is linear and can be solved by the methods developed in this section. This method of solution was discovered by *Gottfried Wilhelm Leibniz* (1646–1716) in 1696.

10. Use the method suggested in Exercise 9(b) to find the general solution to each of the following.

(a) $y' - 4y = 4y^2$ (b) $xy' - 2y = x^3y^2$
(c) $2xyy' + y^2 = 2x$ (d) $\sqrt{y}(3y' + y) = x$
(e) $y' = y^2$ (f) $y' = xy^3$

(g) $y'' = (y')^2$ HINT: First, let $y'(x) = u(x)$.
(h) $y''' + (y'')^2 = 0$ HINT: First, let $y''(x) = u(x)$.

11. (*Riccati equation*) The equation

$$\boxed{y' = p(x)y^2 + q(x)y + r(x)} \tag{11.1}$$

is called **Riccati's equation**, after the Italian mathematician *Jacopo Francesco Riccati* (1676–1754). The Riccati equation is nonlinear if $p(x)$ is not identically zero. Recall from Exercise 9 that the Bernoulli equation can always be reduced to a linear equation by a suitable change of variables. Likewise, for the Riccati equation, provided that any one particular solution can be found.

Let $Y(x)$ be any one particular solution of (11.1), as found by inspection, trial and error, or any other means. [Depending on $p(x)$, $q(x)$, and $r(x)$, finding such a $Y(x)$ may be easy, or it may prove too great a task.] Show that by changing the dependent variable from $y(x)$ to $u(x)$ according to

$$y = Y(x) + \frac{1}{u} \tag{11.2}$$

the Riccati equation (11.1) can be converted to the equation

$$u' + [2p(x)Y(x) + q(x)]\, u = -p(x), \tag{11.3}$$

which is linear and can be solved by the methods developed in this section. This method of solution was discovered by *Leonhard Euler* (1707–1783) in 1760.

12. Use the method suggested in Exercise 11 to find the general solution to each of the following. Nonelementary integrals, such as $\int \exp(ax^2)\, dx$, may be left as is.

(a) $y' - 4y = y^2$ HINT: $Y(x) = -4$
(b) $y' = y^2 - xy + 1$ HINT: $Y(x) = x$
(c) $(\cos x)y' = 1 - y^2$ HINT: $Y(x) = \sin x$
(d) $y' = e^{-x}y^2 - y$ HINT: $Y(x) = 2e^x$
(e) $y' = xy^2 + 2x - x^5$ HINT: See if you can find a $Y(x)$ in the form ax^b.
(f) $y' = (1-y)(2-y)$
(g) $y' = y^2 - 4$
(h) $y' = (2-y)y$

13. (*d'Alembert-Lagrange equation*) The first- order *nonlinear* differential equation

$$\boxed{y = xf(p) + g(p)} \tag{13.1}$$

on $y(x)$, where it will be convenient to denote y' as p, and f and g are given functions of p, is known as a **d'Alembert–Lagrange equation** after the French mathematicians *Jean le Rond d'Alembert* (1717–1783) and *Joseph-Louis Lagrange* (1736–1813).

(a) Differentiating (13.1) with respect to x, show that

$$p - f(p) = [xf'(p) + g'(p)]\,\frac{dp}{dx}. \tag{13.2}$$

Observe that this nonlinear equation on $p(x)$ can be converted to a linear equation if we interchange the roles of x and p by now regarding x as the independent variable and p as the dependent variable. Thus, obtain from (13.2) the linear equation

$$\frac{dx}{dp} - \frac{f'(p)}{p - f(p)}x = \frac{g'(p)}{p - f(p)} \tag{13.3}$$

on $x(p)$. Since we have divided by $p - f(p)$ we must restrict $f(p)$ so that $f(p) \neq p$. Solving the simpler equation (13.3) for $x(p)$, the solution of (13.1) is thereby obtained in *parametric form*: $x = x(p)$ from solving (13.3), and $y = x(p)f(p) + g(p)$ from (13.1). This result is the key idea of this exercise, and is illustrated in parts (b)–(c). In parts (d)–(k) we consider a more specialized result, namely, for the case where $f(p)$ happens to have a "fixed point."

(b) To illustrate part (a), consider the equation $y = 2xy' + 3y'$ [i.e., where $f(p) = 2p$ and $g(p) = 3p$], and derive a parametric solution as discussed in (a).

(c) To illustrate part (a), consider the equation $y = x(y' + y'^2)$ [i.e., $f(p) = p + p^2$ and $g(p) = 0$], and derive the parametric solution discussed in (a).

(d) Suppose that $f(p)$ has a **fixed point** P_0, that is, such that $f(P_0) = P_0$. [A given function f may have none, one, or any number of fixed points. They are found as the solutions of the equation $f(p) = p$.] Show that (13.1) then has the straight line

$$y = P_0x + g(P_0) \tag{13.4}$$

cases where the integrals that occur in the general solution of (13.3) are too difficult to evaluate.]

(e) Show that $f(p) = 3p^2$ has two fixed points, $p = 0$ and $p = 1/3$, and hence show that the equation $y = 3xp^2 + g(p)$ has straight-line solutions $y = g(0)$ and $y = \frac{1}{3}x + g\left(\frac{1}{3}\right)$ for any given function g.

(f) Determine all particular solutions of the form (13.4), if any, for the equation $y = x\left(y'^2 - 2y' + 2\right) + e^{y'}$.

(g) Same as (f), for $y = xe^{y'} - 5\cos y'$.

(h) Same as (f), for $y = x\left(y'^2 - 2y'\right) + 6y'^3$.

(i) Same as (f), for $y = x\left(y'^3 - 3y'\right) - 2\sin y'$.

(j) Same as (f), for $y - x\left(y'^2 + 3\right) = y'^5$.

(k) Same as (f), for $y + x\left(2y' + 3\right) = e^{-y'}$.

14. (*Clairaut equation*) For the special case $f(p) = p$, the d'Alembert–Lagrange equation (13.1) in the preceding exercise becomes

$$\boxed{y = xp + g(p),} \tag{14.1}$$

which is known as the **Clairaut equation**, after the French mathematician *Alexis Claude Clairaut* (1713–1765). (Recall that p denotes y' here.)

(a) Verify, by direct substitution into (14.1), that (14.1) admits the family of solutions

$$y = Cx + g(C), \tag{14.2}$$

where C is an arbitrary constant.

(b) Recall that (13.3) does not hold if $f(p) = p$, but (13.2) does. Letting $f(p) = p$ in (13.2), derive the family of solutions (14.2), as well as the additional particular solution given parametrically by

$$x = -g'(p), \tag{14.3a}$$

$$y = -pg'(p) + g(p). \tag{14.3b}$$

(c) To illustrate, find the parametric solution (14.3) for the equation $y = xy' - y'^2$. Show that in this example (14.3) can be gotten into the explicit form $y = x^2/4$ by eliminating the parameter p between (14.3a) and (14.3b). Plot, by hand, the family (14.2), for $C = 0, \pm 1/2, \pm 1, \pm 2$, together with the solution $y = x^2/4$. (Observe, from that plot, that the particular solution $y = x^2/4$ forms an "envelope" of the family of straight-line solutions. Such a solution is called a **singular solution** of the differential equation.)

(d) Instead of a hand plot, do a computer plot of $y = x^2/4$ and the family (14.2), for $C = 0, \pm 0.25, \pm 0.5, \pm 0.75, \ldots, \pm 3$, on $-8 \le x \le 8$, $-10 \le y \le 12$.

2.3 Applications of the Linear Equation

In this section we consider representative physical applications that are governed by linear first-order equations: electrical circuits, radioactivity, population dynamics, and mixing problems, with additional applications introduced in the exercises.

2.3.1. Electrical circuits. In Section 1.3 we discussed the mathematical modeling of a mechanical oscillator. The relevant physics was Newton's second law of motion, which relates the net force on a body to its resulting motion. Thus, we needed to find sufficiently accurate expressions for the forces contributed by the individual elements within that system – the forces due to the spring, the friction between the block and the table, and the aerodynamic drag.

In the case of electrical circuits the relevant underlying physics, analogous to Newton's second law for mechanical systems, is provided by Kirchhoff's laws. Instead of forces and displacements in a mechanical system comprised of various elements such as masses and springs, we are interested now in voltages and currents

in an electrical system comprised of various elements such as resistors, inductors, and capacitors.

First, by a current we mean a flow of charges: the *current* through a given control surface, such as the cross section of a wire, is the charge per unit time crossing that surface. Each electron carries a negative charge of 1.6×10^{-19} *coulomb*, and each proton carries an equal positive charge. Current is measured in *amperes*, with one ampere being a flow of one coulomb per second. By convention, a current is counted as positive in a given direction if it is the flow of positive charge in that direction. While, in general, currents can involve the flow of positive or negative charges, in an electrical circuit the flow is of negative charges, free electrons. Thus, when one speaks of a current of one ampere in a given direction in an electrical circuit one really means the flow of one coulomb per second of negative charges (electrons) in the opposite direction.

Just as heat flows due to a temperature difference, from one point to another, an electric current flows due to a difference in the electric potential, or *voltage*, measured in *volts*.

We will need to know the relationship between the voltage difference across a given circuit element and the corresponding current flow. The circuit elements of interest here are resistors, inductors, and capacitors.

For a **resistor**, the voltage drop $E(t)$, where t is the time (in seconds), is proportional to the current $i(t)$ through it:

$$E(t) = Ri(t), \tag{1}$$

where the constant of proportionality R is called the *resistance* and is measured in *ohms*; (1) is called **Ohm's law**. By a resistor we usually mean an "off-the-shelf" electrical device, often made of carbon, that offers a specified resistance – such as 100 ohms, 500 ohms, and so on. But even the current-carrying wire in a circuit is itself a resistor, with its resistance directly proportional to its length and inversely proportional to its cross-sectional area, though that resistance is probably negligible compared to that of other resistors in the circuit. The standard symbolic representation of a resistor is shown in Fig. 1.

For an **inductor**, the voltage drop is proportional to the time rate of change of current through it:

$$E(t) = L\frac{di(t)}{dt}, \tag{2}$$

where the constant of proportionality L is called the **inductance** and is measured in *henrys*. Physically, most inductors are coils of wire, hence the symbolic representation shown in Fig. 1.

For a **capacitor**, the voltage drop is proportional to the charge $Q(t)$ on the capacitor:

$$E(t) = \frac{1}{C}Q(t), \tag{3}$$

where C is called the **capacitance** and is measured in *farads*. Physically, a capacitor is normally comprised of two plates separated by a gap across which no current

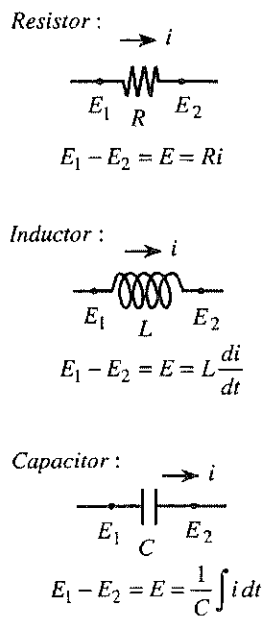


Figure 1. The circuit elements.

flows, and $Q(t)$ is the charge on one plate relative to the other. Though no current flows across the gap, there will be a current $i(t)$ that flows through the circuit that links the two plates and is equal to the time rate of change of charge on the capacitor:

$$i(t) = \frac{dQ(t)}{dt}. \tag{4}$$

From (3) and (4) it follows that the desired voltage/current relation for a capacitor can be expressed as

$$E(t) = \frac{1}{C}\int i(t)\,dt. \tag{5}$$

Now that we have equations (1), (2), and (5) relating the voltage drop to the current, for our various circuit elements, how do we deal with a grouping of such elements within a circuit? The relevant physics that we need, for that purpose, is given by Kirchhoff's laws, named after the German physicist *Gustav Robert Kirchhoff* (1824–1887):

Kirchhoff's current law states that the algebraic sum of the currents approaching (or leaving) any point of a circuit is zero.

Kirchhoff's voltage law states that the algebraic sum of the voltage drops around any loop of a circuit is zero.

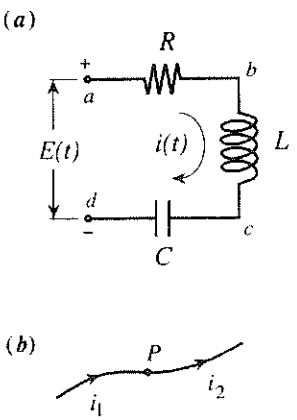


Figure 2. *RLC* circuit.

To apply these ideas, consider the circuit shown in Fig. 2a, consisting of a single loop containing a resistor, an inductor, a capacitor, a voltage source (such as a battery or generator), and the necessary wiring. Let us consider the current $i(t)$ to be positive clockwise; if it actually flows counterclockwise then its numerical value will be negative. In this case Kirchhoff's current law simply says that the current i is a constant from point to point within the circuit and therefore varies only with time. That is, the current law states that at any given point P in the circuit (Fig. 2b), $i_1 + (-i_2) = 0$ or, $i_1 = i_2$. Kirchhoff's voltage law, which is really the self-evident algebraic identity

$$(V_a - V_d) + (V_b - V_a) + (V_c - V_b) + (V_d - V_c) = 0, \tag{6}$$

gives

$$E(t) - Ri - L\frac{di}{dt} - \frac{1}{C}\int i\,dt = 0. \tag{7}$$

The latter is called an **integrodifferential equation** because it contains both derivatives and integrals of the unknown function, but we can convert it to a differential equation in either of two ways.

First, we could differentiate with respect to t to eliminate the integral sign, thereby obtaining

$$\boxed{L\frac{d^2i}{dt^2} + R\frac{di}{dt} + \frac{1}{C}i = \frac{dE(t)}{dt}.} \tag{8}$$

Alternatively, we could use $Q(t)$ instead of $i(t)$ as our dependent variable, for then $\int i\,dt = Q(t)$, and (7) becomes

$$\boxed{L\frac{d^2Q}{dt^2} + R\frac{dQ}{dt} + \frac{1}{C}Q = E(t).} \tag{9}$$

Either way, we obtain a linear second-order differential equation.

Since we are discussing applications of *first*-order linear equations here, let us treat two special cases.

EXAMPLE 1. *RL Circuit.* If we omit the capacitor from our circuit, then (7) reduces to the first-order linear equation*

$$L\frac{di}{dt} + Ri = E(t). \tag{10}$$

If $E(t)$ is a continuous function of time and the current at the initial instant $t = 0$ is $i(0) = i_0$, then the solution to the initial-value problem consisting of (10) plus the initial condition $i(0) = i_0$ is given by (24) in Section 2.2, with "p" $= R/L$ and "q" $= E(t)/L$:

$$i(t) = e^{-\int_0^t \frac{R}{L}\,d\tau}\left(\int_0^t e^{\int_0^\tau \frac{R}{L}\,d\mu}\frac{E(\tau)}{L}\,d\tau + i_0\right),$$

or

$$i(t) = i_0 e^{-Rt/L} + \frac{1}{L}\int_0^t e^{R(\tau-t)/L}E(\tau)\,d\tau \tag{11}$$

over $0 \le t < \infty$, where τ and μ have been used as dummy integration variables.

For instance, if $E(t) = \text{constant} = E_0$, then (11) gives

$$i(t) = i_0 e^{-Rt/L} + \frac{E_0}{R}\left(1 - e^{-Rt/L}\right), \tag{12}$$

or

$$i(t) = \frac{E_0}{R} + \left(i_0 - \frac{E_0}{R}\right)e^{-Rt/L}. \tag{13}$$

As $t \to \infty$, the exponential term in (13) tends to zero, and $i(t) \to E_0/R$. Thus we call the E_0/R term in (13) the **steady-state** solution and the $\left(i_0 - \frac{E_0}{R}\right)e^{-Rt/L}$ term the **transient** part of the solution. The approach to steady state, for several different initial conditions, is shown in Fig. 3.

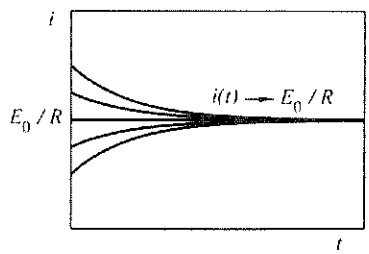


Figure 3. Response $i(t)$ for the case $E(t) = \text{constant} = E_0$; approach to steady state.

As another case, let $E(t) = E_0 \sin\omega t$ and $i_0 = 0$. Then (11) gives

$$i(t) = \frac{E_0\omega L}{R^2 + (\omega L)^2}\left(e^{-Rt/L} + \frac{R}{\omega L}\sin\omega t - \cos\omega t\right). \tag{14}$$

*It may seem curious that if we try deleting the capacitor by setting $C = 0$, then the capacitor term in (7) becomes infinite rather than zero. Physically, however, one can imagine removing the capacitor, in effect, by moving its plates together until they touch. Since the capacitance C varies as the inverse of the gap dimension, then as the gap diminishes to zero $C \to \infty$, and the capacitor term in the differential equation does indeed drop out because of the $1/C$ factor.

As $t \to \infty$, the exponential term in (14) tends to zero, and we are left with the steady-state solution

$$i(t) \to \frac{E_0 \omega L}{R^2 + (\omega L)^2}\left(\frac{R}{\omega L}\sin \omega t - \cos \omega t\right). \qquad (t \to \infty) \tag{15}$$

Observe that by a steady-state solution we mean that which remains after transients have died out; it is not necessarily a constant. For the case where $i(0) = i_0$ and $E(t) = 0$ the steady-state solution is the constant E_0/R, and for the case where $i(0) = 0$ and $E(t) = E_0 \sin \omega t$ the steady-state solution is the oscillatory function given by (15). ∎

EXAMPLE 2. *RC Circuit.* If, instead of removing the capacitor from the circuit shown in Fig. 2, we remove the inductor (so that $L = 0$), then (8) becomes

$$R\frac{di}{dt} + \frac{1}{C}i = \frac{dE(t)}{dt}, \tag{16}$$

which, again, is a first-order linear equation. If we also impose an initial condition $i(0) = i_0$, then

$$i(t) = i_0 e^{-t/RC} + \frac{1}{R}\int_0^t e^{(\tau - t)/RC}\frac{dE(\tau)}{d\tau}\,d\tau \tag{17}$$

gives the solution in terms of i_0 and $E(t)$. ∎

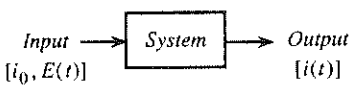


Figure 4. Schematic of the system.

Let us use the electrical circuit problem of Example 1 to make some general remarks. We speak of the initial condition i_0 and the applied voltage $E(t)$ as the **inputs** to the system consisting of the electrical circuit, and the resulting current $i(t)$ as the **output** (or **response**), as denoted symbolically in Fig. 4. From (11), we see that if $i_0 = 0$ and $E(t) = 0$, then $i(t) = 0$: if we put nothing in we get nothing out.*

Consider the inputs and their respective responses separately. If $E(t) = 0$ and $i_0 \neq 0$, then the response

$$i(t) = i_0 e^{-Rt/L}$$

to the input i_0 is seen to be proportional to i_0: if we double i_0 we double its response, if we triple i_0 we triple its response, and so on. Similarly, if $i_0 = 0$ and $E(t)$ is not identically zero, then the response

$$i(t) = \frac{1}{L}\int_0^t e^{R(\tau - t)/L}E(\tau)\,d\tau$$

to the input $E(t)$ is proportional to $E(t)$. This result illustrates an important general property of linear systems: *the response to a particular input is proportional to that input.*

*In contrast with linear initial-value problems, linear *boundary*-value problems can yield nonzero solutions even with zero input – that is, even if the boundary conditions are zero and the equation is homogeneous. These are called **eigensolutions**, and are studied in later chapters.

Further, observe from (11) that the total response $i(t)$ is the sum of the individual responses to i_0 and $E(t)$. This result illustrates the second key property of linear systems: *the response to more than one input is the sum of the responses to the individual inputs.*

In Chapter 3 we prove these two important properties and use them in developing the theory of linear differential equations of second order and higher.

Before closing this discussion of electrical circuits, we wish to emphasize the correspondence, or **analogy**, between the RLC electrical circuit and the mechanical oscillator studied in Section 1.3, and governed by the equation

$$m\frac{d^2x}{dt^2} + c\frac{dx}{dt} + kx = F(t). \tag{18}$$

For we see that both equations (8) (the current formulation) and (9) (the charge formulation) are of exactly the same form as (18). Thus, their mathematical solutions are identical, and hence their physical behavior is identical too. Consider (8), for instance. Comparing it with (18), we note the correspondence

$$L \leftrightarrow m, \quad R \leftrightarrow c, \quad 1/C \leftrightarrow k, \quad i(t) \leftrightarrow x(t), \quad \frac{dE(t)}{dt} \leftrightarrow F(t). \tag{19}$$

Thus, given the values of m, c, k, and the function $F(t)$, we can construct an electrical *analog circuit* by setting $L = m$, $R = c$, $C = 1/k$, and $E(t) = \int F(t)\,dt$. If we also match the initial conditions by setting $i(0) = x(0)$ and $\frac{di}{dt}(0) = \frac{dx}{dt}(0)$, then the resulting current $i(t)$ will be identical to the motion $x(t)$.

Or, we could use (9) to create a different analog, namely,

$$L \leftrightarrow m, \quad R \leftrightarrow c, \quad 1/C \leftrightarrow k, \quad Q(t) \leftrightarrow x(t), \quad E(t) \leftrightarrow F(t). \tag{20}$$

In either case we see that, in mechanical terminology, the inductor provides "inertia" (as does the mass), the resistor provides energy dissipation (as does the friction force), and the capacitor provides a means of energy storage (as does the spring).

Our interest in such analogs is at least twofold. First, to whatever extent we understand the mechanical oscillator, we thereby also understand its electrical analog circuit, and vice versa. Second, if the system is too complex to solve analytically, we may wish to study it experimentally. If so, by virtue of the analogy we have the option of studying whichever is more convenient. For instance, it would no doubt be simpler, experimentally, to study the RLC circuit than the mechanical oscillator.

Finally, just as Hooke's law can be derived theoretically using the governing partial differential equations of the theory of elasticity, our circuit element relations (1)–(5) can be derived using the theory of electromagnetism, the governing equations of which are the celebrated *Maxwell's equations*. We will meet some of the Maxwell's equations later on in this book, when we study scalar and vector field theory.

2.3.2. Radioactive decay; carbon dating. Another important application of first-order linear equations involves radioactive decay and carbon dating.

Radioactive materials, such as carbon–14, einsteinium–253, plutonium–241, radium–226, and thorium–234, are found to decay at a rate that is proportional to the amount of mass present. This observation is consistent with the supposition that the disintegration of a given nucleus, within the mass, is independent of the past or future disintegrations of the other nuclei, for then the number of nuclei disintegrating, per unit time, will be proportional to the total number of nuclei present:

$$\frac{dN}{dt} = -kN, \tag{21}$$

where k is known as the disintegration constant, or decay rate. Actually, the graph of $N(t)$ proceeds in unit steps since $N(t)$ is integer-valued, so $N(t)$ is discontinuous and hence nondifferentiable. However, if N is very large, then the steps are very small compared to N. Thus, we can regard N, approximately, as a continuous function of t and can tolerate the dN/dt derivative in (21). However, it is inconvenient to work with N since one cannot count the number of atoms in a given mass. Thus, we multiply both sides of (21) by the atomic mass, in which case (21) becomes the simple first-order linear equation

$$\frac{dm}{dt} = -km, \tag{22}$$

where $m(t)$ is the total mass, a quantity which is more readily measured. Solving, by means of either (9) or (24) in Section 2.2, we obtain

$$m(t) = m_0 e^{-kt}, \tag{23}$$

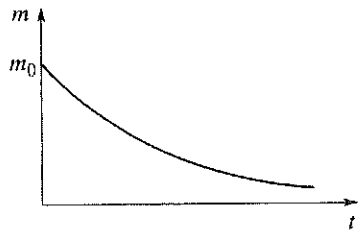


Figure 5. Exponential decay.

where $m(0) = m_0$ is the initial amount of mass (Fig. 5). This result is indeed the exponential decay that is observed experimentally.

Since k gives the rate of decay, it can be expressed in terms of the **half-life** T of the material, the time required for any initial amount of mass m_0 to be reduced by half, to $m_0/2$. Then (23) gives

$$\frac{m_0}{2} = m_0 e^{-kT},$$

so $k = (\ln 2)/T$, and (23) can be re-expressed in terms of T as

$$m(t) = m_0 2^{-t/T}. \tag{24}$$

Thus, if $t = T, 2T, 3T, 4T \ldots$, then $m(t) = m_0, m_0/2, m_0/4, m_0/8$, and so on.

Radioactivity has had an important archeological application in connection with **dating**. The basic idea behind any dating technique is to identify a physical process that proceeds at a known rate. If we measure the state of the system now, and we know its state at the initial time, then from these two quantities together with the known rate of the process, we can infer how much time has elapsed; the mathematics enables us to "travel back in time as easily as a wanderer walks up a frozen river."*

*Ivar Ekeland, *Mathematics and the Unexpected*. Chicago: University of Chicago Press, 1988.

In particular, consider carbon dating, developed by the American chemist *Willard Libby* in the 1950's. The essential idea is as follows. Cosmic rays consisting of high-velocity nuclei penetrate the earth's lower atmosphere. Collisions of these nuclei with atmospheric gases produce free neutrons. These, in turn, collide with nitrogen, thus changing some of the nitrogen to carbon–14, which is radioactive, and which decays to nitrogen–14 with a half-life of around 5,570 years. Thus, some of the carbon dioxide which is formed in the atmosphere contains this radioactive C–14. Plants absorb both radioactive and nonradioactive CO_2, and humans and animals inhale both and also eat the plants. Consequently, the countless plants and animals living today contain both C–12 and, to a much lesser extent, its radioactive isotope C–14, in a ratio that is essentially the same from one plant or animal to another.

EXAMPLE 3. *Carbon Dating.* Consider a wood sample that we wish to date. Since C–14 emits approximately 15 beta particles per minute per gram, we can determine how many grams of C–14 are contained in the sample by measuring the rate of beta particle emission. Suppose that we find that the sample contains 0.2 gram of C–14, whereas if it were alive today it would, based upon its weight, contain around 2.6 grams. Thus, we assume that it contained 2.6 grams of C–14 when it died. That mass of C–14 will have decayed, over the subsequent time span t, to 0.2 gram. Then (24) gives

$$0.2 = (2.6)\,2^{-t/5570},$$

and, solving for t, we determine the sample to be around $t = 2{,}100$ years old.

However, it must be emphasized that this method (and the various others that are based upon radioactive decay) depend critically upon assumptions of uniformity. To date the wood sample studied in this example, for instance, we need to know the amount of C–14 present in the sample when the tree died, and what the decay rate was over the time period in question. To apply the method, we assume, first, that the decay rate has remained constant over the time period in question and, second, that the ratio of the amounts of C–14 to C–12 was the same when the tree died as it is today. Observe that although these assumptions are usually stated as fact they can never be proved, since it is too late for direct observation and the only evidence available now is necessarily circumstantial. ∎

2.3.3. Population dynamics. In this application, we are again interested in the variation of a population $N(t)$ with the time t, not the population of atoms this time, but the population of a particular species such as fruit flies or human beings.

According to the simplest model, the rate of change dN/dt is proportional to the population N:

$$\frac{dN}{dt} = \kappa N, \tag{25}$$

where the constant of proportionality κ is the net birth/death rate, that is, the birth rate minus the death rate. As in our discussion of radioactive decay, we regard $N(t)$ as continuous because the unit steps in N are extremely small compared to N itself.

Solving (25), we obtain the exponential behavior

$$N(t) = N_0 e^{\kappa t}, \tag{26}$$

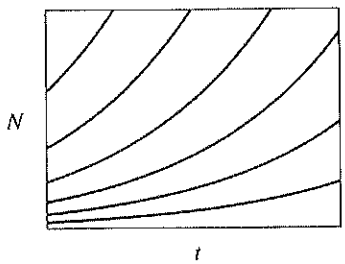


Figure 6. Exponential growth.

where $N(0) = N_0$ is the initial condition. If the death rate exceeds the birth rate, then $\kappa < 0$ and (26) expresses exponential decrease, with $N \to 0$ as $t \to \infty$. That result seems fair enough. However, if $\kappa > 0$, then (26) expresses exponential growth, with $N \to \infty$ as $t \to \infty$, as displayed in Fig. 6 for several different initial conditions N_0. That result is unrealistic because as N becomes sufficiently large other factors will undoubtedly come into play, such as insufficient food or other resources. In other words, we expect that κ will not really be a constant but will vary with N. In particular, we expect it to decrease as N increases. As a simple model of such behavior, suppose that κ varies linearly with N: $\kappa = a - bN$, with a and b positive, so that κ diminishes as N increases, and even becomes negative when N exceeds a/b. Then (25) is to be replaced by the equation

$$\frac{dN}{dt} = (a - bN)N. \tag{27}$$

The latter is known as the **logistic equation**, or the **Verhulst equation**, after the Belgian mathematician *P. F. Verhulst* (1804–1849) who introduced it in his work on population dynamics. Due to the N^2 term, the equation is *nonlinear*, so that the solution that we developed in Section 2.2 does not apply. However, the Verhulst equation is interesting, and we will return to it.

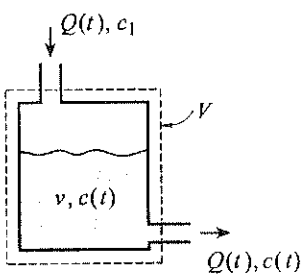


Figure 7. Mixing tank.

2.3.4. Mixing problems. In this final application we consider a mixing tank with an inflow of $Q(t)$ gallons per minute and an equal outflow, where t is the time; see Fig. 7. The inflow is at a constant concentration c_1 of a particular solute (pounds per gallon), and the tank is constantly stirred, so that the concentration $c(t)$ within the tank is uniform. Hence, the outflow is at concentration $c(t)$. Let v denote the volume within the tank, in gallons; v is a constant because the inflow and outflow rates are equal. To keep track of the instantaneous mass of solute $x(t)$ within the tank, let us carry out a mass balance for the "control volume" V (dashed lines in the figure):

$$\begin{array}{c}\text{Rate of increase}\\ \text{of mass of solute}\\ \text{within } V\end{array} = \text{Rate in} - \text{Rate out}, \tag{28}$$

$$\frac{dx}{dt}\frac{\text{lb}}{\text{min}} = \left(Q(t)\frac{\text{gal}}{\text{min}}\right)\left(c_1\frac{\text{lb}}{\text{gal}}\right) - \left(Q(t)\frac{\text{gal}}{\text{min}}\right)\left(c(t)\frac{\text{lb}}{\text{gal}}\right), \tag{29}$$

or, since $c(t) = x(t)/v$,

$$\frac{dx(t)}{dt} + \frac{Q(t)}{v}x(t) = c_1 Q(t), \tag{30}$$

which is a first-order linear equation on $x(t)$. Alternatively, we have the first-order linear equation

$$\frac{dc(t)}{dt} + \frac{Q(t)}{v}c(t) = \frac{c_1 Q(t)}{v} \tag{31}$$

on the concentration $c(t)$.

Recall that in modeling a physical system one needs to incorporate the relevant physics such as Newton's second law or Kirchoff's laws. In the present application, the relevant physics is provided entirely by (28). To better understand (28), suppose we rewrite it with one more term included on the right-hand side:

$$\begin{array}{ccccccc} \text{Rate of increase} & & \text{Rate} & & \text{Rate} & & \text{Rate of creation} \\ \text{of mass of solute} & = & \text{into} & - & \text{out of} & + & \text{of mass} \\ \text{within } V & & V & & V & & \text{within } V. \end{array} \tag{32}$$

The equation (32) is merely a matter of logic, or bookkeeping, not physics. Since (28) follows from (32) only if there is no creation (or destruction) of mass, we can now understand (28) to be a statement of the physical principle of *conservation of mass*, namely, that matter can neither be created nor destroyed (except under exceptional circumstances that are not present in this situation).

Closure. In this section we study applications of first-order linear equations to electrical circuit theory, to radioactivity and population dynamics, and to mixing problems. Although our RLC circuit gives rise to a second-order differential equation, we find that we can work with first-order equations if we omit either the inductor or the capacitor. We will return to the RLC circuit when we discuss second-order equations, so the background provided here, including the expressions for the voltage/current relations and Kirchoff's two laws, will be drawn upon at that time.

The electrical circuit applications also gives us an opportunity to emphasize the extremely important consequences of the linearity of the differential equation upon the relationship between the input and output. The key ideas are that for a linear system: (1) *the response to a particular input is proportional to that input*, and (2) *the response to more than one input is the sum of the responses to the individual inputs*. These ideas are developed and proved in Chapter 3.

EXERCISES 2.3

NOTE: Thus far we have assumed that $p(x)$ and $q(x)$ in $y' + p(x)y = q(x)$ are continuous, yet in applications that may not be the case. In particular, the "input" $q(x)$ may be discontinuous. In Example 1, for instance, $E(t)$ in $L\,di/dt + Ri = E(t)$ may well be discontinuous, such as

$$E(t) = \begin{cases} E_0, & 0 < t < t_1 \\ 0, & t_1 < t < \infty. \end{cases}$$

We state that in such cases, where $E(t)$ has one or more jump discontinuities, the solution (11) [more generally, (24) in Section 2.2] is still valid, and can be used in these exercises.

1. (*RL circuit*) For the RL circuit of Example 1, with $i_0 = 0$ and $E(t) = E_0$, determine the

(a) time required for $i(t)$ to reach 99% of its steady-state value;
(b) resistance R needed to ensure that $i(t)$ will attain 99% of

its steady-state value within 2 seconds, if $L = 0.001$ henry;
(c) inductance L needed to ensure that $i(t)$ will attain 99% of its steady-state value within 0.5 seconds, if $R = 50$ ohm.

2. (*RL circuit*) For the RL circuit of Example 1, suppose that $i(0) = i_0$ and that $E(t)$ is as given below. In each case, determine $i(t)$ and identify the steady-state solution. If a steady state does not exist, then state that. Also, sketch the graph of $i(t)$ and label key values.

(a) $$E(t) = \begin{cases} E_0, & 0 < t < t_1 \\ 0, & t_1 < t < \infty \end{cases}$$

(b) $$E(t) = \begin{cases} 0, & 0 < t < t_1 \\ E_0, & t_1 < t < \infty \end{cases}$$

(c) $$E(t) = \begin{cases} 0, & 0 < t < t_1 \\ E_0, & t_1 < t < t_2 \\ 0, & t_2 < t < \infty \end{cases}$$

3. (*RC circuit*) (a) For the RC circuit of Example 2, suppose that $i_0 = 0$ and that $E(t) = E_0 e^{-Rt/L}$. Solve for $i(t)$ and identify the steady-state solution, treating these cases separately: $R^2C \neq L$, and $R^2C = L$. If there does not exist a steady state, then state that. Sketch the graph of $i(t)$ and label any key values.
(b) Same as (a), but with $R = C = 1$ and $E(t) = E_0 \sin t$.

4. Verify that (14) can be re-expressed as

$$i(t) = \frac{E_0 \omega L}{R^2 + (\omega L)^2} e^{-Rt/L} + \frac{E_0}{\sqrt{R^2 + (\omega L)^2}} \sin(\omega t - \phi),$$

where ϕ is the (unique) angle between 0 and $\pi/2$ such that $\tan \phi = \omega L/R$; ϕ is called the **phase angle.**

5. A seashell contains 90% as much C–14 as a living shell of the same size. How old is it? Approximately how many years did it take for its C–14 content to diminish from its initial value to 99% of that?

6. If 10 grams of some radioactive substance will be reduced to 8 grams in 60 years, in how many years will 2 grams be left? In how many years will 0.1 gram be left?

7. If 20% of a radioactive substance disappears in 70 days, what is its half-life?

8. Show that if m_1 and m_2 grams of a radioactive substance are present at times t_1 and t_2, respectively, then its half-life is

$$T = (t_2 - t_1)\frac{\ln 2}{\ln(m_1/m_2)}.$$

9. (*Verhulst equation*) Solve the Verhulst equation (27), subject to the initial condition $N(0) = N_0$, two ways:

(a) by noting that it is a Bernoulli equation;
(b) by noting that it is (also) a Riccati equation.
NOTE: The Bernoulli and Riccati equations, and their solutions, were discussed in the exercises for Section 2.2. (The Verhulst equation can also be solved by the method of separation of variables, which method is the subject of the next section.)

10. (*Mixing tank*) For the mixing tank governed by (31):

(a) Let $Q(t) = \text{constant} = Q$ and $c(0) = c_0$. Solve for $c(t)$.
(b) Let $Q(t) = 4$ for $0 < t < 1$ and 2 for $t > 1$, and let $v = c_1 = 1$ and $c(0) = 0$. Solve for $c(t)$. HINT: The application of (24) in Section 2.2 is not so hard when $q(x)$ in the differential equation $y' + p(x)y = q(x)$ is defined piecewise (e.g., as in Exercise 2 above), but is tricky when $p(x)$ is defined piecewise. In this exercise we suggest that you use (24) to solve for $c(t)$ first for $0 < t < 1$, with "a"=0 and "b" = $c(0) = 0$. Then, use that solution to evaluate $c(1)$ and use (24) again, for $1 < t < \infty$, this time with "a"= 1 and "b" = $c(1)$, where $c(1)$ has already been determined.
(c) Let $Q(t) = 2$, $c_1 = 0$, $v = 1$, and $c(0) = 0.3$. Solve for $c(t)$ and thus show that although $c(t) \to 0$ as $t \to \infty$, it never actually reduces to zero, so that it is not possible to wash every molecule of solute out of the tank. Does this result make sense? Explain.

11. (*Mass on an inclined plane*) The equation $mx'' + cx' = mg \sin\alpha$ governs the straight-line displacement $x(t)$ of a mass m along a plane that is inclined at an angle α with respect to the horizontal, if it slides under the action of gravity and friction. If $x(0) = 0$ and $x'(0) = 0$, solve for $x(t)$. HINT: First, integrate the equation once with respect to t to reduce it to a first-order linear equation.

12. (*Free fall; terminal velocity*) The equation of motion of a body of mass m falling vertically under the action of a downward gravitational force mg and an upward aerodynamic drag force $f(v)$, is

$$mv' = mg - f(v), \tag{12.1}$$

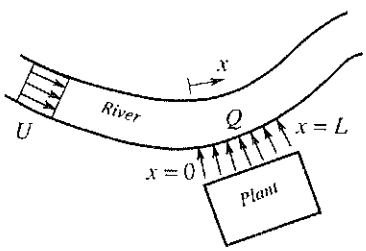
where $v(t)$ is the velocity [so that $v'(t)$ is the acceleration]. The determination of the form of $f(v)$, for the given body shape, would require either careful wind tunnel measurements, or sophisticated theoretical and/or numerical analysis, the result being a plot of the nondimensional drag coefficient versus the nondimensional Reynolds number. All we need to know here is that for a variety of body shapes, the result of such an analysis is the determination that $f(v)$ can be approximated

(over some limited range of velocities) in the form cv^β, for suitable constants c and β. For low velocities (more precisely, for low Reynolds numbers) $\beta \approx 1$, and for high velocities (i.e., for high Reynolds numbers) $\beta \approx 2$.

(a) Solve (12.1), together with the initial condition $v(0) = 0$, for the case where $f(v) \approx cv$. What is the terminal (i.e., steady-state) velocity?
(b) Same as (a), for $f(v) \approx cv^2$. HINT: Read Exercise 11 in Section 2.2.

13. (*Light extinction*) As light passes through window glass some of it is absorbed. If x is a coordinate normal to the glass (with $x = 0$ at the incident face) and $I(x)$ is the light intensity at x, then the fractional loss in intensity, $-dI/I$ (with the minus sign included because dI will be negative), will be proportional to dx: $-dI/I = k\,dx$, where k is a positive constant. Thus, $I(x)$ satisfies the differential equation $I'(x) = -kI(x)$. The problem: If 80% of the light penetrates a 1-inch thick slab of this glass, how thin must the glass be to let 95% penetrate? NOTE: Your answer should be numerical, not in terms of an unknown k.

14. (*Pollution in a river*) Suppose that a pollutant is discharged into a river at a steady rate Q (grams/second) over a distance L, as sketched in the figure, and we wish to



determine the distribution of pollutant in the river – that is, its concentration c (grams/meter^3). Measure x as arc length along the river, positive downstream. The river flows with velocity U (meters/second) and has a cross-sectional area A (meters^2), both of which, for simplicity, we assume to be constant. Also for simplicity, suppose that c is a function of x only. That is, it is a constant over each cross section of the stream. This is evidently a poor approximation near the interval $0 < x < L$, where we expect appreciable across-stream and vertical variations in c, but it should suffice if we are concerned mostly with the far field, that is, more than several river widths upstream or downstream of the interval $0 < x < L$. Then it can be shown that $c(x)$ is governed by the differential equation

$$kc'' - Uc' - \beta c = -\frac{Q(x)}{A}, \quad (-\infty < x < \infty) \qquad (14.1)$$

where k (meters^2/second) is a diffusion constant, β (grams per second per gram) is a chemical decay constant, and $Q(x)$ is the constant Q over $0 < x < L$ and 0 outside that interval. [Physically, (14.1) expresses a mass balance between the *input* $-Q(x)/A$, the transport of pollutant by *diffusion*, kc'', the transport of pollutant by *convection* with the moving stream, Uc', and by disappearance through *chemical decay*, βc.] We assume that the river is clear upstream; that is, we have the initial condition $c(-\infty) = 0$.

(a) Let $L = \infty$. Suppose that k is sufficiently small so that we can neglect the diffusion term. Then (14.1) reduces to the first-order linear equation $Uc' + \beta c = Q(x)/A$. Solve for $c(x)$ and sketch its graph, labeling any key values.
(b) Repeat part (a) for the case where L is finite.

15. (*Newton's law of cooling*) Suppose that a body initially at a uniform temperature u_0 is exposed to a surrounding environment that is at a lower temperature U. Then the outer portion of the body will cool relative to its interior, and this temperature differential within the body will cause heat to flow from the interior to the surface. If the body is a sufficiently good conductor of heat so that the heat transfer within the body is much more rapid than the rate of heat loss to the environment at the outer surface, then it can be assumed, as an approximation, that heat transfer will be so rapid that the interior temperature will adjust to the surface temperature instantaneously, and the body will be at a uniform temperature $u(t)$ at each instant t. **Newton's law of cooling** states that the time rate of change of $u(t)$ will be proportional to the instantaneous temperature difference $U - u$, so that

$$\frac{du}{dt} = k(U - u), \qquad (15.1)$$

where k is a constant.

(a) Solve (15.1) for $u(t)$ subject to the initial condition $u(0) = u_0$. NOTE: Actually, it is not necessary that $U < u_0$; (15.1) is equally valid if $U > u_0$. In most physical applications, however, one is interested in a hot body (such as a cup of coffee or a hot ingot) in a cooler environment.
(b) An interesting application of (15.1) occurs in connection with the determination of the time of death in a homicide. Suppose that a body is discovered at a time T after death and its temperature is measured to be 90° F. We wish to solve for T. Suppose that the ambient temperature is $U = 70°$ F and assume that $u_0 = 98.6°$ F. Putting this information into the solution to (15.1) we can solve for T, provided that we know

k, but we don't. Proceeding indirectly, we can infer the value of k by taking one more temperature reading. Thus, suppose that we wait an hour and again measure the temperature of the body, and find that $u(T+1) = 87°$ F. Use this information to solve for T.

16. (*Compound interest*) Suppose that a sum of money earns interest at a rate k, compounded yearly, monthly, weekly, or even daily. If it is compounded continuously, then $dS/dt = kS$, where $S(t)$ denotes the sum at time t. If $S(0) = S_0$, then the solution is

$$S(t) = S_0 e^{kt}. \tag{16.1}$$

Instead, suppose that interest is compounded yearly. Then after t years

$$S(t) = S_0(1+k)^t,$$

and if the compounding is done n times per year, then

$$S(t) = S_0\left(1+\frac{k}{n}\right)^{nt}. \tag{16.2}$$

(a) Show that if we let $n \to \infty$ in (16.2), then we do recover the continuous compounding result (16.1). HINT: Recall, from the calculus, that

$$\lim_{m\to\infty}\left(1+\frac{1}{m}\right)^m = e.$$

(b) Let $k = 0.05$ (i.e., 5% interest) and compare $S(t)/S_0$ after 1 year ($t = 1$) if interest is compounded yearly, monthly, weekly, daily, and continuously.

2.4 Separable Equations

2.4.1. Separable equations. The general first-order differential equation is of the form

$$F(x, y, y') = 0. \tag{1}$$

If we can solve (1), by algebra, for y', then we can re-express it in the form

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

which form we take as our starting point for this section.

Actually, it is conceivable that we cannot solve (1) for y'. For instance, the equation

$$xy' - y = \sin y' + 4$$

or, equivalently,

$$F(x, y, y') = xy' - y - \sin y' - 4 = 0,$$

cannot be solved, by algebra, for y' in terms of x and y. However, such cases are rarely encountered in applications and will not be considered here. Thus, we assume that the equation can be expressed in the form (2).

If, further, $f(x, y)$ can be expressed as a function of x times a function of y, so that (2) can be written as

$$\boxed{y' = X(x)Y(y),} \tag{3}$$

then we say that the differential equation is **separable**. For instance, $y' = x\exp(x+2y)$ is separable because we can factor $x\exp(x+2y)$ as $x\exp x$ times $\exp(2y)$, but $y' = 3x - y$ is not.

To solve (3), we divide both sides by $Y(y)$ (if $Y(y) \neq 0$) and integrate both sides with respect to x:

$$\int \frac{1}{Y(y)} y' \, dx = \int X(x) \, dx, \tag{4}$$

or, since $y'dx = dy$, from the differential calculus,

$$\boxed{\int \frac{dy}{Y(y)} = \int X(x) \, dx.} \tag{5}$$

We also know from the integral calculus that if $1/Y(y)$ is a continuous function of y (over the relevant y interval) and $X(x)$ is a continuous function of x (over the relevant x interval), then the two integrals in (5) exist, in which case (5) gives the general solution of (2).

EXAMPLE 1. Solve the equation

$$y' = -y^2. \tag{6}$$

Though not linear, (6) is separable. Separating the variables and integrating gives

$$\int \frac{dy}{y^2} = -\int dx, \tag{7}$$

$$-\frac{1}{y} + C_1 = -x + C_2, \tag{8}$$

where C_1 and C_2 are arbitrary. With $C = C_1 - C_2$, we have the general solution

$$y(x) = \frac{1}{x + C}. \tag{9}$$

If we impose an initial condition $y(0) = y_0$ then we can solve for C and obtain the particular solution

$$y(x) = \frac{1}{x + 1/y_0} = \frac{y_0}{1 + y_0 x}, \tag{10}$$

which is plotted in Fig. 1 for the representative values $y_0 = 1$ and $y_0 = 2$. The solution through the initial point $(0, 1)$ exists over $-1 < x < \infty$, the one through (0,2) exists over $-1/2 < x < \infty$. More generally, the one through $(0, y_0)$ exists over $-1/y_0 < x < \infty$ because the denominator in (10) becomes zero at $x = -1/y_0$. We could plot (10) to the left of that point as well, but such extension of the graph would be meaningless because the point $x = -1/y_0$ serves as a "barrier;" y and y' fail to exist there, so the solution cannot be continued beyond that point. ∎

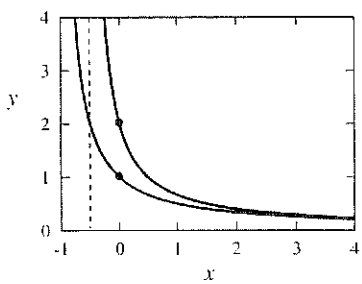


Figure 1. Particular solutions given by (10).

EXAMPLE 2. Solve the initial-value problem

$$y' = \frac{4x}{1 + 2e^y}; \qquad y(0) = 1. \tag{11}$$

Though not linear, the differential equation is separable and can be solved accordingly:

$$\int (1 + 2e^y)\, dy = \int 4x\, dx, \tag{12}$$

$$y + 2e^y = 2x^2 + C. \tag{13}$$

Unfortunately, the latter is a transcendental equation in y, so we cannot solve it for y explicitly as a function of x, as we were able to solve (8). Nevertheless, we can impose the initial condition on (13) to evaluate C: $1 + 2e = 0 + C$, so $C = 1 + 2e$ and the solution is given, in "implicit" form, by

$$y + 2e^y = 2x^2 + 1 + 2e. \tag{14}$$

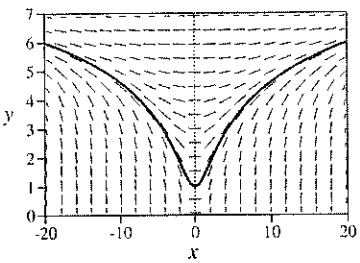


Figure 2. The solution (14) of(11).

The resulting solution is plotted in Fig. 2, along with the direction field. [Actually, we did not plot (14) in Fig. 2; we used the following *Maple* phaseportrait commands to solve (11) and to plot the solution:

with (DEtools):

phaseportrait($4 * x/(1 + 2 * \exp(y)$, $[x, y]$, $x = -20..20$, $\{[0, 1]\}$, stepsize $= 0.05$, arrows=LINE);

where the default stepsize was too large and gave a jagged curve, so we reduced it to 0.05, and where we also included the direction field to give us a feeling for the overall "flow."]

COMMENT 1. Observe that if we use the definite integrals

$$\int_1^y (1 + 2e^y)\, dy = \int_0^x 4x\, dx,$$

with the lower limits dictated by the initial condition $y(0) = 1$, then we bypass the need for an integration constant C and its subsequent evaluation.

COMMENT 2. What is the interval of existence of the solution? In Example 1 we were able to ascertain that interval by direct examination of the solution (10). Here, however, such examination is not possible because the solution (14) is in implicit form. It appears, from Fig. 2, that the solution exists for all x, but of course Fig. 2 covers only $-20 < x < 20$. Equation (14) reveals the asymptotic behavior $2e^y \sim 2x^2$, or $y \sim 2\ln|x|$ as $|x| \to \infty$, so it seems clear that the solution continues to grow smoothly as $|x|$ increases. ∎

2.4.2. Existence and uniqueness. (Optional) In this section we have begun to solve nonlinear differential equations. Before we get too deeply involved in solution techniques, let us return to the more fundamental questions of existence and uniqueness of solutions. For the linear equation

$$y' + p(x)y = q(x) \tag{15}$$

we have Theorem 2.2.1, which tells us that (15) does admit a solution through an initial point $y(a) = b$ if $p(x)$ and $q(x)$ are continuous at a. That solution is unique, and it exists at least on the largest x interval containing $x = a$, over which $p(x)$ and $q(x)$ are continuous. What can be said about existence and uniqueness for the more general equation $y' = f(x, y)$ (which could, of course, be linear, but, in general, is not)?

THEOREM 2.4.1 *Existence and Uniqueness*
If $f(x, y)$ is continuous on some rectangle R in the x, y plane containing the point (a, b), then the problem

$$y' = f(x, y); \qquad y(a) = b \tag{16}$$

has at least one solution defined on some open x interval* containing $x = a$. If, in addition, $\partial f/\partial y$ is continuous on R, then the solution to (16) is unique on some open interval containing $x = a$.

Notice that whereas Theorem 2.2.1 predicts the minimum interval of existence and uniqueness, Theorem 2.4.1 merely ensures existence and uniqueness over *some* interval; it gives no clue as to how broad that interval will be. Thus, we say that Theorem 2.4.1 is a *local* result; it tells us that under the stipulated conditions all is well locally, in some neighborhood of $x = a$. More informative theorems could be cited, but this one will suffice here.

Let us illustrate Theorem 2.4.1 with two examples.

EXAMPLE 3. The equation

$$y' = \frac{y(y-2)}{x(y-1)} \tag{17}$$

is separable, and separating the variables gives

$$\int \frac{y-1}{y(y-2)}\, dy = \int \frac{dx}{x}. \tag{18}$$

By partial fractions (which method is reviewed in Appendix A),

$$\frac{y-1}{y(y-2)} = \frac{1}{2}\frac{1}{y} + \frac{1}{2}\frac{1}{y-2}. \tag{19}$$

With this result, integration of (18) gives

$$\frac{1}{2}\ln|y| + \frac{1}{2}\ln|y-2| = \ln|x| + C, \qquad (-\infty < C < \infty) \tag{20}$$

where C is the arbitrary constant of integration. Equivalently,

$$\ln\left|\frac{y(y-2)}{x^2}\right| = 2C, \tag{21}$$

*By an **open interval** we mean $x_1 < x < x_2$, and by a **closed interval** we mean $x_1 \le x \le x_2$. Thus, a closed interval includes its endpoints, an open interval does not. It is common to use the notation (x_1, x_2) and $[x_1, x_2]$ for open and closed intervals, respectively. Further, $(x_1, x_2]$ means $x_1 < x \le x_2$, and $[x_1, x_2)$ means $x_1 \le x < x_2$.

so

$$\left|\frac{y(y-2)}{x^2}\right| = e^{2C} \equiv B, \qquad (0 \le B < \infty) \tag{22}$$

where B is introduced for convenience and is nonnegative because $\exp(2C)$ is nonnegative. Thus,

$$\frac{y(y-2)}{x^2} = \pm B \equiv A, \qquad (-\infty < A < \infty) \tag{23}$$

where A replaces the "$\pm B$." Finally, (23) gives $y^2 - 2y - Ax^2 = 0$ so, by the quadratic formula, we have the general solution

$$y(x) = 1 \pm \sqrt{1 + Ax^2} \tag{24}$$

of (17).

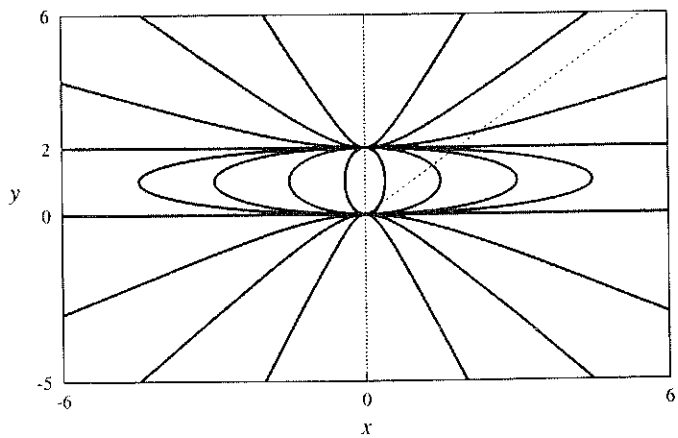


Figure 3. Solution curves corresponding to equation (17).

These solution curves are plotted in Fig. 3. The choice $A = 0$ gives the solution curves $y(x) = 0$ and $y(x) = 2$. As representative of solutions above the line $y = 2$, consider the initial condition $y(1) = 4$. Then (24) gives $y(1) = 4 = 1 \pm \sqrt{1+A}$, which requires that we select the plus sign and $A = 8$, so $y(x) = 1 + \sqrt{1+8x^2}$. As representative of solutions below the line $y = 0$, consider the initial condition $y(1) = -3$. Then (24) gives $y(1) = -3 = 1 \pm \sqrt{1+A}$, which requires that we select the minus sign and $A = 15$, so $y(x) = 1 - \sqrt{1+15x^2}$. Finally, as representative of the solutions between $y = 0$ and $y = 2$, consider the initial condition $y(2) = 3/2$, say. Then $y(2) = 3/2 = 1 \pm \sqrt{1+4A}$, so we choose the plus sign and $A = -3/16$, in which case (24) gives $y(x) = 1 + \sqrt{1 - 3x^2/16}$, namely, the upper branch of the ellipse $\frac{3}{16}x^2 + (y-1)^2 = 1$.

In terms of the Existence and Uniqueness Theorem 2.4.1, observe that the conditions of the theorem are met everywhere in the x, y plane except along the vertical line $x = 0$ and the horizontal line $y = 1$, and indeed we do have breakdowns in existence and uniqueness all along these lines. On $x = 0$ (the y axis) there are no solutions through initial points other than $y = 0$ and $y = 2$ (lack of existence), and through each of those points there

are an infinite number of solutions (lack of uniqueness). Initial points on the line $y = 1$ are a bit more subtle. We do have elliptical solution curves through each such point, yet at the initial point (on $y = 1$) the slope is infinite, so the differential equation (17) cannot be satisfied at that point. Thus, we have a breakdown in existence for each initial point on $y = 1$. Further, realize that for any such ellipse, between $y = 0$ and $y = 2$, the upper and lower halves are separate solutions. For instance, the ellipse $(3x/4)^2 + (y-1)^2 = 1$, mentioned above, really amounts to the separate solutions $y(x) = 1 \pm \sqrt{1-(3x/4)^2}$, each valid over $-4/3 < x < 4/3$.

COMMENT. Observe that the right side of (17) is asymptotic to y/x as $y \to \pm\infty$, so the solutions of (17) should be asymptotic to the solutions of the simpler equation $y' = y/x$ as $y \to \pm\infty$, namely, the family of straight lines through the origin, and that result can be seen, somewhat, in Fig. 3. ∎

EXAMPLE 4. *Free Fall.* This time consider a physical application. Suppose that a body of mass m is dropped, from rest, at time $t = 0$. With its displacement $x(t)$ measured downward from the point of release, the equation of motion is $mx'' = mg$, where g is the acceleration of gravity and t is the time. Thus,

$$x'' = g, \qquad (0 \le t < \infty) \tag{25a}$$

$$x(0) = 0, \tag{25b}$$

$$x'(0) = 0. \tag{25c}$$

Equation (25a) is of second order, whereas this chapter is about first-order equations, but it is readily integrated twice with respect to t. Doing so, and invoking (25b) and (25c) gives the solution

$$x(t) = \frac{g}{2}t^2, \tag{26}$$

which result is probably familiar to you from a first course in physics.

However, instead of multiplying (25a) through by dt and integrating on t, let us multiply it by dx and integrate on x. Then $x''dx = g\,dx$ and since, from the calculus,

$$x''dx = \frac{dx'}{dt}dx = \frac{dx'}{dt}\frac{dx}{dt}dt = x'\frac{dx'}{dt}dt = x'dx', \tag{27}$$

$x''dx = g\,dx$ becomes

$$x'dx' = g\,dx. \tag{28}$$

Integrating (28) gives

$$\frac{1}{2}x'^2 = gx + A, \tag{29}$$

and $x(0) = x'(0) = 0$ imply that $A = 0$. Thus, we have reduced (25) to the *first*-order problem

$$x' = \sqrt{2g}\,x^{1/2}, \qquad (0 \le t < \infty) \tag{30a}$$

$$x(0) = 0, \tag{30b}$$

which shall now be the focus of this example. Equation (30a) is separable and readily solved. The result is the general solution

$$x(t) = \frac{1}{4}\left(\sqrt{2g}\,t + C\right)^2, \tag{31}$$

which is shown, for various values of C, in Fig. 4. Applying (30b) gives $C = 0$, so (31) gives $x(t) = gt^2/2$, in agreement with (26). However, from the figure we can see that although a solution exists over the full t interval of interest ($t \geq 0$), that solution is not unique because other solutions satisfying both (30a) and (30b) are given by the curve $x(t) = 0$ from the origin up to any point Q, followed by the parabola QR. Physically, the solution OQR corresponds to the mass levitating until time Q, then beginning its descent.

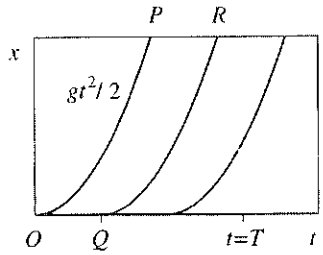


Figure 4. Nonuniqueness of the solution to (30).

Surely that sounds physically impossible, but let us look at the mathematics. We cannot apply Theorem 2.2.1 because (30) is nonlinear, but we can use Theorem 2.4.1 (with x and y replaced by t and x, of course). Since $f(t,x) = \sqrt{2g}\,x^{1/2}$, we see that f is continuous for all $t \geq 0$ and $x \geq 0$, but $f_x(t,x) = \sqrt{g/2}\,x^{-1/2}$ is not continuous over any interval containing the initial point $x = 0$. Thus, the theorem tells us that there does exist a solution over some t interval containing $t = 0$ (which turns out to be the entire positive t axis), but it does not guarantee uniqueness over any such interval, and as it turns out we do not have uniqueness over any such interval.

Next, consider the physics. When we multiply force by distance we get work, and work shows up (in a system without dissipation, as in this example) as energy. Thus, multiplying (25a) by dx and integrating converted the original force equation (Newton's second law) to an energy equation. That is, (29) tells us that the total energy (kinetic plus potential) is conserved; it is constant for all time: $x'^2/2 + (-gx) =$ constant or, equivalently,

$$\frac{1}{2}mx'^2 + (-mgx) = A. \tag{32}$$

$$\text{Kinetic energy} + \text{Potential energy} = \text{Constant.}$$

Since $x(0) = x'(0) = 0$, the total energy A is zero. When the mass falls, its kinetic energy becomes positive and its potential energy becomes negative such that their total remains zero for all $t > 0$. However, the energy equation is also satisfied if the released mass levitates for any amount of time and then falls, or if indeed it levitates for all time [that is $x(t) = 0$ for all $t \geq 0$]. Thus, our additional solutions are indeed physically meaningful in that they do satisfy the requirement of conservation of energy. Observe, however, that they do not satisfy the equation of motion (25a) since the insertion of $x(t) = 0$ into that equation gives $0 = g$. Thus, the spurious additional solution $x(t) = 0$ must have entered somewhere between (25) and (30). In fact, we introduced it inadvertently when we multiplied (25a) by dx because $x''dx = g\,dx$ is satisfied not only by $x'' = g$, but also by $dx = 0$ [i.e., by $x(t) =$ constant].

The upshot is that although the solution to (30) is nonunique, a look at our derivation of (30) shows that we should discount the solution $x(t) = 0$ of (30) since it does not also satisfy the original equation of motion $x'' = g$. In that case we are indeed left with the unique solution $x(t) = gt^2/2$, corresponding to the parabola OP in Fig. 4. ∎

It is important to understand that the solution $x(t) = 0$ of (30) is not contained within the general solution (31), for any finite choice of C. Such an additional solution is known as a **singular solution**, and brief consideration of these will be reserved for the exercises.

2.4.3. Applications. Let us study two physical applications of the method of separation of variables.

EXAMPLE 5. *Gravitational Attraction.* **Newton's law of gravitation** states that the force of attraction F exerted by any one point mass M on any other point mass m is*

$$F = G\frac{Mm}{d^2}, \tag{33}$$

where d is the distance between them and $G(= 6.67 \times 10^{-8}\,\text{cm}^3/\text{g sec}^2)$ is called the **universal gravitational constant**; (33) is said to be an *inverse-square law* since the force varies as the inverse square of the distance. (By M and m being point masses, we mean that their sizes are negligible compared with d.)

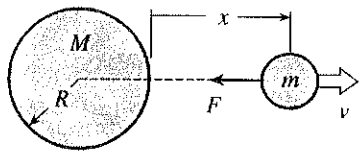


Figure 5. Rocket launch.

Consider the linear motion of a rocket of mass m that is launched from the surface of the earth, as sketched in Fig. 5, where M and R are the mass and radius of the earth, respectively. From Newton's second law of motion and his law of gravitation, it follows that the equation of motion of the rocket is

$$m\frac{d^2x}{dt^2} = -G\frac{Mm}{(x+R)^2}. \tag{34}$$

Although (34) is a second-order equation, we can reduce it to one of first order by noting that

$$\frac{d^2x}{dt^2} = \frac{d}{dt}\left(\frac{dx}{dt}\right) = \frac{dv}{dt} = \frac{dv}{dx}\frac{dx}{dt} = v\frac{dv}{dx}, \tag{35}$$

*Newton derived (33) from *Kepler's laws* of planetary motion which, in turn, were inferred empirically from the voluminous measurements recorded by the Danish astronomer Tycho Brahe (1546–1601). Usually, in applications (not to mention homework assignments in mechanics), one is given the force exerted on a mass and asked to determine the motion by twice integrating Newton's second law of motion. In deriving (33), however, Newton worked "backwards:" the motion of the planets was supplied in sufficient detail by Kepler's laws, and Newton used those laws to infer the force needed to sustain that motion. It turned out to be an inverse-square force directed toward the sun. Being aware of other such forces between masses, for example, the force that kept his shoes on the floor, Newton then proposed the bold generalization that (33) holds not just between planets and the sun, but between any two bodies in the universe; hence the name *universal law of gravitation*. Just as it is difficult to overestimate the importance of Newton's law of gravitation and its impact upon science, it is also difficult to overestimate how the idea of a force acting at a distance, rather than through physical contact, must have been incredible when first proposed.

In fact, such eminent scientists and mathematicians as Huygens, Leibniz, and John Bernoulli referred to Newton's idea of gravitation as absurd and revolting. Imagine Newton's willingness to stand nonetheless upon the results of his mathematics, in inferring the concept of gravitation, even in the absence of any physical mechanism or physical plausibility, and in the face of such opposition.

Remarkably, *Coulomb's law* subsequently stated an inverse-square type of electrical attraction or repulsion between two charges. Why these two types of force field turn out to be of the same mathematical form is not known. Equally remarkable is the fact that although the forms of the two laws are identical, the magnitudes of the forces are staggeringly different. Specifically, the ratio of the electrical force of repulsion to the gravitational force exerted on each other by two electrons (which is independent of the distance of separation) is 4.17×10^{42}.

where v is the velocity, and where the third equality follows from the chain rule. Thus (34) becomes the first-order equation

$$v\frac{dv}{dx} = -\frac{GM}{(x+R)^2}, \tag{36}$$

which is separable and gives

$$\int v\,dv = -GM\int \frac{dx}{(x+R)^2}, \tag{37}$$

$$\frac{v^2}{2} = \frac{GM}{x+R} + C. \tag{38}$$

If the launch velocity is $v(0) = V$, then (38) gives $C = (V^2/2) - GM/R$, so

$$v = \sqrt{V^2 - \frac{2GM}{R}\frac{x}{x+R}} \tag{39}$$

is the desired expression of v as a function of x.

If we wish to know $x(t)$ as well, then we can re-write (39) as

$$\frac{dx}{dt} = \sqrt{V^2 - \frac{2GM}{R}\frac{x}{x+R}}, \tag{40}$$

which once again is variable separable and can be solved for $x(t)$. However, let us be content with (39).

Observe from (39) that v decreases monotonically with increasing x, from its initial value $v = V$ to $v = 0$, the latter occurring at

$$x_{max} = \frac{V^2R^2}{2GM - V^2R}. \tag{41}$$

Subsequently, the rocket will be drawn back toward the earth and will strike it with speed V. [We need to choose the negative square root in (39) to track the return motion.] Equation (41) can be simplified by noting that when $x = 0$, the right-hand side of (34) must be $-mg$, where g is the familiar gravitational acceleration at the earth's surface. Thus, $-mg = -GMm/R^2$, so $GM/R^2 = g$, and (41) becomes

$$x_{max} = \frac{V^2R}{2gR - V^2}. \tag{42}$$

We see from (42) that x_{max} increases as V is increased, as one would expect, and becomes infinite as $V \to \sqrt{2gR}$. Thus, the critical value $V_e = \sqrt{2gR}$ is the *escape velocity*. Numerically, $V_e \approx 6.9$ miles/sec.

COMMENT 1. Recall that the law of gravitation (33) applies to two point masses separated by a distance d, whereas the earth is hardly a point mass. Thus, it is appropriate to question the validity of (34). In principle, to find the correct attractive force exerted on the rocket by the earth we need to consider the earth as a collection of point masses dM, compute the force dF induced by each dM, and add the dF's vectorially to find the resultant force F.

This calculation is carried out later, in Section 15.7, and the result, remarkably, is that the resultant F acting at any point P outside the earth (or *any* homogeneous spherical mass), per unit mass at P, is the same as if its entire mass M were concentrated at a single point, namely, at its center! Thus, the earth might as well be thought of as a point mass, of mass M, located at its center, so (34) is exactly true, if we are willing to approximate the earth as a homogeneous sphere.

COMMENT 2. The steps in (35), whereby we were able to reduce our second-order equation (34) to the first-order equation (36), were not limited to this specific application. They apply whenever the force is a function of x alone, for if we apply (35) to the equation

$$m\frac{d^2x}{dt^2} = f(x), \tag{43}$$

we get the separable first-order equation

$$mv\frac{dv}{dx} = f(x) \tag{44}$$

with solution

$$\frac{mv^2}{2} = \int f(x)\,dx + C \tag{45}$$

or, equivalently,

$$\left.\frac{mv^2}{2}\right|_{x_1}^{x_2} = \int_{x_1}^{x_2} f(\xi)\,d\xi. \tag{46}$$

In the language of mechanics, the right-hand side is the work done by the force $f(x)$ as the body moves from x_1 to x_2, and $mv^2/2$ is the kinetic energy. Thus, the physical significance of (46) is that it is a work-energy statement: the change in the kinetic energy of the body is equal to the work done on it.

COMMENT 3. Observe the change in viewpoint as we progressed from (34) to (36). Until the third equality in (35), we regarded x and v as dependent variables – functions of the independent variable t. But beginning with the right-hand side of that equality, we began to regard v as a function of x. However, once we solved (36) for v in terms of x, in (39), we replaced v by dx/dt, and x changed from independent variable to dependent variable once again. In general, then, which variable is regarded as the independent variable and which is the dependent variable is not so much figured out, as it is a decision that we make, and that decision, or viewpoint, can sometimes change, profitably, over the course of the solution. ∎

EXAMPLE 6. *Verhulst Population Model.* Consider the Verhulst population model

$$N'(t) = (a - bN)N; \qquad N(0) = N_0 \tag{47}$$

that was introduced in Section 2.3.3, where $N(t)$ is the population of a given species. This example emphasizes that a given equation might be solvable by a number of different methods. Though (47) is not a linear equation, it is both a Bernoulli equation and a Riccati equation, which equations were discussed in the exercises of Section 2.2. Now we see that

it is also separable, since the right side is a function of N [namely, $(a - bN)N$] times a function of t (namely, 1). Thus,

$$\int \frac{dN}{(a - bN)N} = \int dt. \tag{48}$$

By partial fractions,

$$\frac{1}{(a - bN)N} = -\frac{1}{b}\frac{1}{(N - \frac{a}{b})N} = -\frac{1}{a}\frac{1}{N - \frac{a}{b}} + \frac{1}{a}\frac{1}{N}$$

so (48) gives

$$-\frac{1}{a}\ln\left|N - \frac{a}{b}\right| + \frac{1}{a}\ln N = t + C, \tag{49}$$

where C is an arbitrary constant $(-\infty < C < \infty)$. [Whether we write $\ln|N|$ or $\ln N$ in (49) is immaterial since $N > 0$.] Equivalently,

$$\left|\frac{N}{N - \frac{a}{b}}\right|^{1/a} = e^{t+C}, \qquad \left|\frac{N}{N - \frac{a}{b}}\right| = e^{at+aC} = Be^{at}, \tag{50}$$

where we have replaced $\exp(aC)$ by B, so $0 \le B < \infty$. Thus

$$\frac{N}{N - a/b} = \pm Be^{at} \equiv Ae^{at}, \tag{51}$$

where A is arbitrary $(-\infty < A < \infty)$. Finally, imposing the initial condition $N(0) = N_0$ gives $A = N_0/(N_0 - a/b)$, and putting that expression into (50) and solving for N gives

$$N(t) = \frac{aN_0}{(a - bN_0)e^{-at} + bN_0}. \tag{52}$$

What can be learned of the behavior of $N(t)$ from (52)? We can see from (52) that for every initial value N_0 (other than $N_0 = 0$), $N(t)$ tends to the constant value a/b as $t \to \infty$. [If $N_0 = 0$, then $N(t) = 0$ for all t, as it should, because if a species starts with no members it can hardly wax or wane.] Beyond observing that asymptotic information, it is an excellent idea to plot the results, especially now that one has such powerful and convenient computer software for that purpose. However, observe that the solution (52) contains the three parameters a, b, and N_0, and to use plotting software we need to choose numerical values for these parameters. If, for instance, we wish to plot $N(t)$ versus t for five values of a, five of b, and five of N_0, then we will be generating $5^3 = 125$ curves! Thus, the point is that if we wish to do a parametric study of the solution (i.e., examine the solution for a range of values of the various parameters), then there is a serious problem with managing all of the needed plots. In Section 2.4.4 below, we offer advice on how to deal with this common and serious predicament. ∎

2.4.4. Nondimensionalization. (Optional) One can usually reduce the number of parameters in a problem, sometimes dramatically, by a suitable scaling of the

independent and dependent variables so that the new variables are nondimensional (i.e., dimensionless).

EXAMPLE 7. *Example 6, Continued.* To begin such a process of nondimensionalization, we list all dependent and independent variables and parameters, and their dimensions:

	Variable	Dimensions	Parameter	Dimensions
Independent:	t	time	a	1/time
Dependent:	N	number	b	1/[(time)(number)]
			N_0	number

(By number we mean the number of living members of the species.) How did we know that a has dimensions of 1/time, and that b has dimensions of 1/[(time)(number)]? From the differential equation $\dfrac{dN}{dt} = aN - bN^2$. That is, the dimensions of the term on the left are number/time, so the dimensions of aN and bN^2 must be the same. Dimensionally, then, aN = number/time, so a = 1/time. Similarly, bN^2 = number/time, so b = 1/[(time)(number)].

Next, we nondimensionalize the independent and dependent variables (t and N) using suitable combinations of the parameters. From the parameter list, observe that $1/a$ has dimensions of time and can therefore be used as a "reference time" to nondimensionalize the independent variable t. That is, we can introduce a nondimensional version of t, say $\bar{t}$, by $\bar{t} = t/(1/a) = at$.

Next, we need to nondimensionalize the dependent variable N. From the parameter list, observe that N_0 has dimensions of number, so let us introduce a nondimensional version of N, say $\overline{N}$, by $\overline{N} = N/N_0$. In case the notion of nondimensionalization still seems unclear, realize that it is merely a change of variables, from t and N to $\bar{t}$ and $\overline{N}$; a rather simple change of variables in fact, since $\bar{t}$ is simply a constant times t, and $\overline{N}$ is simply a constant times N.

Putting $t = \bar{t}/a$ and $N = N_0\overline{N}$ into (47) gives

$$aN_0\frac{d\overline{N}}{d\bar{t}} = \left(a - bN_0\overline{N}\right)N_0\overline{N}; \qquad N_0\overline{N}(0) = N_0, \tag{53}$$

where the left side of the differential equation follows from the chain differentiation $\dfrac{dN}{dt} = \dfrac{dN}{d\overline{N}}\dfrac{d\overline{N}}{d\bar{t}}\dfrac{d\bar{t}}{dt} = (N_0)\left(\dfrac{d\overline{N}}{d\bar{t}}\right)(a) = aN_0\dfrac{d\overline{N}}{d\bar{t}}$. (More simply, but less rigorously, we could merely replace the dN in dN/dt by $N_0 d\overline{N}$ and the dt by $d\bar{t}/a$.) Simplifying (53) gives

$$\frac{d\overline{N}}{d\bar{t}} = \left(1 - \alpha\overline{N}\right)\overline{N}; \qquad \overline{N}(0) = 1, \tag{54}$$

where $\alpha \equiv bN_0/a$. Thus, (54) contains only the single parameter α. The solution of (54) is

$$\overline{N}(\bar{t}) = \frac{1}{\alpha + (1-\alpha)e^{-\bar{t}}}. \tag{55}$$

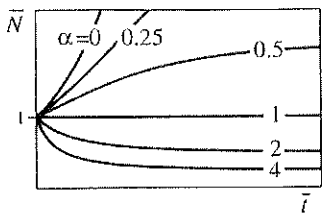


Figure 6. Nondimensional solution of Verhulst problem.

The idea is that if we plot $\overline{N}(\overline{t})$ versus $\overline{t}$ (rather than $N(t)$ versus t), then we have only the *one*-parameter family of solutions given by (55), where the parameter is the nondimensional quantity $\alpha = bN_0/a$. Those solutions are shown in Fig. 6 for several different values of α. As $\overline{t} \to \infty$ (and hence $t \to \infty$), $\overline{N} \to 1/\alpha$, so $N/N_0 \to 1/(bN_0/a)$, or $N \to a/b$, as found in Example 4.

COMMENT. The nondimensionalization of the independent and dependent variables can often be done in more than one way. In the present example, for instance, we used N_0 to nondimensionalize N: $\overline{N} = N/N_0$. However, a/b also has the dimensions of number, so we could have defined $\overline{N} = N/(a/b) = bN/a$ instead. Similarly, we could have nondimensinalized t differently, as $\overline{t} = N_0bt$, because N_0b has dimensions of 1/time. Any nondimensionalization will work, and we leave these other choices for the exercises. ∎

EXAMPLE 8. *Example 5, Continued.* As one more example of the simplifying use of nondimensionalization, consider the initial-value problem

$$\frac{dx}{dt} = \sqrt{V^2 - \frac{2GM}{R}\frac{x}{x+R}}; \qquad x(0) = R \tag{56}$$

from Example 5. As above, we begin by listing all variables and parameters, and their dimensions:

	Variable	Dimensions	Parameter	Dimensions
Independent:	t	time	V	length/time
Dependent:	x	length	R	length

We didn't bother with G and M in the parameter list because V and R are all we need to nondimensionalize t and x. Specifically, R has dimensions of length, so we can choose $\overline{x} = x/R$, and R/V has dimensions of time, so we can choose $\overline{t} = t/(R/V)$. Putting $x = R\overline{x}$ and $t = R\overline{t}/V$ into (56) gives

$$\frac{R}{R/V}\frac{d\overline{x}}{d\overline{t}} = \sqrt{V^2 - \frac{2GM}{R}\frac{R\overline{x}}{R\overline{x}+R}}; \qquad R\overline{x}(0) = R \tag{57}$$

or

$$\frac{d\overline{x}}{d\overline{t}} = \sqrt{1 - \alpha\frac{\overline{x}}{\overline{x}+1}}; \qquad \overline{x}(0) = 1, \tag{58}$$

with the single parameter $\alpha \equiv 2GM/RV^2$. Since all other quantities in the final differential equation are nondimensional, it follows that α must be nondimensional as well, as could be checked from the known dimensions of G, M, R, and V.

Of course, whereas we've used the generic dimensions "time" and "length," we could have used specific dimensions such as seconds and meters. ∎

It is common in engineering and science to nondimensionalize the governing equations and initial or boundary conditions even before beginning the solution, so as to reduce the number of parameters as much as possible. In each of the foregoing two examples we ended up with a single parameter, but the final number of

parameters will vary from case to case. The nondimensional parameters that result (such as α in Example 6) are sometimes well known and of great importance. For instance, if one nondimensionalizes the differential equations governing fluid flow, two nondimensional parameters that arise are the *Reynolds number* Re and the *Mach number* M. Without getting into the fluid mechanics, let it suffice to say that the Reynolds number is a measure of the relative importance of viscous effects to inertial effects: if the Reynolds number is sufficiently large one can neglect the viscous terms in the governing equations of motion, and if it is sufficiently small then one can neglect the inertial terms. Similarly, the Mach number is a measure of the importance of the effects of the compressibility of the fluid: if M is sufficiently small then one can neglect those effects and consider the fluid to be incompressible. In fact, any given approximation that is made in engineering science is probably based upon whether some relevant nondimensional parameter is sufficiently large or small, for one is always neglecting one effect *relative* to others.

Closure. We see that the method of separation of variables is relatively simple: one separates the variables and integrates. Thus, given a specific differential equation, one is well advised to see immediately if the equation is of separable type and, if it is, to solve by separation of variables. Of course, it might turn out that one or both of the integrations are difficult, but the general rule of thumb is that there is a conservation of difficulty, so that if the integrations are difficult, then an equivalent difficulty will show up if one tries a different solution technique.

In the last part of this section we discuss the idea of nondimensionalization. The latter is not central to the topic of this section, separation of variables, but arises tangentially with regard to the efficient management of systems that contain numerous parameters, which situation makes graphical display and general understanding of the results more difficult.

Computer software. A potential difficulty with the method of separation of variables is that the integrations involved may be difficult. Using *Maple*, for instance, integrations can be carried out using the **int** command. For example, to evaluate the integral on the left side of (48), enter

$$\text{int}(1/((a - b * N) * N),\ N);$$

and return. The output is

$$-\frac{\ln(a - bN)}{a} + \frac{\ln(N)}{a}$$

which (to within an additive constant) is the same as the left side of equation (49). That is not to say that all integrals can be evaluated in closed form by computer software. To illustrate the use of the int command for definite integrals, consider the integral of e^{-ax} from $x = 0$ to $x = \infty$. Enter

$$\text{int}(\exp(-x),\ x = 0..\text{infinity});$$

and return. The output is 1.

Of course, if we are going to use *Maple* to evaluate the integrals that arise, then we might as well see if the *Maple* dsolve command will solve the differential equation directly, as is discussed at the end of Section 2.2. For instance, to solve the equation $y' = -3x^2y^5$, enter

$$\text{dsolve}(\text{diff}(y(x), x) = -3 * x\text{^}2 * (y(x))\text{^}5,\ y(x));$$

and return. The output is the general solution

$$\frac{1}{y(x)^4} = 4x^3 + _C1$$

in implicit form, where $_C1$ is the arbitrary constant of integration.

EXERCISES 2.4

NOTE: Solutions should be expressed in explicit form if possible.

1. Use separation of variables to find the general solution. Then, obtain the particular solution satisfying the given initial condition. Sketch the graph of the solution, showing the key features, and label any key values.

(a) $y' - 3x^2e^{-y} = 0;\quad y(0) = 0$
(b) $y' = 6x^2 + 5;\quad y(0) = 0$
(c) $y' + 4y = 0;\quad y(-1) = 0$
(d) $y' = 1 + y^2;\quad y(2) = 5$
(e) $y' = (y^2 - y)e^x;\quad y(0) = 2$
(f) $y' = y^2 + y - 6;\quad y(5) = 10$
(g) $y' = y(y + 3);\quad y(0) = -4$
(h) $y' = 6\dfrac{y \ln y}{x};\quad y(1) = e$
(i) $y' = e^{x+2y};\quad y(0) = 1$
(j) $y' = \dfrac{y}{2x};\quad y(3) = -1$
(k) $y' + 3y(y + 1)\sin 2x = 0;\quad y(0) = 1$
(l) $y = \ln y';\quad y(0) = 5$
(m) $y' = y \ln y;\quad y(0) = 5$
(n) $y' + 2y = y^2 + 1;\quad y(-3) = 0$

2.(a)–(n) For the equation given in Exercise 1, use computer software to solve for $y(x)$. Verify, by direct substitution, that your solution does satisfy the given differential equation and initial condition.

3. The problem $du/dt = k(U - u)$; $u(0) = u_0$, where k and U are constants, occurred in the exercises for Section 2.3 in connection with Newton's law of cooling. Solve by separation of variables.

4. The Verhulst population problem

$$N'(t) = (a - bN)N; \qquad N(0) = N_0 \qquad (a > 0,\ b > 0)$$

was studied in Section 2.3 and solved as a Bernoulli equation, and also as a Riccati equation. Here we ask you to solve it by separation of variables.

5. The Bernoulli equation $y' + p(x)y = q(x)y^n$ is not variable separable, in general, but it is if p and q are constants, if one of the functions $p(x)$ and $q(x)$ is zero, or if one is a constant times the other. Obtain the general solution for the case where each is a nonzero constant, for any real number n. HINT: A difficult integral will occur. Our discussion of the Bernoulli equation in the exercises for Section 2.2 should help you to find a change of variables that will simplify that integration.

6. Solve $y' = (6x^2 + 1)/(y - 1)$, subject to the given initial condition.

(a) $y(0) = -2$ (b) $y(0) = 4$ (c) $y(0) = 0$
(d) $y(1) = 3$ (e) $y(2) = 4$ (f) $y(-1) = 0$

7. Solve $y' = (3x^2 - 1)/2y$, subject to the given initial condition.

(a) $y(0) = -3$ (b) $y(0) = -1$ (c) $y(4) = 5$
(d) $y(-1) = 0$ (e) $y(-2) = -4$ (f) $y(1) = -6$

8. (*Homogeneous functions*) A function $f(x_1, \ldots, x_n)$ is said to be **homogeneous of degree k** if $f(\lambda x_1, \ldots, \lambda x_n) = \lambda^k f(x_1, \ldots, x_n)$ for any λ. For example,

$$f(x, y, z) = \frac{4x^5}{y^2 + 3z^2} \sin\left(\frac{y}{z}\right)$$

is homogeneous of degree 3 because

$$f(\lambda x, \lambda y, \lambda z) = \frac{4(\lambda x)^5}{(\lambda y)^2 + 3(\lambda z)^2} \sin\left(\frac{\lambda y}{\lambda z}\right) = \lambda^3 f(x, y, z).$$

State whether f is homogeneous or not. If it is, determine its degree.

(a) $f(x, y) = x^2 + 4y^2 - 7$
(b) $f(x, y, z) = \cos\left(\dfrac{x - y}{5z}\right)$
(c) $f(x, y) = x^2 - y^2 + 7xz - 3xy$
(d) $f(x, y) = \sin\left(x^2 + y^2\right)$

9. (*Homogeneous equation*) The equation

$$\boxed{y' = f\left(\frac{y}{x}\right)} \tag{9.1}$$

is said to be **homogeneous** because $f(y/x)$ is homogeneous (of degree zero); see the preceding exercise. CAUTION: The term homogeneous is also used to describe a linear differential equation that has zero as its "forcing function" on the right-hand side, as defined in Section 1.2. Thus, one needs to use the context to determine which meaning is intended.

(a) Show, by examples, that (9.1) may, but need not, be separable.
(b) In any case, show that the change of dependent variable $w = y/x$, from $y(x)$ to $w(x)$, reduces (9.1) to the separable form

$$w' = \frac{f(w) - w}{x} \tag{9.1}$$

10. Use the idea contained in the preceding exercise, to find the general solution to each of the following equations.

(a) $y' = \dfrac{y}{x} + 3\sqrt{\dfrac{x}{y}}$ (b) $y' = \dfrac{2y - x}{y - 2x}$
(c) $y' = \dfrac{xy + 2y^2}{x^2}$ (d) $y' = -\dfrac{2x + y}{x}$
(e) $y' = e^{y/x} + \dfrac{y}{x}$

11. (*Almost-homogeneous equation*) (a) Show that

$$y' = \frac{ax + by + c}{dx + ey + f} \qquad (a, b, \ldots, f \text{ constants}) \tag{11.1}$$

can be reduced to homogeneous form by the change of variables $x = u + h$, $y = v + k$, where h and k are suitably chosen constants, provided that $ae - bd \neq 0$.
(b) Thus, find the general solution of $y' = (2x - y - 6)/(x - y - 3)$.
(c) Similarly, for $y' = (1 - y)/(x + 4y - 3)$.
(d) Similarly, for $y' = (x + y)/(x - y + 1)$.
(e) Similarly, for $y' = (x - y - 4)/(x + y - 4)$.
(f) Devise a method of solution that will work in the exceptional case where $ae - bd = 0$, and apply it to the case $y' = (x + 2y - 1)/(2x + 4y - 1)$.

12. (*Algebraic, exponential, and explosive growth*) We saw, in Section 2.3.3, that the population model

$$\frac{dN}{dt} = \kappa N \qquad (\kappa > 0) \tag{12.1}$$

gives exponential growth, whereby $N \to \infty$ as $t \to \infty$. More generally, consider the model

$$\frac{dN}{dt} = \kappa N^p, \qquad (\kappa > 0) \tag{12.2}$$

where p is a positive constant. Solve (12.2) and show that if $0 < p < 1$ then the solution exhibits **algebraic growth** [i.e., $N(t) \sim \alpha t^\beta$ as $t \to \infty$]. Show that as $p \to 0$ the exponent β tends to unity, and as $p \to 1$ the exponent β tends to infinity. (Of course, when $p = 1$ we then have **exponential growth**, as mentioned above, so we can think – crudely – of exponential growth as a limiting case of algebraic growth, in the limit as the exponent β becomes infinite. Thus, exponential growth is powerful indeed.) If p is increased beyond 1 then we expect the growth to be even more spectacular. Show that if $p > 1$ then the solution exhibits **explosive growth**, explosive in the sense that $N \to \infty$ in *finite* time, as $t \to T$, where

$$T = \frac{1}{\kappa(p - 1)N_0^{p-1}}; \tag{12.3}$$

N_0 denotes the initial value $N(0)$. Observe that T diminishes as p increases.

13. (*Nondimensionalization*) In Example 7 we nondimensionalized according to $\bar{t} = at$ and $\overline{N} = N/N_0$. Instead, nondimensionalize (47) according to $\bar{t} = at$ and $\overline{N} = bN/a$, and thus derive the solution

$$\overline{N}(\bar{t}) = \frac{\beta}{\beta + (1 - \beta)e^{-\bar{t}}},$$

where $\beta \equiv bN_0/a$. Sketch the graph of $\overline{N}(\bar{t})$ versus $\bar{t}$, for several different values of β, labeling any key value(s).

14. The initial-value problem

$$mx'' + cx' + kx = F \sin \omega t; \qquad x(0) = x_0, \quad x'(0) = x_0' \tag{14.1}$$

corresponding to a damped mechanical oscillator driven by the force $F \sin \omega t$, contains seven parameters: $m, c, k, F, \omega, x_0, x_0'$. Nondimensionalize (14.1). How many parameters are present in the nondimensionalized system?

2.5 Exact Equations and Integrating Factors

Thus far we have developed solution techniques for first-order differential equations that are linear or separable. In addition, Bernoulli, Riccati, Clairaut, homogeneous, and almost-homogeneous equations were discussed in the exercises. In this section we consider one more important case, equations that are "exact," and ones that are not exact but can be made exact.

First, let us review some information, from the calculus, about partial derivatives. Specifically, recall that the symbol $\dfrac{\partial^2 f}{\partial x \partial y}$ is understood to mean $\dfrac{\partial}{\partial x}\left(\dfrac{\partial f}{\partial y}\right)$. If we use the standard subscript notation instead, then this quantity would be expressed as f_{yx}, that is, $(f_y)_x$. Does the order of differentiation matter? That is, is $f_{yx} = f_{xy}$? It is shown in the calculus that a sufficient condition for f_{xy} to equal f_{yx} is that f_x, f_y, f_{yx}, and f_{xy} all be continuous within the region in question. These conditions are met so typically in applications, that in textbooks on engineering and science f_{xy} and f_{yx} are generally treated as indistinguishable. Here, however, we will treat them as equal only if we explicitly assume the continuity of f_x, f_y, f_{yx}, and f_{xy}.

2.5.1. Exact differential equations. To motivate the idea of exact equations, consider the equation

$$\frac{dy}{dx} = \frac{\sin y}{2y - x\cos y} \tag{1}$$

or, rewritten in differential form,

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

If we notice that the left-hand side is the differential of $F(x,y) = x\sin y - y^2$, then (2) is simply $dF = 0$, which can be integrated to give $F =$ constant; that is,

$$F(x,y) = x\sin y - y^2 = C, \tag{3}$$

where C is an arbitrary constant of integration. Equation (3) gives the general solution to (1), in implicit form.

Really, our use of the differential form (2) begs justification since we seem to have thereby treated dy/dx as a fraction of computable quantities dy and dx, whereas it is actually the limit of a difference quotient. Such justification is possible, but it may suffice to note that the use of differentials is a matter of convenience and is not essential to the method. For instance, observe that if we write

$$\sin y + (x\cos y - 2y)\frac{dy}{dx} = 0 \tag{4}$$

in place of (2), to avoid any questionable use of differentials, then the left-hand side of (4) is the x derivative (total, not partial) of $F(x,y) = x\sin y - y^2$:

$$\frac{d}{dx}F(x,y(x)) = \frac{d}{dx}\left(x\sin y - y^2\right) = \sin y + (x\cos y - 2y)\frac{dy}{dx},$$

so $dF/dx = 0$. Integrating the latter gives $F(x, y) = x \sin y - y^2 = C$, just as before.

Thus, let us continue, without concern about manipulating dy/dx as though it were a fraction. Seeking to generalize the method outlined above, we consider the differential equation

$$\frac{dy}{dx} = -\frac{M(x, y)}{N(x, y)}, \tag{5}$$

where the minus sign is included so that when we re-express (5) in the differential form

$$M(x, y)dx + N(x, y)dy = 0, \tag{6}$$

then both signs on the left will be positive. It is important to be aware that in equation (5) y is regarded as a function of x, as is clear from the presence of the derivative dy/dx. That is, there is a hierarchy whereby x is the independent variable and y is the dependent variable. But upon re-expressing (5) in the form (6) we change our viewpoint and now consider x and y as having the same status; now they are both independent variables.

We observe that integration of (6) is simple if $Mdx + Ndy$ happens to be the differential of some function $F(x, y)$, for if there does exist a function $F(x, y)$ such that

$$dF(x, y) = M(x, y)dx + N(x, y)dy, \tag{7}$$

then (6) is

$$dF(x, y) = 0, \tag{8}$$

which can be integrated to give the general solution

$$F(x, y) = C, \tag{9}$$

where C is an arbitrary constant.

Given $M(x, y)$ and $N(x, y)$, suppose that there does exist an $F(x, y)$ such that $Mdx + Ndy = dF$. Then we say that $Mdx + Ndy$ is an **exact differential**, and that (6) is an **exact differential equation**. That case is of great interest because its general solution is given immediately, in implicit form, by (9).

Two questions arise. How do we determine if such an F exists and, if it does, then how do we find it? The first is addressed by the following theorem.

THEOREM 2.5.1 *Test for Exactness*
Let $M(x, y)$, $N(x, y)$, $\partial M/\partial y$, and $\partial N/\partial x$ be continuous within a rectangle R in the x, y plane. Then $Mdx + Ndy$ is an exact differential, in R, if and only if

$$\boxed{\frac{\partial M}{\partial y} = \frac{\partial N}{\partial x}} \tag{10}$$

everywhere in R.

Partial Proof: Let us suppose that $Mdx + Ndy$ is exact, so that there is an F satisfying (7). Then it must be true, according to the chain rule of the calculus, that

$$M = \frac{\partial F}{\partial x} \tag{11a}$$

and

$$N = \frac{\partial F}{\partial y} \tag{11b}$$

Differentiating (11a) partially with respect to y, and (11b) partially with respect to x, gives

$$M_y = F_{xy}, \tag{12a}$$

and

$$N_x = F_{yx}. \tag{12b}$$

Since M, N, M_y, and N_x have been assumed continuous in R, it follows from (11) and (12) that F_x, F_y, F_{xy}, and F_{yx} are too, so $F_{xy} = F_{yx}$. Then it follows from (12) that $M_y = N_x$, which is equation (10). Because of the "if and only if" wording in the theorem, we also need to prove the reverse: that the truth of (10) implies the existence of F. That part of the proof can be carried out using results established in Section 16.12, and will not be given here. ■

Actually R need not be a rectangle; it merely needs to be "simply connected," that is, a region without holes. Simple connectedness will be defined and used extensively in Chapter 16 on Field Theory.

Assuming that the conditions of the theorem are met, so that we are assured that such an F exists, how do we *find* F? We can find it by integrating (11a) with respect to x, and (11b) with respect to y. Let us illustrate the method by reconsidering the example given above.

EXAMPLE 1. Consider equation (1) once again, or, in differential form,

$$\sin y\, dx + (x \cos y - 2y)dy = 0. \tag{13}$$

First, we identify $M = \sin y$, and $N = x \cos y - 2y$. Clearly, M, N, M_y, and N_x are continuous in the whole plane, so we turn to the exactness condition (10): $M_y = \cos y$, and $N_x = \cos y$, so (10) is satisfied, and it follows from Theorem 2.5.1 that there does exist an $F(x, y)$ such that the left-hand side of (13) is dF. Next, we find F from (11):

$$\frac{\partial F}{\partial x} = \sin y, \tag{14a}$$

$$\frac{\partial F}{\partial y} = x \cos y - 2y. \tag{14b}$$

Integrating (14a) partially, with respect to x, gives

$$F(x, y) = \int \sin y\, \partial x = x \sin y + A(y), \tag{15}$$

where the $\sin y$ integrand was treated as a constant in the integration since it was a "partial integration" on x, holding y fixed [just as y was held fixed in computing $\partial F/\partial x$ in (14a)]. The constant of integration A must therefore be allowed to depend upon y since y was held fixed and was therefore constant. If you are not convinced of this point, observe that taking a partial x-derivative of (15) does indeed recover (14a).

Observe that initially $F(x, y)$ was unknown. The integration of (14a) reduced the problem from an unknown function F of x and y to an unknown function A of y alone. $A(y)$, in turn, can now be determined from (14b). Specifically, we put the right-hand side of (15) into the left-hand side of (14b) and obtain

$$x \cos y + A'(y) = x \cos y - 2y, \tag{16}$$

where the prime denotes d/dy. Cancelling terms gives $A'(y) = -2y$, so

$$A(y) = -\int 2y\, dy = -y^2 + B, \tag{17}$$

where this integration was not a "partial integration," it was an ordinary integration on y since $A'(y)$ was an ordinary derivative of A. Combining (17) and (15) gives

$$F(x, y) = x \sin y - y^2 + B = \text{constant}. \tag{18}$$

Finally, absorbing B into the constant, and calling the result C, gives the general solution

$$x \sin y - y^2 = C \tag{19}$$

of (1), in implicit form.

COMMENT 1. Be aware that the partial integration notation $\int(\ )\partial x$ and $\int(\ )\partial y$ is not standard; we use it here because we find it reasonable, and helpful in reminding us that any y's in the integrand of $\int(\ )\partial x$ are to be treated as constants, and likewise any for any x's in $\int(\ )\partial y$.

COMMENT 2. From (13) all the way through (19), x and y have been regarded as independent variables. With (19) in hand, we can now return to our original viewpoint of y being a function of x. We can, if possible, solve (19) by algebra for $y(x)$ [in this case it is not because (19) is transcendental], plot the result, and so on. Even failing to solve (19) for $y(x)$, we can nevertheless verify that $x \sin y - y^2 = C$ satisfies (1) by differentiating with respect to x. That step gives $\sin y + x(\cos y)y' - 2yy' = 0$ or $y' = (\sin y)/(2y - x \cos y)$, which does agree with (1).

COMMENT 3. It would be natural to wonder how this method can *fail* to work. That is, whether or not $M_y = N_x$, why can't we always successfully integrate (11) to find F? The answer is to be found in (16). For suppose (16) were $2x \cos y + A'(y) = x \cos y - 2y$ instead. Then the $x \cos y$ terms would not cancel, as they did in (16), and we would have $A'(y) = -x \cos y - 2y$, which is impossible because it expresses a relationship between x and y, whereas x and y are regarded here as independent variables. Thus, the cancellation of the $x \cos y$ terms in (16) was crucial and was not an accident; it was a consequence of the fact that M and N satisfied the exactness condition (10).

COMMENT 4. Though we used (14a) first, then (14b), the order is immaterial and could

have been reversed. ∎

2.5.2. Integrating factors. It may be discouraging to realize that for any given pair of functions M and N it is unlikely that the exactness condition (10) will be satisfied. However, there is power available to us that we have not yet tapped, for even if M and N fail to satisfy (10), so that the equation

$$M(x,y)dx + N(x,y)dy = 0 \tag{20}$$

is not exact, it *may* be possible to find a multiplicative factor $\sigma(x,y)$ so that

$$\sigma(x,y)M(x,y)dx + \sigma(x,y)N(x,y)dy = 0 \tag{21}$$

is exact. That is, we seek a function $\sigma(x,y)$ so that the revised exactness condition

$$\boxed{\frac{\partial}{\partial y}(\sigma M) = \frac{\partial}{\partial x}(\sigma N)} \tag{22}$$

is satisfied. Of course, we need $\sigma(x,y) \neq 0$ for (21) to be equivalent to (20).

If we can find a $\sigma(x,y)$ satisfying (22), then we call it an **integrating factor** of (20) because then (21) is equivalent to $dF = 0$, for some $F(x,y)$, and $dF = 0$ can be integrated immediately to give the solution of the original differential equation as $F(x,y) =$ constant.

How do we find such a σ? It is any (nonzero) solution of (22), that is, of

$$\sigma_y M + \sigma M_y = \sigma_x N + \sigma N_x. \tag{23}$$

Of course, (23) is a first-order partial differential equation on σ, so we have made dubious headway: to solve our original first-order ordinary differential equation on $y(x)$, we now need to solve the first-order partial differential equation (23) on $\sigma(x,y)$!

However, perhaps an integrating factor σ can be found that is a function of x alone: $\sigma(x)$. Then (23) reduces to the differential equation

$$\sigma M_y = \frac{d\sigma}{dx}N + \sigma N_x$$

or

$$\frac{d\sigma}{dx} = \sigma\left(\frac{M_y - N_x}{N}\right), \tag{24}$$

which is separable. This idea succeeds if and only if the $(M_y - N_x)/N$ ratio on the right-hand side of (24) is a function of x only, for if it did contain any y dependence then (24) would amount to the impossible situation of a function of x equalling a function of x and y, where x and y are independent variables. Thus, if

$$\boxed{\frac{M_y - N_x}{N} = \text{function of } x \text{ alone,}} \tag{25}$$

then integration of (24) gives

$$\boxed{\sigma(x) = e^{\int \frac{M_y - N_x}{N}\,dx}.} \tag{26}$$

Actually, the general solution of (24) includes an arbitrary constant factor, but that factor is inconsequential and can be taken to be 1. Also, remember that we need σ to be nonzero and we are pleased to see, a posteriori, that the σ given in (26) cannot equal zero because it is an exponential function.

If $(M_y - N_x)/N$ is not a function of x alone, then an integrating factor $\sigma(x)$ does not exist, but we can try to find σ as a function of y alone: $\sigma(y)$. Then (23) reduces to

$$\frac{d\sigma}{dy}M + \sigma M_y = \sigma N_x$$

or

$$\frac{d\sigma}{dy} = -\sigma\left(\frac{M_y - N_x}{M}\right),$$

which, again, is separable. If

$$\boxed{\frac{M_y - N_x}{M} = \text{function of } y \text{ alone},} \tag{27}$$

then

$$\boxed{\sigma(y) = e^{-\int \frac{M_y - N_x}{M}\,dy}.} \tag{28}$$

EXAMPLE 2. Consider the equation (already expressed in differential form)

$$dx + \left(3x - e^{-2y}\right)dy = 0. \tag{29}$$

Then $M = 1$ and $N = 3x - e^{-2y}$, so (10) is not satisfied and (29) is not exact. Seeking an integrating factor that is a function of x alone, we find that

$$\frac{M_y - N_x}{N} = \frac{0-3}{3x - e^{-2y}} \neq \text{function of } x \text{ alone}, \tag{30}$$

and conclude that $\sigma(x)$ is not possible. Seeking instead an integrating factor that is a function of y alone,

$$\frac{M_y - N_x}{M} = \frac{0-3}{1} = -3 = \text{function of } y \text{ alone}, \tag{31}$$

so that $\sigma(y)$ is possible, and is given by

$$\sigma(y) = e^{-\int \frac{M_y - N_x}{M}\,dy} = e^{\int 3\,dy} = e^{3y}. \tag{32}$$

Multiply (29) through by the integrating factor $\sigma = e^{3y}$ and obtain

$$e^{3y}dx + e^{3y}\left(3x - e^{-2y}\right)dy = 0,$$

which is now exact. Thus,

$$\frac{\partial F}{\partial x} = e^{3y}$$

and

$$\frac{\partial F}{\partial y} = e^{3y}\left(3x - e^{-2y}\right)$$

so

$$F(x,y) = \int e^{3y}\,\partial x = xe^{3y} + A(y),$$

and

$$\frac{\partial F}{\partial y} = e^{3y}\left(3x - e^{-2y}\right) = 3xe^{3y} + A'(y). \tag{33}$$

The latter gives

$$A'(y) = -e^{y}$$

so

$$A(y) = -e^{y} + B.$$

Thus,

$$F(x,y) = xe^{3y} + A(y) = xe^{3y} + (-e^{y} + B) = \text{constant}$$

or

$$xe^{3y} - e^{y} = C, \tag{34}$$

where C is an arbitrary constant; (34) is the general solution of (29), in implicit form.

COMMENT. Can we solve (34) for y? If we let $e^{y} \equiv z$, then (34) is the cubic equation $xz^3 - z = C$ in z, and there is a known solution to cubic equations. If we can solve for z, then we have y as $y = \ln z$. However, the solution of that cubic equation (as can be obtained using the *Maple* solve command) is quite a messy expression. ∎

EXAMPLE 3. *First-Order Linear Equation.* We've already solved the general first-order linear equation

$$\frac{dy}{dx} + p(x)y = q(x) \tag{35}$$

in Section 2.2, but let us see if we can solve it again, using the ideas of this section. First, express (35) in the form

$$[p(x)y - q(x)]\,dx + dy = 0. \tag{36}$$

Thus, $M = p(x)y - q(x)$ and $N = 1$, so $M_y = p(x)$ and $N_x = 0$. Hence $M_y \neq N_x$, so (36) is not exact [except in the trivial case when $p(x) = 0$]. Since

$$\frac{M_y - N_x}{N} = \frac{p(x) - 0}{1} = \text{function of } x \text{ alone},$$

$$\frac{M_y - N_x}{M} = \frac{p(x) - 0}{p(x)y - q(x)} \neq \text{function of } y \text{ alone},$$

we can find an integrating factor that is a function of x alone, but not one that is a function of y alone. We leave it for the exercises to show that the integrating factor is

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

and that the final solution (this time obtainable in explicit form) is

$$y(x) = e^{-\int p\,dx}\left(\int e^{\int p\,dx} q\,dx + C\right), \tag{37}$$

as found earlier, in Section 2.2. ∎

Closure. Let us summarize the main results. Given a differential equation $dy/dx = f(x, y)$, the first step in using the method of exact differentials is to re-express it in the differential form $M(x, y)dx + N(x, y)dy = 0$. If M, N, M_y, and N_x are all continuous in the region of interest, check to see if the exactness condition (10) is satisfied. If it is, then the equation is exact, and its general solution is $F(x, y) = C$, where F is found by integrating (11a) and (11b). As a check on your work, a differential of $F(x, y) = C$ should give you back the original equation $Mdx + Ndy = 0$.

If it is not exact, see if $(M_y - N_x)/N$ is a function of x alone. If it is, then an integrating factor $\sigma(x)$ can be found from (26). Multiplying the given equation $Mdx + Ndy = 0$ through by that $\sigma(x)$, the new equation is exact, and you can proceed as outlined above for an exact equation.

If $(M_y - N_x)/N$ is not a function of x alone, check to see if $(M_y - N_x)/M$ is a function of y alone. If it is, then an integrating factor $\sigma(y)$ can be found from (28). Multiplying $Mdx + Ndy = 0$ through by that $\sigma(y)$, the new equation is exact, and you can proceed as outlined above for an exact equation.

If $M_y \neq N_x$, $(M_y - N_x)/N$ is not a function of x alone, and $(M_y - N_x)/M$ is not a function of y alone, then the method is of no help unless an integrating factor σ can be found that is a function of both x and y.

EXERCISES 2.5

NOTE: Solutions should be expressed in explicit form if possible.

1. Show that the equation is exact, and obtain its general solution. Also, find the particular solution corresponding to the given initial condition as well.

(a) $3dx - dy = 0$; $y(0) = 6$
(b) $x^2dx + y^2dy = 0$; $y(9) = -1$
(c) $x\,dx + 2ydy = 0$; $y(1) = 2$
(d) $4\cos 2u\,du - e^{-5v}dv = 0$; $v(0) = -6$
(e) $e^y dx + (xe^y - 1)dy = 0$; $y(-5) = 6$
(f) $(e^y + z)dy - (\sin z - y)dz = 0$; $z(0) = 0$
(g) $(x - 2z)dx - (2x - z)dz = 0$; $z(-3) = 5$
(h) $(\sin y + y\cos x)dx + (\sin x + x\cos y)dy = 0$; $y(2) = 3$
(i) $(\sin xy + xy\cos xy)dx + x^2\cos xy\,dy = 0$; $y(0) = -1$
(j) $(3x^2\sin 2y - 2xy)dx + (2x^3\cos 2y - x^2)dy = 0$; $y(0.5) = 3.1$
(k) $(4x^3y^5\sin 3x + 3x^4y^5\cos 3x)dx + 5x^4y^4\sin 3x\,dy = 0$; $y(0) = 1$
(l) $3x^2y\ln y\,dx + (x^3\ln y + x^3 - 2y)dy = 0$; $y(8) = -3$
(m) $(2ye^{2xy}\sin x + e^{2xy}\cos x + 1)dx + 2xe^{2xy}\sin x\,dy = 0$; $y(2.3) = -1.25$

2.(a)–(m) Find the general solution of the equation given in Exercise 1 using computer software, and also the particular solution corresponding to the given initial condition.

3. Make up three different examples of exact equations.

4. Determine whatever conditions, if any, are needed on the constants $a, b, \ldots, f, A, B, \ldots, F$ for the equation to be exact.

(a) $(ax + by + c)dx + (Ax + By + C)dy = 0$
(b) $(ax^2 + by^2 + cxy + dx + ey + f)dx + (Ax^2 + By^2 + Cxy + Dx + Ey + F)dy = 0$

5. Find a suitable integrating factor $\sigma(x)$ or $\sigma(y)$, and use it to find the general solution of the differential equation.

(a) $3ydx + dy = 0$
(b) $ydx + x\ln x\, dy = 0$
(c) $y\ln y\, dx + (x+y)dy = 0$
(d) $dx + (x - e^{-y})dy = 0$
(e) $dx + xdy = 0$
(f) $(ye^{-x} + 1)dx + (xe^{-x})dy = 0$
(g) $\cos y\, dx - [2(x-y)\sin y + \cos y]dy = 0$
(h) $(1 - x - z)dx + dz = 0$
(i) $(2 + \tan^2 x)(1 + e^{-y})dx - e^{-y}\tan x\, dy = 0$
(j) $(3u^2 \sinh 3v - 2u)du + 3u^3 \cosh 3v\, dv = 0$
(k) $\cos x\, dx + (3\sin x + 3\cos y - \sin y)dy = 0$
(l) $(y\ln y + 2xy^2)dx + (x + x^2 y)dy = 0$
(m) $(3x - 2p)dx - x\, dp = 0$
(n) $y\, dx + (x^2 - x)dy = 0$
(o) $2xy\, dx + (y^2 - x^2)dy = 0$

6. (*First-order linear equation*) Verify that $\sigma(x) = e^{\int p(x)\,dx}$ is an integrating factor for the general linear first-order equation (35), and use it to derive the general solution (37).

7. Show that the given equation is not exact and that an integrating factor depending on x alone or y alone does not exist. If possible, find an integrating factor in the form $\sigma(x,y) = x^a y^b$, where a and b are suitably chosen constants. If such a σ can be found, then use it to obtain the general solution of the differential equation; if not, state that.

(a) $(3xy - 2y^2)dx + (2x^2 - 3xy)dy = 0$
(b) $(3xy + 2y^2)dx + (3x^2 + 4xy)dy = 0$
(c) $(x + y^2)dx + (x - y)dy = 0$
(d) $ydx - (x^2 y - x)dy = 0$

8. Show that the equation is not exact and that an integrating factor depending on x alone or y alone does not exist. Nevertheless, find a suitable integrating factor by inspection, and use it to obtain the general solution.

(a) $e^y dx + e^x dy = 0$ (b) $y^2 dx - e^{3x} dy = 0$

(c) $e^{2y} dx - \tan x\, dy = 0$

9. Obtain the general solution, using the methods of this section.

(a) $\dfrac{dy}{dx} = \dfrac{x-y}{x+y}$ (b) $\dfrac{dr}{d\theta} = -\dfrac{r^2\cos\theta}{2r\sin\theta + 1}$

(c) $\dfrac{dy}{dx} = \dfrac{2xy - e^y}{x(e^y - x)}$ (d) $\dfrac{dy}{dx} = \dfrac{y(2x - \ln y)}{x}$

(e) $\dfrac{dy}{dx} = -\dfrac{\sin y + y\cos x}{\sin x + x\cos y}$

10. What do the integrating factors defined by (26) and (28) turn out to be if the equation is exact to begin with?

11.(a) Show that $(x^3 + y)dx + (y^3 + x)dy = 0$ is exact.
(b) More generally, is $M(x,y)dx + M(y,x)dy$ exact? Explain.

12. If $F(x,y) = C$ is the general solution (in implicit form) of a given first-order equation, then what is the particular solution (in implicit form) satisfying the initial condition $y(a) = b$?

13. If $Mdx + Ndy = 0$ and $Pdx + Qdy = 0$ are exact, is $(M + P)dx + (N + Q)dy = 0$ exact? Explain.

14. Show that for $[p(x) + q(y)]dx + [r(x) + s(y)]dy = 0$ to be exact, it is necessary and sufficient that $q(y)dx + r(x)dy$ be an exact differential.

15. Show that for $p(x)dx + q(x)r(y)dy = 0$ to be exact, it is necessary and sufficient that $q(x)$ be a constant.

Chapter 2 Review

Following is a listing of the types of equations covered in this chapter.

SECTION 2.2

First-order linear: $y' + p(x)y = q(x)$.

This equation can be solved by the integrating factor method or by solving the homogeneous equation and using variation of parameters. Its general solution is

$$y(x) = e^{-\int p(x)\,dx}\left(\int e^{\int p(x)\,dx} q(x)\,dx + C\right).$$

A particular solution satisfying $y(a) = b$ is

$$y(x) = e^{-\int_a^x p(\xi)\,d\xi}\left(\int_a^x e^{\int_a^\xi p(\zeta)\,d\zeta} q(\xi)\,d\xi + b\right).$$

Bernoulli: $y' + p(x)y = q(x)y^n$. $(n \neq 0, 1)$.

This equation can be converted to the first-order *linear* equation $v' + (1-n)p(x)v = (1-n)q(x)$ by the change of variables $v = y^{1-n}$ (Exercise 9).

Riccati: $y' = p(x)y^2 + q(x)y + r(x)$.

This equation can be solved by setting $y = Y(x) + \dfrac{1}{v}$, if a particular solution $Y(x)$ of the Riccati equation can be found (Exercise 11).

d'Alembert–Lagrange: $y' = xf(y') + g(y')$. $[f(y') \neq y']$

By letting $y' = p$ be a new independent variable, one can obtain a *linear* first-order equation on $x(p)$ (Exercise 13).

Clairaut: $y' = xy' + g(y')$.

This equation admits the family of straight-line solutions $y = Cx + g(C)$ and, in general, a singular solution as well (Exercise 14).

SECTION 2.4

Separable: $y' = X(x)Y(y)$.

General solution obtained by integrating

$$\int \frac{dy}{Y(y)} = \int X(x)dx.$$

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

Can be made separable by setting $v = y/x$ (Exercise 9).

Almost Homogeneous: $y' = \dfrac{ax + by + c}{dx + ey + f}$. $(ae - bd \neq 0)$

Can be made homogeneous by setting $x = u + h$, $y = v + k$ (Exercise 11).

SECTION 2.5

Exact: $M(x, y)dx + N(x, y)dy = 0$. $(M_y = N_x)$

General solution $F(x, y) = C$ found by integrating $F_x = M$, $F_y = N$. If $M_y \neq N_x$, can make exact by means of an integrating factor $\sigma(x)$ if $(M_y - N_x)/N$ is a function of x only, or by an integrating factor $\sigma(y)$ if $(M_y - N_x)/M$ is a function of y only.

Chapter 3

Linear Differential Equations of Second Order and Higher

PREREQUISITES: In this chapter on linear differential equations, we encounter systems of linear algebraic equations, and it is presumed that the reader is familiar with the theory of the existence and uniqueness of solutions to such equations, especially as regards the role of the determinant of the coefficient matrix. That material is covered in Chapters 8–10, but the essential results that are needed for the present chapter are summarized briefly in Appendix B. Thus, either Sections 8.1–10.6 or Appendix B is a prerequisite for this chapter. Also presumed is a familiarity with the complex plane and the algebra of complex numbers. That material is covered in Section 21.2 which, likewise, is a prerequisite for Chapter 3.

3.1 Introduction

As we prepare to move from first-order equations to those of higher order, this is a good time to pause for an overview that looks back to Chapter 2 and ahead to Chapters 3–7. If, as you proceed through Chapters 3–7, you lose sight of the forest for the trees, we urge you to come back to this overview.

LINEAR EQUATIONS

First order:

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

General solution found [(2.1) in Section 2.2] in explicit form. Existence and uniqueness of solution of initial-value problem [with $y(a) = b$] guaranteed over a predetermined interval, based upon the continuity of $p(x)$ and $q(x)$. Solution of initial-value problem expressible as a superposition of responses to the two inputs [the initial value b and the forcing function $q(x)$] with each

response being proportional to that input: for example, if we double the input we double the output.

Second order and higher:

$$a_0(x)\frac{d^n y}{dx^n} + a_1(x)\frac{d^{n-1}y}{dx^{n-1}} + \cdots + a_{n-1}(x)\frac{dy}{dx} + a_n(x)y = f(x). \tag{2}$$

Constant coefficients (the a_j's are constants) and homogeneous ($f = 0$):

This is the simplest case. We will see (Section 3.4) that the general solution can be found in terms of exponential functions, and perhaps powers of x times exponential functions.

Constant coefficients and nonhomogeneous:

Additional solution is needed due to the forcing function $f(x)$ and can be found by the method of undetermined coefficients (Section 3.7.2) or the method of variation of parameters (Sections 3.7.3 and 3.7.4). Still simple. An alternative approach, the Laplace transform, is given in Chapter 5.

Nonconstant coefficients:

Essentially, the only simple case is the Cauchy – Euler equation (Section 3.6.1). Other cases are so much more difficult that we give up on finding closed form solutions and use power series methods (Chapter 4). Two particularly important cases are the Legendre (Section 4.4) and Bessel (Section 4.6) equations, which will be needed later in the chapters on partial differential equations.

NONLINEAR EQUATIONS

First order:

$$y' = f(x, y). \tag{3}$$

No solution available for the general case. Need to identify subcategories that are susceptible to special solution techniques. The most important of these subcategories are separable equations (Section 2.4) and exact equations (Section 2.5), and these methods give solutions in implicit form. Several important but more specialized cases are given in the exercises: the Bernoulli, Riccati, d'Alembert – Lagrange, and Clairaut equations in Section 2.2, and "homogeneous" equations in Section 2.4. The idea of the response being a superposition of responses, as it is for the linear equation, is not applicable for nonlinear equations. The subcategories and special cases mentioned above by no means cover all possible equations of the form $y' = f(x, y)$, so that many first-order nonlinear equations simply are not solvable by any known means. A powerful alternative to analytical methods, [i.e., methods

designed to obtain an analytical expression for $y(x)$], is to seek a solution in numerical form, with the help of a computational algorithm and a computer, and these methods are discussed in Chapter 6.

Second order and higher:

Some nonlinear equations of first order can be solved analytically, as we have seen, but for nonlinear equations of higher order analytical solution is generally out of the question, and we rely instead upon a blend of numerical solution (Chapter 6) and qualitative methods, such as the phase plane method described in Chapter 7.

To get started, we limit our attention in the next several sections to the **homogeneous** version of the linear equation (2), namely, where $f(x) = 0$, because that case is simpler and because to solve the nonhomogeneous case we will need to solve the homogeneous version first, anyhow.

To attach physical significance to the distinction between homogeneous and nonhomogeneous equations, it may help to recall from Section 1.3 that the differential equation governing a mechanical oscillator is

$$m\frac{d^2x}{dt^2} + c\frac{dx}{dt} + kx = F(t), \tag{4}$$

where m, c, k are the mass, damping coefficient, and spring stiffness, respectively, and $F(t)$ is the applied force. (In this case, of course, the variables happen to be x and t rather than y and x.) If $F(t) = 0$, then (4) governs the unforced, or "free," vibration of the mass m. Likewise, for any linear differential equation, if all terms containing the unknown and its derivatives are moved to the left-hand side, then whatever is left on the right-hand side is regarded as a "forcing function." From a physical point of view then, when we consider the homogeneous case in the next several sections, we are really limiting our attention to unforced systems.

A brief outline of this chapter follows:

3.2 *Linear Dependence and Linear Independence.* The concept of a general solution to a linear differential equation requires the idea of linear dependence and linear independence, so these ideas are introduced first.

3.3 *Homogeneous Equation: General Solution.* Here we establish the concept of a general solution to the homogeneous equation (2), but do not yet show how to obtain it.

3.4 *Solution of Homogeneous Equation: Constant Coefficients.* It is shown how to find the general solution in the form of a linear combination of solutions that are either exponentials or powers of x times exponentials.

3.5 *Application to Harmonic Oscillator: Free Oscillation.* The foregoing concepts and methods are applied to an extremely important physical application: the free oscillation of a harmonic oscillator.

3.6 *Solution of Homogeneous Equation: Nonconstant Coefficients.* Nonconstant-coefficient equations can be solved in closed form only in exceptional cases. The most important such case is the Cauchy–Euler equation, and that case occupies most of this section.

3.7 *Solution of Nonhomogeneous Equation.* It is shown how to find the additional solution, due to the forcing function, by the methods of undetermined coefficients and variation of parameters.

3.8 *Application to Harmonic Oscillator: Forced Oscillation.* We return to the example of the harmonic oscillator, begun in Section 3.5, and obtain and discuss the solution for the forced oscillation.

3.9 *Systems of Linear Differential Equations.* We consider linear systems of n coupled first-order differential equations on n unkowns and show how to obtain uncoupled nth-order differential equations on each of the n unknowns, which equations can then be solved by the methods described in the preceding sections of this chapter.

3.2 Linear Dependence and Linear Independence

Asked how many different paints he had, a painter replied five: red, blue, green, yellow, and purple. However, it could be argued that the count was inflated since only three (for instance red, blue, and yellow) are independent: the green can be obtained from a certain proportion of the blue and the yellow, and the purple can be obtained from the red and the blue. Similarly, in studying linear differential equations, we will need to determine how many "different," or "independent," functions are contained within a given set of functions. The concept is made precise as follows. We begin by defining a **linear combination** of a set of functions $f_1, \ldots, f_k$ as any function of the form $\alpha_1 f_1 + \cdots + \alpha_k f_k$, where the α_j's are constants. For instance, $2\sin x - \pi \cos x$ is a linear combination of $\sin x$ and $\cos x$.

DEFINITION 3.2.1 *Linear Dependence and Linear Independence*
A set of functions $\{u_1, \ldots, u_n\}$ is said to be **linearly dependent** on an interval I if at least one of them can be expressed as a linear combination of the others on I. If none can be so expressed, then the set is **linearly independent**.

If we do not specify the interval I, then it will be understood to be the entire x axis. NOTE: Since the terms linearly dependent and linearly independent will appear repeatedly, it will be convenient to abbreviate them in this book as **LD** and

LI, respectively, but be aware that this notation is not standard outside of this text.

EXAMPLE 1. The set $\{x^2, e^x, e^{-x}, \sinh x\}$ is seen to be LD (linearly dependent) because we can express $\sinh x$ as a linear combination of the others:

$$\sinh x = \frac{e^x - e^{-x}}{2} = \frac{1}{2}e^x - \frac{1}{2}e^{-x} + 0\,x^2. \tag{1}$$

In fact, we could express e^x as a linear combination of the others too, for solving (1) for e^x gives $e^x = 2\sinh x + e^{-x} + 0\,x^2$. Likewise, we could express $e^{-x} = e^x - 2\sinh x + 0\,x^2$. We cannot express x^2 as a linear combination of the others [since we cannot solve (1) for x^2], but the set is LD nonetheless, because we only need to be able to express "at least one" member as a linear combination of the others. NOTE: The hyperbolic sine and cosine functions, $\sinh x$ and $\cosh x$, were studied in the calculus, but if these functions and their graphs and properties are not familiar to you, you may wish to turn to the review in Section 3.4.1. ∎

The foregoing example was simple enough to be worked by inspection. In more complicated cases, the following theorem provides a test for determining whether a given set is LD or LI.

THEOREM 3.2.1 *Test for Linear Dependence/Independence*
A finite set of functions $\{u_1, \ldots, u_n\}$ is LD on an interval I if and only if there exist scalars α_j, not all zero, such that

$$\alpha_1 u_1(x) + \alpha_2 u_2(x) + \cdots + \alpha_n u_n(x) = 0 \tag{2}$$

identically on I. If (2) is true only if all the α's are zero, then the set is LI on I.

Proof: Because of the "if and only if" we need to prove the statement in both directions. First, suppose that the set is LD. Then, according to the definition of linear dependence, one of the functions, say u_j, can be expressed as a linear combination of the others:

$$u_j(x) = \alpha_1 u_1(x) + \cdots + \alpha_{j-1} u_{j-1}(x) + \alpha_{j+1} u_{j+1}(x) + \cdots + \alpha_n u_n(x), \tag{3}$$

which equation can be rewritten as

$$\alpha_1 u_1(x) + \cdots + \alpha_{j-1} u_{j-1}(x) + (-1) u_j(x) + \alpha_{j+1} u_{j+1}(x) + \cdots + \alpha_n u_n(x) = 0. \tag{4}$$

Even if all the other α's are zero, the coefficient α_j of $u_j(x)$ in (4) is nonzero, namely, -1, so there do exist scalars $\alpha_1, \ldots, \alpha_n$ not all zero such that (2) holds.

Conversely, suppose that (2) holds with the α's not all zero. If α_k, for instance, is nonzero, then (2) can be divided by α_k and solved for $u_k(x)$ as a linear combination of the other u's, in which case $\{u_1, \ldots, u_n\}$ is LD. ■

EXAMPLE 2. To determine if the set $\{1, x, x^2\}$ is LD or LI using Theorem 3.2.1, write equation (2),

$$\alpha_1 + \alpha_2 x + \alpha_3 x^2 = 0, \tag{5}$$

and see if the truth of (5) requires all the α's to be zero. Since (5) needs to hold for all x's in the interval (which we take to be $-\infty < x < \infty$), let us write it for $x = 0, 1, 2$, say, to generate three equations on the three α's:

$$\begin{aligned} \alpha_1 &= 0, \\ \alpha_1 + \alpha_2 + \alpha_3 &= 0, \\ \alpha_1 + 2\alpha_2 + 4\alpha_3 &= 0. \end{aligned} \tag{6}$$

Solution of (6) gives $\alpha_1 = \alpha_2 = \alpha_3 = 0$, so the set is LI.

In fact, (5) really amounts to an *infinite* number of linear algebraic equations on the three α's since there is no limit to the number of x values that could be chosen. However, three different x values sufficed to establish that all of the α's must be zero. ■

Alternative to writing out (2) for n specific x values, to generate n equations on $\alpha_1, \ldots, \alpha_n$, it is more common to generate n such equations by writing (2) and its first $n-1$ derivatives (assuming, of course, that $u_1, \ldots, u_n$ are $n-1$ times differentiable on I),

$$\begin{aligned} \alpha_1 u_1(x) + \cdots + \alpha_n u_n(x) &= 0, \\ \alpha_1 u_1'(x) + \cdots + \alpha_n u_n'(x) &= 0, \\ &\vdots \\ \alpha_1 u_1^{(n-1)}(x) + \cdots + \alpha_n u_n^{(n-1)}(x) &= 0. \end{aligned} \tag{7}$$

Let us denote the determinant of the coefficients as

$$\boxed{W[u_1, \ldots, u_n](x) = \begin{vmatrix} u_1(x) & \cdots & u_n(x) \\ u_1'(x) & \cdots & u_n'(x) \\ \vdots & & \vdots \\ u_1^{(n-1)}(x) & \cdots & u_n^{(n-1)}(x) \end{vmatrix},} \tag{8}$$

which is known as the **Wronskian determinant** of $u_1, \ldots, u_n$, or simply the **Wronskian** of $u_1, \ldots, u_n$, after the Polish mathematician *Josef M. H. Wronski* (1778–1853). The Wronskian W is itself a function of x.

From the theory of linear algebraic equations, we know that if there is any value of x in I, say x_0, such that $W[u_1, \ldots, u_n](x_0) \neq 0$, then it follows from (7) with x set equal to x_0, that all the α's must be zero, so the set $\{u_1, \ldots, u_n\}$ is LI.

THEOREM 3.2.2 *Wronskian Condition for Linear Independence*
If, for a set of functions $\{u_1, \ldots, u_n\}$ having derivatives through order $n-1$ on an interval I, $W[u_1, \ldots, u_n](x)$ is not identically zero on I, then the set is LI on I.

Be careful not to read into Theorem 3.2.2 a converse, namely, that if $W[u_1, \ldots, u_n](x)$ *is* identically zero on I (which we write as $W \equiv 0$), then the set is LD on I. In fact, the latter is not true, as shown by the following example.

EXAMPLE 3. Consider the set $\{u_1, u_2\}$, where

$$u_1(x) = \begin{cases} x^2, & x \le 0 \\ 0, & x \ge 0, \end{cases} \qquad u_2(x) = \begin{cases} 0, & x \le 0 \\ x^2, & x \ge 0. \end{cases} \tag{9}$$

(Sketch their graphs.) Then (2) becomes

$$\begin{aligned} \alpha_1 x^2 + \alpha_2(0) &= 0 \qquad \text{for } x \le 0 \\ \alpha_1(0) + \alpha_2 x^2 &= 0 \qquad \text{for } x \ge 0. \end{aligned}$$

The first implies that $\alpha_1 = 0$, and the second implies that $\alpha_2 = 0$. Hence $\{u_1, u_2\}$ is LI. Yet, $W[u_1, u_2](x) = \begin{vmatrix} x^2 & 0 \\ 2x & 0 \end{vmatrix} = 0$ on $x \le 0$, and $W[u_1, u_2](x) = \begin{vmatrix} 0 & x^2 \\ 0 & 2x \end{vmatrix} = 0$ on $x \ge 0$, so $W[u_1, u_2](x) \equiv 0$ for all x. ∎

However, our interest in linear dependence and independence, in this chapter, is not going to be in connection with sets of randomly chosen functions, but with sets of functions which have in common that they are solutions of a given linear homogeneous differential equation. In that case, it can be shown that the inverse of Theorem 3.2.2 *is* true: that is, if $W \equiv 0$, then the set is LD. Thus, for that case we have the following stronger theorem which, for our subsequent purposes, will be more important to us than Theorem 3.2.2.

THEOREM 3.2.3 *A Necessary and Sufficient Condition for Linear Dependence*
If $u_1, \ldots, u_n$ are solutions of an nth-order linear homogeneous differential equation

$$\frac{d^n y}{dx^n} + p_1(x)\frac{d^{n-1}y}{dx^{n-1}} + \cdots + p_{n-1}(x)\frac{dy}{dx} + p_n(x)y = 0, \tag{10}$$

where the coefficients $p_j(x)$ are continuous on an interval I, then $W[u_1, \ldots, u_n](x) \equiv 0$ on I is both necessary and sufficient for the linear dependence of the set $\{u_1, \ldots, u_n\}$ on I.

EXAMPLE 4. It is readily verified that each of the functions $1, e^x, e^{-x}$ satisfies the equation $y''' - y' = 0$. Since their Wronskian is

$$W\left[1, e^x, e^{-x}\right](x) = \begin{vmatrix} 1 & e^x & e^{-x} \\ 0 & e^x & -e^{-x} \\ 0 & e^x & e^{-x} \end{vmatrix} = 2 \neq 0,$$

it follows from Theorem 3.2.3 that the set $\{1, e^x, e^{-x}\}$ is LI. Another set of solutions of $y''' - y' = 0$ is $e^x, e^{-x}, \cosh x$. Their Wronskian is

$$W\left[e^x, e^{-x}, \cosh x\right](x) = \begin{vmatrix} e^x & e^{-x} & \cosh x \\ e^x & -e^{-x} & \sinh x \\ e^x & e^{-x} & \cosh x \end{vmatrix} = 0,$$

so the set $\{e^x, e^{-x}, \cosh x\}$ is LD. ∎

In connection with Theorem 3.2.3, it would be natural to wonder if W could be zero for some x's and nonzero for others. Subject to the conditions of that theorem, it can be shown (Exercise 5) that

$$W(x) = W(\xi) \exp\left[-\int_{\xi}^{x} p_1(t)\, dt\right], \tag{11}$$

where ξ is any point in the interval and p_1 is the coefficient of the next-to-highest derivative in (10), and where we have written $W[u_1, \ldots, u_n](x)$ as $W(x)$, and $W[u_1, \ldots, u_n](\xi)$ as $W(\xi)$, for brevity. Due to the French mathematican *Joseph Liouville* (1809–1882), and known as **Liouville's formula**, (11) shows that under the conditions of Theorem 3.2.3 the Wronskian is either everywhere zero or everywhere nonzero, for the exponential function is positive for all finite values of its argument and the constant $W(\xi)$ is either zero or not. This fact is illustrated by the two Wronskians in Example 4.

Finally, it is useful to cite the following three simple results, proofs of which are left for the exercises.

THEOREM 3.2.4 *Linear Dependence/Independence of Two Functions*
A set of two functions, $\{u_1, u_2\}$, is LD if and only if one is expressible as a scalar multiple of the other.

THEOREM 3.2.5 *Linear Dependence of Sets Containing the Zero Function*
If a set $\{u_1, \ldots, u_n\}$ contains the zero function [that is, $u_j(x) = 0$ for some j], then the set is LD.

THEOREM 3.2.6 *Equating Coefficients*
Let $\{u_1, \ldots, u_n\}$ be LI on an interval I. Then, for

$$a_1u_1(x) + \cdots + a_nu_n(x) = b_1u_1(x) + \cdots + b_nu_n(x)$$

to hold on I, it is necessary and sufficient that $a_j = b_j$ for each $j = 1, \ldots, n$. That is, the coefficients of corresponding terms on the left- and right-hand sides must match.

EXAMPLE 5. The set $\{x, \sin x\}$ is LI on $-\infty < x < \infty$ according to Theorem 3.2.4 because x is surely not expressible as a constant times $\sin x$ (for $x/\sin x$ is not a constant), nor is $\sin x$ expressible as a constant times x. ∎

EXAMPLE 6. We've seen that $\{1, e^x, e^{-x}\}$ is LI on $-\infty < x < \infty$. Thus, if we meet the equation

$$a + be^x + ce^{-x} = 6 - 2e^{-x}, \tag{12}$$

then it follows from Theorem 3.2.6 that we must have $a = 6$, $b = 0$, $c = -2$, for if we rewrite (12) as

$$(a - b)(1) + be^x + (c + 2)e^{-x} = 0,$$

then it follows from the linear independence of $1, e^x, e^{-x}$ that $a - 6 = 0$, $b = 0$, $c + 2 = 0$; that is, $a = 6$, $b = 0$, $c = -2$. ∎

Closure. We have introduced the concept of linear dependence and linear independence as preliminary to our development of the theory of linear differential equations, which follows next. Following the definitions of these terms, we gave three theorems for the testing of a given set of functions to determine if they are LI or LD. Of these, Theorem 3.2.3 will be most useful to us in the sections to follow because it applies to sets of functions that arise as solutions of a given differential equation.

In case you have trouble remembering which of the conditions $W = 0$ and $W \neq 0$ corresponds to linear dependence and which to linear independence, think of it this way. If we randomly make up a determinant, the chances are that its value is nonzero; that is the generic case. Likewise, if we randomly select a set of functions out of the set of all possible functions, the generic case is for them to be unrelated – namely, LI. The generic cases go together ($W \neq 0$ corresponding to linear independence) and the nongeneric cases go together ($W = 0$ corresponding to linear dependence).

The concept of linear dependence and independence will prove to be important to us later as well, when we study n-dimensional vector spaces and the expansion of a given vector in terms of a set of base vectors.

EXERCISES 3.2

1. (a) Can a set be neither LD nor LI? Explain.
(b) Can a set be both LD and LI? Explain.

2. Show that the following sets are LD by expressing one of the functions as a linear combination of the others.

(a) $\{1,\ x+2,\ 3x-5\}$
(b) $\{x^2,\ x^2+x,\ x^2+x+1,\ x-1\}$
(c) $\{x^4+x^2+1,\ x^4-x^2+1,\ x^4-x^2-1\}$
(d) $\{e^x,\ e^{2x},\ \sinh x,\ \cosh x\}$
(e) $\{\sinh 3x,\ e^x,\ e^{3x},\ e^{5x},\ e^{-3x}\}$
(f) $\{e^x,\ e^{2x},\ xe^x,\ (7x-2)e^x\}$
(g) $\{0, x, x^3\}$
(h) $\{x, 2x, x^2\}$

3. Show whether the given set is LD or LI. HINT: In most cases, the brief discussion of determinants given in Appendix B will suffice; in others, you will need to use known properties of determinants given in Section 10.4. Also, note that the *Maple* command for calculating determinants (the elements of which need not be constants) is given at the end of Section 10.4.

(a) $\{1,\ x,\ x^2, \ldots, x^n\}$
(b) $\{e^{a_1x},\ e^{a_2x}, \ldots, e^{a_nx}\}$
(c) $\{1,\ 1+x,\ 1+x^2\}$
(d) $\{e^x,\ e^{2x}\}$
(e) $\{\sin x,\ \cos x,\ \sinh x\}$
(f) $\{x,\ x^2\}$
(g) $\{1,\ \sin 3x\}$
(h) $\{1-x,\ 1+x,\ x^3\}$
(i) $\{x,\ 1+x,\ e^x\}$
(j) $\{x,\ e^x,\ \cos x\}$

4. Verify that each of the given functions is a solution of the given differential equation, and then use Theorem 3.2.3 to determine if the set is LD or LI. As a check, use Theorem 3.2.4 if that theorem applies.

(a) $y''' - 6y'' + 11y' - 6y = 0, \quad \{e^x, e^{2x}, e^{3x}\}$
(b) $y'' + 4y = 0, \quad \{\sin 2x, \cos 2x\}$
(c) $y''' - 6y'' + 9y' - 4y = 0, \quad \{e^x, xe^x, e^{4x}\}$
(d) $y''' - 6y'' + 9y' - 4y = 0, \quad \{e^x, xe^x, (1-x)e^x\}$
(e) $y''' - y'' - 2y' = 0, \quad \{1, e^{-x}, e^{2x}\}$
(f) $y'''' - 5y'' + 4y = 0, \quad \{e^x, e^{-x}, e^{2x}, e^{-2x}\}$
(g) $x^2y'' - 3xy' + 3y = 0, \quad \{x, x^3\}, \quad$ on $x > 0$
(h) $x^2y'' - 3xy' + 4y = 0, \quad \{x^2, x^2 \ln x\}, \quad$ on $x > 0$
(j) $y'' - 4y' + 4y = 0, \quad \{e^{2x}, xe^{2x}\}$
(j) $x^3y''' + xy' - y = 0, \quad \left\{x, x\ln x, x(\ln x)^2\right\}, \quad$ on $x > 0$

5. (*Liouville's formula*) (a) Derive Liouville's formula, (11), for the special case where $n = 2$, by writing out $W'(x)$, showing that

$$W'(x) = -p_1(x)W(x), \tag{5.1}$$

and integrating the latter to obtain (11).
(b) Derive (11) for the general case (i.e., where n need not equal 2), by showing that $W'(x)$ is the sum of n determinants where the jth one is obtained from the W determinant by differentiating the jth row and leaving the other rows unchanged. Show that each of these n determinants, except the nth one, has two identical rows and hence vanishes, so that

$$W'(x) = \begin{vmatrix} u_1(x) & \cdots & u_n(x) \\ \vdots & & \vdots \\ u_1^{(n-2)}(x) & \cdots & u_n^{(n-2)}(x) \\ u_1^{(n)}(x) & \cdots & u_n^{(n)}(x) \end{vmatrix}. \tag{5.2}$$

In the last row, substitute $u^{(n)}(x) = -p_1(x)u^{(n-1)}(x) - \cdots - p_n(x)u(x)$ from (10), again omit vanishing determinants, and again obtain (5.1) and hence the solution (11). HINT: You may use the various properties of determinants, given in Section 10.4.

6. (a) Prove Theorem 3.2.4.
(b) Prove Theorem 3.2.5.
(c) Prove Theorem 3.2.6.

7. If u_1 and u_2 are LI, u_1 and u_3 are LI, and u_2 and u_3 are LI, does it follow that $\{u_1, u_2, u_3\}$ is LI? Prove or disprove. HINT: If a proposition is false it can be disproved by a single counterexample, but if it is true then a single example does not suffice as proof.

8. Verify that x^2 and x^3 are solutions of $x^2y'' - 4xy' + 6y = 0$ on $-\infty < x < \infty$. Also verify, from Theorem 3.2.4, that they are LI on that interval. Does the fact that their Wronskian $W[x^2, x^3](x) = x^4$ vanishes at $x = 0$, together with their linear independence on $-\infty < x < \infty$ violate Theorem 3.2.3? Explain.

3.3 Homogeneous Equation: General Solution

3.3.1. General solution. We studied the first-order linear homogeneous equation

$$y' + p(x)y = 0 \tag{1}$$

in Chapter 2, where $p(x)$ is continuous on the x interval of interest, I, and found the solution to be

$$y(x) = Ce^{-\int p(x)\,dx}, \tag{2}$$

where C is an arbitrary constant. If we append to (1) an initial condition $y(a) = b$, where a is a point in I, then we obtain, from (2),

$$y(x) = be^{-\int_a^x p(\xi)\,d\xi}, \tag{3}$$

as was shown in Section 2.2.

The solution (2) is really a family of solutions because of the arbitrary constant C. We showed that (2) contains *all* solutions of (1), so we called it a general solution of (1). In contrast, (3) was only one member of that family, so we called it a particular solution.

Now we turn to the nth-order linear equation

$$\frac{d^n y}{dx^n} + p_1(x)\frac{d^{n-1}y}{dx^{n-1}} + \cdots + p_{n-1}(x)\frac{dy}{dx} + p_n(x)y = 0, \tag{4}$$

and once again we are interested in general and particular solutions. By a **general solution** of (4), on an interval I, we mean a family of solutions that contains every solution of (4) on that interval, and by a **particular solution** of (4), we mean any one member of that family of solutions.

We begin with a fundamental existence and uniqueness theorem.*

THEOREM 3.3.1 *Existence and Uniqueness for Initial-Value Problem*
If $p_1(x), \ldots, p_n(x)$ are continuous on a closed interval I, then the initial-value problem consisting of the differential equation

$$\frac{d^n y}{dx^n} + p_1(x)\frac{d^{n-1}y}{dx^{n-1}} + \cdots + p_{n-1}(x)\frac{dy}{dx} + p_n(x)y = 0, \tag{5a}$$

together with initial conditions

$$y(a) = b_1,\ y'(a) = b_2,\ \ldots,\ y^{(n-1)}(a) = b_n, \tag{5b}$$

*For a more complete sequence of theorems, and their proofs, we refer the interested reader to the little book by J. C. Burkill, *The Theory of Ordinary Differential Equations* (Edinburgh: Oliver and Boyd, 1956) or to William E. Boyce and Richard C. DiPrima, *Elementary Differential Equations and Boundary Value Problems*, 5th ed. (New York: Wiley, 1992).

where the initial point a is in I, has a solution on I, and that solution is unique.

Notice how the initial conditions listed in (5b) are perfect – not too few and not too many – in narrowing the general solution of (5a) down to a unique particular solution, for (5a) gives $y^{(n)}(x)$ as a linear combination of $y^{(n-1)}(x), \ldots, y(x)$, the derivative of (5a) gives $y^{(n+1)}(x)$ as a linear combination of $y^{(n)}(x), \ldots, y(x)$, and so on. Thus, knowing $y(a), \ldots, y^{(n-1)}(a)$ we can use the differential equation (5a) and its derivatives to compute $y^{(n)}(a)$, $y^{(n+1)}(a)$, and so on, and therefore to develop the Taylor series of $y(x)$ about the point a; that is, to determine $y(x)$.

Let us leave the initial-value problem (5) now, and turn our attention to determining the nature of the general solution of the nth-order linear homogeneous equation (4). We begin by re-expressing (4) in the compact form

$$L[y] = 0, \tag{6}$$

where

$$L = \frac{d^n}{dx^n} + p_1(x)\frac{d^{n-1}}{dx^{n-1}} + \cdots + p_{n-1}(x)\frac{d}{dx} + p_n(x) \tag{7}$$

is called an nth-order **differential operator** and

$$\begin{aligned} L[y] &= \left(\frac{d^n}{dx^n} + p_1(x)\frac{d^{n-1}}{dx^{n-1}} + \cdots + p_{n-1}(x)\frac{d}{dx} + p_n(x)\right)[y] \\ &\equiv \frac{d^n}{dx^n}y(x) + p_1(x)\frac{d^{n-1}}{dx^{n-1}}y(x) + \cdots + p_n(x)y(x) \end{aligned} \tag{8}$$

defines the action of L on any n-times differentiable function y. $L[y]$ is itself a function of x, with values $L[y](x)$. For instance, if $n = 2$, $p_1(x) = \sin x$, $p_2(x) = 5x$, and $y(x) = x^2$, then $L[y](x) = (x^2)'' + (\sin x)(x^2)' + 5x(x^2) = 2 + 2x\sin x + 5x^3$.

The key property of the operator L defined by (8) is that

$$\boxed{L[\alpha u + \beta v] = \alpha L[u] + \beta L[v]} \tag{9}$$

for any (n-times differentiable) functions u, v and any constants α, β. Let us verify (9) for the representative case where L is of second order:

$$\begin{aligned} L[\alpha u + \beta v] &= \left(\frac{d^2}{dx^2} + p_1\frac{d}{dx} + p_2\right)(\alpha u + \beta v) \\ &= (\alpha u + \beta v)'' + p_1(\alpha u + \beta v)' + p_2(\alpha u + \beta v) \\ &= \alpha u'' + \beta v'' + p_1\alpha u' + p_1\beta v' + p_2\alpha u + p_2\beta v \\ &= \alpha\left(u'' + p_1u' + p_2u\right) + \beta\left(v'' + p_1v' + p_2v\right) \\ &= \alpha L[u] + \beta L[v]. \end{aligned} \tag{10}$$

Similarly for $n \geq 3$.

Recall that the differential equation (4) was classified as linear. Likewise, the corresponding operator L given by (8) is said to be a **linear differential operator**. The key and defining feature of a linear differential operator is the linearity property (9), which will be of great importance to us.

In fact, (9) holds not just for two functions u and v, but for any finite number of functions, say $u_1, \ldots, u_k$. That is,

$$\boxed{L[\alpha_1 u_1 + \cdots + \alpha_k u_k] = \alpha_1 L[u_1] + \cdots + \alpha_k L[u_k]} \tag{11}$$

for any functions $u_1, \ldots, u_k$ and any constants $\alpha_1, \ldots, \alpha_k$. (Of course it should be understood, whether we say so explicitly or not, that $u_1, \ldots, u_k$ must be n times differentiable since they are being operated on by the nth-order differential operator L.) To prove (11) we apply (9) step by step. For instance, if $k = 3$ we have

$$\begin{aligned} L[\alpha_1 u_1 + \alpha_2 u_2 + \alpha_3 u_3] &= L[\alpha_1 u_1 + 1(\alpha_2 u_2 + \alpha_3 u_3)] \\ &= \alpha_1 L[u_1] + 1L[\alpha_2 u_2 + \alpha_3 u_3] && \text{from (9)} \\ &= \alpha_1 L[u_1] + \alpha_2 L[u_2] + \alpha_3 L[u_3] && \text{from (9) again.} \end{aligned}$$

From (11) we have the following superposition theorem:

THEOREM 3.3.2 *Superposition of Solutions of (4)*
If $y_1 \ldots, y_k$, are solutions of (4), then $C_1 y_1 + \cdots + C_k y_k$ is too, for any constants $C_1, \ldots, C_k$.

Proof: It follows from (11) that

$$\begin{aligned} L[C_1 y_1 + \cdots + C_k y_k] &= C_1 L[y_1] + \cdots + C_k L[y_k] \\ &= C_1(0) + \cdots + C_k(0). \quad \blacksquare \end{aligned}$$

EXAMPLE 1. *Superposition.* It is readily verified, by direct substitution, that $y_1 = e^{3x}$ and $y_2 = e^{-3x}$ are solutions of the equation

$$y'' - 9y = 0. \tag{12}$$

(We are not yet concerned with how to find such solutions; we will come to that in the next section.) Thus $y = C_1 e^{3x} + C_2 e^{-3x}$ is also a solution, as can be verified by substituting it into (12), for any constants C_1, C_2. ∎

To emphasize that the theorem does not hold for nonlinear or nonhomogeneous equations, we offer two counter-examples:

EXAMPLE 2. It can be verified that $y_1 = 1$ and $y_2 = x^2$ are solutions of the nonlinear

equation $x^3y'' - yy' = 0$, yet their linear combination $4 + 3x^2$ is not. ∎

EXAMPLE 3. It can be verified that $y_1 = 4e^{3x} - 2$ and $y_2 = e^{3x} - 2$ are solutions of the nonhomogeneous equation $y'' - 9y = 18$, yet their sum $5e^{3x} - 4$ is not. ∎

We can now prove the following fundamental result.

THEOREM 3.3.3 *General Solution of (4)*
Let $p_1(x), \ldots, p_n(x)$ be continuous on an open interval I. Then the nth-order linear homogeneous differential equation (4) admits exactly n LI solutions; that is, at least n and no more than n. If $y_1(x), \ldots, y_n(x)$ is such a set of LI solutions on I, then the general solution of (4), on I, is

$$y(x) = C_1y_1(x) + \cdots + C_ny_n(x), \tag{13}$$

where $C_1, \ldots, C_n$ are arbitrary constants.

Proof: To show that (4) has no more than n LI solutions, suppose that $y_1(x), \ldots, y_m(x)$ are solutions of (4), where $m > n$. Let ξ be some point in I. The n linear algebraic equations

$$\begin{aligned} c_1y_1(\xi) + \cdots + c_my_m(\xi) &= 0 \\ &\vdots \\ c_1y_1^{(n-1)}(\xi) + \cdots + c_my_m^{(n-1)}(\xi) &= 0 \end{aligned} \tag{14}$$

in the m unknown c's have nontrivial solutions because $m > n$. Choosing such a nontrivial set of c's, define

$$v(x) \equiv c_1y_1(x) + \cdots + c_my_m(x), \tag{15}$$

and observe first that

$$\begin{aligned} L[v] &= L[c_1y_1 + \cdots + c_my_m] \\ &= c_1L[y_1] + \cdots + c_mL[y_m] = c_1(0) + \cdots + c_m(0) = 0, \end{aligned} \tag{16}$$

where L is the differential operator in (4) and, second, that $v(\xi) = v'(\xi) = \cdots = v^{(n-1)}(\xi) = 0$. One function $v(x)$ that satisfies $L[v] = 0$ and $v(\xi) = \cdots = v^{(n-1)}(\xi) = 0$ is simply $v(x) = 0$. By the uniqueness part of Theorem 3.3.1 it is the only such function, so $v(x) = 0$. Recalling that the c's in $v(x) = c_1y_1(x) + \cdots + c_my_m(x) = 0$ are not all zero, it follows that $y_1(x), \ldots, y_m(x)$ must be LD. Thus, (4) cannot have more than n LI solutions.

To show that there are indeed n LI solutions of (4), let us put forward n such solutions. According to the existence part of Theorem 3.3.1, there must be solutions $y_1(x), \ldots, y_n(x)$ of (4) satisfying the initial conditions

$$\begin{array}{llll} y_1(a) = \alpha_{11}, & y_1'(a) = \alpha_{12}, & \cdots & y_1^{(n-1)}(a) = \alpha_{1n}, \\ \vdots & \vdots & & \vdots \\ y_n(a) = \alpha_{n1}, & y_n'(a) = \alpha_{n2}, & \cdots & y_n^{(n-1)}(a) = \alpha_{nn}, \end{array} \tag{17}$$

where a is any chosen point in I and the α's are any chosen numbers such that their determinant is nonzero. (For instance, one such set of α's is given by $\alpha_{ii} = 1$ for each $i = 1$ through n and $\alpha_{ij} = 0$ whenever $i \neq j$.) According to Theorem 3.2.3, $y_1(x), \ldots, y_n(x)$ must be LI since their Wronskian is nonzero at $x = a$. Thus, there are indeed n LI solutions of (4).

Finally, every solution of (4) must be expressible as a linear combination of those n LI solutions, as in (13), for otherwise there would be more than n LI solutions of (4). ■

Any such set of n LI solutions is called a **basis**, or **fundamental set**, of solutions of the differential equation.

EXAMPLE 4. Suppose we begin writing solutions of the equation $y'' - 9y = 0$, from Example 1: $e^{3x}, 5e^{3x}, e^{-3x}, 2e^{3x} + e^{-3x}, \sinh 3x, \cosh 3x, e^{3x} - 4\cosh 3x$, and so on. (That each is a solution is easily verified.) From among these we can indeed find two that are LI, but no more than two. For instance, $\{e^{3x}, e^{-3x}\}$, $\{e^{3x}, 2e^{3x} + e^{-3x}\}$, $\{e^{3x}, \sinh 3x\}$, $\{\sinh 3x, \cosh 3x\}$, $\{\sinh 3x, e^{-3x}\}$ are bases, so the general solution can be expressed in any of these ways:

$$y(x) = C_1 e^{3x} + C_2 e^{-3x}, \tag{18a}$$
$$y(x) = C_1 e^{3x} + C_2 \left(2e^{3x} + e^{-3x}\right), \tag{18b}$$
$$y(x) = C_1 e^{3x} + C_2 \sinh 3x, \tag{18c}$$
$$y(x) = C_1 \sinh 3x + C_2 \cosh 3x, \tag{18d}$$

and so on. Each of these is a general solution of $y'' - 9y = 0$, and all of them are equivalent. For instance, (18a) implies (18d) because

$$\begin{aligned} y(x) &= C_1 e^{3x} + C_2 e^{-3x} \\ &= C_1 (\cosh 3x + \sinh 3x) + C_2 (\cosh 3x - \sinh 3x) \\ &= (C_1 + C_2) \cosh 3x + (C_1 - C_2) \sinh 3x \\ &= C_1' \cosh 3x + C_2' \sinh 3x, \end{aligned}$$

where C_1', C_2' are arbitrary constants (Exercise 15). ■

EXAMPLE 5. Solve the initial-value problem

$$y''' + y' = 0 \tag{19a}$$

$$y(0) = 3, \quad y'(0) = 5, \quad y''(0) = -4. \tag{19b}$$

A general solution of (19a) is

$$y(x) = C_1 \cos x + C_2 \sin x + C_3, \tag{20}$$

because $\cos x$, $\sin x$, and 1 are LI solutions of (19a). Imposing (19b) gives

$$\begin{aligned} y(0) &= 3 = C_1 + C_3, \\ y'(0) &= 5 = C_2, \\ y''(0) &= -4 = -C_1, \end{aligned}$$

which equations admit the unique solution $C_1 = 4, C_2 = 5, C_3 = -1$. Thus,

$$y(x) = 4 \cos x + 5 \sin x - 1$$

is the unique solution to the initial-value problem (19). ∎

3.3.2. Boundary-value problems. It must be remembered that the existence and uniqueness theorem, Theorem 3.3.1, is for *initial*-value problems. Though most of our interest is in initial-value problems, one also encounters problems of boundary-value type, where conditions are specified at *two* points, normally the ends of the interval I. Not only are boundary-value problems not covered by Theorem 3.3.1, it is striking that *boundary-value problems need not have unique solutions*. In fact, they may have *no solution, a unique solution, or a nonunique solution*, as shown by the following example.

EXAMPLE 6. *Boundary-Value Problem.* It is readily verified that the differential equation

$$y'' + y = 0 \tag{21}$$

admits a general solution

$$y(x) = C_1 \cos x + C_2 \sin x. \tag{22}$$

Consider three different sets of boundary values.

Case 1: $y(0) = 2, y(\pi) = 1$. Then

$$\begin{aligned} y(0) &= 2 = C_1 + 0, \\ y(\pi) &= 1 = -C_1 + 0, \end{aligned}$$

which has no solution for C_1, C_2, so the boundary-value problem has no solution for $y(x)$.

Case 2: $y(0) = 2, y(\pi/2) = 3$. Then

$$\begin{aligned} y(0) &= 2 = C_1 + 0, \\ y(\pi/2) &= 3 = 0 + C_2, \end{aligned}$$

so $C_1 = 2, C_2 = 3$, and the boundary-value problem has the unique solution $y(x) = 2\cos x + 3\sin x$.

Case 3: $y(0) = 2$, $y(\pi) = -2$. Then

$$y(0) = 2 = C_1 + 0,$$
$$y(\pi) = -2 = -C_1 + 0,$$

so $C_1 = 2$, and C_2 is arbitrary, and the boundary-value problem has the nonunique solution (indeed, the infinity of solutions) $y(x) = 2\cos x + C_2 \sin x$, where C_2 is an arbitrary constant. ∎

Boundary-value problems are studied further in Sections 6.4, 17.7, and 17.8.

Closure. In this section we have considered the nth-order linear homogeneous differential equation $L[y] = 0$. The principal result was that a general solution

$$y(x) = C_1 y_1(x) + \cdots + C_n y_n(x)$$

can be built up by the superposition of n LI solutions $y_1(x), \ldots, y_n(x)$, thanks to the linearity of L, namely, the property of L that

$$L[\alpha u + \beta v] = \alpha L[u] + \beta L[v]$$

for any n-times differentiable functions u, v and any constants α, β. Any such set of n LI solutions $\{y_1, \ldots, y_k\}$ of $L[y] = 0$ is called a **basis of solutions** (or **basis**, for brevity) for that equation.

For the initial-value problem consisting of the differential equation $L[y] = 0$ together with the initial conditions (5b), we found that a solution exists and is unique, but for the boundary-value version we found that a solution need not exist, it may exist and be unique, or there may exist a nonunique solution.

Theorems 3.3.1 and 3.3.3 are especially important.

EXERCISES 3.3

NOTE: If not specified, the interval I is understood to be the entire x axis.

1. Show whether or not each of the following is a general solution to the equation given.

(a) $y'' - 3y' + 2y = 0$; $C_1 e^x + C_2 e^{2x}$
(b) $y'' - 3y' + 2y = 0$; $C_1(e^x - e^{2x}) + C_2 e^x$
(c) $y'' - y' - 2y = 0$; $C_1(e^{-x} + e^{2x})$
(d) $y'' - y' - 2y = 0$; $C_1 e^{-x} + C_2 e^{2x}$
(e) $y''' + 4y' = 0$; $C_1 \cos 2x + C_2 \sin 2x$
(f) $y''' + 4y' = 0$; $C_1 + C_2 \cos 2x + C_2 \sin 2x$
(g) $y''' - 2y'' + y' = 0$; $(C_1 + C_2 x + C_3 x^2)\, e^x$
(h) $y''' - 2y'' + y' = 0$; $(C_1 + C_2 x)\, e^x + C_3$
(i) $y''' - y'' - y' + y = 0$; $C_1 e^x + C_2 e^{-x} + C_3 x e^x$

2. Show whether or not each of the following is a basis for the given equation.

(a) $y'' - 9y = 0$; $\{e^{3x}, \cosh 3x, \sinh 3x\}$
(b) $y'' - 9y = 0$; $\{e^{3x}, \cosh 3x\}$
(c) $y'' - y = 0$; $\{\sinh 3x, 2\cosh 3x\}$
(d) $y''' - 3y'' + 3y' - y = 0$; $\{e^x, xe^x, x^2 e^x\}$
(e) $y''' - 3y'' = 0$; $\{1, x, e^{3x}\}$

(f) $y'''' + 2y'' + y = 0$; $\{\cos x, \sin x, x\cos x, x\sin x\}$

3. Are the following general solutions of $x^2y'' + xy' - 4y = 0$ on $0 < x < \infty$? On $-\infty < x < \infty$? Explain.

(a) C_1x^2 (b) $C_1x^2 + C_2x^{-2}$
(c) $C_1(x^2 + x^{-2}) + C_2(x^2 - x^{-2})$

4. Are the following bases for the equation $x^2y'' - xy' + y = 0$ on $0 < x < \infty$? On $-\infty < x < 0$? On $-\infty < x < \infty$? On $6 < x < 10$? Explain.

(a) $\{x, x^2\}$
(b) $\{e^x, e^{-x}\}$
(c) $\{x, x\ln|x|\}$
(d) $\{x + x\ln|x|, x - x\ln|x|\}$

5. Show whether or not the following is a general solution of $y^{(vii)} - 4y^{(vi)} - 14y^{(v)} + 56y^{(iv)} + 49y''' - 196y'' - 36y' + 144y = 0$.

(a) $C_1e^x + C_2e^{-x} + C_3e^{2x} + C_4e^{-2x} + C_5e^{3x} + C_6e^{-3x}$
(b) $C_1e^x + C_2e^{-x} + C_3e^{2x} + C_4e^{-2x} + C_5e^{3x} + C_6e^{-3x} + C_7\sinh x + C_8\cosh 2x$

6. Show that $y_1 = 1$ and $y_2 = 2$ are solutions of $(y^3 - 6y^2 + 11y - 6)\,y'' = 0$. Is $y = y_1 + y_2 = 1 + 2 = 3$ a solution as well? Does your result contradict the sentence preceding Example 2? Explain.

7. Show that each of the functions $y_1 = 3x^2 - x$ and $y_2 = x^2 - x$ is a solution of the equation $x^2y'' - 2y = 2x$. Is the linear combination $C_1y_1 + C_2y_2$ a solution as well, for all choices of the constants C_1 and C_2?

8. (*Taylor series method*) Use the Taylor series method described below Theorem 3.3.1 to solve each of the following initial-value problems for $y(x)$, up to and including terms of fifth order. NOTE: The term $f^{(n)}(a)(x-a)^n/n!$ in the Taylor series of $f(x)$ about $x = a$ is said to be of ***n*th-order**.

(a) $y'' + y = 0$; $y(0) = 4,\ y'(0) = 3$
(b) $y'' - 4y = 0$; $y(0) = -1,\ y'(0) = 0$
(c) $y'' + 5y' + 6y = 0$; $y(0) = 2,\ y'(0) = -5$
(d) $y'' + xy = 0$; $y(0) = 1,\ y'(0) = 0$
(e) $y'' + x^2y = 0$; $y(0) = 2,\ y'(0) = -3$
(f) $y'' - 3y = 0$; $y(5) = 4,\ y'(5) = 6$ HINT: Expand about $x = 5$.
(g) $y'' + 3y' - y = 0$; $y(1) = 2,\ y'(1) = 0$ HINT: Expand about $x = 1$.
(h) $y''' - y' + 2y = 0$; $y(0) = 0,\ y'(0) = 0,\ y''(0) = 1$
(i) $y''' - xy = 0$; $y(0) = 0,\ y'(0) = 3,\ y''(0) = -2$

9. Does the problem stated have a unique solution? No solution? A nonunique solution? Explain.

(a) $y'' + 2y' + 3y = 0$; $y(0) = 5,\ y'(0) = -1$
(b) $y'' + 2y' + 3y = 0$; $y(3) = 2,\ y'(3) = 37$
(c) $y'' + xy' - y = 0$; $y(3) = y'(3) = 0$
(d) $xy''' + xy' - y = 0$; $y(-1) = y'(-1) = 0,\ y''(-1) = 4$
(e) $x^2y'' - y' - y = 0$; $y(6) = 0,\ y'(6) = 1$
(f) $(\sin x)y'''' + xy''' = 0$; $y(2) = y'(2) = y''(2) = 0$, $y'''(2) = -9$

10. Verify that (22) is indeed a general solution of (21).

11. Consider the boundary-value problem consisting of the differential equation $y'' + y = 0$ plus the boundary conditions given. Does the problem have any solutions? If so, find them. Is the solution unique? HINT: A general solution of the differential equation is $y = C_1\cos x + C_2\sin x$.

(a) $y(0) = 0,\ y(2) = 0$
(b) $y(0) = 0,\ y(2\pi) = -3$
(c) $y(1) = 1,\ y(2) = 2$
(d) $y'(0) = 0,\ y(5) = 1$
(e) $y'(0) = 0,\ y'(\pi) = 0$
(f) $y'(0) = 0,\ y'(6\pi) = 0$
(g) $y'(0) = 0,\ y'(2\pi) = 38$

12. Consider the boundary-value problem consisting of the differential equation $y'''' + 2y'' + y = 0$ plus the boundary conditions given. Does the problem have any solutions? If so, find them. Is the solution unique? HINT: A general solution of the differential equation is $y = (C_1 + C_2x)\cos x + (C_3 + C_4x)\sin x$.

(a) $y(0) = y'(0) = 0,\ y(\pi) = 0,\ y'(\pi) = 2$
(b) $y(0) = y'(0) = y''(0) = 0,\ y(\pi) = 1$
(c) $y(0) = y''(0) = 0,\ y(\pi) = 0 = y''(\pi) = 0$
(d) $y(0) = y''(0) = 0,\ y(\pi) = y''(\pi) = 3$

13. Prove that the linearity property (10) is equivalent to the two properties

$$L[u + v] = L[u] + L[v], \tag{13.1a}$$

$$L[\alpha u] = \alpha L[u]. \tag{13.1b}$$

That is, show that the truth of (10) implies the truth of (13.1), and conversely.

14. We showed that (11) holds for the case $k = 3$, but did not prove it in general. Here, we ask you to prove (11) for any integer $k \geq 1$. HINT: It is suggested that you use **mathematical induction**, whereby a proposition $P(k)$, for $k \geq 1$, is proved by first showing that it holds for $k = 1$, and then showing that if it holds for k then it must also hold for $k + 1$. In the present example, the proposition $P(k)$ is the equation (11).

15. (*Example 4, Continued*) (a) Verify that each of (18a)

through (18d) is a general solution of $y'' - 9y = 0$.
(b) It seems reasonable that if C_1, C_2 are arbitrary constants, and if we call

$$C_1 + C_2 = C_1' \quad \text{and} \quad C_1 - C_2 = C_2', \tag{15.1}$$

then C_1', C_2' are arbitrary too, as we claimed at the end of Example 4. Actually, for that claim to be true we need to be able to show that corresponding to any chosen values C_1', C_2' the equations (15.1) on C_1, C_2 are consistent – that is, that they admit one or more solutions for C_1, C_2. Show that (15.1) is indeed consistent.
(c) Show that if, instead, we had $C_1 + C_2 = C_1'$ and $2C_1 + 2C_2 = C_2'$, where C_1, C_2 are arbitrary constants, then it is *not* true that C_1', C_2' are arbitrary too.

3.4 Solution of Homogeneous Equation: Constant Coefficients

Knowing that the general solution of an nth-order homogeneous differential equation is an arbitrary linear combination of n LI (linearly independent) solutions, the question is: How do we find those solutions? That question will occupy us for the remainder of this chapter and for the next three chapters as well. In this section we consider the constant-coefficient case,

$$\frac{d^n y}{dx^n} + a_1 \frac{d^{n-1} y}{dx^{n-1}} + \cdots + a_{n-1} \frac{dy}{dx} + a_n y = 0; \tag{1}$$

that is, where the a_j coefficients are constants, not functions of x. This case is said to be "elementary" in the sense that the solutions will always be found among the *elementary functions* (powers of x, trigonometric functions, exponentials, and logarithms), but it is also elementary in the sense that it is the simplest case: nonconstant-coefficient equations are generally much harder, and nonlinear equations are much harder still.

Fortunately, the constant-coefficient case is not only the simplest, it is also of great importance in science and engineering. For instance, the equations

$$mx'' + cx' + kx = 0$$

and

$$Li'' + Ri' + \frac{1}{C} i = 0,$$

governing mechanical and electrical oscillators, where primes denote derivatives with respect to the independent variable t, are both of the type (1) because m, c, k and L, R, C are constants; they do not vary with t.

3.4.1. Euler's formula and review of the circular and hyperbolic functions. We are going to be involved with the exponential function e^z, where $z = x + iy$ is complex and $i = \sqrt{-1}$. The first point to appreciate is that we cannot figure out how to evaluate e^{x+iy} from our knowledge of the function e^x where x is real. That

is, e^{x+iy} is a "new object," and its values are a matter of definition, not a matter of figuring out. To motivate that definition, let us proceed as follows:

$$\begin{aligned} e^z = e^{x+iy} &= e^x e^{iy} \\ &= e^x \left[1 + iy + \frac{(iy)^2}{2!} + \frac{(iy)^3}{3!} + \frac{(iy)^4}{4!} + \cdots\right] \\ &= e^x \left[\left(1 - \frac{y^2}{2!} + \frac{y^4}{4!} - \cdots\right) + i\left(y - \frac{y^3}{3!} + \frac{y^5}{5!} - \cdots\right)\right]. \end{aligned} \tag{2}$$

Recognizing the two series as the Taylor series representations of $\cos y$ and $\sin y$, respectively, we obtain

$$\boxed{e^{x+iy} = e^x (\cos y + i \sin y),} \tag{3}$$

which is known as **Euler's formula**, after the great Swiss mathematician *Leonhard Euler* (1707–1783), whose many contributions to mathematics included the systematic development of the theory of linear constant-coefficient differential equations.

We say that (3) defines e^{x+iy} since it gives e^{x+iy} in the standard Cartesian form $a + ib$, where the real part a is $e^x \cos y$ and the imaginary part b is $e^x \sin y$. Observe carefully that we cannot defend certain steps in (2). Specifically, the second equality seems to be the familiar formula $e^{a+b} = e^a e^b$, but the latter is for real numbers a and b, whereas iy is not real. Likewise, the third equality rests upon the Taylor series formula $e^u = 1 + u + \frac{u^2}{2!} + \cdots$ that is derived in the calculus for the case where u is real, but iy is not real. The point to understand, then, is that the steps in (2) are merely heuristic; trying to stay as close to real-variable theory as possible, we arrive at (3). Once (3) is obtained, we throw out (2) and take (3) as our (i.e., Euler's) *definition* of e^{x+iy}. Of course, there are an infinite number of ways one can define a given quantity, but some are more fruitful than others. To fully appreciate why Euler's definition is the perfect one for e^{x+iy}, one needs to study complex-variable theory, as we will in later chapters. For the present, we merely propose that the steps in (2) make (3) a *reasonable* choice as a definition of e^z.

As a special case of (3), let $x = 0$. Then (3) becomes

$$e^{iy} = \cos y + i \sin y. \tag{4a}$$

For instance, $e^{\pi i} = \cos \pi + i \sin \pi = -1 + 0i = -1$, and $e^{2-3i} = e^2 (\cos 3 - i \sin 3) = 7.39(-0.990 - 0.141i) = -7.32 - 1.04i$. Since (4a) holds for all y, it must hold also with y changed to $-y$:

$$e^{-iy} = \cos(-y) + i \sin(-y),$$

and since $\cos(-y) = \cos y$ and $\sin(-y) = -\sin y$, it follows that

$$e^{-iy} = \cos y - i \sin y. \tag{4b}$$

Conversely, we can express $\cos y$ and $\sin y$ as linear combinations of the complex exponentials e^{iy} and e^{-iy}, for adding (4a) and (4b) and subtracting them gives $\cos y = \left(e^{iy} + e^{-iy}\right)/2$ and $\sin y = \left(e^{iy} - e^{-iy}\right)/(2i)$. Let us frame these formulas, for emphasis and reference:

$$\boxed{\begin{aligned} e^{iy} &= \cos y + i \sin y \\ e^{-iy} &= \cos y - i \sin y \end{aligned}} \tag{5a,b}$$

and

$$\boxed{\begin{aligned} \cos y &= \frac{e^{iy} + e^{-iy}}{2} \\ \sin y &= \frac{e^{iy} - e^{-iy}}{2i}. \end{aligned}} \tag{6a,b}$$

Observe that all four of these formulas come from the single formula (4a). (Of course there is nothing essential about the name of the variable in these formulas. For instance, $e^{ix} = \cos x + i \sin x$, $e^{i\theta} = \cos\theta + i \sin\theta$, and so on.)

There is a similarity between (5) and (6), relating the cosine and sine to the complex exponentials, to analogous formulas relating the hyperbolic cosine and hyperbolic sine to real exponentials. If we recall the definitions

$$\boxed{\begin{aligned} \cosh y &= \frac{e^{y} + e^{-y}}{2}, \\ \sinh y &= \frac{e^{y} - e^{-y}}{2} \end{aligned}} \tag{7a,b}$$

of the hyperbolic cosine and hyperbolic sine, we find, by addition and subtraction of these formulas, that

$$\boxed{\begin{aligned} e^{y} &= \cosh y + \sinh y, \\ e^{-y} &= \cosh y - \sinh y. \end{aligned}} \tag{8a,b}$$

Compare (5) with (8), and (6) with (7). The graphs of $\cosh x$, $\sinh x$, e^x, and e^{-x} are given in Fig. 1.

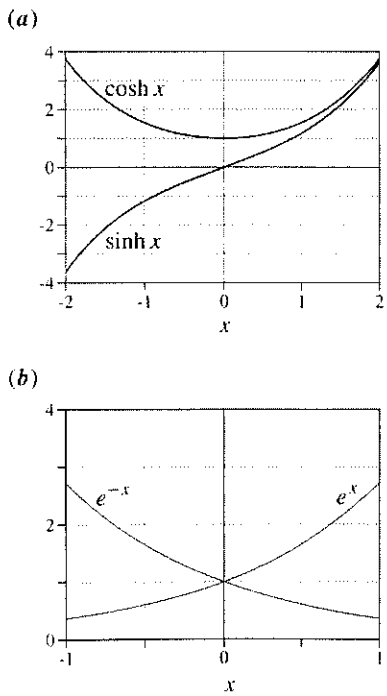


Figure 1. $\cosh x$, $\sinh x$, e^x, and e^{-x}.

Using (6) and (7) we obtain the properties

$$\cos^2 y + \sin^2 y = 1, \tag{9}$$

$$\cosh^2 y - \sinh^2 y = 1. \tag{10}$$

From a geometric point of view, if we parametrize a curve C by the relations

$$x = \cos\tau, \quad y = \sin\tau \tag{11}$$

over $0 \le \tau < 2\pi$, say, then it follows from (9) that $x^2 + y^2 = 1$, so that C is a circle. And if we parametrize C by

$$x = \cosh\tau, \quad y = \sinh\tau \tag{12}$$

instead, then it follows from (10) that $x^2 - y^2 = 1$, so C is a hyperbola. Thus, one refers to $\cos x$ and $\sin x$ as *circular functions* and to $\cosh x$ and $\sinh x$ as *hyperbolic functions*, the *hyperbolic cosine* and the *hyperbolic sine*, respectively.

Besides (9) and (10), various useful identities, such as

$$\sin(A+B) = \sin A \cos B + \sin B \cos A, \tag{13a}$$
$$\cos(A+B) = \cos A \cos B - \sin A \sin B, \tag{13b}$$
$$\sinh(A+B) = \sinh A \cosh B + \sinh B \cosh A, \tag{13c}$$
$$\cosh(A+B) = \cosh A \cosh B + \sinh A \sinh B, \tag{13d}$$

can be derived from (6) and (7), as well as the derivative formulas

$$\begin{aligned} \frac{d}{dx}\cos x &= -\sin x, & \frac{d}{dx}\sin x &= \cos x, \\ \frac{d}{dx}\cosh x &= \sinh x, & \frac{d}{dx}\sinh x &= \cosh x. \end{aligned} \tag{14}$$

We shall be interested specifically in the function $e^{\lambda x}$ and its derivatives with respect to x, where λ is a constant that may be complex, say $\lambda = a + ib$. We know from the calculus that

$$\frac{d}{dx}e^{\lambda x} = \lambda e^{\lambda x} \tag{15}$$

when λ is a real constant. Does (15) hold when λ is complex? To answer the question, use Euler's formula (3) to express

$$e^{\lambda x} = e^{(a+ib)x} = e^{ax}(\cos bx + i \sin bx).$$

Thus,

$$\begin{aligned} \frac{d}{dx}e^{\lambda x} &= \frac{d}{dx}\left[e^{ax}(\cos bx + i \sin bx)\right] \\ &= \frac{d}{dx}(e^{ax}\cos bx) + i\frac{d}{dx}(e^{ax}\sin bx) \\ &= (ae^{ax}\cos bx - be^{ax}\sin bx) + i(ae^{ax}\sin bx + be^{ax}\cos bx) \\ &= e^{ax}(a+ib)(\cos bx + i \sin bx) \\ &= \lambda e^{ax}(\cos bx + i \sin bx) = \lambda e^{\lambda x}, \end{aligned}$$

so the familiar formula (15) does hold even for complex λ.

There is one more fact about the exponential function that we will be needing, namely, that the exponential function e^z cannot be zero for any choice of z; that is, *it has no zeros*, for

$$\begin{aligned} |e^z| &= \left|e^{x+iy}\right| = \left|e^x(\cos y + i \sin y)\right| \\ &= |e^x|\,|\cos y + i \sin y| = e^x\,|\cos y + i \sin y| = e^x. \end{aligned}$$

The fourth equality follows from the fact that the real exponential is everywhere positive, and the fifth equality from the fact that $|a + ib|$ is the square root of the sum of the squares of a and b, and $\cos^2 y + \sin^2 y = 1$. Finally, we know that $e^x > 0$ for all x, so $|e^z| > 0$ for all z, and hence $e^z \neq 0$ for all z, as claimed.

3.4.2. Exponential solutions. To guide our search for solutions of (1), it is a good idea to begin with the simplest case, $n = 1$:

$$\frac{dy}{dx} + a_1 y = 0, \tag{16}$$

the general solution of which is

$$y(x) = Ce^{-a_1 x}, \tag{17}$$

where C is an arbitrary constant. One can derive (17) by noticing that (16) is a first-order linear equation and using the general solution developed in Section 2.2, or by using the fact that (16) is separable.

Observing that (17) is of exponential form, it is natural to wonder if higher-order equations admit exponential solutions too. Consider the second-order equation

$$y'' + a_1 y' + a_2 y = 0, \tag{18}$$

where a_1 and a_2 are real numbers, and let us seek a solution in the form

$$y(x) = e^{\lambda x}. \tag{19}$$

If (19) is to be a solution of (18), then it must be true that

$$\lambda^2 e^{\lambda x} + a_1 \lambda e^{\lambda x} + a_2 e^{\lambda x} = 0, \tag{20}$$

or

$$\left(\lambda^2 + a_1 \lambda + a_2\right) e^{\lambda x} = 0, \tag{21}$$

where (20) holds, according to (15), even if the not-yet-determined constant λ turns out to be complex. For (19) to be a solution of (18) on some interval I, we need (21) to be satisfied on I. That is, we need the left side of (21) to vanish identically on I. Since $e^{\lambda x}$ is not identically zero on any interval I for any choice of λ, we need λ to be such that

$$\lambda^2 + a_1 \lambda + a_2 = 0. \tag{22}$$

This equation and its left-hand side are called the **characteristic equation** and **characteristic polynomial**, respectively, corresponding to the differential equation (18). In general, (22) gives two distinct roots, say λ_1 and λ_2, which can be found from the quadratic formula as

$$\lambda = \frac{-a_1 \pm \sqrt{a_1^2 - 4a_2}}{2}.$$

(The nongeneric case of repeated roots, which occurs if $a_1^2 - 4a_2$ vanishes, is discussed separately, below.)

Thus, our choice of the exponential form (19) has been successful. Indeed, we have found *two* solutions of that form, $e^{\lambda_1 x}$ and $e^{\lambda_2 x}$. Next, from Theorem 3.3.2 it follows [thanks to the linearity of (18)] that if $e^{\lambda_1 x}$ and $e^{\lambda_2 x}$ are solutions of (18) then so is any linear combination of them,

$$y(x) = C_1 e^{\lambda_1 x} + C_2 e^{\lambda_2 x}. \tag{23}$$

Theorem 3.3.3 guarantees that (23) is a general solution of (18) if $e^{\lambda_1 x}$ and $e^{\lambda_2 x}$ are LI on I, and Theorem 3.2.4 tells us that they are indeed LI since neither one is expressible as a scalar multiple of the other. Thus, by seeking solutions in the form (19) we were successful in finding the general solution (23) of (18).

EXAMPLE 1. For the equation

$$y'' - y' - 6y = 0, \tag{24}$$

the characteristic equation is $\lambda^2 - \lambda - 6 = 0$, with roots $\lambda = -2, 3$, so

$$y(x) = C_1 e^{-2x} + C_2 e^{3x} \tag{25}$$

is a general solution of (24). ∎

EXAMPLE 2. For the equation

$$y'' - 9y = 0, \tag{26}$$

the characteristic equation is $\lambda^2 - 9 = 0$, with roots $\lambda = \pm 3$, so a general solution of (26) is

$$y(x) = C_1 e^{3x} + C_2 e^{-3x}. \tag{27}$$

COMMENT 1. As discussed in Example 4 of Section 3.3, an infinite number of forms of the general solution to (26) are equivalent to (27), such as

$$y(x) = C_1 \cosh 3x + C_2 \sinh 3x, \tag{28}$$

$$y(x) = C_1 \sinh 3x + C_2 \left(5e^{-3x} - 2\cosh 3x\right), \tag{29}$$

$$y(x) = C_1 \left(e^{3x} + 4\sinh 3x\right) + C_2 \left(\cosh 3x - \sqrt{\pi}\sinh 3x\right), \tag{30}$$

and so on. Of these, one would normally choose either (27) or (28). What is wrong with (29) and (30)? Nothing, except that they are ugly; e^{3x} and e^{-3x} make a "handsome couple," and $\cosh 3x$ and $\sinh 3x$ do too, but the choices in (29) and (30) seem ugly and purposeless.

COMMENT 2. If (27) and (28) are equivalent, does it matter whether we choose one or the other? No, since they are equivalent. However, one may be more convenient than the other insofar as the application of initial or boundary conditions.

For instance, suppose we append to (26) the initial conditions $y(0) = 4$, $y'(0) = -5$. Applying these to (27) gives

$$\begin{aligned} y(0) &= 4 = C_1 + C_2, \\ y'(0) &= -5 = 3C_1 - 3C_2, \end{aligned} \tag{31}$$

so $C_1 = 7/6$, $C_2 = 17/6$, and $y(x) = (7e^{3x} + 17e^{-3x})/6$. Applying these initial conditions to (28), instead, gives

$$\begin{aligned} y(0) &= 4 = C_1, \\ y'(0) &= -5 = 3C_2, \end{aligned} \tag{32}$$

so $C_1 = 4$, $C_2 = -5/3$, and $y(x) = 4\cosh 3x - (5/3)\sinh 3x$. Whereas our final results are equivalent, we see that (32) was more readily solved than (31). Thus, $\cosh 3x$ and $\sinh 3x$ make a slightly better choice in this case than e^{3x} and e^{-3x} – namely, when initial conditions are given at $x = 0$.

Or, suppose we consider I to be $0 \le x < \infty$ and impose the boundary conditions that $y(0) = 6$, and that $y(x)$ is to be bounded as $x \to \infty$. That is, rather than impose a numerical value on y at infinity, we impose a **boundedness condition**, that $|y(x)| < M$ for all x, for some constant M. Applying these conditions to (27) we see, from the boundedness condition, that we need $C_1 = 0$ since otherwise the e^{3x} will give unbounded growth. Next, $y(0) = 6 = C_2$, and hence the solution is $y(x) = 6e^{-3x}$. Notice how easily the boundedness condition was applied to (27).

If we use (28) instead, the solution is harder since both $\cosh 3x$ and $\sinh 3x$ grow unboundedly as $x \to \infty$. We can't afford to set both $C_1 = 0$ and $C_2 = 0$, in (28) since then we would have $y(x) = 0$, which does indeed satisfy both the equation (26) and the boundedness condition, but cannot satisfy the remaining initial condition $y(0) = 6$. However, perhaps the growth in $\cosh 3x$ and $\sinh 3x$ can be made to cancel. That is, write

$$\begin{aligned} y(x) &= C_1 \cosh 3x + C_2 \sinh 3x \\ &= C_1 \left(\frac{e^{3x} + e^{-3x}}{2} \right) + C_2 \left(\frac{e^{3x} - e^{-3x}}{2} \right) \\ &= \frac{C_1 + C_2}{2} e^{3x} + \frac{C_1 - C_2}{2} e^{-3x}, \end{aligned} \tag{33}$$

so for boundedness we need $C_1 + C_2 = 0$ (and hence $C_2 = -C_1$). Then (33) gives $y(x) = C_1 e^{-3x}$ and $y(0) = 6$ gives $C_1 = 6$ and $y(x) = 6e^{-3x}$, as before. Thus, in the case of a boundedness boundary condition at infinity we see that the exponential form (27) is more convenient than the hyperbolic form (28).

To summarize, when confronted with a choice, such as between (27) and (28), look ahead to the application of any initial or boundary conditions to see if one form will be more convenient than the other. ∎

EXAMPLE 3. For

$$y'' + 9y = 0, \tag{34}$$

the characteristic equation is $\lambda^2 + 9 = 0$, with roots $\lambda = \pm 3i$, so a general solution of (34) is

$$y(x) = C_1 e^{i3x} + C_2 e^{-i3x}. \tag{35}$$

COMMENT 1. Just as the general solution of $y'' - 9y = 0$ was expressible in terms of the real exponentials e^{3x}, e^{-3x} or the hyperbolic functions $\cosh 3x, \sinh 3x$, the general solution of (34) is expressible in terms of the complex exponentials e^{i3x}, e^{-i3x} or in terms of the circular functions $\cos 3x, \sin 3x$, for we can use Euler's formula to re-express (35) as

$$\begin{aligned} y(x) &= C_1 (\cos 3x + i \sin 3x) + C_2 (\cos 3x - i \sin 3x) \\ &= (C_1 + C_2) \cos 3x + i (C_1 - C_2) \sin 3x. \end{aligned} \tag{36}$$

Since C_1 and C_2 are arbitrary constants, we can simplify this result by letting $C_1 + C_2$ be a new constant A, and letting $i(C_1 - C_2)$ be a new constant B, so we have, from (36), the form

$$y(x) = A \cos 3x + B \sin 3x, \tag{37}$$

where A, B are arbitrary constants. As in Example 1, we note that (35) and (37) are but two out of an infinite number of equivalent forms.

COMMENT 2. You may be concerned that if $y(x)$ is a physical quantity such as the displacement of a mass or the current in an electrical circuit, then it should be real, whereas the right side of (35) seems to be complex. To explore this point, let us solve a complete problem, the differential equation (34) plus a representative set of initial conditions, say $y(0) = 7, y'(0) = 3$, and see if the final answer is real or not. Imposing the initial conditions on (35),

$$\begin{aligned} y(0) &= 7 = C_1 + C_2, \\ y'(0) &= 3 = i3C_1 - i3C_2, \end{aligned}$$

so $C_1 = (7-i)/2$ and $C_2 = (7+i)/2$. Putting these values into (35), we see from (36) that $y(x) = \frac{1}{2}[(7-i)+(7+i)] \cos 3x + \frac{1}{2} i[(7-i)-(7+i)] \sin 3x = 7 \cos 3x + \sin 3x$, which is indeed real. Put differently, if the differential equation and initial conditions represent some physical system, then the mathematics "knows all about" the physics; it is built in, and we need not be anxious. ∎

Having already made the point that the general solution can always be expressed in various different (but equivalent) forms, we will generally adopt the exponential form when the exponentials are real, and the circular function form when they are complex. This decision is one of personal preference.

EXAMPLE 4. The equation

$$y'' + 4y' + 7y = 0 \tag{38}$$

has the characteristic equation $\lambda^2 + 4\lambda + 7 = 0$, with distinct roots $\lambda = -2 \pm i\sqrt{3}$, so a general solution of (38) is

$$\begin{aligned} y(x) &= C_1 e^{(-2+i\sqrt{3})x} + C_2 e^{(-2-i\sqrt{3})x} \\ &= e^{-2x} \left(C_1 e^{i\sqrt{3}x} + C_2 e^{-i\sqrt{3}x} \right) \\ &= e^{-2x} \left(A \cos \sqrt{3}x + B \sin \sqrt{3}x \right). \end{aligned} \tag{39}$$

That is, first we factor out the common factor e^{-2x}, then we re-express the complex exponentials in terms of the circular functions.

If we impose initial conditions $y(0) = 1, y'(0) = 0$, say, we find that $A = 1$ and $B = 2/\sqrt{3}$, so $y(x) = e^{-2x}\left(\cos\sqrt{3}x + \frac{2}{\sqrt{3}}\sin\sqrt{3}x\right)$. According to Theorem 3.3.1, that solution is unique. ∎

3.4.3. Higher-order equations $(n > 2)$. Examples 1–4 are representative of the four possible cases for second-order equations having distinct roots of the characteristic equation: if the roots are both real then the solution is expressible as a linear combination of two real exponentials (Example 1); if they are both real and equal and opposite in sign, then the solution is expressible either as exponentials or as a hyperbolic cosine and a hyperbolic sine (Example 2); if they are not both real then they will be complex conjugates. If those complex conjugates are purely imaginary, then the solution is expressible as a linear combination of two complex exponentials or as a sine and a cosine (Example 3); if they are not purely imaginary, then the solution is expressible as a real exponential times a linear combination of complex exponentials or a sine and a cosine (Example 4).

Turning to higher-order equations $(n > 2)$, our attention focuses on the characteristic equation

$$\lambda^n + a_1\lambda^{n-1} + \cdots + a_{n-1}\lambda + a_n = 0. \tag{40}$$

If $n = 1$, then (40) becomes $\lambda + a_1 = 0$ which, of course, has the root $\lambda = -a_1$ on the real axis. If $n = 2$, then (40) becomes $\lambda^2 + a_1\lambda + a_2 = 0$, and to be assured of the existence of solutions we need to extend our number system from a real axis to a complex plane. If the roots are indeed complex (and both a_1 and a_2 are real) they will necessarily occur as a complex conjugate pair, as in Example 4.

One might wonder if a further extension of the number system, beyond the complex plane, is required to assure the existence of solutions to (40) for $n \geq 3$. However, it turns out that the complex plane continues to suffice. The characteristic equation (40) necessarily admits n roots. As for the case $n = 2$, they need not be distinct and they need not be real, but if there are complex roots then they necessarily occur in complex conjugate pairs (if all of the a_j's are real).

In this subsection we limit attention to the case where there are n distinct roots of (40), which we denote as $\lambda_1, \lambda_2, \ldots, \lambda_n$. Then each of the exponentials $e^{\lambda_1 x}, \ldots, e^{\lambda_n x}$ is a solution of (1) and, by Theorem 3.3.3,

$$y(x) = C_1 e^{\lambda_1 x} + \cdots + C_n e^{\lambda_n x} \tag{41}$$

is a general solution of (1) if and only if the set of exponentials is LI.

THEOREM 3.4.1 *Linear Independence of a Set of Exponentials*
Let $\lambda_1, \ldots, \lambda_n$ be any numbers, real or complex. The set $\{e^{\lambda_1 x}, \ldots, e^{\lambda_n x}\}$ is LI (on any given interval I) if and only if the λ's are distinct.

Proof: Recall from Theorem 3.2.2 that if the Wronskian determinant

$$W\left[e^{\lambda_1 x}, \ldots, e^{\lambda_n x}\right](x) = \begin{vmatrix} e^{\lambda_1 x} & \cdots & e^{\lambda_n x} \\ \lambda_1 e^{\lambda_1 x} & \cdots & \lambda_n e^{\lambda_n x} \\ \vdots & & \vdots \\ \lambda_1^{n-1} e^{\lambda_1 x} & \cdots & \lambda_n^{n-1} e^{\lambda_n x} \end{vmatrix} \tag{42}$$

is not identically zero on I, then the set is LI on I. According to the properties of determinants (Section 10.4), we can factor $e^{\lambda_1 x}$ out of the first column, $e^{\lambda_2 x}$ out of the second, and so on, so that we can re-express W as

$$W\left[e^{\lambda_1 x}, \ldots, e^{\lambda_n x}\right](x) = e^{(\lambda_1 + \cdots + \lambda_n)x} \begin{vmatrix} 1 & \cdots & 1 \\ \lambda_1 & \cdots & \lambda_n \\ \vdots & & \vdots \\ \lambda_1^{n-1} & \cdots & \lambda_n^{n-1} \end{vmatrix}. \tag{43}$$

The exponential function on the right-hand side is nonzero. Further, the determinant is of Vandermonde type (Section 10.4), a key property of which is that it is nonzero if the λ's are distinct, as indeed has been assumed here. Thus, W is nonzero (on any interval), so the given set is LI.

Conversely, if the λ's are not distinct, then surely the set is LD because at least two of its members are identical. ■

Consider the following examples.

EXAMPLE 5. The equation

$$y''' - 8y' + 8y = 0 \tag{44}$$

has the characteristic equation $\lambda^3 - 8\lambda + 8 = 0$. Trial and error reveals that $\lambda = 2$ is one root. Hence we can factor $\lambda^3 - 8\lambda + 8$ as $(\lambda - 2)p(\lambda)$, where $p(\lambda)$ is a quadratic function of λ. To find $p(\lambda)$ we divide $\lambda - 2$ into $\lambda^3 - 8\lambda + 8$, by long division, and obtain $p(\lambda) = \lambda^2 + 2\lambda - 4$ which, in turn, can be factored as $[\lambda - (-1 + \sqrt{5})][\lambda - (-1 - \sqrt{5})]$. Thus, λ equals 2 and $-1 \pm \sqrt{5}$, so

$$\begin{aligned} y(x) &= C_1 e^{2x} + C_2 e^{(-1+\sqrt{5})x} + C_3 e^{(-1-\sqrt{5})x} \\ &= C_1 e^{2x} + e^{-x}\left(C_2 e^{\sqrt{5}x} + C_2 e^{-\sqrt{5}x}\right) \end{aligned} \tag{45}$$

is a general solution of (44).

COMMENT. Alternative to long division, we can find $p(\lambda)$ by writing $\lambda^3 - 8\lambda + 8 = (\lambda - 2)(a\lambda^2 + b\lambda + c) = a\lambda^3 + (b - 2a)\lambda^2 + (c - 2b)\lambda - 2c$ and determining a, b, c so that coefficients of like powers of λ match on both sides of the equation. ■

EXAMPLE 6. The equation

$$y''' - y = 0 \tag{46}$$

has the characteristic equation $\lambda^3 - 1 = 0$, which surely has the root $\lambda = 1$. Thus, $\lambda^3 - 1$ is $\lambda - 1$ times a quadratic function of λ, which function can be found, as above, by long division. Thus, we obtain

$$(\lambda - 1)\left(\lambda^2 + \lambda + 1\right) = 0,$$

so λ equals 1 and $(-1 \pm \sqrt{3}i)/2$. Hence

$$\begin{aligned} y(x) &= C_1 e^x + C_2 e^{(-1+\sqrt{3}i)x/2} + C_3 e^{(-1-\sqrt{3}i)x/2} \\ &= C_1 e^x + e^{-x/2}\left(C_2 e^{i\sqrt{3}x/2} + C_3 e^{-i\sqrt{3}x/2}\right) \\ &= C_1 e^x + e^{-x/2}\left(C_2' \cos\frac{\sqrt{3}}{2}x + C_3' \sin\frac{\sqrt{3}}{2}x\right), \end{aligned} \tag{47}$$

where C_1', C_2', C_3' are arbitrary constants. (Of course, we don't really need the primes in the final answer.) ∎

EXAMPLE 7. The equation

$$y^{(v)} - 7y''' + 12y' = 0 \tag{48}$$

has the characteristic equation $\lambda^5 - 7\lambda^3 + 12\lambda = 0$ or, $\lambda(\lambda^4 - 7\lambda^2 + 12) = 0$. The λ factor gives the root $\lambda = 0$. The quartic factor is actually a quadratic in λ^2, so the quadratic equation gives $\lambda^2 = 4$ and $\lambda^2 = 3$. Thus, λ equals $0, \pm 2, \pm\sqrt{3}$, so

$$y(x) = C_1 + C_2 e^{2x} + C_3 e^{-2x} + C_4 e^{\sqrt{3}x} + C_5 e^{-\sqrt{3}x} \tag{49}$$

is a general solution of (48). ∎

EXAMPLE 8. The equation

$$y^{(iv)} + ky = 0 \tag{50}$$

arises in studying the deflected shape $y(x)$ of a beam on an elastic foundation, where k is a known positive physical constant. Since the characteristic equation $\lambda^4 + k = 0$ gives $\lambda^4 = -k$, to find λ we need to evaluate $(-k)^{1/4}$. The general result is that $z^{1/n}$, for any complex number $z = a + ib$ and any integer $n > 1$, has n values in the complex plane. These values are equally spaced on a circle of radius $r = \sqrt{a^2 + b^2}$ centered at the origin of the complex plane, as is explained in Section 22.4. For our present purpose, let it suffice to merely give the result: $\lambda = (-k)^{1/4} = \pm k^{1/4}\dfrac{1+i}{\sqrt{2}}$ and $\pm k^{1/4}\dfrac{1-i}{\sqrt{2}}$, so

$$\begin{aligned} y(x) &= C_1 e^{k^{1/4}(1+i)x/\sqrt{2}} + C_2 e^{k^{1/4}(1-i)x/\sqrt{2}} \\ &\quad + C_3 e^{k^{1/4}(-1+i)x/\sqrt{2}} + C_4 e^{k^{1/4}(-1-i)x/\sqrt{2}} \\ &= e^{k^{1/4}x/\sqrt{2}}\left(C_1' \cos\frac{k^{1/4}}{\sqrt{2}}x + C_2' \sin\frac{k^{1/4}}{\sqrt{2}}x\right) \\ &\quad + e^{-k^{1/4}x/\sqrt{2}}\left(C_3' \cos\frac{k^{1/4}}{\sqrt{2}}x + C_4' \sin\frac{k^{1/4}}{\sqrt{2}}x\right) \end{aligned} \tag{51}$$

is a general solution of (50). ∎

3.4.4. Repeated roots. Thus far we have considered only the generic case, where the nth-order characteristic equation (40) admits n distinct roots $\lambda_1, \ldots, \lambda_n$. To complete our discussion, we need to consider the case where one or more of the roots is repeated. We say that a root λ_j of (40) is **repeated** if (40) contains the factor $\lambda - \lambda_j$ more than once. More specifically, we say that λ_j is a **root of order k** if (40) contains the factor $\lambda - \lambda_j$ k times. For instance, if the characterisitic equation for some given sixth-order equation can be factored as $(\lambda + 2)(\lambda - 5)^3(\lambda - 1)^2 = 0$, then the roots $\lambda = 5$ and $\lambda = 1$ are repeated; $\lambda = 5$ is a root of order 3 and $\lambda = 1$ is a root of order 2. We can say that

$$y(x) = C_1 e^{-2x} + C_2 e^{5x} + C_3 e^{x}$$

is a solution for any constants C_1, C_2, C_3, but the latter falls short of being a *general* solution of the sixth-order differential equation since it is not a linear combination of six LI solutions. The problem, in such a case of repeated roots, is how to find the missing solutions. Evidently, they will not be of the form $e^{\lambda x}$, for if they were then we would have found them when we sought $y(x)$ in that form.

We will use a simple example to show how to obtain such "missing solutions," and will then state the general result as a theorem.

EXAMPLE 9. *Reduction of Order.* The equation

$$y'' + 2y' + y = 0 \tag{52}$$

has the characteristic equation $\lambda^2 + 2\lambda + 1 = (\lambda + 1)^2 = 0$, so $\lambda = -1$ is a root of order 2. Thus, we have the solution Ae^{-x} but are missing a second linearly independent solution, which is needed if we are to obtain a general solution of (52).

To find the missing solution, we use Lagrange's method of **reduction of order**, which works as follows. Suppose that we know one solution, say $y_1(x)$, of a given linear homogeneous differential equation, and we seek one or more other linearly independent solutions. If $y_1(x)$ is a solution then, of course, so is $Ay_1(x)$, where A is an arbitrary constant. According to the method of reduction of order, we let A vary and seek $y(x)$ in the form $y(x) = A(x)y_1(x)$. Putting that form into the given differential equation results in another differential equation on the unknown $A(x)$, but that equation inevitably will be simpler than the original differential equation on y, as we shall see.

In the present example, $y_1(x)$ is e^{-x}, so to find the missing solution we seek

$$y(x) = A(x)e^{-x}. \tag{53}$$

From (53), $y' = (A' - A)e^{-x}$ and $y'' = (A'' - 2A' + A)e^{-x}$, and putting these expressions into (52) gives

$$(A'' - 2A' + A + 2A' - 2A + A)\, e^{-x} = 0, \tag{54}$$

so that $A(x)$ must satisfy the second-order differential equation obtained by equating the coefficient of e^{-x} in (54) to zero, namely, $A'' - 2A' + A + 2A' - 2A + A = 0$. The

cancellation of the three A terms in that equation is not a coincidence, for if $A(x)$ were a constant [in which case the A' and A'' terms in (54) would drop out] then the terms on the left-hand side of (54) would have to cancel to zero because Ae^{-x} is a solution of the original homogeneous differential equation if A is a constant. Thanks to that (inevitable) cancellation, the differential equation governing $A(x)$ will be of the form

$$A'' + \alpha A' = 0, \tag{55}$$

for some constant α, and this second-order equation can be reduced to the first-order equation $v' + \alpha v = 0$ by setting $A' = v$; hence the name reduction of order for the method. In fact, not only do the A terms cancel, as they must, the A' terms happen to cancel as well, so in place of (55) we have the even simpler equation

$$A'' = 0 \tag{56}$$

on $A(x)$. Integration gives $A(x) = C_1 + C_2 x$, so that (53) becomes

$$y(x) = C_1 e^{-x} + C_2 x e^{-x}. \tag{57}$$

The $C_1 e^{-x}$ term merely reproduces that which was already known (recall the second sentence of this example), and the $C_2 x e^{-x}$ term is the desired missing solution. Since the two are LI, (57) is a general solution of (52). ∎

Similarly, suppose we have an eighth-order equation, the characteristic equation of which can be factored as $(\lambda - 2)^3(\lambda + 1)^4(\lambda + 5)$, say, so that 2 is a root of order 3 and -1 is a root of order 4. If we take the solution Ae^{2x} associated with the root $\lambda = 2$, and apply reduction of order by seeking y in the form $A(x)e^{2x}$, then we obtain $A''' = 0$ and $A(x) = C_1 + C_2 x + C_3 x^2$ and hence the "string" of solutions $C_1 e^{2x}, C_2 x e^{2x}, C_3 x^2 e^{2x}$ coming from the root $\lambda = 2$. Likewise, if we take the solution Ae^{-x} associated with the root $\lambda = -1$, and apply reduction of order, we obtain $A(x) = C_4 + C_5 x + C_6 x^2 + C_7 x^3$ and hence the string of solutions $C_4 e^{-x}, C_5 x e^{-x}, C_6 x^2 e^{-x}, C_7 x^3 e^{-x}$ coming from the root $\lambda = -1$, so that we have a general solution

$$\begin{aligned} y(x) = &\left(C_1 + C_2 x + C_3 x^2\right) e^{2x} \\ &+ \left(C_4 + C_5 x + C_6 x^2 + C_7 x^3\right) e^{-x} + C_8 e^{-5x} \end{aligned} \tag{58}$$

of the original differential equation. [To verify that this is indeed a general solution one would need to show that the eight solutions contained within (58) are LI, as could be done by working out the Wronskian W and showing that $W \neq 0$.]

EXAMPLE 10. For

$$y'''' - y'' = 0 \tag{59}$$

the characteristic equation $\lambda^4 - \lambda^2 = 0$ gives $\lambda = 0, 0, 1, -1$ and hence the solution $y(x) = A + Be^x + Ce^{-x}$. The latter falls short of being a general solution of (59) because the repeated root $\lambda = 0$ gave the single solution A. To find the missing solution by reduction

of order we could vary the parameter A and seek $y(x) = A(x)$, but surely there can be no gain in that step since it merely amounts to a name change, from $y(x)$ to $A(x)$. This situation will always occcur when the repeated root is zero, but in that case we can achieve a reduction of order more directly. In the case of (59) we can set $y'' = p$. Then the fourth-order equation (59) is reduced to the second-order equation $p'' - p = 0$, so

$$p(x) = Ae^{x} + Be^{-x}.$$

But $y'' = p$, so

$$y'(x) = \int p(x)\,dx = Ae^{x} - Be^{-x} + C.$$

Hence

$$y(x) = \int \left(Ae^{x} - Be^{-x} + C\right) dx = Ae^{x} + Be^{-x} + Cx + D$$

is the general solution of (59). Observe that the pattern is the same: the repeated root $\lambda = 0$ gives the solution $(C_1 + C_2 x)e^{0x}$, where C_1 is D and C_2 is C. ∎

We organize these results as the following theorem.

THEOREM 3.4.2 *Repeated Roots of Characteristic Equation*
If λ_1 is a root of order k, of the characteristic equation (40), then $e^{\lambda_1 x}, xe^{\lambda_1 x}, \ldots, x^{k-1}e^{\lambda_1 x}$ are k LI solutions of the differential equation (1).

Proof: Denote (1) in operator form as $L[y] = 0$, where

$$L = \frac{d^{(n)}}{dx^{(n)}} + a_1 \frac{d^{(n-1)}}{dx^{(n-1)}} + \cdots + a_{n-1}\frac{d}{dx} + a_n. \tag{60}$$

Then

$$L\left[e^{\lambda x}\right] = \left(\lambda^n + a_1\lambda^{n-1} + \cdots + a_{n-1}\lambda + a_n\right) e^{\lambda x},$$

or

$$L\left[e^{\lambda x}\right] = (\lambda - \lambda_1)^k\, p(\lambda) e^{\lambda x}, \tag{61}$$

where $p(\lambda)$ is a polynomial in λ, of degree $n - k$. Since (61) holds for all λ, we can set $\lambda = \lambda_1$ in that formula. Doing so, the right-hand side of (61) vanishes, so that $L\left[e^{\lambda_1 x}\right] = 0$ and hence $e^{\lambda_1 x}$ is a solution of $L[y] = 0$. Our object, now, is to show that $xe^{\lambda_1 x}, \ldots, x^{k-1}e^{\lambda_1 x}$ are solutions as well.

To proceed, differentiate (61) with respect to λ (λ, not x):

$$\frac{d}{d\lambda} L\left[e^{\lambda x}\right] = k\,(\lambda - \lambda_1)^{k-1}\, p(\lambda) e^{\lambda x} + (\lambda - \lambda_1)^k \frac{d}{d\lambda}\left(p(\lambda) e^{\lambda x}\right). \tag{62}$$

The left-hand side of (62) calls for $e^{\lambda x}$ to be differentiated first with respect to x, according to the operator L defined in (60) and then with respect to λ. Since we can

interchange the order of these differentiations, we can express the left-hand side as $L\left[\frac{d}{d\lambda}e^{\lambda x}\right]$, that is, as $L\left[xe^{\lambda x}\right]$. Thus, one differentiation of (61) with respect to λ gives

$$L\left[xe^{\lambda x}\right] = k\left(\lambda-\lambda_1\right)^{k-1}p(\lambda)e^{\lambda x} + \left(\lambda-\lambda_1\right)^k \frac{d}{d\lambda}\left(p(\lambda)e^{\lambda x}\right). \tag{63}$$

Setting $\lambda = \lambda_1$ in (63) gives $L\left[xe^{\lambda_1 x}\right] = 0$. Hence, not only is $e^{\lambda_1 x}$ a solution, so is $xe^{\lambda_1 x}$. Repeated differentiation with respect to λ reveals that $x^2e^{\lambda_1 x}, \ldots, x^{k-1}e^{\lambda_1 x}$ are solutions as well, as was to be proved.

That's as far as we can go because at that point one more differentiation would give a leading term of $k!\left(\lambda-\lambda_1\right)^0 p(\lambda_1)e^{\lambda_1 x}$ plus terms with factors of $\left(\lambda-\lambda_1\right)^1, \left(\lambda-\lambda_1\right)^2, \ldots, \left(\lambda-\lambda_1\right)^k$ on the right-hand side. The latter terms vanish for $\lambda = \lambda_1$, but the leading term does not because $p(\lambda_1) \neq 0$ (because $\lambda - \lambda_1$ is not among the factors of p) and $e^{\lambda_1 x} \neq 0$.

Verification that the solutions $e^{\lambda_1 x}, xe^{\lambda_1 x}, \ldots, x^{k-1}e^{\lambda_1 x}$ are LI is left for the exercises. ■

EXAMPLE 11. As a final example, consider the equation

$$y^{(iv)} - 8y''' + 26y'' - 40y' + 25y = 0 \tag{64}$$

with characteristic equation $\lambda^4 - 8\lambda^3 + 26\lambda^2 - 40\lambda + 25 = 0$ and repeated complex roots $\lambda = 2+i,\ 2+i,\ 2-i,\ 2-i$. It follows that

$$\begin{aligned} y(x) &= \left(C_1 + C_2x\right)e^{(2+i)x} + \left(C_3 + C_4x\right)e^{(2-i)x} \\ &= e^{2x}\left[\left(C_1e^{ix} + C_3e^{-ix}\right) + x\left(C_2e^{ix} + C_4e^{-ix}\right)\right] \\ &= e^{2x}\left[\left(A\cos x + B\sin x\right) + x\left(C\cos x + D\sin x\right)\right] \end{aligned}$$

is a general solution of (64). ■

3.4.5. Stability. An important consideration in applications, especially feedback control systems, is whether or not a system is "stable." Normally, stability has to do with the behavior of a system over time, so let us change the name of the independent variable from x to t in (1):

$$\frac{d^ny}{dt^n} + a_1\frac{d^{n-1}y}{dt^{n-1}} + \cdots + a_{n-1}\frac{dy}{dt} + a_ny = 0, \tag{65}$$

and let us denote the general solution of (65) as $y(t) = C_1y_1(t) + \cdots + C_ny_n(t)$. We say that the system described by (65) (be it mechanical, electrical, economic, or whatever) is **stable** if all of its solutions are bounded – that is, if there exists a constant M_j for each solution $y_j(t)$ such that $|y_j(t)| < M_j$ for all $t > 0$. If the system is not stable, then it is **unstable**.

THEOREM 3.4.3 *Stability*
For the system described by (65) to be stable, it is necessary and sufficient that the characteristic equation of (65) have no roots to the right of the imaginary axis in the complex plane and that any roots on the imaginary axis be nonrepeated.

Proof: Let $\lambda = a + ib$ be any nonrepeated root of the characteristic equation; we call a *the real part of* λ and write $\mathrm{Re}\lambda = a$, and b *the imaginary part of* λ and write $\mathrm{Im}\lambda = b$. Such a root will contribute a solution $e^{(a+ib)t} = e^{at}(\cos bt + i \sin bt)$. Since the magnitude (modulus, to be more precise) of a complex number $x + iy$ is defined as $|x + iy| = \sqrt{x^2 + y^2}$, and the magnitude of the product of complex numbers is the product of their magnitudes, we see that $\left|e^{(a+ib)t}\right| = \left|e^{at}(\cos bt + i \sin bt)\right| = \left|e^{at}\right| |\cos bt + i \sin bt| = e^{at}\sqrt{\cos^2 bt + \sin^2 bt} = e^{at}$ so that solution will be bounded if and only if $a \leq 0$, that is, if λ does not lie to the right of the imaginary axis.

Next, let $\lambda = a + ib$ be a repeated root of order k, with $a \neq 0$. Such a root will contribute solutions of the form $t^p e^{(a+ib)t} = t^p e^{at}(\cos bt + i \sin bt)$, for $p = 0, \ldots, k-1$, with magnitude $t^p e^{at}$. Surely the latter grows unboundedly if $a > 0$ because both factors do, but its behavior is less obvious if $a < 0$ since then the t^p factor grows and the e^{at} decays. To see which one "wins," one can rewrite the product as t^p/e^{-at} and then apply l'Hôpital's rule p times. Doing so, one finds that the ratio tends to zero as $t \to \infty$. [Recall that l'Hôpital's rule applies to indeterminate forms of the type $0/0$ or ∞/∞, not $(\infty)(0)$; that is why we first rewrite $t^p e^{at}$ in the form t^p/e^{-at}.] The upshot is that such solutions are bounded if $\lambda = a + ib$ lies in the left half plane ($a < 0$), and unbounded if it lies in the right half plane ($a > 0$). If λ lies *on* the imaginary axis ($a = 0$), then $\left|t^p e^{(a+ib)t}\right| = \left|t^p e^{ibt}\right| = |t^p(\cos bt + i \sin bt)| = t^p$, which grows unboundedly.

Our conclusion is that all solutions are bounded if and only if no roots lie to the right of the imaginary axis and no repeated roots lie on the imaginary axis, as was to be proved. ■

One is often interested in being able to determine whether the system is stable or not without actually evaluating the n roots of the characteristic equation (40). There are theorems that provide information about stability based directly upon the a_j coefficients in (40). One such theorem is stated below. Another, the Routh–Hurwitz criterion, is given in the exercises to Section 10.4.

THEOREM 3.4.4 *Coefficients of Mixed Sign*
If the coefficients in (40) are real and of mixed sign (there is at least one positive and at least one negative), then there is at least one root λ with $\mathrm{Re}\lambda > 0$, so the system is unstable.

These theorems are not as important as they were before the availability of computer software that can determine the roots of (40) numerically and with great ease. For instance, using *Maple*, one can obtain all roots of the equation

$$x^5 + 3x^4 - 2x^3 + x^2 + x + 5 = 0$$

simply by using the **fsolve** command. Enter

$$\text{fsolve}(x\hat{}5 + 3 * x\hat{}4 - 2 * x\hat{}3 + x\hat{}2 + x + 5 = 0,\ x,\ \text{complex});$$

and return. This gives the following printout of the five solutions:

$$-3.6339286,\ -.58045036 - .79731249I,\ -.58045036 + .79731249I,$$
$$.89741468 - .78056850I,\ .89741468 + .78056850I$$

In this example, observe that there are, indeed, roots with positive real parts, as predicted by Theorem 3.4.4, so the system is unstable.

For equations of fourth degree or lower, such software works even if one or more of the coefficients are unspecified, in which case the roots are given in terms of those parameters.

Closure. In this section we limited our attention to linear homogeneous differential equations with constant coefficients, a case of great importance in applications. Seeking solutions in exponential form, we found the characteristic equation to be central. According to the fundamental theorem of algebra, such equations always have at least one root, so we are guaranteed of finding at least one exponential solution of the differential equation. If the n roots $\lambda_1, \ldots, \lambda_n$ are distinct, then each root λ_j contributes a solution $e^{\lambda_j x}$, and their superposition gives a general solution of (1) in the form

$$y(x) = C_1 e^{\lambda_1 x} + \cdots + C_n e^{\lambda_n x}. \tag{66}$$

If any root λ_j is repeated, say of order k, then it contributes not only the solution $e^{\lambda_j x}$, but the k LI solutions $e^{\lambda_j x}, xe^{\lambda_j x}, \ldots, x^{k-1}e^{\lambda_j x}$ to the general solution. Thus, in the generic case of distinct roots, the general solution of (1) is of the form (66); in the nongeneric case of repeated roots, the solution also contains one or more terms which are powers of x times exponentials.

It should be striking how simple is the solution process for linear constant-coefficient homogeneous equations, with the only difficulty being algebraic – the need to find the roots of the characteristic equation. The reason for this simplicity is that most of the work was done simply in deciding to look for solutions in the right place, within the set of exponential functions. Also, observe that although in a fundamental sense the solving of a differential equation in some way involves integration, the methods discussed in this section required no integrations, in contrast to most of the methods of solution of first-order equations in Chapter 1.

In the final (optional) section we introduced the concept of stability, and in Theorem 3.4.3 we related the stability of the physical system to the placement of the roots of the characteristic equation in the complex plane.

Computer software. To obtain a general solution of $y''' - 9y' = 0$ using *Maple*, use the command

$$\text{dsolve}(\{\text{diff}(y(x), x, x, x) - 9 * \text{diff}(y(x), x) = 0\}, y(x));$$

and to solve the ODE subject to the initial conditions $y(0) = 5,\ y'(0) = 2,\ y''(0) = -4$, use the command

$$\text{dsolve}(\{\text{diff}(y(x), x, x, x) - 9 * \text{diff}(y(x), x) = 0,\ y(0) = 5,\ D(y)(0) = 2,\ D(D(y))(0) = -4\}, y(x));$$

In place of diff$(y(x), x, x, x)$ we could use diff$(y(x), x\$3)$, for brevity.

EXERCISES 3.4

1. Use whichever of equations (5)–(8) are needed to derive these relations between the circular and hyperbolic functions:

(a) $\cos(ix) = \cosh x$ (b) $\sin(ix) = i \sinh x$
(c) $\cosh(ix) = \cos x$ (d) $\sinh(ix) = i \sin x$

2. Use equations (6) and/or (7) to derive or verify

(a) equation (9) (b) equation (10)
(c) equation (13a) (d) equation (13b)
(e) equation (13c) (f) equation (13d)

3. Theorem 3.4.2 states that $e^{\lambda_1 x}, xe^{\lambda_1 x}, \ldots, x^{k-1}e^{\lambda_1 x}$ are LI. Prove that claim.

4. (*Nonrepeated roots*) Find a general solution of each of the following equations, and a particular solution satisfying the given conditions, if such conditions are given.

(a) $y'' + 5y' = 0$
(b) $y'' - y' = 0$
(c) $y'' + y' = 0;\quad y(0) = 3,\quad y'(0) = 0$
(d) $y'' - 3y' + 2y = 0;\quad y(1) = 1,\quad y'(1) = 0$
(e) $y'' - 4y' - 5y = 0;\quad y(1) = 1,\ y'(1) = 0$
(f) $y'' + y' - 12y = 0;\quad y(-1) = 2,\quad y'(-1) = 5$
(g) $y'' - 4y' + 5y = 0;\quad y(0) = 2,\quad y'(0) = 5$
(h) $y'' - 2y' + 3y = 0;\quad y(0) = 4,\quad y'(0) = -1$
(i) $y'' - 2y' + 2y = 0;\quad y(0) = 0,\quad y'(0) = -5$
(j) $y'' + 2y' + 3y = 0;\quad y(0) = 0,\quad y'(0) = 3$
(k) $y''' + 3y' - 4y = 0;\quad y(0) = 0,\quad y'(0) = 0,\quad y''(0) = 6$
(l) $y''' - y'' + 2y' = 0;\quad y(0) = 1,\quad y'(0) = 0,\quad y''(0) = 0$
(m) $y''' + y'' - 2y = 0$
(n) $y^{(iv)} - y = 0$
(o) $y^{(iv)} - 2y'' - 3y = 0$
(p) $y^{(iv)} + 6y'' + 8y = 0$
(q) $y^{(iv)} + 7y'' + 12y = 0$
(r) $y^{(iv)} - 2y''' - y'' + 2y' = 0$

5. (a)–(r) Solve the corresponding problem in Exercise 4 using computer software.

6. (*Repeated roots*) Find a general solution of each of the following equations, and a particular solution satisfying the given conditions, if such conditions are given.

(a) $y'' = 0;\quad y(-3) = 5,\quad y'(-3) = -1$
(b) $y'' + 6y' + 9y = 0;\quad y(1) = e,\quad y'(1) = -2$
(c) $y''' = 0;\quad y(0) = 3,\quad y'(0) = -5,\quad y''(0) = 1$
(d) $y''' + 5y'' = 0;\quad y(0) = 1,\quad y'(0) = 0,\quad y''(0) = 0$
(e) $y''' + 3y'' + 3y' + y = 0$
(f) $y''' - 3y'' + 3y' - y = 0$
(g) $y''' - y'' - y' + y = 0$
(h) $y^{(iv)} + 3y''' = 0$
(i) $y^{(iv)} + y''' + y'' = 0$
(j) $y^{(iv)} + 8y'' + 16y = 0$
(k) $y^{(vi)} = 0;\quad y(0) = y'(0) = y''(0) = y'''(0) = y^{(iv)}(0) = 0,\quad y^{(v)}(0) = 3$

7. (a)–(k) Solve the corresponding problem in Exercise 6 using computer software.

8. If the roots of the characteristic equation are as follows, then find the original differential equation and also a general solution of it:

(a) $2, 6$ (b) $2i, -2i$
(c) $4 - 2i, 4 + 2i$ (d) $-2, 3, 5$

(e) $2, 3, -1$ (f) $1, 1, -2$
(g) $4, 4, 4, i, -i$ (h) $1, -1, 2+i, 2-i$
(i) $0, 0, 0, 0, 7, 9$ (j) $1+i, 1+i, 1-i, 1-i$

9. (*Complex a_j's*) Find a general solution of each of the following equations. NOTE: Normally, the a_j coefficients in (1) are real, but the results of this section hold even if they are not (except for Theorem 3.4.4, which explicitly requires that the coefficients be real). However, be aware that if the a_j coefficients are not all real, then complex roots do not necessarily occur in complex conjugate pairs. For instance, $\lambda^2 + 2i\lambda + 1 = 0$ has the roots $\lambda = (\sqrt{2}-1)i, -(\sqrt{2}+1)i$.

(a) $y'' - 2iy' + y = 0$ (b) $y'' - 3iy' - 2y = 0$
(c) $y'' + iy' - y = 0$ (d) $y'' - 2iy' - y = 0$

(e) $y'' - iy = 0$ HINT: Verify, and use, the fact that $\sqrt{i} = \pm(1+i)/\sqrt{2}$.
(f) $y''' + 4iy'' - y' = 0$
(g) $y''' + iy' = 0$ HINT: Verify, and use, the fact that $\sqrt{-i} = \pm(1-i)/\sqrt{2}$
(h) $y''' - (1+2i)y'' + (1+i)y' - 2(1+i)y = 0$ HINT: One root is found, by inspection, to be $\lambda = -i$.

10. (a)–(h) Solve the corresponding problem in Exercise 9 using computer software.

11. (*Solution by factorization of the operator*) We motivated the idea of seeking solutions to (1) in the form $e^{\lambda x}$ by observing that the general solution of the first-order equation $y' + a_1 y = 0$ is an exponential, $Ce^{-a_1 x}$, and wondering if higher-order equations might admit exponential solutions too. A more compelling approach is as follows. Having already seen that the first-order equation admits an exponential solution, consider the second-order equation (18).

(a) Show that (18) can be written, equivalently, as

$$(D - \lambda_1)(D - \lambda_2)y = 0, \tag{11.1}$$

where D denotes d/dx, and λ_1 and λ_2 are the two roots of $\lambda^2 + a_1\lambda + a_2 = 0$. NOTE: In (11.1) we accomplish a **factorization** of the original differential operator $L = D^2 + a_1 D + a_2$ as $(D-\lambda_1)(D-\lambda_2)$. By the left-hand side of (11.1), we mean $(D - \lambda_1)((D - \lambda_2)y)$. That is, first let the operator to the left of y (namely, $D - \lambda_2$) act on y, then let the operator to the left of $(D - \lambda_2)y$ (namely, $D - \lambda_1$) act on that.
(b) To solve (11.1), let $(D - \lambda_2)y = u$, so that (18) reduces to the first-order equation

$$\frac{du}{dx} - \lambda_1 u = 0. \tag{11.2}$$

Solve (11.2) for u, put that u on the right-hand side of $\frac{dy}{dx} - \lambda_2 y = u$, which is again of first order, and solve the latter for y. Show that if λ_1, λ_2 are distinct, then the result is given by (23), whereas if they are repeated, then the result is $y(x) = (C_1 + C_2 x)e^{\lambda_1 x}$.
(c) Solve $y'' - 3y' + 2y = 0$ by factoring the operator as $(D-1)(D-2)y = 0$. Solve the latter by the method outlined in (b): Setting $(D-2)y \equiv u$, solve $(D-1)u = u' - u = 0$ for $u(x)$. Then, knowing $u(x)$, solve $(D-2)y = u$, namely, $y' - 2y = u(x)$, for $y(x)$.
(d) Same as (c), for $y'' - 4y = 0$.
(e) Same as (c), for $y'' + 4y' + 3y = 0$.
(f) Same as (c), for $y'' + 2y' + y = 0$.
(g) Same as (c), for $y'' + 4y' + 4y = 0$.
NOTE: Similarly for higher-order equations. For instance, $y''' - 2y'' - y' + 2y = (D-2)(D+1)(D-1)y = 0$ can be solved by setting $(D+1)(D-1)y \equiv u$ and solving $(D-2)u = 0$ for $u(x)$; then set $(D-1)y \equiv v$ and solve $(D+1)v = u$ for $v(x)$; finally, solve $(D-1)y = v$ for $y(x)$. The upshot is that the solution of an nth-order linear homogeneous differential equation with constant coefficients can be reduced to the solution of a sequence of n first-order linear equations.

12. Use computer software to obtain the roots of the given characteristic equation, state whether the system is stable or unstable, and explain why. If Theorem 3.4.4 applies, then show that your results are consistent with the predictions of that theorem.

(a) $\lambda^3 - 3\lambda^2 + 26\lambda - 2 = 0$
(b) $\lambda^3 + 3\lambda^2 + 2\lambda + 2 = 0$
(c) $\lambda^4 + \lambda^3 + 3\lambda^2 + 2\lambda + 2 = 0$
(d) $\lambda^4 + \lambda^3 + 5\lambda^2 + \lambda + 4 = 0$
(e) $\lambda^6 + \lambda^5 + 5\lambda^4 + 2\lambda^3 - \lambda^2 + \lambda + 3 = 0$
(f) $\lambda^6 + 9\lambda^5 + 5\lambda^4 + 2\lambda^3 + 7\lambda^2 + \lambda + 3 = 0$
(g) $\lambda^6 + \lambda^5 + 5\lambda^4 + 4\lambda^3 + 4\lambda^2 + 8\lambda + 4 = 0$
(h) $\lambda^6 + \lambda^5 + 5\lambda^4 + 2\lambda^3 + 7\lambda^2 + \lambda + 3 = 0$
(i) $\lambda^8 - \lambda^6 + \lambda^5 + 5\lambda^4 + 2\lambda^3 + 7\lambda^2 + \lambda + 3 = 0$
(j) $\lambda^8 + \lambda^7 + \lambda^6 + \lambda^5 + 5\lambda^4 + 21\lambda^3 + 7\lambda^2 + \lambda + 3 = 0$

3.5 Application to Harmonic Oscillator: Free Oscillation

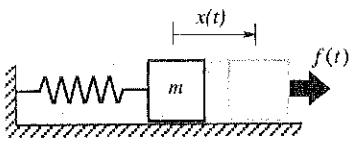


Figure 1. Mechanical oscillator.

In Section 1.3 we discussed the modeling of the mechanical oscillator reproduced here in Fig. 1. Neglecting air resistance, the block of mass m is subjected to a restoring force due to the spring, a "drag" force due to the friction between the block and the lubricated table top, and an applied force $f(t)$. (By a restoring force, we mean that the force opposes the stretch or compression in the spring.) Most of that discussion focused on the modeling of the spring force and friction force, and we derived the approximate equation of motion

$$mx'' + cx' + kx = f(t), \tag{1}$$

where c is the *damping coefficient*, k is the *spring stiffness*. Besides the differential equation, let us regard the initial displacement and initial velocity,

$$x(0) = x_0 \quad \text{and} \quad x'(0) = x_0', \tag{2}$$

respectively, as specified values.

In this section we consider the solution for the case where $f(t) = 0$:

$$mx'' + cx' + kx = 0. \tag{3}$$

This is the so-called *unforced*, or *free*, oscillation. According to Theorem 3.3.1, the solution $x(t)$ to (3) and (2) does exist and is unique. To find it, we seek $x(t) = e^{\lambda t}$ and obtain the characteristic equation $m\lambda^2 + c\lambda + k = 0$, with roots

$$\lambda = \frac{-c \pm \sqrt{c^2 - 4mk}}{2m}. \tag{4}$$

Consider first the case where there is no damping, so $c = 0$ and (3) becomes

$$\boxed{mx'' + kx = 0.} \tag{5}$$

That is, the friction is small enough so that it can be neglected altogether. Then (4) gives $\lambda = \pm i\sqrt{k/m}$, and the solution of (5) is

$$\boxed{x(t) = Ae^{i\omega t} + Be^{-i\omega t},} \tag{6}$$

where $\omega = \sqrt{k/m}$ is the so-called **natural frequency** of the system, in rad/sec. Or, equivalent to (6) and favored in this text,

$$\boxed{x(t) = C\cos\omega t + D\sin\omega t.} \tag{7}$$

In fact, there is another useful form of the general solution, namely,

$$\boxed{x(t) = E\sin(\omega t + \phi),} \tag{8}$$

where the integration constants E and ϕ can be determined in terms of C and D as follows. To establish the equivalence of (8) and (7), recall the trigonometric identity $\sin(A+B) = \sin B \cos A + \sin A \cos B$. Then

$$E \sin(\omega t + \phi) = E \sin\phi \cos\omega t + E \cos\phi \sin\omega t,$$

which is identical to $C \cos\omega t + D \sin\omega t$ if

$$C = E \sin\phi \quad \text{and} \quad D = E \cos\phi. \tag{9a,b}$$

Squaring and adding equations (9), and also dividing one by the other, gives

$$\boxed{E = \sqrt{C^2 + D^2} \quad \text{and} \quad \phi = \tan^{-1}\frac{C}{D},} \tag{10a,b}$$

respectively, as the connection between the equivalent forms (7) and (8). It will be important to be completely comfortable with the equivalence between the solution forms (6), (7), and (8). Both the square root and the $\tan^{-1}$ in (10) are multi-valued. We will understand the square root as the positive one and the $\tan^{-1}$ to lie between $-\pi$ and π. Specifically, it follows from (9), with $E > 0$, that if $C > 0$ and $D > 0$ then $0 < \phi < \pi/2$, if $C > 0$ and $D < 0$ then $\pi/2 < \phi < \pi$, if $C < 0$ and $D > 0$ then $-\pi/2 < \phi < 0$, and if $C < 0$ and $D < 0$ then $-\pi < \phi < -\pi/2$.

For instance, consider $6\cos t - 2\sin t$. Then $E = \sqrt{36+4} = \sqrt{40}$ and $\phi = \tan^{-1}\left(\frac{+6}{-2}\right)$. A calculator or typical computer software will interpret $\tan^{-1}(\ )$ as $-\pi/2 < \tan^{-1}(\ ) < \pi/2$, namely, in the first or fourth quadrant. Not able to distinguish $(+6)/(-2)$ from $(-6)/(+2)$, it will give $\tan^{-1}\left(-\frac{6}{2}\right) = -1.25$ rad, which is incorrect. The correct value is in the second quadrant, namely, $\phi = \pi - 1.25 = 1.89$ rad. Thus, $6\cos t - 2\sin t = \sqrt{40}\sin(t + 1.89)$.

Whereas C and D in (7) have no special physical significance, E and ϕ in (8) are the **amplitude** and **phase angle** of the vibration, respectively (Fig. 2a).

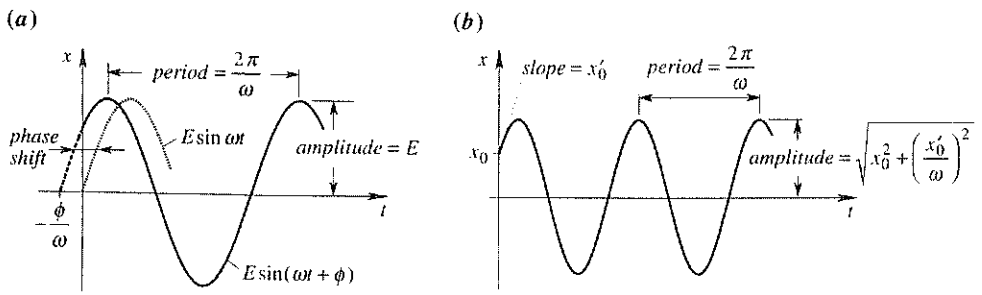


Figure 2. (a) Graphical significance of ω, ϕ. (b) Undamped free oscillation.

Although (8) is advantageous conceptually, in that the amplitude E and phase angle ϕ are physically and graphically meaningful, it is a bit easier to apply the initial conditions to (7):

$$x(0) = x_0 = C, \quad x'(0) = x'_0 = \omega D$$

so $C = x_0, D = x_0'/\omega$, and the solution is

$$x(t) = x_0 \cos \omega t + \frac{x_0'}{\omega} \sin \omega t, \tag{11}$$

a plot of which is shown in Fig. 2b for representative initial conditions x_0 and x_0'.

Before continuing, consider the relationship between the mathematics and the physics. For example, the frequency $\omega = \sqrt{k/m}$ increases with k, decreases with m, and is independent of the initial conditions x_0 and x_0', and hence the amplitude which, according to (11) and (10), is $\sqrt{x_0^2 + (x_0'/\omega)^2}$. Do these results make sense? Probably the increase of ω with k fits with our experience with springs, and its decrease with m makes sense as well. However, one might well expect the frequency to vary with the amplitude of the vibration. We will come back to this point in Chapter 7, where we consider more realistic *non*linear models.

Now suppose there is some damping, $c > 0$. From (4) we see that there are three cases of interest. If we define the **critical damping** as $c_{cr} = \sqrt{4mk} = 2\sqrt{mk}$, then the solution is qualitatively different depending upon whether $c < c_{cr}$ (the "underdamped" case), $c = c_{cr}$ (the "critically damped" case), or $c > c_{cr}$ (the "overdamped" case).

Underdamped vibration ($c < c_{cr}$). In this case (4) gives two complex conjugate roots

$$\begin{aligned}\lambda &= \frac{1}{2m}\left(-c \pm \sqrt{c^2 - c_{cr}^2}\right) = \frac{1}{2m}\left(-c \pm i\sqrt{c_{cr}^2 - c^2}\right) \\ &= -\frac{c}{2m} \pm i\sqrt{\omega^2 - \left(\frac{c}{2m}\right)^2}\end{aligned}$$

so a general solution of (1) is

$$x(t) = e^{-\frac{c}{2m}t}\left[A \cos \sqrt{\omega^2 - \left(\frac{c}{2m}\right)^2}\, t + B \sin \sqrt{\omega^2 - \left(\frac{c}{2m}\right)^2}\, t\right], \tag{12}$$

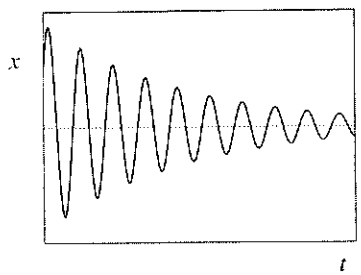


Figure 3. Underdamped free oscillation.

where A and B can be determined from the initial conditions (2). Of course, we could express the bracketed part in the form (8) if we like.

Comparing (7) and (12), observe that the damping has two effects. First, it introduces the $e^{-(c/2m)t}$ factor, which causes the oscillation to "damp out" as $t \to \infty$, as illustrated in Fig. 3. That is, the amplitude tends to zero as $t \to \infty$. Second, it reduces the frequency from the natural frequency ω to $\sqrt{\omega^2 - (c/2m)^2}$; that is, it makes the system more sluggish, as seems intuitively reasonable. (It might appear from Fig. 2b and 3 that the damping *in*creases the frequency, but that appearance is only because we have compressed the t scale in Fig. 3.)

Critically damped vibration ($c = c_{cr}$). As c is increased further, the system

becomes so sluggish that when c attains the critical value c_{cr} the oscillation ceases altogether. In this case (4) gives the repeated root $\lambda = -c/2m$, of order two, so

$$x(t) = (A + Bt)e^{-\frac{c}{2m}t}. \tag{13}$$

Although the t in $A+Bt$ grows unboundedly, the exponential function decays more powerfully (as discussed within the proof of Theorem 3.4.3) and the solution (13) decays without oscillation, as shown in the $c = c_{cr}$ part of Fig. 4.

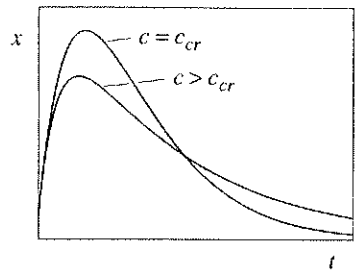


Figure 4. Critically damped and overdamped cases.

Overdamped vibration ($c > c_{cr}$). As c increases beyond c_{cr}, (4) once again gives two distinct roots, but now they are both real and negative (because the $\sqrt{c^2 - 4mk}$ is smaller than c), so

$$x(t) = e^{-\frac{c}{2m}t}\left[A\cosh\sqrt{\left(\frac{c}{2m}\right)^2 - \omega^2}\,t + B\sinh\sqrt{\left(\frac{c}{2m}\right)^2 - \omega^2}\,t\right], \tag{14}$$

where A and B can be determined from the initial conditions (2). Indeed, if one or both roots were positive then we could have exponential growth, which would make no sense, physically. If that did happen we should expect that either there is an error in our mathematics or that our mathematical modeling of the phenomenon is grossly inaccurate.

A representative plot of that solution is shown in the $c > c_{cr}$ part of Fig. 4. For the sake of comparison we have used the same initial conditions to generate the three plots in Figures 3 and 4. Though one can use positive and negative exponentials within the parentheses in (14), in place of the hyperbolic cosine and sine, the latter are more convenient for the application of the initial conditions since the sinh is zero at $t = 0$ and so is the derivative of the cosh.

This completes our solution of equation (3), governing the free oscillation of the mechanical oscillator shown in Fig. 1. It should be emphasized that Fig. 1 is intended only as a schematic equivalent of the actual physical system. For instance, suppose the actual system consists of a beam cantilevered downward, with a mass m at its end, as shown in Fig. 5a. We assume the mass of the beam to be negligible compared to m. It is known from Euler beam theory that if we apply a constant force F, as shown in Fig. 5b, then the end deflection x is given by $x = FL^3/(3EI)$, where L is the length of the beam and EI is its "stiffness" (E is Young's modulus of the material and I is a cross-sectional moment of inertia). Re-expressing the latter as $F = (3EI/L^3)x$, we see that it is of the form $F = kx$, as for a linear spring of stiffness k. Thus, insofar as the modeling and analysis is concerned, the physical beam system is equivalent to the mass-spring arrangement shown in Fig. 6c, where $k_{eq} = 3EI/L^3$ is the stiffness of the equivalent spring and where there is no friction between the block and the table top. The governing equation of motion is

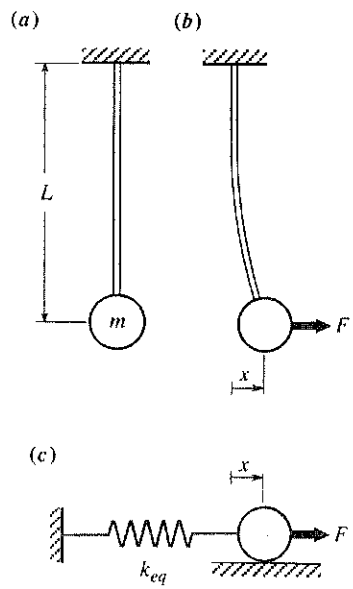


Figure 5. Equivalent mechanical systems.

$$mx'' + k_{eq}x = 0. \tag{15}$$

Just as we neglected the mass of the beam, compared to m, likewise let us neglect

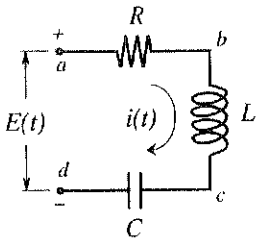


Figure 6. Electrical oscillator; *RLC* circuit.

the mass of the spring compared to m. (How to account for that mass, approximately, is discussed in the exercises.) It should be noted that, in addition, we are neglecting the rotational motion of the mass, in Fig. 5b, since we have already limited ourselves to the case of small deflections of the beam.

Finally, it has already been pointed out, in Section 2.3, that the force-driven mechanical oscillator is analogous to the voltage-driven RLC electrical circuit reproduced here in Fig. 6, under the equivalence

$$L \leftrightarrow m,\ R \leftrightarrow c,\ \frac{1}{C} \leftrightarrow k,\ i(t) \leftrightarrow x(t),\ \frac{dE(t)}{dt} \leftrightarrow F(t), \tag{16}$$

so whatever results we have obtained in this section for the mechanical oscillator apply equally well to the electrical oscillator shown in Fig. 6, according to the equivalence given above.

Closure. In this section we have considered the free oscillations of the mechanical harmonic oscillator. We found that for the undamped case the solution is a pure sine wave with an amplitude and phase shift that depend upon the initial conditions – that is, the solution is "harmonic." In the presence of light damping (i.e., for $c < c_{cr}$), the solution suffers exponential decay and a reduction in frequency, these effects becoming more pronounced as c is increased. When c reaches a critical value c_{cr} the oscillation ceases altogether, and as c is increased further the exponential decay is increasingly pronounced.

It should be emphasized that by the damped harmonic oscillator we mean a system that can be modeled by a linear equation of the form $mx'' + cx' + kx = 0$. In most applications, however, the restoring force can be regarded as a linear function of x (namely, kx) only for motions that are sufficiently small departures from an equilibrium configuration; if the motion is not sufficiently small, then one must deal with a more difficult nonlinear differential equation. Thus, for the harmonic oscillator, damped or not, we are able to generate simple closed form solutions, as we have done in this section. For nonlinear oscillators one often gives up on the possibility of finding closed form analytical solutions and relies instead on numerical simulation, as will be discussed in Chapter 6. To illustrate how such nonlinear oscillators arise in applications, we have included several such examples in the exercises that follow.

In terms of formulas, the equivalence of the three forms (6), (7), and (8) should be clearly understood and remembered. In a given application we will use whichever of these seems most convenient, usually (7).

EXERCISES 3.5

1. Re-express each expression in the form $E\sin(\omega t + \phi)$; that is, evaluate E, ϕ, ω.

(a) $6\cos t + \sin t$

(b) $3\cos 6t - 4\sin 6t$

(c) $5\cos 2t - 12\sin 2t$

(d) $-2\cos 3t + 2\sin 3t$

(e) $\cos 5t - \sin 5t$

(f) $x_0 \cos \omega t + \frac{x_0'}{\omega} \sin \omega t$, from (11)

2. We emphasized the equivalence of the solution forms (6), (7), and (8), and discussed the equations (10a,b) that relate C and D in (7) to E and ϕ in (8). Of course, we could have used the cosine in place of the sine, and expressed

$$x(t) = G \cos(\omega t + \psi) \tag{2.1}$$

instead. Derive formulas analogous to (10a,b), expressing G and ψ in terms of C and D.

3. Apply the initial conditions (2) to the general solution (12), and solve for the integration constants A and B in terms of m, c, ω, x_0 and x_0'.

4. Apply the initial conditions (2) to the general solution (14), and solve for the integration constants A and B in terms of m, c, ω, x_0 and x_0'.

5. Consider an undamped harmonic oscillator governed by the equation $mx'' + kx = 0$, with initial conditions $x(0) = x_0, x'(0) = x_0'$. One might expect the frequency of oscillation to depend on the initial displacement x_0. Does it? Explain.

6. We mentioned in the text that the oscillation ceases altogether when c is increased to c_{cr} or beyond. Let us make that statement more precise: for $c \geq c_{cr}$ the graph of $x(t)$ has at most one "flat spot" (on $0 \leq t < \infty$), that is, where $x' = 0$.

(a) Prove that claim.
(b) Make up a case (i.e., give numerical values of m, c, k, x_0, x_0') where there is *no* flat spot on $0 \leq t < \infty$.
(c) Make up a case where there is one flat spot on $0 \leq t < \infty$.

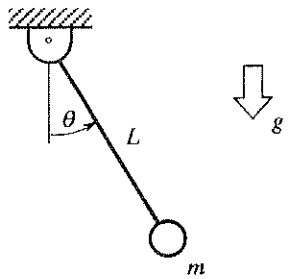
7. (*Logarithmic decrement*) For the underdamped case ($c < c_{cr}$), let x_n and x_{n+1} denote any two successive maxima of $x(t)$.

(a) Show that the ratio $r_n = x_n/x_{n+1}$ is a constant, say r; that is, $x_1/x_2 = x_2/x_3 = \cdots = r$.
(b) Further, show that the natural logarithm of r, called the **logarithmic decrement** δ, is given by $\delta = \frac{c}{m} \frac{\pi}{\sqrt{\omega^2 - (c/2m)^2}}$.

8. (*Grandfather clock*) Consider a pendulum governed by the equation of motion $mL\theta'' + mg \sin\theta = 0$, or

$$\theta'' + \frac{g}{L} \sin\theta = 0, \tag{8.1}$$

where g is the acceleration of gravity. (See the figure.) If



$|\theta| \ll 1$ (where $\ll$ means much smaller than), then $\sin\theta \approx \theta$, and the nonlinear equation of motion (8.1) can be simplified to the linear equation

$$\theta'' + \frac{g}{L}\theta = 0, \tag{8.2}$$

or, if we allow for some inevitable amount of damping due to friction and air resistance,

$$\theta'' + \epsilon\theta' + \frac{g}{L}\theta = 0, \tag{8.3}$$

where $0 < \epsilon \ll 1$. Now imagine the pendulum to be part of a grandfather's clock. If a ratchet mechanism converts each oscillation to one second of recorded time, how does the clock maintain its accuracy even as it runs down, that is, even when its amplitude of oscillation has diminished to a small fraction of its initial value? Explain.

9. (*Correction for the mass of the spring*) Recall that our model of the mechanical oscillator neglects the effect of the mass of the spring on the grounds that it is sufficiently small compared to that of the mass m. In this exercise we seek to improve our model so as to account, if only approximately, for the mass of the spring. In doing so, we consider the undamped case, for which the equation of motion is $mx'' + kx = 0$.

(a) Multiplying that equation by dx and integrating, derive the "first integral"

$$\frac{1}{2}mx'^2 + \frac{1}{2}kx^2 = C, \tag{9.1}$$

which states that the total energy, the kinetic energy of the mass plus the potential energy of the spring, is a constant.
(b) Let the mass of the spring be m_s. Suppose that the velocity of the elements within the spring at any time t varies linearly from 0 at the fixed end to $x'(t)$ at its attachment to the mass m. Show, subject to that assumption, that the kinetic energy in the spring is $\frac{1}{6}m_s x'^2(t)$. Improving (9.1) to the form

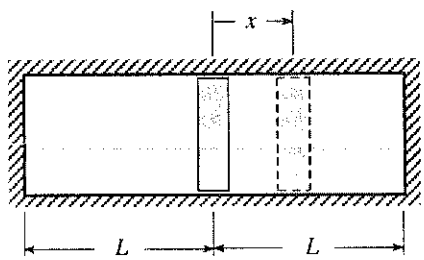
$$\frac{1}{2}\left(m + \frac{1}{3}m_s\right)x'^2 + \frac{1}{2}kx^2 = C, \tag{9.2}$$

obtain, by differentiation with respect to t, the improved equation of motion

$$\left(m + \frac{1}{3}m_s\right)x'' + kx = 0. \tag{9.3}$$

Thus, as a correction, to take into account the mass of the spring, we merely replace the mass m in $mx'' + kx = 0$ by an "effective mass" $m + \frac{1}{3}m_s$, which incorporates the effect of the mass of the spring. NOTE: This analysis is approximate in that it assumes the velocity distribution within the spring, whereas that distribution itself needs to be determined, which determination involves the solution of a partial differential equation of wave type, as studied in a later chapter.
(c) In obtaining an effective mass of the form $m + \alpha m_s$, why is it reasonable that α turns out to be less than 1?

10. (*Piston oscillator*) Let a piston of mass m be place at the midpoint of a closed cylinder of cross-sectional area A and length $2L$, as sketched. Assume that the pressure p on either



side of the piston satisfies Boyle's law (namely, that the pressure times the volume is constant), and let p_0 be the pressure on both sides when $x = 0$.

(a) If the piston is disturbed from its equilibrium position $x = 0$, show that the governing equation of motion is

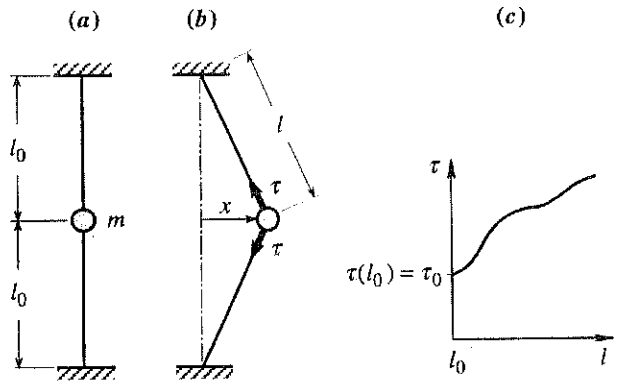
$$mx'' + 2p_0AL\frac{x}{L^2 - x^2} = 0. \tag{10.1}$$

(b) Is (10.1) linear or nonlinear? Explain.
(c) Expand the $x/(L^2 - x^2)$ term in a Taylor series about $x = 0$, up to the third-order term. Keeping only the leading term, derive the linearized version

$$mx'' + \frac{2p_0A}{L}x = 0 \tag{10.2}$$

of (10.1), which is restricted to the case of small oscillations – that is, where the amplitude of oscillation is small compared to L.
(d) From (10.2), determine the frequency of oscillation, in cycles per second.
(e) Is the resulting linearized model equivalent to the vibration of a mass/spring system, with an equivalent spring stiffness of $k_{eq} = 2p_0A/L$? Explain.

11. (*Lateral vibration of a bead on a string*) Consider a mass m, such as a bead, restrained by strings (of negligible mass), in each of which there is a tension τ_0, as shown in Fig. *a*.



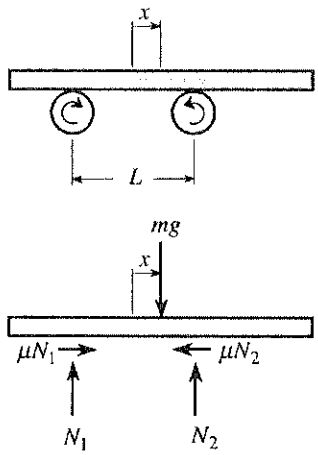
We seek the frequency of small lateral oscillations of m. A lateral displacement x (Fig. *b*) will cause the length of each string to increase from l_0 to $l(x) = \sqrt{l_0^2 + x^2}$. Suppose that the tension τ is found, empirically, to increase with l, from its initial value τ_0, as shown in Fig. *c*.

(a) Show that the governing equation of lateral motion is

$$mx'' + 2\frac{\tau\left(\sqrt{l_0^2 + x^2}\right)}{\sqrt{l_0^2 + x^2}}x = 0, \tag{11.1}$$

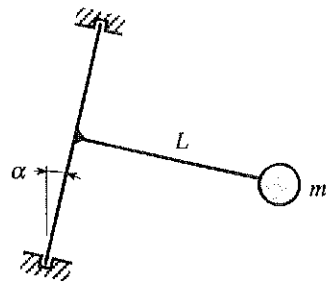
where $\tau\left(\sqrt{l_0^2 + x^2}\right)$ is a function, not a product.
(b) Is (11.1) linear or nonlinear? Explain.
(c) Expand the $\tau\left(\sqrt{l_0^2 + x^2}\right)x/\sqrt{l_0^2 + x^2}$ term in a Taylor series about $x = 0$, up to the third-order term. [You should find that the coefficients of these terms involve l_0, τ_0, and $\tau'(l_0)$.]
(d) Linearize the equation of motion by retaining only the leading term of that Taylor series, show that the equivalent spring stiffness is $k_{eq} = 2\tau_0/l_0$, and that the frequency of small oscillations is $\frac{1}{2\pi}\sqrt{\frac{2\tau_0}{ml_0}}$ cycles/sec.

12. (*Oscillating platform*) A uniform horizontal platform of mass m is supported by counter-rotating cylinders a distance L apart (see figure). The friction force f exerted on the



platform by each cylinder is proportional to the normal force N between the platform and the cylinder, with constant of proportionality (coefficient of sliding friction) μ: $f = \mu N$. Show that if the cylinder is disturbed from its equilibrium position ($x = 0$), then it will undergo a lateral oscillation of frequency $\omega = \sqrt{2\mu g/L}$ rad/sec, where g is the acceleration of gravity. HINT: Derive the equation of motion governing the lateral displacement x of the midpoint of the platform relative to a point midway between the cylinders.

13. (*Tilted pendulum*) Consider a rod of length L with a point mass m at its end, where the mass of the rod is negligible compared to m. The rod is welded at a right angle to another, which rotates without friction about an axis that is tilted by an angle of α with respect to the vertical (see figure). Let θ denote



the angle of rotation of the pendulum, with respect to its equilibrium position (where m is at its lowest possible point, namely, in the plane of the paper).

(a) Derive the governing equation of motion

$$\theta'' + \frac{g}{L}\sin\alpha\sin\theta = 0. \tag{13.1}$$

As a partial check of this result, observe that for $\alpha = \pi/2$ (14.1) does reduce to the equation of motion of the ordinary pendulum (see Exercise 8). HINT: Write down an equation of conservation of energy (kinetic plus potential energy equal a constant), and differentiate it with respect to the time t.

(b) What is the frequency of small amplitude oscillations, in rad/sec? In cycles/sec?

3.6 Solution of Homogeneous Equation: Nonconstant Coefficients

We return to the nth-order linear homogeneous equation

$$a_0(x)\frac{d^n y}{dx^n} + a_1(x)\frac{d^{n-1}y}{dx^{n-1}} + \cdots + a_{n-1}(x)\frac{dy}{dx} + a_n(x)y = 0, \tag{1}$$

and this time allow the a_j coefficients to be nonconstant. The theory of the homogeneous equation, given in Section 3.3, holds whether the coefficients are constant or not. However, the task of finding solutions is generally much more difficult if the coefficients in (1) are not all constants. Only in special cases are we able to find solutions in terms of the elementary functions and in closed form (as opposed, say, to infinite series). This section is devoted to those special cases – most notably, to equations of Cauchy–Euler type. In other cases we generally give up on finding closed-form solutions and, instead, seek solutions in the form of infinite series

(Chapter 4) or pursue a numerical approach (Chapter 6).

3.6.1. Cauchy–Euler equation. If (1) is of the special form

$$\boxed{x^n \frac{d^n y}{dx^n} + c_1 x^{n-1} \frac{d^{n-1} y}{dx^{n-1}} + \cdots + c_{n-1} x \frac{dy}{dx} + c_n y = 0,} \tag{2}$$

where the c_j's are constants, it is called a **Cauchy–Euler equation**, and is also called an **equidimensional equation**.

Of most importance to us will be the case where $n = 2$, so let us consider that case first, namely,

$$x^2 y'' + c_1 x y' + c_2 y = 0, \tag{3}$$

and let us consider the x interval to be $0 < x < \infty$; the case of negative x's will be treated separately, below. Suppose we try to solve (3) by seeking y in the form $y = e^{\lambda x}$, where λ is a yet-to-be-determined constant, which form proved successful for the constant-coefficient case. Then $y' = \lambda e^{\lambda x}$ and $y'' = \lambda^2 e^{\lambda x}$, so (3) becomes

$$\lambda^2 x^2 e^{\lambda x} + \lambda c_1 x e^{\lambda x} + c_2 e^{\lambda x} = 0. \tag{4}$$

If we cancel the (nonzero) exponentials we obtain a quadratic equation for λ, solution of which gives λ as a function of x. However, λ was supposed to be a constant, so we have a contradiction, and the method fails. (Specifically, if λ turns out to be a function of x, then $y' = \lambda e^{\lambda x}$ and $y'' = \lambda^2 e^{\lambda x}$, above, were incorrect.) Said differently, the $x^2 e^{\lambda x}, x e^{\lambda x}, e^{\lambda x}$ terms in (4) are LI and cannot be made to cancel identically to zero on any given x interval.

The reason we have discussed this fruitless approach is to emphasize that it is incorrect, and to caution against using it. By contrast, if the equation were of constant-coefficient type, say $y'' + c_1 y' + c_2 y = 0$, then $y = e^{\lambda x}$ would work because $y = e^{\lambda x}$, $y' = \lambda e^{\lambda x}$, $y'' = \lambda^2 e^{\lambda x}$ are LD, so the combination $y'' + c_1 y' + c_2 y$ could be made to cancel to zero by suitable choice of λ.

Although the form $e^{\lambda x}$ will not work for Cauchy–Euler equations, the form

$$\boxed{y = x^\lambda} \tag{5}$$

will, because $y = x^\lambda, xy' = \lambda x^\lambda, x^2 y'' = \lambda(\lambda - 1) x^\lambda, \ldots$ are LD since each is a constant times x^λ. Putting (5) into the second-order equation (3) gives

$$[\lambda(\lambda - 1) + c_1 \lambda + c_2]\, x^\lambda = 0.$$

Since $x^\lambda \neq 0$, we require of λ that

$$\lambda^2 - (1 - c_1)\lambda + c_2 = 0,$$

so

$$\lambda = \frac{1 - c_1 \pm \sqrt{(1 - c_1)^2 - 4c_2}}{2}. \tag{6}$$

We distinguish three cases, depending upon whether the discriminant $\Delta \equiv (1 - c_1)^2 - 4c_2$ is positive, zero, or negative:

$\Delta > 0$: **Distinct real roots.** If $\Delta > 0$, then (6) gives two distinct real roots, say λ_1 and λ_2, so we have the general solution to (3) as

$$\boxed{y(x) = Ax^{\lambda_1} + Bx^{\lambda_2}.} \tag{7}$$

EXAMPLE 1. To solve

$$x^2y'' - 2xy' - 10y = 0, \tag{8}$$

seek $y = x^\lambda$. That form gives $\lambda^2 - 3\lambda - 10 = 0$, with roots $\lambda = -2$ and 5, so the general solution of (8) is

$$y(x) = \frac{A}{x^2} + Bx^5. \quad \blacksquare$$

$\Delta = 0$: **Repeated real roots.** In this case (6) gives the repeated root $\lambda = \dfrac{1 - c_1}{2} \equiv \lambda_1$. Thus we have the solution Ax^{λ_1}, but are missing a second linearly independent solution, which is needed if we are to obtain a general solution of (3). Evidently, the missing solution is not of the form x^λ, or we would have found it when we sought $y(x) = x^\lambda$.

To find the missing solution we use Lagrange's method of **reduction of order**, as we did in Section 3.3.3 for constant-coefficient differential equations with repeated roots of the characteristic equation. That is, we let A vary, and seek

$$y(x) = A(x)x^{\lambda_1}. \tag{9}$$

Putting (9) into (3) gives (we leave the details to the reader, as Exercise 3)

$$xA'' + A' = 0.$$

Next, set $A' \equiv p$, say, to reduce the order:

$$x\frac{dp}{dx} + p = 0, \tag{10}$$

so $p = D/x$ and $A(x) = D \ln x + C$, where C, D are arbitrary constants (Exercise 4). Finally, putting the latter back into (9) gives the general solution of (3) as

$$\boxed{y(x) = (C + D \ln x)\, x^{\lambda_1}.} \tag{11}$$

EXAMPLE 2. To solve

$$x^2y'' + 7xy' + 9y = 0, \tag{12}$$

seek $y = x^\lambda$. That form gives $\lambda^2 + 6\lambda + 9 = 0$, with the repeated root $\lambda = -3$, so the general solution of (12) is

$$y(x) = (A + B\ln x)\,x^{-3}. \quad \blacksquare$$

$\Delta < 0$: **Complex roots.** In this case (6) gives the distinct complex-conjugate roots

$$\lambda = \frac{1-c_1}{2} \pm i\frac{\sqrt{4c_2 - (1-c_1)^2}}{2} \equiv \alpha \pm i\beta, \tag{13}$$

so we have the general solution of (3) as

$$\begin{aligned} y(x) &= Ax^{\alpha+i\beta} + Bx^{\alpha-i\beta} \\ &= x^\alpha\left(Ax^{i\beta} + Bx^{-i\beta}\right). \end{aligned} \tag{14}$$

However, since we normally prefer real solution forms, let us use the identity $u = e^{\ln u}$ to re-express (14) as*

$$\begin{aligned} y(x) &= x^\alpha\left(Ae^{\ln x^{i\beta}} + Be^{\ln x^{-i\beta}}\right) = x^\alpha\left(Ae^{i\beta\ln x} + Be^{-i\beta\ln x}\right) \\ &= x^\alpha\left\{A\left[\cos(\beta\ln x) + i\sin(\beta\ln x)\right] + B\left[\cos(\beta\ln x) - i\sin(\beta\ln x)\right]\right\} \\ &= x^\alpha\left[(A+B)\cos(\beta\ln x) + i(A-B)\sin(\beta\ln x)\right]. \end{aligned} \tag{15}$$

Or, letting $A + B \equiv C$ and $i(A - B) \equiv D$,

$$\boxed{y(x) = x^\alpha\left[C\cos(\beta\ln x) + D\sin(\beta\ln x)\right].} \tag{16}$$

EXAMPLE 3. To solve

$$x^2y'' - 2xy' + 4y = 0, \tag{17}$$

*It is important to appreciate that the $x^{\alpha\pm i\beta}$ quantities, in (14), are "new objects" for us, for we have not yet (in this book) defined a real number x raised to a complex power (unless x happens to be e, in which case the already-discussed Euler's formulas apply). Staying as close as possible to familiar *real* variable results, let us write

$$x^{\alpha+i\beta} = x^\alpha x^{i\beta} = x^\alpha\left(e^{\ln x^{i\beta}}\right) = x^\alpha e^{i\beta\ln x},$$

and similarly for $x^{\alpha-i\beta}$. None of these three equalities are justifiable, since they rely on the formulas $x^{a+b} = x^a x^b, u = e^{\ln u}$, and $\ln x^c = c\ln x$, which assume x, a, b, c to be real, but we hereby understand them to hold by definition. Observe that complex quantities and complex functions keep forcing themselves upon us. Therefore, it behooves us to establish a general theory of complex functions, rather than deal with these issues one by one. We will do exactly that, but not until much later in the text.

seek $y = x^\lambda$. That form gives $\lambda^2 - 3\lambda + 4 = 0$, so $\lambda = \frac{3}{2} \pm i\frac{\sqrt{7}}{2}$. Hence

$$\begin{aligned} y(x) &= Ax^{3/2+i\sqrt{7}/2} + Bx^{3/2-i\sqrt{7}/2} = x^{3/2}\left(Ax^{i\sqrt{7}/2} + Bx^{-i\sqrt{7}/2}\right) \\ &= x^{3/2}\left(Ae^{i\frac{\sqrt{7}}{2}\ln x} + Be^{-i\frac{\sqrt{7}}{2}\ln x}\right) \\ &= x^{3/2}\left[C\cos\left(\frac{\sqrt{7}}{2}\ln x\right) + D\sin\left(\frac{\sqrt{7}}{2}\ln x\right)\right]. \quad \blacksquare \end{aligned}$$

Recall that we have limited our discussion of (3) to the case where $x > 0$. The reason for that limitation is as follows. For a function $y(x)$ to be a solution of (3) on an x interval I, we first of all need each of y, y', and y'' to exist on I; that is, to be defined there. The function $\ln x$ and its derivatives are not defined at $x = 0$, nor is $\ln x$ defined (as a real-valued function) for $x < 0$. The functions $x^{\lambda_1}, x^{\lambda_2}$ in (7), x^{λ_1} in (11), and x^α in (16) cause similar problems. To deal with the case where $x < 0$, it is more convenient to make the change of variable $x = -\xi$ in (3), so that ξ will be positive. Letting $y(x) = y(-\xi) \equiv Y(\xi)$,*

$$\frac{dy}{dx} = \frac{dY}{d\xi}\frac{d\xi}{dx} = -\frac{dY}{d\xi}, \qquad \frac{d^2y}{dx^2} = \frac{d}{d\xi}\left(-\frac{dY}{d\xi}\right)\frac{d\xi}{dx} = \frac{d^2Y}{d\xi^2}, \tag{18}$$

so (3) becomes

$$\xi^2\frac{d^2Y}{d\xi^2} + c_1\xi\frac{dY}{d\xi} + c_2Y = 0, \qquad (\xi > 0) \tag{19}$$

which is the same as (3)! Thus, its solutions are the same, but with x changed to ξ. For the case of distinct real roots, for instance,

$$y(x) = Ax^{\lambda_1} + Bx^{\lambda_2}$$

for $x > 0$, and

$$y(x) = Y(\xi) = Y(-x) = A(-x)^{\lambda_1} + B(-x)^{\lambda_2}$$

for $x < 0$. Observe that both of these forms are accounted for by the single expression

$$y(x) = A\,|x|^{\lambda_1} + B\,|x|^{\lambda_2}.$$

Similarly for the other cases (repeated real roots and complex roots). Let us state these results, for reference, as a theorem.

*Why do we change the name of the dependent variable from y to Y? Because they are different functions. To illustrate, suppose $y(x) = 5 + x^3$. Then $Y(\xi) = 5 + (-\xi)^3 = 5 - \xi^3$. For instance, if the argument of y is 2, then y is 13, but when the argument of Y is 2, then Y is -3.

THEOREM 3.6.1 *Second-Order Cauchy–Euler Equation*
The general solution of the second-order Cauchy–Euler equation

$$x^2y'' + c_1xy' + c_2y = 0, \tag{20}$$

on any x interval not containing the origin, is

$$y(x) = \begin{cases} A\,|x|^{\lambda_1} + B\,|x|^{\lambda_2} \\ (A + B\ln|x|)\,|x|^{\lambda_1} \\ |x|^{\alpha}\left[A\cos(\beta\ln|x|) + B\sin(\beta\ln|x|)\right] \end{cases} \tag{21}$$

if the roots λ_1, λ_2 of $\lambda^2 + (c_1 - 1)\lambda + c_2 = 0$ are real and distinct, real and repeated, or complex ($\lambda = \alpha \pm i\beta$), respectively.

Of course, if the x interval is to the right of the origin, then the absolute value signs in (21) can be dropped.

To close our discussion of the Cauchy–Euler equation, consider the higher-order case, $n > 2$. For simplicity, we consider $x > 0$; as for the second-order case treated above, $x < 0$ can be handled simply by changing all x's in the solution to $|x|$.

EXAMPLE 4. Consider the third-order Cauchy–Euler equation

$$x^3y''' - 3x^2y'' + 7xy' - 8y = 0. \tag{22}$$

Seeking $y(x) = x^\lambda$ gives

$$\lambda^3 - 6\lambda^2 + 12\lambda - 8 = 0,$$

with the roots $\lambda = 2, 2, 2$. Thus we have the solution $y(x) = Ax^2$, but we need two more linearly independent solutions. To find them, we use reduction of order and seek $y(x) = A(x)x^2$. Putting that form into (22) gives the equation

$$x^2A''' + 3xA'' + A' = 0$$

on $A(x)$, which can be reduced to the second-order equation

$$x^2p'' + 3xp' + p = 0 \tag{23}$$

by letting $A' = p$. The latter is again of Cauchy–Euler type, and letting $p(x) = x^\lambda$ gives $\lambda = -1, -1$, so that

$$p(x) = (B + C\ln x)\frac{1}{x}.$$

Since $A' = p$,

$$A(x) = \int p\,dx = B\ln x + C\frac{(\ln x)^2}{2} + D,$$

and

$$y(x) = \left[C_1 + C_2 \ln x + C_3(\ln x)^2\right] x^2 \tag{24}$$

is the desired general solution of (22). ∎

Comparing the latter result with the solution (11) for the second-order Cauchy–Euler equation with repeated root λ_1, we might well suspect that if any Cauchy–Euler equation has a repeated root λ_1 of order k, then that root contributes the form

$$\left[C_1 + C_2 \ln x + C_3(\ln x)^2 + \cdots + C_k(\ln x)^{k-1}\right] x^{\lambda_1} \tag{25}$$

to the general solution. We state, without proof, that that is indeed the case.

EXAMPLE 5. As a summary example, suppose that upon seeking solutions of a given eighth-order Cauchy–Euler equation in the form $y(x) = x^\lambda$ we obtain the roots

$$\lambda = -2.4,\ 1.7,\ 1.7,\ 1.7,\ -3+4i,\ -3+4i,\ -3-4i,\ -3-4i.$$

Then the general solution is

$$\begin{aligned} y(x) = {} & C_1 x^{-2.4} + \left[C_2 + C_3(\ln x) + C_4(\ln x)^2\right] x^{1.7} \\ & + (C_5 + C_6 \ln x)\, x^{-3+4i} + (C_7 + C_8 \ln x)\, x^{-3-4i}, \end{aligned} \tag{26}$$

or (Exercise 5),

$$\begin{aligned} y(x) = {} & C_1 x^{-2.4} + \left[C_2 + C_3(\ln x) + C_4(\ln x)^2\right] x^{1.7} \\ & + \{[C_9 \cos(4 \ln x) + C_{10} \sin(4 \ln x)] \\ & \quad + \ln x\, [C_{11} \cos(4 \ln x) + C_{12} \sin(4 \ln x)]\}\, x^{-3}. \end{aligned} \tag{27}$$

Although such high-order Cauchy–Euler equations are uncommon, we include this example to illustrate the general solution pattern for any order $n \geq 2$. ∎

This concludes our discussion of the Cauchy–Euler equation. We will meet Cauchy–Euler equations again in the next chapter, in connection with power series solutions of differential equations, and again when we study the partial differential equations governing such phenomena as heat conduction, electric potential, and certain types of fluid flow, in later chapters.

3.6.2. Reduction of order. (Optional) We have already used Lagrange's method of reduction of order to find "missing solutions," for constant-coefficient equations and for Cauchy–Euler equations as well. Here, we focus not on constant-coefficient or Cauchy–Euler equations, but on the method itself and indicate its more general application to any linear homogeneous differential equation.

For definiteness, consider the second-order case,

$$y'' + a_1(x)y' + a_2(x)y = 0. \tag{28}$$

Suppose that one solution is known to us, say $Y(x)$, and that a second linearly independent solution is sought. If $Y(x)$ is a solution, then so is $AY(x)$ for any constant A. The idea behind Lagrange's method of **reduction of order** is to seek the missing solution in the form

$$y(x) = A(x)Y(x), \tag{29}$$

where $A(x)$ is to be determined.

The method is similar to Lagrange's method of *variation of parameters*, introduced in Section 2.2, but its purpose is different. The latter was used to find the general solution of the *non*homogeneous equation $y' + p(x)y = q(x)$ from the solution $y_h(x) = Ae^{-\int p(x)\,dx}$ of the homogeneous equation $y' + p(x)y = 0$, by varying the parameter A and seeking $y(x) = A(x)e^{-\int p(x)\,dx}$. Reduction of order is similar in that we vary the parameter A in $y = AY(x)$, but different in that it is used to find a missing solution of a homogeneous equation from a known solution $Y(x)$ of that homogeneous equation.

We begin by emphasizing that at first glance the form (29) seems to be without promise. To explain that statement, observe that the search for a pair of lost glasses can be expected to be long and arduous if we merely know that they are somewhere in North America, but shorter and easier to whatever extent we are able to narrow the domain of the search. If, for instance, we know that they are somewhere on our desk, then the search is short and simple. Likewise, when we solve a constant-coefficient equation by seeking y in the form $e^{\lambda x}$ then the search is short and simple since, first, solutions will indeed be found within that subset and, second, because that subset is tiny compared to the set of all possible functions, just as one's desk is tiny compared to North America. Similarly, when we solve a Cauchy–Euler equation by seeking y in the form x^{λ}.

With this idea in mind, observe that the form (29) does not narrow our search in the slightest, since it includes *all* functions! That is, any given function $f(x)$ can be expressed as $A(x)Y(x)$ simply by choosing $A(x)$ to be $f(x)/Y(x)$.

Proceeding nonetheless, we put (29) into (28) and obtain the differential equation

$$A''Y + (2Y' + a_1Y)A' + (Y'' + a_1Y' + a_2Y)A = 0 \tag{30}$$

on $A(x)$. At first glance it appears that this differential equation on $A(x)$ is probably even harder than the original equation, (28), on $y(x)$. However, and this is the heart of Lagrange's idea, all of the undifferentiated A terms must cancel, because if A were a constant (in which case the A' and A'' terms would all drop out), then the remaining terms would have to cancel to zero because $AY(x)$ is a solution of (28). Thus, the coefficient of A in (30) is zero, so (30) becomes

$$A''Y + (2Y' + a_1Y)A' = 0, \tag{31}$$

the order of which can now be reduced from two to one by letting $A' = p$:

$$\frac{dp}{dx} + \left(\frac{2Y' + a_1Y}{Y}\right)p = 0. \tag{32}$$

Integrating the latter gives

$$\begin{aligned} p(x) &= Be^{-\int \frac{2Y'+a_1Y}{Y}\,dx} = Be^{-2\int \frac{dY}{Y} - \int a_1\,dx} \\ &= BY(x)^{-2}e^{-\int a_1(x)\,dx}. \end{aligned}$$

Finally, integration of $A' = p$ gives

$$A(x) = \int p(x)\,dx = B\int Y(x)^{-2}e^{-\int a_1(x)\,dx}dx + C,$$

so (29) becomes

$$y(x) = \left[B\int Y(x)^{-2}e^{-\int a_1(x)\,dx}dx + C\right]Y(x). \tag{33}$$

The $CY(x)$ term merely reproduces the solution that was already known; the missing solution is provided by the other term, $BY(x)\int Y(x)^{-2}e^{-\int a_1(x)\,dx}dx$. That this solution and the original solution $Y(x)$ are necessarily LI is left for Exercise 6.

Incidentally, the result (33) could also be written using definite integrals if one prefers, as

$$y(x) = \left[B\int_\alpha^x Y(\xi)^{-2}e^{-\int_\beta^\xi a_1(\eta)\,d\eta}d\xi + C\right]Y(x), \tag{34}$$

where the lower limits α and β are arbitrary numbers, for the effect of changing α is simply to add some constant to the ξ integral, and that constant times B can be absorbed by the arbitrary constant C. Likewise, changing β simply adds some constant, say P, to the η integral, and the resulting e^{-P} factor can be absorbed by the arbitrary constant B.

EXAMPLE 6. *Legendre's equation.* The equation

$$(1-x^2)y'' - 2xy' + 2y = 0, \qquad (-1 < x < 1) \tag{35}$$

is known as **Legendre's equation**, after the French mathematician *Adrien Marie Legendre* (1752–1833). It is studied in Chapter 4, and used in later chapters when it arises in physical applications.

Observe that (35) admits the simple solution $y(x) = x$. To find a second solution, and hence a general solution, we can seek $y(x) = A(x)x$ and follow the steps outlined above. Rather, let us simply use the derived result (33). First, we divide (35) by $1-x^2$ to reduce it to the form (28), so that we can identify $a_1(x)$. Thus, with $a_1(x) = -\dfrac{2x}{1-x^2}$ and $Y(x) = x$, (33) gives

$$\begin{aligned} y(x) &= \left[B\int \frac{e^{\int 2x\,dx/(1-x^2)}}{x^2}\,dx + C\right]x = \left[B\int \frac{dx}{x^2(1-x^2)} + C\right]x \\ &= \left[B\int\left(\frac{1}{x^2} + \frac{1}{2}\frac{1}{1-x} + \frac{1}{2}\frac{1}{1+x}\right)dx + C\right]x \\ &= \left[B\left(-\frac{1}{x} + \frac{1}{2}\ln\frac{1+x}{1-x}\right) + C\right]x, \end{aligned}$$

or, equivalently,

$$y(x) = C_1 x + C_2 \left(1 - \frac{x}{2} \ln \frac{1+x}{1-x}\right). \quad \blacksquare$$

In this example we were able to evaluate the integrals that occurred. In other cases we may not be able to, even with the help of computer software, and may therefore need to leave the answer in integral form.

3.6.3. Factoring the operator. (Optional) We have been considering the nth-order linear homogeneous equation

$$L[y] = \left[\frac{d^n}{dx^n} + a_1(x)\frac{d^{n-1}}{dx^{n-1}} + \cdots + a_n(x)\right] y = 0,$$

or

$$\left(D^n + a_1 D^{n-1} + \cdots + a_n\right) y = 0, \tag{36}$$

where $D = \dfrac{d}{dx}$, $D^2 = DD = \dfrac{d}{dx}\dfrac{d}{dx} = \dfrac{d^2}{dx^2}$, and so on.

Suppose, first, that (36) is of constant-coefficient type (i.e., each a_j is a constant), and that the characteristic polynomial $\lambda^n + a_1\lambda^{n-1} + \cdots + a_n$ can be factored as $(\lambda - \lambda_1)(\lambda - \lambda_2)\cdots(\lambda - \lambda_n)$, where one or more of the roots λ_j may be repeated. Then the differential operator $L = D^n + a_1 D^{n-1} + \cdots + a_n$ can be factored in precisely the same way, as $(D - \lambda_1)(D - \lambda_2)\cdots(D - \lambda_n)$, where we understand $(D - \lambda_1)(D - \lambda_2)\cdots(D - \lambda_n)y$ to mean that first y is operated on by $D - \lambda_n$, then the result of that step is operated on by $D - \lambda_{n-1}$, and so on. That is, we begin at the right and proceed to the left. Further, it is readily verified that the sequential order of the $D - \lambda_j$ factors is immaterial, that is, they commute. If $n = 2$, for instance,

$$\begin{aligned}(D - \lambda_1)(D - \lambda_2)y &= (D - \lambda_1)(y' - \lambda_2 y)\\ &= D(y' - \lambda_2 y) - \lambda_1(y' - \lambda_2 y)\\ &= y'' - (\lambda_2 + \lambda_1)y' + \lambda_1\lambda_2 y\end{aligned}$$

and

$$\begin{aligned}(D - \lambda_2)(D - \lambda_1)y &= (D - \lambda_2)(y' - \lambda_2 y)\\ &= D(y' - \lambda_1 y) - \lambda_2(y' - \lambda_1 y)\\ &= y'' - (\lambda_2 + \lambda_1)y' + \lambda_2\lambda_1 y\end{aligned}$$

are the same.

By factoring L, we are able to reduce the solution of

$$(D - \lambda_1)(D - \lambda_2)\cdots(D - \lambda_n)y = 0 \tag{37}$$

to the solution of a sequence of n first-order equations, each of which is of the form $y' - py = q$ or

$$(D - p)y = q, \tag{38}$$

where p is a constant and $q(x)$ is known. From Section 2.2, we know that the solution of (38) is

$$y(x) = e^{px}\left(\int e^{-px}q(x)\,dx + A\right), \tag{39}$$

where A is an arbitrary constant.

Let us illustrate with an example.

EXAMPLE 7. The equation

$$y''' - 3y'' + 4y = 0 \tag{40}$$

admits the characteristic roots $\lambda = -1, 2, 2$, so we can factor (40) as

$$(D + 1)(D - 2)(D - 2)y = 0. \tag{41}$$

We begin the solution procedure by setting

$$(D - 2)(D - 2)y = u, \tag{42}$$

so that (41) becomes

$$(D + 1)u = 0,$$

with the solution

$$u(x) = Ae^{-x}.$$

Putting the latter into (42) gives

$$(D - 2)(D - 2)y = Ae^{-x}, \tag{43}$$

in which we set

$$(D - 2)y = v. \tag{44}$$

Then (43) becomes

$$(D - 2)v = Ae^{-x},$$

with the solution

$$v(x) = e^{2x}\left(\int e^{-2x}Ae^{-x}dx + B\right) = -\frac{A}{3}e^{-x} + Be^{2x}.$$

Finally, putting the latter into (44) gives

$$(D - 2)y = -\frac{A}{3}e^{-x} + Be^{2x},$$

with the solution

$$\begin{aligned} y(x) &= e^{2x}\left[\int e^{-2x}\left(-\frac{A}{3}e^{-x} + Be^{2x}\right)dx + C\right] \\ &= \frac{A}{9}e^{-x} + Bxe^{2x} + Ce^{2x}, \end{aligned}$$

or, equivalently,

$$y(x) = C_1e^{-x} + (C_2 + C_3x)\,e^{2x},$$

which is the same solution as obtained by methods discussed in earlier sections. Notice, in particular, that the presence of the repeated root $\lambda = 2, 2$ presented no additional difficulty. ∎

Although the factorization method reduces an nth-order equation to a sequence of n first-order equations, it is quite different from the method of reduction of order described above in Section 3.6.2.

Thus far we have limited our discussion of factorization to the constant-coefficient case. The nonconstant-coefficient case is more difficult. To appreciate the difficulty, consider the equation

$$y'' - x^2y = (D^2 - x^2)y = 0. \tag{45}$$

If we can factor $D^2 - x^2$ as $(D+x)(D-x)$, then we can solve (45) by the method outlined above. However,

$$\begin{aligned} (\underline{D} + x)(D - \underline{x})y &= (D + x)(y' - xy) = D(y' - xy) + x(y' - xy) \\ &= y'' - xy' - y + xy' - x^2y = y'' - (x^2 + 1)y, \end{aligned} \tag{46}$$

so $(D+x)(D-x) = D^2 - (x^2+1)$ is not the same as $D^2 - x^2$. The problem is that the differential operator on the left-hand side of (46) acts not only on y but also on itself, in the sense that an additional term is contributed to the final result, namely $-y$, through the action of the underlined D on the underlined x. Observe, further, that $D + x$ and $D - x$ do not commute since $(D + x)(D - x) = D^2 - (x^2 + 1)$, whereas $(D - x)(D + x) = D^2 - (x^2 - 1)$.

Thus, the following practical question arises: given a nonconstant coefficient operator, can it be factored and, if so, how?

Limiting our attention to equations of second order (which, arguably, is the most important case in applications), suppose that $a_1(x)$ and $a_2(x)$ are given, and that we seek $a(x)$ and $b(x)$ so that

$$y'' + a_1(x)y' + a_2(x)y = [D - a(x)][D - b(x)]y. \tag{47}$$

Writing out the right-hand side,

$$\begin{aligned} y'' + a_1y' + a_2y &= (D - a)(y' - by) \\ &= y'' - (a + b)y' + (ab - b')y. \end{aligned} \tag{48}$$

Since this equation needs to hold for all (twice-differentiable) functions y, a and b must satisfy the conditions (Exercise 13)

$$a + b = -a_1, \tag{49a}$$
$$ab - b' = a_2, \tag{49b}$$

or, isolating a and b (Exercise 14),

$$a' = a^2 + (a_1)a + (a_2 - a_1'), \tag{50a}$$
$$b' = -b^2 - (a_1)b - (a_2). \tag{50b}$$

Each of these equations is a special case of the nonlinear **Riccati equation**

$$\boxed{y' = p(x)y^2 + q(x)y + r(x),} \tag{51}$$

which was discussed in Exercise 11 of Section 2.2.

Thus, from a global point of view, it is interesting to observe that *the class of second-order equations with nonconstant coefficients is, in a sense, equivalent in difficulty to the class of nonlinear first-order Riccati equations.* We saw, in Exercise 11 of Section 2.2 that in the exceptional case where a particular solution $Y(x)$ of (51) can be found, perhaps by inspection, the nonlinear equation (51) can be converted to the linear equation

$$v' + [2p(x)Y(x) + q(x)]\, v = -p(x) \tag{52}$$

by the change of variables

$$y = Y(x) + \frac{1}{v}. \tag{53}$$

Thus, just as we are able to solve the Riccati equation only in exceptional cases, we are able to factor second-order nonconstant coefficient equations (and solve them readily) only in exceptional cases. In general, then, nonconstant-coefficient differential equations are hard in that we are unable to find closed form solutions.

EXAMPLE 8. Consider the equation

$$y'' - (x^2 + 1)y = 0. \tag{54}$$

Here $a_1(x) = 0$ and $a_2(x) = -(x^2 + 1)$, so (50a,b) are

$$a' = a^2 - x^2 - 1, \tag{55a}$$
$$b' = -b^2 + x^2 + 1. \tag{55b}$$

In this case we are lucky enough to notice the particular solution $a(x) = -x$ of (55a). Putting this result into (49a) then gives $b(x) = x$. [Equivalently, we could have noticed the particular solution $b(x) = x$ of (55b) and then obtained $a(x) = -x$ from (49a).] Thus, we have the factorization

$$y'' - (x^2 + 1)y = (D + x)(D - x)y = 0. \tag{56}$$

Proceeding as outlined above, we are able (Exercise 15) to derive the general solution

$$y(x) = Ae^{x^2/2} + Be^{x^2/2} \int e^{-x^2} dx. \tag{57}$$

Going one step further, suppose that initial conditions $y(0) = 0$ and $y'(0) = 1$ are prescribed and that we wish to evaluate A and B. First, we re-express (57) in the equivalent and more convenient form

$$y(x) = Ae^{x^2/2} + Be^{x^2/2} \int_0^x e^{-\xi^2} d\xi. \tag{58}$$

We could have used any lower integration limit, but 0 will be most convenient because the initial conditions are at $x = 0$. Then

$$\begin{aligned} y(0) &= 0 = (A)(1) + (B)(0), \\ y'(0) &= 1 = (A)(0) + (B)(1), \end{aligned}$$

where we have used the fundamental theorem of the calculus (Section 2.2) to differentiate the integral term. Thus, $A = 0$ and $B = 1$, so

$$y(x) = e^{x^2/2} \int_0^x e^{-\xi^2} d\xi \tag{59}$$

is the desired particular solution. ∎

The integral in (59) is nonelementary in that it cannot be evaluated in closed form in terms of the elementary functions. But it arises often enough so that it has been used to define a new function, the so-called **error function**

$$\boxed{\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-\xi^2} d\xi,} \tag{60}$$

where the $2/\sqrt{\pi}$ is included to normalize $\operatorname{erf}(x)$ to unity as $x \to \infty$ since (as will be shown in a later chapter)

$$\int_0^\infty e^{-\xi^2} d\xi = \frac{\sqrt{\pi}}{2}. \tag{61}$$

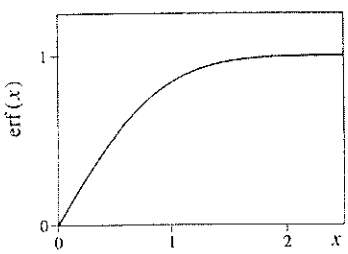


Figure 1. The error function $\operatorname{erf}(x)$.

The graph of $\operatorname{erf}(x)$ is shown in Fig. 1 for $x > 0$. For $x < 0$ we rely on the fact that $\operatorname{erf}(-x) = -\operatorname{erf}(x)$ (Exercise 18); for instance, $\operatorname{erf}(-\infty) = -\operatorname{erf}(\infty) = -1$. Since $e^{-\xi^2}$ is (to within a scale factor) a *normal probability distribution*, one way in which the error function arises is in the study of phenomena that are governed by normal distributions. For instance, we will encounter the error function when we study the movement of heat by the physical process of conduction.

Thus, our solution (59) can be re-expressed as $y(x) = \sqrt{\pi/2}\, e^{x^2/2} \operatorname{erf}(x)$. Just as we know the values of $\sin x$, its Taylor series, and its various properties, likewise we know the values of $\operatorname{erf}(x)$, its Taylor series, and its various properties, so

we should feel comfortable with $\operatorname{erf}(x)$ and regard it henceforth as a known function. Though not included among the so-called "elementary functions," it is one of many "special functions" that are now available in the engineering science and mathematics literature.

Closure. We have seen, in this section, that nonconstant-coefficient equations can be solved in closed form only in exceptional cases. The most important of these is the case of the Cauchy–Euler equation

$$x^n \frac{d^n y}{dx^n} + c_1 x^{n-1} \frac{d^{n-1} y}{dx^{n-1}} + \cdots + c_{n-1} x \frac{dy}{dx} + c_n y = 0.$$

Recall that a constant-coefficient equation necessarily admits at least one solution in the form $e^{\lambda x}$, and that in the case of a repeated root of order k the solutions corresponding to that root can be found by reduction of order to be $\left(C_1 + C_2 x + \cdots + C_k x^{k-1}\right) e^{\lambda x}$. Analogously, a Cauchy–Euler equation necessarily admits at least one solution in the form x^λ, and in the case of a repeated root of order k the solutions corresponding to that root can be found by reduction of order to be $\left[C_1 + C_2 \ln x + \cdots + C_k (\ln x)^{k-1}\right] x^\lambda$.

In fact, it turns out that the connection between constant-coefficient equations and Cauchy–Euler equations is even closer than that in as much as *any given Cauchy–Euler equation can be reduced to a constant-coefficient equation* by a change of independent variable according to $x = e^t$. Discussion of that point is reserved for the exercises.

Beyond our striking success with the Cauchy–Euler equation, other successes for nonconstant-coefficient equations are few and far between. For instance, we might be able to obtain one solution by inspection and others, from it, by reduction of order. Or, we might, in exceptional cases, be successful in factoring the differential operator but, again, such successes are exceptional. Thus, other lines of approach will be needed for nonconstant-coefficient equations, and they are developed in Chapters 4 and 6.

EXERCISES 3.6

1. Derive a general solution for the given Cauchy–Euler equation by seeking $y(x) = x^\lambda$. That is, derive the solution, rather than merely use any stated result such as Theorem 3.6.1. In addition, find the particular solution corresponding to the initial conditions, if such conditions are given, and state the interval of validity of that solution.

(a) $xy' + y = 0$
(b) $xy' - y = 0; \quad y(2) = 5$
(c) $xy'' + y' = 0$
(d) $xy'' - 4y' = 0; \quad y(1) = 0, \quad y'(1) = 3$
(e) $x^2y'' + xy' - 9y = 0; \quad y(2) = 1, \quad y'(2) = 2$
(f) $x^2y'' + xy' + y = 0; \quad y(1) = 1, \quad y'(1) = 0$
(g) $x^2y'' + 3xy' + 2y = 0; \quad y(1) = 0, \quad y'(1) = 2$
(h) $x^2y'' - 2y = 0; \quad y(-5) = 3, \quad y'(-5) = 0$
(i) $4x^2y'' + 5y = 0; \quad y(1) = 0, \quad y'(1) = 1$
(j) $x^2y'' + xy' + 4y = 0$
(k) $x^2y'' + 2xy' - 2y = 0; \quad y(3) = 2, \quad y'(3) = 2$
(l) $(x+2)^2y'' - y = 0$ HINT: Let $x + 2 = t$.
(m) $x^2y''' - 2y' = 0; \quad y(1) = 2, \quad y'(1) = y''(1) = 0$
(n) $xy''' - y'' = 0; \quad y(0) = 1, \quad y'(0) = y''(0) = 0$
(o) $x^2y'' + xy' - \kappa^2 y = 0$ (κ a constant)
(p) $x^3y''' + xy' - y = 0$

(q) $x^3y''' + 2xy' - 2y = 0$
(r) $x^2y''' + xy'' - y' = 0$
(s) $x^3y''' + 6x^2y'' + 7xy' + y = 0$
(t) $x^4y'''' + 6x^3y''' + 3x^2y'' - 3xy' + 4y = 0$
(u) $x^4y'''' + 6x^3y''' + 3x^2y'' - 3xy' + y = 0$
(v) $xy'''' - 2y''' = 0; \quad y(1) = 5,$
$y'(1) = y''(1) = y'''(1) = 0$

2. (a)–(v) For the corresponding problem in Exercise 1, use computer software to obtain a general solution, as well as a particular solution if initial conditions are given.

3. Putting (9) into (3), show that the equation $xA'' + A' = 0$ results, as claimed below equation (9).

4. Solve (10), and derive the general solution $A(x) = D \ln x + C$ stated below equation (10).

5. Fill in the steps between (26) and (27).

6. Prove that the two solutions within (33) [or, equivalently, (34)] are necessarily LI. You may assume that $a_1(x)$ is continuous on the x interval of interest. HINT: Recall the fundamental theorem of the calculus (given in Section 2.2).

7. It was stated in the Closure that *any given Cauchy–Euler equation can be reduced to a constant-coefficient equation by the change of variables* $x = e^t$. In this exercise we ask you to try that idea for some specific cases; in the next exercise we ask for a general proof of the italicized claim. Let $y(x(t)) \equiv Y(t)$, and let y' and Y' denote dy/dx and dY/dt, respectively.

(a) Show that the change of variables $x = e^t$ reduces the Cauchy–Euler equation $x^2y'' - xy' - 3y = 0$ to the constant-coefficient equation $Y'' - 2Y' - 3Y = 0$. Thus, show that $Y(t) = Ae^{-t} + Be^{3t}$. Since $t = \ln x$, show that $y(x) = Ax^{-1} + Bx^3$.
(b) Same as (a), for $x^2y'' + xy' - 4y = 0$.
(c) Same as (a), for $x^2y'' + xy' + 4y = 0$.
(d) Same as (a), for $x^2y'' + 3xy' + y = 0$.
(e) Same as (a), for $x^2y'' + xy' - 9y = 0$.
(f) Same as (a), for $x^2y'' + y = 0$.
(g) Same as (a), for $x^2y'' + 2xy' - 2y = 0$.
(h) Same as (a), for $4x^2y'' - y = 0$.

8. First, read the introduction to Exercise 7. Consider the general Cauchy–Euler equation

$$\left(x^nD^n + c_1x^{n-1}D^{n-1} + \cdots + c_{n-1}xD + c_n\right)y = 0, \tag{8.1}$$

where $D = d/dx$. Let $x = e^t$, and define $y(x(t)) \equiv Y(t)$.

(a) Using chain differentiation, show that

$$\begin{aligned} xDy &= DY, \\ x^2D^2y &= D(D-1)Y, \\ x^3D^3y &= D(D-1)(D-2)Y, \end{aligned} \tag{8.2}$$

where D acting on $y(x)$ means d/dx and D acting on $Y(t)$ means d/dt.
(b) The results (8.2) suggest that the formula

$$x^kD^ky = D(D-1)\cdots(D-k+1)Y, \tag{8.3}$$

holds for all positive integers k. Prove that (8.3) is indeed correct. HINT: Use mathematical induction. That is, assume that (8.3) holds for any given positive k and, by differentiating both sides with respect to x, show that

$$x^{k+1}D^{k+1}y = D(D-1)\cdots(D-k)Y, \tag{8.4}$$

which is the same as (8.3) but with k changed to $k+1$. Thus, it must be true that (8.3) holds for all positive integers k.
(c) Finally, replacing each x^kD^ky in (8.1) by the corresponding right-hand side of (8.3), state why the resulting differential equation on $Y(t)$ will be of constant-coefficient type.

9. (*Electric potential*) The electric potential Φ within an annular region such as a long metal pipe of inner radius r_1 and outer radius r_2, satisfies the differential equation

$$r\frac{d^2\Phi}{dr^2} + \frac{d\Phi}{dr} = 0. \qquad (r_1 < r < r_2)$$

Solve for the potential distribution $\Phi(r)$ subject to these boundary conditions:

(a) $\Phi(r_1) = \Phi_1, \quad \Phi(r_2) = \Phi_2$
(b) $\dfrac{d\Phi}{dr}(r_1) = 0, \quad \Phi(r_2) = \Phi_2$

10. (*Steady-state temperature distribution*) The steady-state temperature distribution u within a hollow sphere, of inner radius r_1 and outer radius r_2, is governed by the differential equation

$$r\frac{d^2u}{dr^2} + 2\frac{du}{dr} = 0.$$

Solve for $u(r)$ subject to these boundary conditions:

(a) $u(r_1) = u_1, \quad u(r_2) = u_2$
(b) $\dfrac{du}{dr}(r_1) = 3, \quad u(r_2) = 0$
(c) $u(r_1) = u_1, \quad \dfrac{du}{dr}(r_2) = 0$

EXERCISES FOR OPTIONAL SECTION 3.6.2

11. Use the given solution $y_1(x)$ of the differential equation to find the general solution by the method of reduction of order (leaving the second solution in integral form if necessary).

(a) $y'' + xy' - y = 0$; $y_1(x) = x$
(b) $xy'' + xy' - y = 0$; $y_1(x) = x$
(c) $3xy'' - xy' + y = 0$; $y_1(x) = x$
(d) $(x^2 - 1)y'' - 2y = 0$; $y_1(x) = x^2 - 1$
(e) $2y'' + xy' - 2y = 0$; $y_1(x) = x^2 + 2$

12. (a)–(e) Obtain a general solution of the corresponding differential equation in Exercise 11, using computer software.

EXERCISES FOR OPTIONAL SECTION 3.6.3

13. State, clearly and convincingly, logic by which (49a,b) follow from (48).

14. Fill in the steps between (49a,b) and (50a,b).

15. Provide the steps that are missing between the equation (56) and its solution (57).

16. If $a_1(x)$ and $a_2(x)$ are constants, then the factorization (47) should be simple. Show that the Riccati equations (50a,b) on a and b do indeed give the same results for a and b as can be obtained by more elementary means.

17. In general, the Riccati-type equations (50a,b) are hard. However, we should be able to solve them if the given nonconstant-coefficient equation $y'' + a_1(x)y' + a_2(x)y = 0$ is a Cauchy–Euler equation because that case is simple. Thus, use the method of factoring the operator for these equations:

(a) $x^2y'' - 2xy' + 2y = 0$
(b) $x^2y'' + xy' + 9y = 0$
(c) $x^2y'' + xy' - 9y = 0$
(d) $x^2y'' + 5xy' + 4y = 0$

18. From its integral definition, (60), show that $\text{erf}(-x) = -\text{erf}(x)$.

19. (*Integral representations*) The notion of an integral representation of a function, as used in (60) to define the error function $\text{erf}(x)$, might be unfamiliar to you. If so, it might help to point out that even the elementary functions can be introduced in that manner. For example, one can define the logarithm $\ln x$ as

$$\ln x = \int_1^x \frac{dt}{t}, \qquad (x > 0) \tag{19.1}$$

from which formula the values of $\ln x$ can be derived (by numerical integration), and its various properties derived as well.

(a) To illustrate the latter claim, use (19.1) to derive the well known property $\ln x^a = a \ln x$ of the logarithm.
(b) Likewise, use (19.1) to derive the property $\ln(xy) = \ln x + \ln y$.

3.7 Solution of Nonhomogeneous Equation

Thus far, for differential equations of second-order and higher, we have studied only the homogeneous equation $L[y] = 0$. In this section we turn to the *non*homogeneous case

$$L[y] = f(x), \tag{1}$$

where L is an nth-order linear differential operator. That is, this time we include a nonzero forcing function $f(x)$.

Before proceeding with solution techniques, let us reiterate that the function $f(x)$ (that is, what's left on the right-hand side when the terms involving y and its derivatives are put on the left-side side) is, essentially, a **forcing function**, and we will call it that, in this text, as a reminder of its physical significance.

For instance, we have already met (Section 1.3) the equation

$$m\frac{d^2x}{dt^2} + c\frac{dx}{dt} + kx = F(t) \tag{2}$$

governing the displacement $x(t)$ of a mechanical oscillator. Here, the forcing function is the applied force $F(t)$. For the analogous electrical oscillator (Section 2.3),

governed by the equations

$$L\frac{d^2i}{dt^2} + R\frac{di}{dt} + \frac{1}{C}i = \frac{dE(t)}{dt} \tag{3}$$

on the current $i(t)$, and

$$L\frac{d^2Q}{dt^2} + R\frac{dQ}{dt} + \frac{1}{C}Q = E(t) \tag{4}$$

on the charge $Q(t)$ on the capacitor, the forcing functions are the time derivative of the applied voltage $E(t)$, and the applied voltage $E(t)$, respectively.

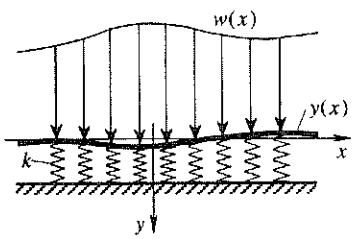


Figure 1. Beam on elastic foundation.

As one more example, we give (without derivation) the differential equation

$$EI\frac{d^4y}{dx^4} + ky = w(x) \tag{5}$$

governing the deflection $y(x)$ of a beam that rests upon an elastic foundation, under a load distribution $w(x)$ (i.e., load per unit x length), as sketched in Fig. 1. E, I, and k are known physical constants: E is the Young's modulus of the beam material, I is the inertia of the beam's cross section, and k is the spring stiffness per unit length (force per unit length per unit length) of the foundation. Thus, in this case the forcing function is $w(x)$, the applied load distribution. [Derivation of (5) involves the so-called Euler beam theory and is part of a first course in solid mechanics.]

3.7.1. General solution. Remember that the general solution of the homogeneous equation $L[y] = 0$ is a family of solutions that contains every solution of that equation, over the interval of interest. Likewise, by the general solution of the nonhomogeneous equation $L[y] = f$, we mean a family of solutions that contains every solution of that equation, over the interval of interest.

Like virtually all of the concepts and methods developed in this chapter, the concepts that follow rest upon the assumed linearity of the differential equation (1), in particular, upon the fact that if L is linear then

$$L\left[\alpha u(x) + \beta v(x)\right] = \alpha L[u(x)] + \beta L[v(x)] \tag{6}$$

for any two functions u, v (n-times differentiable, of course, if L is an nth-order operator) and any constants α, β. Indeed, recall the analogous result for any number of functions:

$$L\left[\alpha_1 u_1(x) + \cdots + \alpha_k u_k(x)\right] = \alpha_1 L[u_1(x)] + \cdots + \alpha_k L[u_k(x)] \tag{7}$$

for any functions $u_1, \ldots, u_k$ and constants $\alpha_1, \ldots, \alpha_k$.

To begin, we suppose that $y_h(x)$ is a general solution of the homogeneous version of (1), $L[y] = 0$, and that $y_p(x)$ is any particular solution of (1): $L[y_p(x)] = f(x)$. That is, $y_p(x)$ is any function which, when put into the left-hand side of (1), gives $f(x)$. We will refer to $y_h(x)$ and $y_p(x)$ as **homogeneous** and **particular solutions** of (1), respectively. [Some authors write $y_c(x)$ in place of $y_h(x)$, and call it the **complementary solution**.]

THEOREM 3.7.1 *General Solution of* $L[y] = f$
If $y_h(x)$ and $y_p(x)$ are homogeneous and particular solutions of (1), respectively, on an interval I, then a general solution of (1), on I, is

$$y(x) = y_h(x) + y_p(x). \tag{8}$$

Proof: That (8) satisfies (1) follows from the linearity of (1):

$$\begin{aligned} L[y_h(x) + y_p(x)] &= L[y_h(x)] + L[y_p(x)] \\ &= 0 + f(x) = f(x), \end{aligned}$$

where the first equality follows from (6), with $\alpha = \beta = 1$, and u, v equal to y_h and y_p, respectively.

To see that it is a general solution, let y be any solution of (1). Again using the linearity of L, we have

$$L[y - y_p] = L[y] - L[y_p] = f - f = 0,$$

so that the most general $y - y_p$ is a linear combination of a fundamental set of solutions of the homogeneous version of (1), namely y_h. Hence $y = y_h + y_p$ is a general solution of (1). ■

Thus, to solve the nonhomogeneous equation (1) we need to augment the homogeneous solution $y_h(x)$ by adding to it any particular solution $y_p(x)$.

Often, in applications, $f(x)$ is not a single term but a linear combination of terms: $f(x) = f_1(x) + \cdots + f_k(x)$. In the equation $L[y] = 5x^2 - 2\sin x + 6$, for instance, we can identify $f_1(x) = 5x^2$, $f_2 = -2\sin x$, and $f_3(x) = 6$.

THEOREM 3.7.2 *General Solution of* $L[y] = f_1 + \cdots + f_k$
If $y_h(x)$ is a general solution of $L[y] = 0$ on an interval I, and $y_{p1}(x), \ldots, y_{pk}(x)$ are particular solutions of $L[y] = f_1, \ldots, L[y] = f_k$ on I, respectively, then a general solution of $L[y] = f_1 + \cdots + f_k$ on I is

$$y(x) = y_h(x) + y_{p1}(x) + \cdots + y_{pk}(x). \tag{9}$$

Proof: That (9) satisfies (1) follows from (7), with all the α's equal to 1:

$$\begin{aligned} L[y_h + y_{p1} + \cdots + y_{pk}] &= L[y_h] + L[y_{p1}] + \cdots + L[y_{pk}] \\ &= 0 + f_1 + \cdots + f_k \\ &= f_1 + \cdots + f_k, \end{aligned} \tag{10}$$

as was to be verified.

To see that it is a *general* solution of (1), let y be any solution of (1). Then

$$\begin{aligned} L[y - y_{p1} - \cdots - y_{pk}] &= L[y] - L[y_{p1}] - \cdots - L[y_{pk}] \\ &= f - f_1 - \cdots - f_k = 0, \end{aligned}$$

so the most general $y - y_{p1} - \cdots - y_{pk}$ is a general solution y_h of the homogeneous version of (1). ■

This result is a **superposition principle**. It tells us that the response y_p to a superposition of inputs (the forcing functions $f_1, \ldots, f_k$) is the superposition of their individual outputs $(y_{p1}, \ldots, y_{pk})$.

The upshot is that to solve a nonhomogeneous equation we need to find both the homogeneous solution y_h and a particular solution y_p. Having already developed methods for determining y_h – at least for certain types of equations – we now need to present methods for determining particular solutions y_p, and that is the subject of Sections 3.7.2 and 3.7.3 that follow.

3.7.2. Undetermined coefficients. The method of undetermined coefficients is a procedure for determining a particular solution to the linear equation

$$\begin{aligned} L[y] &= f(x) \\ &= f_1(x) + \cdots + f_k(x), \end{aligned} \tag{11}$$

subject to two conditions:

(i) Besides being linear, L is of constant-coefficient type.

(ii) Repeated differentiation of each $f_j(x)$ term in (11) produces only a finite number of LI (linearly independent) terms.

To explain the latter condition, consider $f_j(x) = 2xe^{-x}$. The sequence consisting of this term and its successive derivatives is

$$2xe^{-x} \longrightarrow \left\{2xe^{-x},\ 2e^{-x} - 2xe^{-x},\ -4e^{-x} + 2xe^{-x},\ \ldots\right\},$$

and we can see that this sequence contains only the two LI functions e^{-x} and xe^{-x}. Thus, $f_j(x) = 2xe^{-x}$ satisfies condition (ii).

As a second example, consider $f_j(x) = x^2$. This term generates the sequence

$$x^2 \longrightarrow \left\{x^2, 2x, 2, 0, 0, \ldots\right\},$$

which contains only the three LI functions x^2, x, and 1. Thus, $f_j(x) = x^2$ satisfies condition (ii).

The term $f_j(x) = 1/x$, however, generates the sequence

$$1/x \longrightarrow \left\{1/x, -1/x^2, 2/x^3, -6/x^4, \ldots\right\},$$

which contains an infinite number of LI terms $(1/x, 1/x^2, 1/x^3, \ldots)$. Thus, $f_j(x) = 1/x$ does not satisfy condition (ii).

If the term $f_j(x)$ does satisfy condition (ii), then we will call the finite set of LI terms generated by it, through repeated differentiation, the **family generated by $f_j(x)$**. (That terminology is for convenience here and is not standard.) Thus, the family generated by $2xe^{-x}$ is comprised of e^{-x} and xe^{-x}, and the family generated by $3x^2$ is comprised of x^2, x, and 1.

Let us now illustrate the method of undetermined coefficients.

EXAMPLE 1. Consider the differential equation

$$y'''' - y'' = 3x^2 - \sin 2x. \tag{12}$$

First, we see that L is linear, with constant coefficients, so condition (i) is satisfied. Next, we identify $f_1(x)$, $f_2(x)$, and their generated sequences as

$$f_1(x) = 3x^2 \longrightarrow \left\{3x^2, 6x, 6, 0, 0, \ldots\right\}, \tag{13a}$$
$$f_2(x) = -\sin 2x \longrightarrow \left\{-\sin 2x, -2\cos 2x, 4\sin 2x, \ldots\right\}. \tag{13b}$$

Thus, f_1 and f_2 do generate the finite families

$$f_1(x) = 3x^2 \longrightarrow \left\{x^2, x, 1\right\}, \tag{14a}$$
$$f_2(x) = -\sin 2x \longrightarrow \left\{\sin 2x, \cos 2x\right\}, \tag{14b}$$

so condition (ii) is satisfied.

To find a particular solution y_{p1} corresponding to f_1, tentatively seek it as a linear combination of the terms in (14a):

$$y_{p1}(x) = Ax^2 + Bx + C, \tag{15}$$

where A, B, C are the so-called **undetermined coefficients**. Next, we write down the homogeneous solution of (12),

$$y_h(x) = C_1 + C_2x + C_3e^x + C_4e^{-x}, \tag{16}$$

and check each term in y_{p1} [i.e., in (15)] for duplication with terms in y_h. Doing so, we find that the Bx and C terms in (15) duplicate (to within constant scale factors) the C_2x and C_1 terms, respectively, in (16). The method then calls for multiplying the entire family, involved in the duplication, by the lowest positive integer power of x needed to eliminate all such duplication. Thus, we revise (15) as

$$y_{p1}(x) = x\left(Ax^2 + Bx + C\right) = Ax^3 + Bx^2 + Cx, \tag{17}$$

but find that the Cx term in (17) is still "in violation" in that it duplicates the C_2x term in (16). Thus, try

$$y_{p1}(x) = x^2\left(Ax^2 + Bx + C\right) = Ax^4 + Bx^3 + Cx^2. \tag{18}$$

This time we are done, since all duplication has now been eliminated.

Next, we put the final revised form (18) into the equation $y'''' - y'' = 3x^2$ [i.e., $L[y] = f_1(x)$] and obtain

$$24A - 12Ax^2 - 6Bx - 2C = 3x^2. \tag{19}$$

Finally, equating coefficients of like terms gives

$$\begin{aligned} x^2 : &\quad -12A = 3 \\ x : &\quad -6B = 0 \\ 1 : &\quad 24A - 2C = 0, \end{aligned} \tag{20}$$

so that $A = -1/4$, $B = 0$, $C = -3$. Thus

$$y_{p1}(x) = -\frac{1}{4}x^4 - 3x^2. \tag{21}$$

Next, we need to find y_{p2} corresponding to f_2. To do so, we seek it as a linear combination of the terms in (14b):

$$y_{p2}(x) = D\sin 2x + E\cos 2x. \tag{22}$$

Checking each term in (22) for duplication with terms in y_h, we see that there is no such duplication. Thus, we accept the form (22), put it into the equation $y'''' - y'' = -\sin 2x$ [i.e., $L[y] = f_2(x)$], and obtain

$$20D\sin 2x + 20E\cos 2x = -\sin 2x. \tag{23}$$

Equating coefficients of like terms gives $20D = -1$, and $20E = 0$, so that $D = -1/20$ and $E = 0$. Thus,

$$y_{p2}(x) = -\frac{1}{20}\sin 2x. \tag{24}$$

Finally, a general solution of (12) is, according to Theorem 3.7.2,

$$y(x) = y_h(x) + y_p(x) = y_h(x) + y_{p1}(x) + y_{p2}(x),$$

namely,

$$y(x) = C_1 + C_2x + C_3e^x + C_4e^{-x} - \frac{1}{4}x^4 - 3x^2 - \frac{1}{20}\sin 2x. \tag{25}$$

COMMENT 1. We obtained (20) by "equating coefficients of like terms" in (19). That step amounted to using the concept of linear independence – namely, noting that $1, x, x^2$ are LI (on any given x interval) and then using Theorem 3.2.6. Alternatively, we could have rewritten (19) as

$$(24A - 2C)1 + (-6B)x + (-12A - 3)x^2 = 0 \tag{26}$$

and used the linear independence of $1, x, x^2$ to infer that the coefficient of each term must be zero, which step once again gives (20).

COMMENT 2. The key point in the analysis is that the system (20), consisting of three linear algebraic equations in the three unknowns A, B, C, is consistent. Similarly for the system $20D = -1$, $20E = 0$ governing D and E. The guarantee provided by the method

of undetermined coefficients is that the resulting system of linear algebraic equations on the unknown coefficients will indeed be consistent, so that we can successfully solve for the undetermined coefficients. What would have happened if we had used the form (15) for $y_p(x)$ instead of (18) – that is, if we had not adhered to the prescribed procedure? We would have obtained, in place of (20), the equations

$$\begin{aligned} x^2: &\quad 0 = 3 \\ x: &\quad 0 = 0 \\ 1: &\quad -2A = 0, \end{aligned}$$

which are *in*consistent because the "equation" $0 = 3$ cannot be satisfied by any choice of A, B, C. That is, (15) would not have worked. ∎

Let us summarize.

STEPS IN THE METHOD OF UNDETERMINED COEFFICIENTS:

1. Verify that condition (i) is satisfied.
2. Identify the $f_j(x)$'s and verify that each one satisfies condition (ii).
3. Determine the finite family corresponding to each $f_j(x)$ [(14a,b) in Example 1].
4. Seek $y_{p1}(x)$, tentatively, as a linear combination of the terms in the family corresponding to $f_1(x)$ [(15) in Example 1].
5. Obtain the general solution $y_h(x)$ of the homogeneous equation [(16) in Example 1].
6. If such duplication is found, multiply the entire linear combination of terms by the lowest positive integer power of x necessary to remove all such duplication between those terms and the terms in y_h [(18) in Example 1].
7. Substitute the final version of the form assumed for $y_{p1}(x)$ into the left-hand side of the equation $L[y] = f_1$, and equate coefficients of like terms.
8. Solve the resulting system of linear algebraic equations for the undetermined coefficients. That step completes our determination of $y_{p1}(x)$.
9. Repeat steps 4–8 for $y_{p2}, \ldots, y_{pk}$.
10. Then the general solution of $L[y] = f_1 + \cdots + f_k$ is given, according to Theorem 3.7.2, by $y(x) = y_h(x) + y_p(x) = y_h(x) + y_{p1}(x) + \cdots + y_{pk}(x)$.

EXAMPLE 2. As a final example, consider

$$y'' - 9y = 4 + 5\sinh 3x, \tag{27}$$

which is indeed linear and of constant-coefficient type. Since

$$f_1(x) = 4 \longrightarrow \{4, 0, 0, \ldots\},$$
$$f_2(x) = 5\sinh 3x \longrightarrow \{5\sinh 3x, 15\cosh 3x, 45\sinh 3x, \ldots\},$$

we see that these terms generate the finite families

$$f_1(x) = 4 \longrightarrow \{1\},$$
$$f_2(x) = 5\sinh 3x \longrightarrow \{\sinh 3x, \cosh 3x\},$$

so we tentatively seek

$$y_{p1}(x) = A. \tag{28}$$

Since

$$y_h(x) = C_1 e^{3x} + C_2 e^{-3x}, \tag{29}$$

there is no duplication between the term in (28) and those in (29). Putting (28) into $y'' - 9y = 4$ gives $-9A = 4$, so $A = -4/9$. Thus,

$$y_{p1}(x) = -\frac{4}{9}. \tag{30}$$

Next, we tentatively seek

$$y_{p2}(x) = B\sinh 3x + C\cosh 3x. \tag{31}$$

At first glance, it appears that there is no duplication between any of the terms in (31) and those in (29). However, since the $\sinh 3x$ and $\cosh 3x$ are linear combinations of e^{3x} and e^{-3x}, we do indeed have duplication. Said differently, each of the $\sinh 3x$ and $\cosh 3x$ terms are solutions of the homogeneous equation. Thus, we need to multiply the right-hand side of (31) by x and revise y_p as

$$y_{p2}(x) = x\left(B\sinh 3x + C\cosh 3x\right). \tag{32}$$

Now that y_{p2} is in a satisfactory form we put that form into $y'' - 9y = 5\sinh 3x$ [i.e., $L[y] = f_2(x)$] and obtain the equation

$$(3C + 3C)\sinh 3x + (3B + 3B)\cosh 3x$$
$$+(9B - 9B)x\sinh 3x + (9C - 9C)x\cosh 3x = 5\sinh 3x.$$

Equating coefficients of like terms gives $B = 5/6$ and $C = 0$, so

$$y_{p2}(x) = \frac{5}{6}x\sinh 3x. \tag{33}$$

It follows then, from Theorem 3.7.2, that a general solution of (27) is

$$y(x) = C_1 e^{3x} + C_2 e^{-3x} - \frac{4}{9} + \frac{5}{6}x\sinh 3x. \tag{34}$$

Naturally, one's final result can (and should) be checked by direct substitution into the original differential equation.

COMMENT. Suppose that in addition to the differential equation (27), initial conditions $y(0) = 0, y'(0) = 2$ are specified. Imposing these conditions on (34) gives $C_1 = 5/9, C_2 = -1/9$, and hence the particular solution

$$y(x) = \frac{5}{9}e^{3x} - \frac{1}{9}e^{-3x} - \frac{4}{9} + \frac{5}{6}x \sinh 3x. \tag{35}$$

Do not be concerned that we call (35) a particular solution even though each of the two exponential terms in (35) is a homogeneous solution because if we put (35) into the left-hand side of (27) it does give the right-hand side of (27); thus, it is a particular solution of (27). ∎

As a word of caution, suppose the differential equation is

$$y'' - 3y' + 2y = 2\sinh x,$$

with homogeneous solution $C_1e^x + C_2e^{2x}$. Observe that $2\sinh x = e^x - e^{-x}$ contains an e^x term, which corresponds to one of the homogeneous solutions. To bring this duplication into the light, we should re-express the differential equation as

$$y'' - 3y' + 2y = e^x - e^{-x}$$

before beginning the method of undetermined coefficients. Then, the particular solution due to $f_1(x) = e^x$ will be $y_{p1}(x) = Axe^x$ and the assumed particular solution due to $f_2(x) = e^{-x}$ will be $y_{p2}(x) = Be^{2x}$. We find that $A = -1$ and $B = -1/6$ so the general solution is

$$y(x) = C_1e^x + C_2e^{2x} - xe^x - \frac{1}{6}e^{-x}.$$

Closing this discussion of the method of undetermined coefficients, let us reconsider condition (ii), that repeated differentiation of each term in the forcing function must produce only a finite number of LI terms. How broad is the class of functions that satisfy that condition? If a forcing function f satisfies that condition, then it must be true that coefficients a_j exist, not all of them zero, such that

$$a_0f^{(N)} + a_1f^{(N-1)} + \cdots + a_{N-1}f' + a_Nf = 0 \tag{36}$$

over the x interval under consideration. From our discussion of the solution of such constant-coefficient equations we know that solutions f of (36) must be of the form $Cx^me^{(\alpha+i\beta)x}$, or a linear combination of such terms. Such functions are so common in applications that condition (ii) is not as restrictive as it may seem.

3.7.3. Variation of parameters. Although easy to apply, the method of undetermined coefficients is limited by the two conditions cited above – that L be of

constant-coefficient type, and that repeated differentiation of each $f_j(x)$ forcing term produces only a finite number of LI terms.

The method of variation of parameters, due to Lagrange, is more powerful in that it is not subject to those restrictions. As with automobile engines, we can expect more power to come at a higher price and, as we shall see, Lagrange's method is indeed the more difficult to apply.

In fact, we have already presented the method, in Section 2.2, for the general linear first-order equation

$$y' + p(x)y = q(x), \tag{37}$$

and we urge you to review that discussion. The idea was to seek a particular solution y_p by varying the parameter A (i.e., the constant of integration) in the homogeneous solution

$$y_h(x) = Ae^{-\int p(x)\,dx}.$$

Thus, we sought

$$y_p(x) = A(x)e^{-\int p(x)\,dx},$$

put that form into (37) and solved for $A(x)$.

Likewise, if an nth-order linear differential equation $L[y] = f$ has a homogeneous solution

$$y_h(x) = C_1y_1(x) + \cdots + C_ny_n(x), \tag{38}$$

then according to the method of variation of parameters we seek a particular solution in the form

$$y_p(x) = C_1(x)y_1(x) + \cdots + C_n(x)y_n(x); \tag{39}$$

that is, we "vary the parameters" (constants of integration) $C_1, \ldots, C_n$ in (38).

Let us carry out the procedure for the linear second-order equation

$$L[y] = y'' + p_1(x)y' + p_2(x)y = f(x), \tag{40}$$

where the coefficients $p_1(x)$ and $p_2(x)$ are assumed to be continuous on the x interval of interest, say I. We suppose that

$$y_h(x) = C_1y_1(x) + C_2y_2(x) \tag{41}$$

is a known general solution of the homogeneous equation on I, and seek

$$y_p(x) = C_1(x)y_1(x) + C_2(x)y_2(x). \tag{42}$$

Needing y_p' and y_p'', to substitute into (40), we differentiate (42):

$$y_p' = C_1y_1' + C_2y_2' + C_1'y_1 + C_2'y_2. \tag{43}$$

Looking ahead, y_p'' will include $C_1, C_2, C_1', C_2', C_1'', C_2''$ terms, so that (40) will become a nonhomogeneous second-order differential equation in C_1 and C_2, which can hardly be expected to be simpler than the original equation (40)! However, it

will be only one equation in the two unknowns C_1, C_2, so we are free to impose another condition on C_1, C_2 to complete, and simplify, the system.

An especially convenient condition to impose will be

$$C_1'y_1 + C_2'y_2 = 0, \tag{44}$$

for this condition will knock out the C_1', C_2' terms in (43), so that y_p'' will contain only *first*-order derivatives of C_1 and C_2. Then (43) reduces to $y_p' = C_1y_1' + C_2y_2'$, so

$$y_p'' = C_1y_1'' + C_2y_2'' + C_1'y_1' + C_2'y_2', \tag{45}$$

and (40) becomes

$$C_1\left(y_1'' + p_1y_1' + p_2y_1\right) + C_2\left(y_2'' + p_1y_2' + p_2y_2\right) + C_1'y_1' + C_2'y_2' = f. \tag{46}$$

The two parenthetic groups vanish by virtue of y_1 and y_2 being solutions of the homogeneous equation $L[y] = 0$, so (46) simplifies to $C_1'y_1' + C_2'y_2' = f$. That result, together with (44), gives us the equations

$$\begin{aligned} y_1C_1' + y_2C_2' &= 0, \\ y_1'C_1' + y_2'C_2' &= f \end{aligned} \tag{47}$$

on C_1', C_2'. The latter will be solvable for C_1', C_2', uniquely, if the determinant of the coefficients does not vanish on I. In fact, we recognize that determinant as the Wronskian of y_1 and y_2,

$$W[y_1, y_2](x) = \begin{vmatrix} y_1(x) & y_2(x) \\ y_1'(x) & y_2'(x) \end{vmatrix}, \tag{48}$$

and the latter is necessarily nonzero on I by Theorem 3.2.3 because y_1 and y_2 are LI solutions of $L[y] = 0$.

Solving (47) by Cramer's rule gives

$$C_1'(x) = \frac{\begin{vmatrix} 0 & y_2 \\ f & y_2' \end{vmatrix}}{\begin{vmatrix} y_1 & y_2 \\ y_1' & y_2' \end{vmatrix}} = \frac{W_1(x)}{W(x)}, \quad C_2'(x) = \frac{\begin{vmatrix} y_1 & 0 \\ y_1' & f \end{vmatrix}}{\begin{vmatrix} y_1 & y_2 \\ y_1' & y_2' \end{vmatrix}} = \frac{W_2(x)}{W(x)},$$

where W_1, W_2 simply denote the determinants in the numerators. Integrating these equations and putting the results into (42) gives

$$\boxed{y_p(x) = \left[\int^x \frac{W_1(\xi)}{W(\xi)}\,d\xi\right] y_1(x) + \left[\int^x \frac{W_2(\xi)}{W(\xi)}\,d\xi\right] y_2(x),} \tag{49}$$

or, more compactly,

$$y_p(x) = \int^x \frac{W_1(\xi)y_1(x) + W_2(\xi)y_2(x)}{W(\xi)}\, d\xi. \tag{50}$$

EXAMPLE 3. To solve

$$y'' - 4y = 8e^{2x}, \tag{51}$$

we note the general solution

$$y_h(x) = C_1e^{2x} + C_2e^{-2x} \tag{52}$$

of the homogeneous equation, so that we may take $y_1(x) = e^{2x}, y_2(x) = e^{-2x}$. Then $W(x) = y_1y_2' - y_1'y_2 = -4, W_1(x) = -f(x)y_2(x) = -8e^{2x}e^{-2x} = -8$, and $W_2(x) = f(x)y_1(x) = 8e^{2x}e^{2x} = 2e^{4x}$, so (49) gives

$$\begin{aligned} y_p(x) &= \left(\int^x 2\, d\xi\right) e^{2x} + \left(\int^x -2e^{4\xi}\, d\xi\right) e^{-2x} \\ &= (2x + A)e^{2x} + \left(-\frac{e^{4x}}{2} + B\right) e^{-2x} = 2xe^{2x} - \frac{e^{2x}}{2} + Ae^{2x} + Be^{-2x}, \end{aligned} \tag{53}$$

where A, B are the arbitrary constants of integration. We can omit the A, B terms in (53) because they give terms (Ae^{2x} and Be^{-2x}) that merely duplicate those already present in the homogeneous solution y_h. That will always be the case: we can omit the constants of integration in the two integrals in (49). In the present example we can even drop the $-e^{2x}/2$ term in the right side of (53) since it too is a homogeneous solution [and can be absorbed in the C_1e^{2x} term in (55)]. Thus, we write

$$y_p(x) = 2xe^{2x}. \tag{54}$$

Finally,

$$y(x) = y_h(x) + y_p(x) = C_1e^{2x} + C_2e^{-2x} + 2xe^{2x} \tag{55}$$

gives a general solution of (51). ∎

3.7.4. Variation of parameters for higher-order equations. (Optional) For higher-order equations the idea is essentially the same. For the third-order equation

$$L[y] = y''' + p_1(x)y'' + p_2(x)y' + p_3(x)y = f(x), \tag{56}$$

for instance, if

$$y_h(x) = C_1y_1(x) + C_2y_2(x) + C_3y_3(x) \tag{57}$$

is known, then we seek

$$y_p(x) = C_1(x)y_1(x) + C_2(x)y_2(x) + C_3(x)y_3(x). \tag{58}$$

Looking ahead, when we put (58) into (56) we will have one equation in the three unknown functions C_1, C_2, C_3, so we can impose two additional conditions

on C_1, C_2, C_3 to complete, and simplify, the system. Proceeding, differentiation of (58) gives

$$y_p'(x) = C_1 y_1' + C_2 y_2' + C_3 y_3' + C_1' y_1 + C_2' y_2 + C_3' y_3, \tag{59}$$

so we set

$$C_1' y_1 + C_2' y_2 + C_3' y_3 = 0 \tag{60}$$

to suppress higher-order derivatives of C_1, C_2, C_3. Then (59) reduces to

$$y_p' = C_1 y_1' + C_2 y_2' + C_3 y_3', \tag{61}$$

and another differentiation gives

$$y_p'' = C_1 y_1'' + C_2 y_2'' + C_3 y_3'' + C_1' y_1' + C_2' y_2' + C_3' y_3'. \tag{62}$$

Again, to suppress higher-order derivatives of C_1, C_2, C_3, set

$$C_1' y_1' + C_2' y_2' + C_3' y_3' = 0. \tag{63}$$

Then (62) reduces to

$$y_p'' = C_1 y_1'' + C_2 y_2'' + C_3 y_3'', \tag{64}$$

so

$$y_p''' = C_1 y_1''' + C_2 y_2''' + C_3 y_3''' + C_1' y_1'' + C_2' y_2'' + C_3' y_3''. \tag{65}$$

Finally, putting (65), (64), (61), and (58) into (56) gives

$$\begin{aligned} C_1 \left(y_1''' + p_1 y_1'' + p_2 y_1' + p_3 y_1\right) + C_2 \left(y_2''' + p_1 y_2'' + p_2 y_2' + p_3 y_2\right) \\ + C_1 \left(y_3''' + p_1 y_3'' + p_2 y_3' + p_3 y_3\right) + C_1' y_1'' + C_2' y_2'' + C_3' y_3'' = f, \end{aligned} \tag{66}$$

or

$$C_1' y_1'' + C_2' y_2'' + C_3' y_3'' = f \tag{67}$$

since each of the three parenthetic groups in (66) vanishes because y_1, y_2, y_3 are homogeneous solutions.

Equations (60), (63), and (67) are three linear algebraic equations in C_1', C_2', C_3':

$$\begin{aligned} y_1 C_1' + y_2 C_2' + y_3 C_3' &= 0, \\ y_1' C_1' + y_2' C_2' + y_3' C_3' &= 0, \\ y_1'' C_1' + y_2'' C_2' + y_3'' C_3' &= f. \end{aligned} \tag{68}$$

That system is uniquely solvable for C_1', C_2', C_3' because the determinant of the coefficients is the Wronskian determinant of y_1, y_2, y_3,

$$W[y_1, y_2, y_3](x) = \begin{vmatrix} y_1 & y_2 & y_3 \\ y_1' & y_2' & y_3' \\ y_1'' & y_2'' & y_3'' \end{vmatrix}, \tag{69}$$

which is nonzero on the interval I because y_1, y_2, y_3 are LI on I by the assumption that (57) is a general solution of the homogeneous equation $L[y] = 0$. Solving (68) by means of Cramer's rule gives

$$C_1' = \frac{\begin{vmatrix} 0 & y_2 & y_3 \\ 0 & y_2' & y_3' \\ f & y_2'' & y_3'' \end{vmatrix}}{W(x)} = \frac{W_1(x)}{W(x)}, \qquad C_2' = \frac{\begin{vmatrix} y_1 & 0 & y_3 \\ y_1' & 0 & y_3' \\ y_1'' & f & y_3'' \end{vmatrix}}{W(x)} = \frac{W_2(x)}{W(x)},$$
$$C_3' = \frac{\begin{vmatrix} y_1 & y_2 & 0 \\ y_1' & y_2' & 0 \\ y_1'' & y_2'' & f \end{vmatrix}}{W(x)} = \frac{W_3(x)}{W(x)}. \tag{70}$$

Finally, integrating these equations and putting the results into (58) gives the particular solution

$$\boxed{\begin{aligned} y_p(x) = &\left[\int^x \frac{W_1(\xi)}{W(\xi)}\,d\xi\right] y_1(x) + \left[\int^x \frac{W_2(\xi)}{W(\xi)}\,d\xi\right] y_2(x) \\ &+ \left[\int^x \frac{W_3(\xi)}{W(\xi)}\,d\xi\right] y_3(x). \end{aligned}} \tag{71}$$

EXAMPLE 4. Consider the nonhomogeneous Cauchy–Euler equation

$$x^3 y''' + x^2 y'' - 2xy' + 2y = \frac{2}{x}. \qquad (0 < x < \infty) \tag{72}$$

Observe that we cannot use the method of undetermined coefficients in this case because the differential operator is not of constant-coefficient type, and also because the forcing function does not generate only a finite number of linearly-independent derivatives.

To use (71), we need to know y_1, y_2, y_3, and f. It is readily found that the homogeneous solution is

$$y_h(x) = C_1 \frac{1}{x} + C_2 x + C_3 x^2, \tag{73}$$

so we can take $y_1 = 1/x$, $y_2 = x$, $y_3 = x^2$. But be careful: $f(x)$ is not $2/x$ because (72) is not yet in the form of (56). That is, (72) must be divided by x^3 so that the coefficient of y''' becomes 1, as in (56). Doing so, it follows that $f(x) = 2/x^4$.

We can now evaluate the determinants needed in (71):

$$W(x) = \begin{vmatrix} x^{-1} & x & x^2 \\ -x^{-2} & 1 & 2x \\ 2x^{-3} & 0 & 2 \end{vmatrix} = \frac{6}{x}, \qquad W_1(x) = \begin{vmatrix} 0 & x & x^2 \\ 0 & 1 & 2x \\ 2x^{-4} & 0 & 2 \end{vmatrix} = \frac{2}{x^2},$$

$$W_2(x) = \begin{vmatrix} x^{-1} & 0 & x^2 \\ -x^{-2} & 0 & 2x \\ 2x^{-3} & 2x^{-4} & 2 \end{vmatrix} = -\frac{2}{x^4}, \; W_3(x) = \begin{vmatrix} x^{-1} & x & 0 \\ -x^{-2} & 1 & 0 \\ 2x^{-3} & 0 & 2x^{-4} \end{vmatrix} = \frac{4}{x^5}, \tag{74}$$

so (71) gives

$$\begin{aligned} y_p(x) &= \left(\int^x \frac{1}{3\xi}\, d\xi\right)\frac{1}{x} + \left(\int^x -\frac{1}{3\xi^3}\, d\xi\right) x + \left(\int^x \frac{2}{3\xi^4}\, d\xi\right) x^2 \\ &= \frac{\ln x}{3x} - \frac{1}{18x}. \end{aligned} \tag{75}$$

The $-1/(18x)$ term can be dropped because it is a homogeneous solution, so

$$y_p(x) = \frac{1}{3}\frac{\ln x}{x}, \tag{76}$$

as can be verified by direct substitution into (56). ∎

Generalization of the method to the nth-order equation

$$y^{(n)} + p_1(x)y^{(n-1)} + \cdots + p_{n-1}(x)y' + p_n(x)y = f(x) \tag{77}$$

is straightforward and the result is

$$\boxed{y_p(x) = \left[\int^x \frac{W_1(\xi)}{W(\xi)}\, d\xi\right] y_1(x) + \cdots + \left[\int^x \frac{W_n(\xi)}{W(\xi)}\, d\xi\right] y_n(x),} \tag{78}$$

where the y_j's are n LI homogeneous solutions, W is the Wronskian of $y_1, \ldots, y_n$, and W_j is identical to W, but with the jth column replaced by a column of zeros – except for the bottom element, which is f.

Closure. In this section we have discussed the nonhomogeneous equation $L[y] = f$, where L is an nth-order linear differential operator.

In Section 3.7.1 we provided the theoretical framework, which was based upon the linearity of L. Specifically, Theorem 3.7.1 showed that the general solution of $L[y] = f$ can be formed as the sum of a homogeneous solution $y_h(x)$, and a particular solution $y_p(x)$; $y_h(x)$ is a general solution of $L[y] = 0$, so it contains the n constants of integration, and $y_p(x)$ is any particular solution of the full equation $L[y] = f$. Further, we showed that if f is broken down as $f = f_1 + \cdots + f_k$,

then y_p is the sum of particular solutions $y_{p1}, \ldots, y_{pk}$ corresponding to $f_1, \ldots, f_k$, respectively.

Having studied homogeneous equations earlier, we considered y_h as known and focused our attention, in Sections 3.7.2 and 3.7.3, on methods for finding particular solutions y_p: undetermined coefficients and variation of parameters. Of these, the former is the easier to implement, but it is limited to cases where L is of constant-coefficient type and where each $f_j(x)$ has only a finite number of LI derivatives. The latter is harder to apply since it requires integrations, but is more powerful since it works even if L has nonconstant coefficients, and for any functions $f_j(x)$.

EXERCISES 3.7

1. Show whether or not the given forcing function satisfies condition (ii), below equation (11). If so, give a finite family of LI functions generated by it.

(a) $x^2 \cos x$ (b) $\cos x \sinh 2x$
(c) $\ln x$ (d) $x^2 \ln x$
(e) $\sin x/x$ (f) e^{x^2}
(g) e^{9x} (h) $(x-1)/(x+2)$
(i) $\tan x$ (j) $e^x \cos 3x$
(k) $x^3 e^{-x} \sinh x$ (l) $\cos x \cos 2x$
(m) $\sin x \sin 2x \sin 3x$ (n) $e^x/(x+1)$

2. Obtain a general solution using the method of undetermined coefficients.

(a) $y' - 3y = xe^{2x} + 6$
(b) $y' + y = x^4 + 2x$
(c) $y' + 2y = 3e^{2x} + 4\sin x$
(d) $y' - 3y = xe^{3x} + 4$
(e) $y' + y = 5 - e^{-x}$
(f) $y' - y = x^2 e^x$
(g) $y'' - y' = 5\sin 2x$
(h) $y'' + y' = 4xe^x + 3\sin x$
(i) $y'' + y = 3\sin 2x - 5 + 2x^2$
(j) $y'' + y' - 2y = x^3 - e^{-x}$
(k) $y'' + y = 6\cos x + 2$
(l) $y'' + 2y' = x^2 + 4e^{2x}$
(m) $y'' - 2y' + y = x^2 e^x$
(n) $y'' - 4y = 5(\cosh 2x - x)$
(o) $y'' - y' = 2xe^x$
(p) $y''' - y' = 25\cos 2x$
(q) $y''' - y'' = 6x + 2\cosh x$
(r) $y'''' + y'' - 2y = 3x^2 - 1$
(s) $y'''' - y = 5(x + \cos x)$

3. (a)–(s) Use computer software to solve the corresponding problem in Exercise 2.

4. Obtain a general solution using the method of variation of parameters.

(a) $y' + 2y = 4e^{2x}$
(b) $y' - y = xe^x + 1$
(c) $xy' - y = x^3$
(d) $xy' + y = 1/x \quad (x > 0)$
(e) $x^3 y' + x^2 y = 1 \quad (x > 0)$
(f) $y'' - y = 8x$
(g) $y'' - y = 8e^x$
(h) $y'' - 2y' + y = 6x^2$
(i) $y'' - 2y' + y = 2e^x$
(j) $y'' + y = 4\sin x$
(k) $y'' + 4y' + 4y = 2e^{-2x}$
(l) $6y'' - 5y' + y = x^2$
(m) $x^3 y'' + x^2 y' - 4xy = 1 \quad (x > 0)$
(n) $x^2 y'' - xy' - 3y = 4x \quad (x < 0)$
(o) $y''' + y'' - y' - y = x$
(p) $y''' - 6y'' + 11y' - 6y = e^{4x}$

5. (a)–(p) Use computer software to solve the corresponding problem in Exercise 4.

6. In the method of variation of parameters we used indefinite integrals [in (49) and (78)]. However, we could use definite integrals in those formulas, instead, if we choose. Specifically, show that, in place of (49),

$$y_p(x) = \left[\int_{\alpha_1}^{x} \frac{W_1(\xi)}{W(\xi)}\, d\xi\right] y_1(x) + \left[\int_{\alpha_2}^{x} \frac{W_2(\xi)}{W(\xi)}\, d\xi\right] y_2(x)$$

is also correct, for any choice of the constants α_1, α_2 (although normally one would choose α_1 and α_2 to be the same).

7. We chose (44) as "an especially convenient" condition to impose on C_1 and C_2. The condition

$$C_1'y_1 + C_2'y_2 = 6,$$

say, is not as simple inasmuch as 6 is not as simple as 0, but would it work? If so, use it in place of (44) and derive the final result for y_p, in place of (49). If not, explain why it would not work.

3.8 Application to Harmonic Oscillator: Forced Oscillation

The free oscillation of the harmonic oscillator (Fig. 1) was studied in Section 3.5. Now that we know how to find particular solutions, we can return to the harmonic oscillator and consider the case of forced oscillations, governed by the second-order, linear, constant-coefficient, nonhomogeneous equation

$$mx'' + cx' + kx = f(t). \tag{1}$$

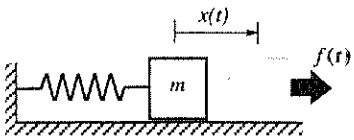


Figure 1. Mechanical oscillator.

In particular, we consider the important case where the forcing function is harmonic,

$$f(t) = F_0 \cos \Omega t. \tag{2}$$

3.8.1. Undamped case. To begin, consider the undamped case ($c = 0$),

$$mx'' + kx = F_0 \cos \Omega t. \tag{3}$$

The homogeneous solution of (3) is

$$x_h(t) = A \cos \omega t + B \sin \omega t, \tag{4}$$

where $\omega = \sqrt{k/m}$ is the natural frequency (i.e., the frequency of the free oscillation), and the forcing function $F_0 \cos \Omega t$ generates the family $\{\cos \Omega t, \sin \Omega t\}$. Thus, to find a particular solution of (3) by the method of undetermined coefficients, seek

$$x_p(t) = C \cos \Omega t + D \sin \Omega t. \tag{5}$$

Two cases present themselves. In the generic case, the driving frequency Ω is different from the natural frequency ω, so the terms in (5) do not duplicate any of those in (4) and we can accept (5) without modification. In the exceptional, or "singular," case where Ω is equal to ω, the terms in (5) repeat those in (4), so we need to modify (5) by multiplying the right side of (5) by t. For reasons that will become clear below, these cases are known as *nonresonance* and *resonance*, respectively.

Nonresonant oscillation. Putting (5) into the left side of (3) gives

$$\left(\omega^2 - \Omega^2\right) C \cos \Omega t + \left(\omega^2 - \Omega^2\right) D \sin \Omega t = \frac{F_0}{m} \cos \Omega t. \tag{6}$$

Since $\Omega \neq \omega$ by assumption, it follows from (6), by equating the coefficients of $\cos \Omega t$ and $\sin \Omega t$ on the left and right sides, that $C = (F_0/m)/(\omega^2 - \Omega^2)$ and $D = 0$. Thus

$$x_p(t) = \frac{F_0/m}{\omega^2 - \Omega^2} \cos \Omega t, \tag{7}$$

so a general solution of (3) is

$$\begin{aligned} x(t) &= x_h(t) + x_p(t) \\ &= A \cos \omega t + B \sin \omega t + \frac{F_0/m}{\omega^2 - \Omega^2} \cos \Omega t. \end{aligned} \tag{8}$$

In a sense we are done, and if we wish to impose any prescribed initial conditions $x(0)$ and $x'(0)$, then we could use those conditions to evaluate the constants A and B in (8). Then, for any desired numerical values of m, k, F_0, and Ω we could plot $x(t)$ versus t and see what the solution looks like. However, in science and engineering one is interested not only in obtaining answers, but also in understanding phenomena, so the question is: How can we extract, from (8), an understanding of the phenomenon? To answer that question, let us first rewrite (8) in the equivalent form

$$x(t) = E \sin(\omega t + \phi) + \frac{F_0/m}{\omega^2 - \Omega^2} \cos \Omega t \tag{9}$$

since then we can see it more clearly as a superposition of two harmonic solutions, of different amplitude, frequency, and phase.

The homogeneous solution $E \sin(\omega t + \phi)$ in (9), the "free vibration," was already discussed in Section 3.5. [Alternative to $E \sin(\omega t + \phi)$, we could use the form $E \cos(\omega t + \phi)$, whichever one prefers; it doesn't matter.] Thus, consider the particular solution, or "forced response," given by (7) and the last term in (9). It is natural to regard m and k (and hence ω) as fixed, and F_0 and Ω as controllable quantities or parameters. That the response (7) is merely proportional to F_0 is no surprise, for it follows from the linearity of the differential operator in (3). We also see, from (7), that the response is at the same frequency as the forcing function, Ω. More interesting is the variation of the amplitude $(F_0/m)/(\omega^2 - \Omega^2)$ with Ω, which is sketched in Fig. 2. The change in sign, as Ω increases through ω, is awkward since it prevents us from interpreting the plotted quantity as a pure magnitude. Thus, let us re-express (7) in the equivalent form

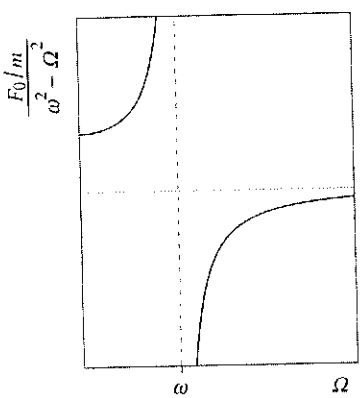


Figure 2. Magnitude of response (undamped case).

$$x_p(t) = \frac{F_0/m}{|\omega^2 - \Omega^2|} \cos(\Omega t + \Phi), \tag{10}$$

where the phase angle Φ is 0 for $\Omega < \omega$ and π for $\Omega > \omega$ [since $\cos(\Omega t + \pi) =$

$-\cos\Omega t$ gives the desired sign change for $\Omega > \omega$]. The resulting amplitude- and phase-response curves are shown in Fig. 3. From Fig. 3a, observe that as the driving frequency approaches the natural frequency the amplitude tends to infinity! [Of course, we must remember that right *at* $\Omega = \omega$ our particular solution is invalid since (6) is then $(0)\cos\Omega t + (0)\sin\Omega t = (F_0/m)\cos\Omega t$, which cannot be satisfied.] Further, as $\Omega \to \infty$ the amplitude tends to zero. Finally, we see from Fig. 3b that the response is in-phase ($\Phi = 0$) with the forcing function for $\Omega < \omega$, but for all $\Omega > \omega$ it is 180° out-of-phase. This discontinuous jump is striking since only an infinitesimal change in Ω (from just below ω to just above it) produces a discontinuous change in the response.

Also of considerable interest phenomenologically is the possibility of what is known as *beats*, but we will postpone that dicussion until we have had a look at the special case of resonance.

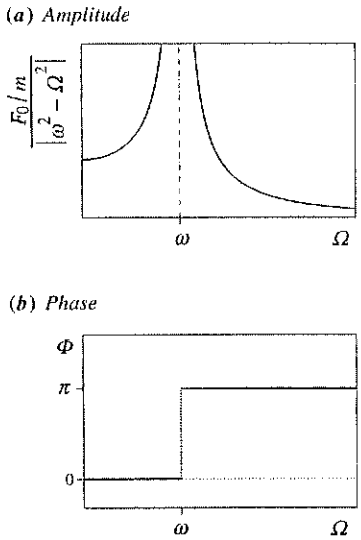


Figure 3. Amplitude- and phase-response curves (undamped case).

Resonant oscillation. For the special case where $\Omega = \omega$ (that is, where we force the system precisely at its natural frequency), the terms in (5) duplicate those in (4) so, according to the method of undetermined coefficients, we need to revise x_p as

$$x_p(t) = t\,(C\cos\omega t + D\sin\omega t)\,. \tag{11}$$

Since the duplication has thereby been removed, we accept (11). Putting that form into (3), we find that $C = 0$ and $D = F_0/(2m\omega)$, so

$$x_p(t) = \frac{F_0}{2m\omega}t\sin\omega t, \tag{12}$$

which is shown in Fig. 4. In this special case the response is not a harmonic oscillation but a harmonic function times t, which factor causes the magnitude to tend to infinity as $t \to \infty$. This result is known as **resonance**. Of course, the magnitude does not grow unboundedly in a real application since the mathematical model of the system (the governing differential equation) will become inaccurate for sufficiently large amplitudes, parts will break, and so on.

Resonance is sometimes welcome and sometimes unwelcome. That is, sometimes we wish to amplify a given input, and can do so by "tuning" the system to be at or near resonance, as when we tune a radio circuit to a desired broadcast frequency. And other times we wish to suppress inputs, as a well designed automobile suspension suppresses, rather than amplifies, the inputs from a bumpy road.

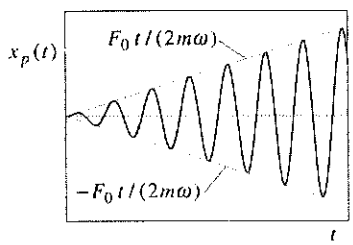


Figure 4. Resonant oscillation.

Beats. Isn't it striking that the response $x(t)$ is the sum of two harmonics [given by (5)] for all $\Omega \neq \omega$, yet it is of the different form (12) for the single case $\Omega = \omega$? One might wonder whether the resonant case is really of any importance at all since one can never get Ω to exactly equal ω. It is therefore of interest to look at the solution $x(t)$ as Ω *approaches* ω. To do so, let us use the simple initial conditions $x(0) = 0$ and $x'(0) = 0$, for definiteness, in which case we can evaluate A and B in (8), and obtain

$$x(t) = -\frac{F_0/m}{\omega^2 - \Omega^2}\,(\cos\omega t - \cos\Omega t)\,, \tag{13}$$

or, recalling the trigonometric identity $\cos A - \cos B = 2\sin\dfrac{B+A}{2}\sin\dfrac{B-A}{2}$,

$$x(t) = \frac{2F_0/m}{\omega^2 - \Omega^2}\sin\left(\frac{\omega+\Omega}{2}\right)t\ \sin\left(\frac{\omega-\Omega}{2}\right)t. \tag{14}$$

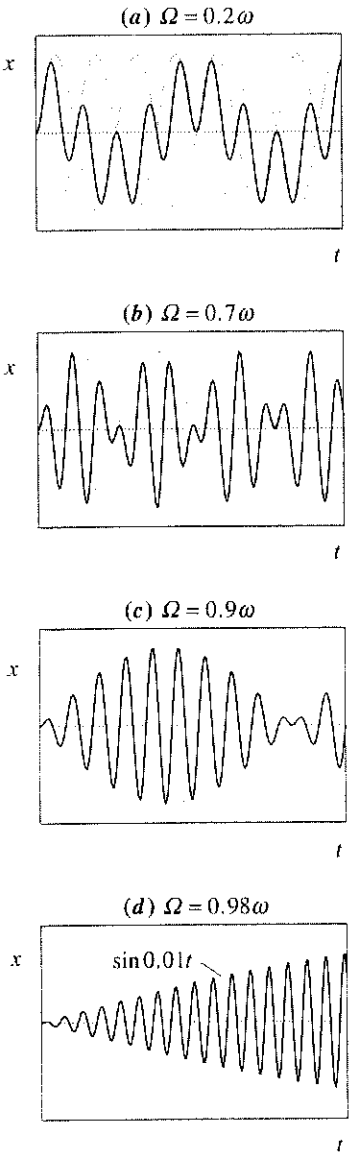


Figure 5. Beats, and approach to resonance.

Now, suppose that Ω is close to (but not equal to) the natural frequency ω. Then the frequency of the second sinusoid in (14) is very small compared to that of the first, so the $\sin\left(\frac{\omega-\Omega}{2}\right)t$ factor amounts, essentially, to a slow "amplitude modulation" of the relatively high frequency $\sin\left(\frac{\omega+\Omega}{2}\right)t$ factor. This phenomenon is known as **beats**, and is seen in Fig. 5, where we have plotted the solution (14) for four representative cases: in Fig. 5a Ω is not close to ω, and there is no discernible beat phenomenon, but as Ω is increased the beat phenomenon becomes well established, as seen in Fig. 5b, 5c, and 5d. [We have shown the "envelope" $\sin\left(\frac{\omega-\Omega}{2}\right)t$ as dotted.]

We can now see that the resonance phenomenon at $\Omega = \omega$ is not an isolated behavior but is a limiting case as $\Omega \to \omega$. That is, resonance (Fig. 4) is actually a limit of the sequence shown in Fig. 5, as $\Omega \to \omega$. Rather than depend only on these suggestive graphical results, we can proceed analytically as well. Specifically, we can take the limit of the response (13) as $\Omega \to \omega$ and, with the help of l'Hôpital's rule, we do obtain (12)!

With our foregoing discussion of the undamped forced harmonic oscillator in mind, *we cannot overstate that we are by no means dealing only with the solving of equations but with the phenomena thereby being described. To understand phenomena, we normally need to do several things: we do need to solve the equations that model the phenomena (analytically or, if that is too hard, numerically), but we also need to study, interpret, and understand the results.* Such study normally includes the generation of suitably chosen graphical displays (such as our Fig. 2, 3, and 4), the isolation of special cases [such as our initial consideration of the case where there is no damping; $c = 0$ in (1)], and perhaps the examination of various limiting cases (such as the limit $\Omega \to \omega$ in the present example). Emphasis in this book is on the mathematics, with the detailed study of the relevant physics left for applications courses such as Fluid Mechanics, Electromagnetic Field Theory, and so on, but we will occasionally try to show not only the connections between the mathematics and the physics but also the process whereby we determine those connections.

3.8.2. Damped case. We now reconsider the harmonically driven oscillator, this time with a cx' damping term included ($c > 0$):

$$mx'' + cx' + kx = F_0\cos\Omega t. \tag{15}$$

Recall from Section 3.5 that the homogeneous solution is

$$x_h(t) = \begin{cases} e^{-\frac{c}{2m}t}\left[A\cos\sqrt{\omega^2-\left(\frac{c}{2m}\right)^2}\,t + B\sin\sqrt{\omega^2-\left(\frac{c}{2m}\right)^2}\,t\right] \\ e^{-\frac{c}{2m}t}(A+Bt) \\ e^{-\frac{c}{2m}t}\left[A\cosh\sqrt{\left(\frac{c}{2m}\right)^2+\omega^2}\,t + B\sinh\sqrt{\left(\frac{c}{2m}\right)^2+\omega^2}\,t\right] \end{cases} \tag{16}$$

for the underdamped ($c < c_{cr}$), critically damped ($c = c_{cr}$), and overdamped ($c > c_{cr}$) cases, respectively, and where $\omega = \sqrt{k/m}$ and $c_{cr} = 2\sqrt{mk}$.

This time, when we write

$$x_p(t) = C\cos\Omega t + D\sin\Omega t, \tag{17}$$

according to the method of undetermined coefficients, there is no duplication between terms in (17) and (16), even if $\Omega = \omega$, because of the $\exp(-ct/2m)$ factors in (16), so we can accept (17) without modification. Putting (17) into (15) and equating coefficients of the $\cos\Omega t$ terms on both sides of the equation, and similarly for the coefficients of the $\sin\Omega t$ terms, enables us to solve for C and D. The result (Exercise 3a) is that

$$\begin{aligned} x_p(t) = &\frac{(F_0/m)\left(\omega^2-\Omega^2\right)}{\left(\omega^2-\Omega^2\right)^2+\left(c\Omega/m\right)^2}\cos\Omega t \\ &+\frac{F_0 c\Omega/m^2}{\left(\omega^2-\Omega^2\right)^2+\left(c\Omega/m\right)^2}\sin\Omega t, \end{aligned} \tag{18}$$

or (Exercise 3b), equivalently,

$$x_p(t) = E\cos(\Omega t + \Phi), \tag{19a}$$

where the amplitude E and phase Φ are

$$E = \frac{F_0/m}{\sqrt{\left(\omega^2-\Omega^2\right)^2+\left(c\Omega/m\right)^2}}, \tag{19b}$$

$$\Phi = \tan^{-1}\frac{c\Omega/m}{\Omega^2-\omega^2}, \tag{19c}$$

with the $\tan^{-1}$ understood to lie between 0 and π.

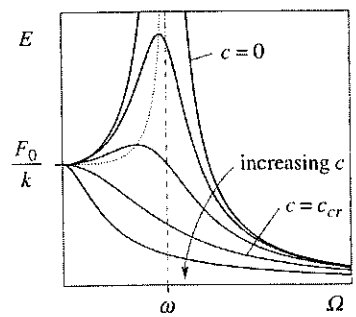


Figure 6. Amplitude response curves.

As for the undamped case, we have great interest in the amplitude- and frequency-response curves, the graphs of the amplitude E, and the phase Φ with respect to the driving frequency Ω. The former is given in Fig. 6 for various values of the damping coefficient c, and the latter is left for the exercises.

From Fig. 6 we see that true resonance is possible only in the case of no damping ($c = 0$), which case is an idealization since in reality there is inevitably some damping present. Analytically, we see the same thing: (19a) shows that the amplitude E can become infinite only if $c = 0$, and that occurs only for $\Omega = \omega$. However, for $c > 0$ there is still a peaking of the amplitude, even if that peak is now finite, at a driving frequency Ω which diminishes from ω as c increases, and which is 0 for all $c \geq c_{cr}$. Further, the peak magnitude (located by the dotted curve) diminishes from ∞ to F_0/k as c is increased from 0 to c_{cr}, and remains F_0/k for all $c > c_{cr}$.

What is the significance of the F_0/k value? For $\Omega = 0$ the differential equation becomes $mx'' + cx' + kx = F_0$, and the method of undetermined coefficients gives $x_p(t) = \text{constant} = F_0/k$, which is merely the static deflection of the mass under the steady force F_0.

Even if true resonance is possible only for the undamped case ($c = 0$), the term resonance is often used to refer to the dramatic peaking of the amplitude response curves if c is not too large.

The general solution, of course, is the sum

$$\begin{aligned} x(t) &= x_h(t) + x_p(t) \\ &= x_h(t) + E\cos(\Omega t + \Phi). \end{aligned} \tag{20}$$

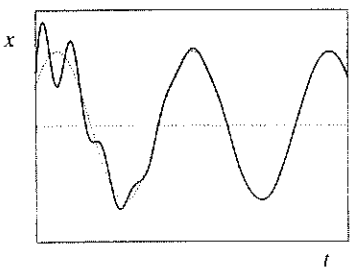


Figure 7. A representative response $x(t)$ (solid); approach to the steady-state oscillation $x_p(t)$ (dotted).

where E and Φ are given by (19b,c) and $x_h(t)$ is given by the suitable right-hand side of (16), according to whether the system is underdamped, critically damped, or overdamped. If we impose initial conditions $x(0)$ and $x'(0)$ on (20), then we can solve for the integration constants A and B within $x_h(t)$.

Notice carefully that the $x_h(t)$ part of the solution inevitably tends to zero as $t \to \infty$ because of the $\exp(-ct/2m)$ factor, no matter how small c is, as long as $c > 0$. Thus, we call $x_h(t)$ in (20) the **transient** part of the solution and we call $x_p(t)$ the **steady-state** part since $x(t) \to E\cos(\Omega t + \Phi)$ as $t \to \infty$. The transient part depends upon the initial conditions, whereas the steady-state part does not. A representative underdamped case is shown in Fig. 7, where we see the approach to the steady-state oscillation $x_p(t)$.

Closure. In this section we considered the forced vibration of a harmonic oscillator – that is, a system governed by the differential equation $mx'' + cx' + kx = f(t)$, for the case of the harmonic excitation $f(t) = F_0 \cos \Omega t$. Thus, besides a homogeneous solution we needed to find a particular solution, and that was done by the method of undetermined coefficients. The particular solution is especially important physically since even an infinitesimal amount of damping will cause the homogeneous solution to tend to zero as $t \to \infty$, so that the particular solution becomes the steady-state response. To understand the physical significance of that response we attached importance to the amplitude- and phase-response curves and discussed the phenomena of resonance and beats. Our discussion in this section has been limited in that we have considered only the case of harmonic excitation, whereas in applications $f(t)$ surely need not be harmonic. However, that case is important enough to deserve this special section. When we study the Laplace transform method in Chapter 5, we will be able to return to problems such as

$mx'' + cx' + kx = f(t)$ and, using the convenient Laplace transform methodology, obtain solutions for virtually any forcing function $f(t)$.

EXERCISES 3.8

1. Applying the initial conditions $x(0) = 0$ and $x'(0) = 0$ to (8), derive (13). Show that the same result can be obtained if we start with the form (9) instead of (8).

2. Derive (12) from (11).

3. (a) Derive (18). (b) Derive (19a,b,c).

4. The amplitude- and phase-response curves shown in Fig. 3 correspond to the equation $mx'' + kx = F_0 \cos \Omega t$. Obtain the equations of the analogous response curves for the equation $mx'' + kx = F_0 \sin \Omega t$, and give labeled sketches of the two curves.

5. Figure 6 shows the amplitude-response curves (E versus Ω) corresponding to (19b), for various values of c.

(a) What happens to the graph as $c \to \infty$? Is $E(\Omega)$ continuous on $0 \le \Omega < \infty$ for $c = \infty$? Explain.
(b) From (19c), obtain the phase-response curves (Φ versus Ω), either by a careful freehand sketch or using a computer, for various values of c, being sure to include the important case $c = 0$. What happens to the graph as $c \to \infty$?

6. In Fig. 7 we show the approach of a representative response curve (solid) to the steady-state oscillation (dotted), for an underdamped system.

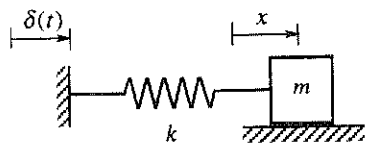
(a) Do the same (with a computer plot) for a critically damped case. The values of $m, c, k, F_0, \Omega, x(0), x'(0)$ are up to you, but the idea is to demonstrate graphically the approach to $x_p(t)$ clearly, as we have in Fig. 7.
(b) Same as (a), for an overdamped system, where $c = 4c_{cr}$, say.

7. Show that taking the limit of the response (13) as $\Omega \to \omega$, with the help of l'Hôpital's rule, does give (12), as claimed two paragraphs below (14).

8. Observe from Fig. 6 that the amplitude E tends to zero as $\Omega \to \infty$. Explain (physically, mathematically, or both) why that result makes sense.

9. (a) What choice of initial conditions $x(0)$ and $x'(0)$ will reduce the solution (20) to just the particular solution, $x(t) = E \cos(\Omega t + \Phi)$?
(b) Using a sketch of a representative $x_p(t)$ such as the dotted curve in Fig. 7, show the graphical significance of those special values of $x(0)$ and $x'(0)$.

10. Imagine the experimental means that would be required to apply a force $F_0 \cos \Omega t$ to a mass. It doesn't sound so hard if the mass is stationary, but imagine trying to apply such a force to a moving mass! In many physical applications, such as earthquake-induced vibration, the driving force is applied indirectly, by "shaking" the wall, rather than being applied directly to the mass. Specifically, for the system shown in the figure, use Newton's second law to show that if the wall is



displaced laterally according to $\delta(t) = \delta_0 \cos \Omega t$, then the equation of motion of the mass m is $mx'' + kx = F_0 \cos \Omega t$, where $F_0 = k\delta_0$. Here, x and δ are measured relative to fixed points in space. NOTE: Observe that such an experiment is more readily performed since it is easier to apply a harmonic displacement $\delta(t)$ than a harmonic force; for instance, one could use a slider-crank mechanism (which converts circular motion to harmonic linear motion). Note further that a displacement input is precisely what an automobile suspension is subjected to when we drive over a bumpy road.

11. For the mechanical oscillator governed by the differential equation $mx'' + cx' + kx = F(t)$, obtain computer plots of the amplitude- and phase-response curves (E versus Ω and Φ versus Ω), for the case where $F(t) = 25 \sin \Omega t$, for these six values of the damping coefficient c: $0, 0.25c_{cr}, 0.5c_{cr}, c_{cr}, 2c_{cr}, 4c_{cr}$, where

(a) $m = 1, k = 1$
(b) $m = 2, k = 5$
(c) $m = 2, k = 10$
(d) $m = 4, k = 2$
(e) $m = 4, k = 10$

12. (*Complex function method*) Let L be a linear constant-coefficient differential operator, and consider the equation

$$L[x] = F_0 \cos \Omega t, \tag{12.1}$$

According to the method of undetermined coefficients, we can find a particular solution $x_p(t)$ by seeking $x_p(t) = A\cos\Omega t + B\sin\Omega t$ (or, in exceptional cases, t to an integer power times that). A slightly simpler line of approach that is sometimes used is as follows. Consider, in place of (12.1),

$$L[w] = F_0 e^{i\Omega t}, \tag{12.2}$$

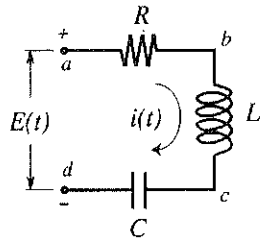
Equation (12.2) is simpler than (12.1) in that to find a particular solution we need only one term, $w_p(t) = Ae^{i\Omega t}$. (If $z = a + ib$ is any complex number, it is standard to call $\operatorname{Re} z = a$ and $\operatorname{Im} z = b$ the **real part** and the **imaginary part** of z, respectively.) Because, according to Euler's formula, $e^{i\Omega t} = \cos\Omega t + i\sin\Omega t$, it follows that $\operatorname{Re} e^{i\Omega t} = \cos\Omega t$ and $\operatorname{Im} e^{i\Omega t} = \sin\Omega t$. Since the forcing function in (12.1) is the real part of the forcing function in (12.2), it seems plausible that $x_p(t)$ should be the real part of $w_p(t)$. Thus, we have the following method: to find a particular solution to (12.1) consider instead the simpler equation (12.2). Solve for $w_p(t)$ by seeking $w_p(t) = Ae^{i\Omega t}$, and then recover the desired $x_p(t)$ from $x_p(t) = \operatorname{Re} w_p(t)$.

(a) Prove that the method described above works. HINT: The key is the linearity of L, so that if $w = u + iv$, then $L[w] = L[u + iv] = L[u] + iL[v]$.
(b)–(k) Use the method to obtain a particular solution to the given equation:

(b) $mx'' + cx' + kx = F_0\cos\Omega t$
(c) $mx'' + cx' + kx = F_0\sin\Omega t$
(d) $x' + 3x = 5\cos 2t$
(e) $x' - x = 4\sin 3t$
(f) $x'' - x' + x = \cos 2t$
(g) $x'' + 5x' + x = 3\sin 4t$
(h) $x'' - 2x' + x = 6\cos 5t$
(i) $x''' + x'' + x' + x = 3\sin t$
(j) $x'''' + x' + x = 3\cos t$
(k) $x'''' + 2x''' + 4x = 9\sin 6t$

13. (*Electrical circuit*) Recall from Section 2.3 that the equations governing the current $i(t)$ in the circuit shown, and the charge $Q(t)$ on the capacitor are

$$L\frac{d^2i}{dt^2} + R\frac{di}{dt} + \frac{1}{C}i = \frac{dE(t)}{dt}, \tag{13.1}$$



and

$$L\frac{d^2Q}{dt^2} + R\frac{dQ}{dt} + \frac{1}{C}Q = E(t), \tag{13.2}$$

respectively, where L, R, C, E, i, and Q are measured in henrys, ohms, farads, volts, amperes, and coulombs, respectively.

(a) Let $L = 2$, $R = 4$, and $C = 0.05$. Solve for $Q(t)$ subject to the initial conditions $Q(0) = Q'(0) = 0$, where $E(t) = 100$. Identify the steady-state solution. Give a computer plot of the solution for $Q(t)$ over a sufficiently long time period to clearly show the approach of Q to its steady state. (Naturally, all plots should be suitably labeled.)
(b) Same as (a), but for $C = 0.08$.
(c) Same as (a), but for $C = 0.2$.
(d) Same as (a), but for $E(t) = 10e^{-t}$.
(e) Same as (a), but for $E(t) = 10\left(1 - e^{-t}\right)$.
(f) Same as (a), but for $E(t) = 50\left(1 + e^{-0.5t}\right)$.

3.9 Systems of Linear Differential Equations

Thus far we have considered the case where there is a single dependent variable such as the current $i(t)$ in a circuit, where t is the time. However, many problems involve two or more dependent variables. In the combustion of fossil fuels, for instance, there are many interacting chemical species (e.g., OH, CH_4, H, CO, H_2O, and so on) whose generation and demise, as a function of time, are governed by

a large set of differential equations. A realistic model could easily contain 100 differential equations on 100 unknowns.

If there are two or more unknowns, then we are involved not with a single differential equation but with a system of such equations. For instance, according to the well known Lotka–Volterra model of predator-prey population dynamics, the populations $x(t)$ and $y(t)$ of predator and prey are governed by the system of two equations

$$\begin{aligned} x' &= -\alpha x + \beta xy, \\ y' &= \gamma y - \delta xy, \end{aligned} \tag{1}$$

where $\alpha, \beta, \gamma, \delta$ are empirical constants and t is the time. This particular system happens to be nonlinear because of the xy products; we will return to it in Chapter 7 when we study nonlinear systems. The present chapter is devoted exclusively to linear differential equations.

By definition, a **linear** first-order system of n equations in the n unknowns $x_1(t), \ldots, x_n(t)$ is of the form

$$\begin{aligned} a_{11}(t)x_1' + \cdots + a_{1n}(t)x_n' + b_{11}(t)x_1 + \cdots + b_{1n}(t)x_n &= f_1(t) \\ &\vdots \\ a_{n1}(t)x_1' + \cdots + a_{nn}(t)x_n' + b_{n1}(t)x_1 + \cdots + b_{nn}(t)x_n &= f_n(t), \end{aligned} \tag{2}$$

where the forcing functions $f_j(t)$ and the coefficients $a_{jk}(t)$ and $b_{jk}(t)$ are prescribed, and where it is convenient to use a double-subscript notation: $a_{jk}(t)$ denotes the coefficient of $x_k'(t)$ in the jth equation, and $b_{jk}(t)$ denotes the coefficient of $x_k(t)$ in the jth equation. We call (2) a first-order system because the highest derivatives are of first order. If the highest derivatives were of second order, we would call it a second-order system, and so on. A linear second-order system of n equations in the n unknowns $x_1(t), \ldots, x_n(t)$ would be of the same form as (2), but with each left-hand side being a linear combination of the second-, first-, and zeroth-order derivatives of the unkowns.

The system (2) is a generalization of the linear first-order equation $y' + p(x)y = q(x)$ in the one unknown $y(x)$ studied in Chapter 2. There, and in most of Chapters 1–3, we favored x as the generic independent variable and y as the generic dependent variable, but in this section the independent variable in most of our applications happens to be the time t, so we will use t as the independent variable.

As in the case of a single differential equation, by a **solution** of a system of differential equations (be they linear or not), in the unknowns $x_1(t), \ldots, x_n(t)$ over some t interval I, we mean a set of functions $x_1(t), \ldots, x_n(t)$ that reduce those equations to identities over I.

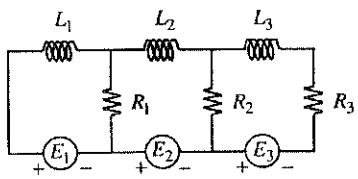


Figure 1. Circuit of Example 1.

3.9.1. Examples. let us begin by giving a few examples of how such systems arise in applications.

EXAMPLE 1. *RL Circuit.* Consider the circuit shown in Fig. 1, comprised of three

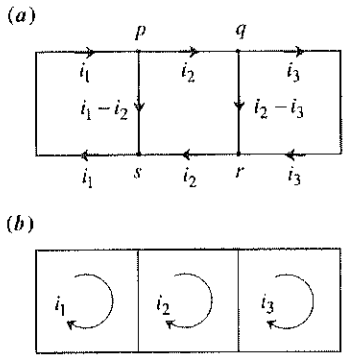


Figure 2. Current designations.

loops. We wish to obtain the differential equations governing the various currents in the circuit. There are two ways to proceed that are different but equivalent, and which correspond to the current labeling shown in Fig. 2a and 2b (in which we have omitted the circuit elements, for simplicity). First consider the former. If the current approaching the junction p from the "west" is designated as i_1 and the current leaving to the east is i_2, then it follows from Kirchoff's current law (namely, that the algebraic sum of the currents approaching or leaving any point of a circuit is zero) that the current to the south must be $i_1 - i_2$. Similarly, if we designate the current leaving the junction q to the east as i_3, then the current leaving to the south must be $i_2 - i_3$. With the current approaching r from the north and east being $i_2 - i_3$ and i_3, it follows that the current leaving to the west must be i_2. Similarly, the current leaving s to the west must be i_1.

Next, apply Kirchoff's voltage law (namely, that the algebraic sum of the voltage drops around each loop of the circuit must be zero) to each loop, recalling from Section 2.3 that the voltage drops across inductors, resistors, and capacitors (of which there are none in this particular circuit) are $L\dfrac{di}{dt}$, Ri, and $\dfrac{1}{C}\displaystyle\int i\,dt$, respectively. For the left-hand loop that step gives $L_1\dfrac{di_1}{dt} + R_1(i_1 - i_2) - E_1(t) = 0$, where the last term (corresponding to the applied voltage E_1) is counted as negative because it amounts to a voltage rise (according to the polarity denoted by the $\pm$ signs in Fig. 1) rather than a drop. Thus, we have for the left, middle, and right loops,

$$\begin{aligned} L_1 i_1' + R_1\,(i_1 - i_2) &= E_1(t), \\ L_2 i_2' + R_2\,(i_2 - i_3) + R_1\,(i_2 - i_1) &= E_2(t), \\ L_3 i_3' + R_3 i_3 + R_2\,(i_3 - i_2) &= E_3(t), \end{aligned} \tag{3}$$

respectively, or,

$$\begin{aligned} L_1 i_1' + R_1 i_1 - R_1 i_2 &= E_1(t), \\ L_2 i_2' - R_1 i_1 + (R_1 + R_2)\,i_2 - R_2 i_3 &= E_2(t), \\ L_3 i_3' - R_2 i_2 + (R_2 + R_3)\,i_3 &= E_3(t), \end{aligned} \tag{4}$$

where $E_1(t), E_2(t), E_3(t)$ are prescribed. It must be remembered that the currents do not need to flow in the directions assumed by the arrows; after all, they are the unknowns. If any of them turn out to be negative (at any given instant t), that merely means that they are flowing in the direction opposite to that tentatively assumed in Fig. 2a.

Alternatively, one can use the idea of "loop currents," as denoted in Fig. 2b. In that case the south-flowing currents in R_1 and R_2 are the net currents $i_1 - i_2$ and $i_2 - i_3$, respectively, just as in Fig. 2a. Either way, the result is the linear first-order system (4). ∎

It is important to see that the system (4) is *coupled*. That is, the individual equations contain more than one unknown so that we cannot separate them and solve the first for i_1 (for instance), the second for i_2, and the third for i_3. Put differently, the currents i_1, i_2, i_3 are interrelated. It is only natural for systems of differential equations to be coupled since the coupling is the mathematical expression of the

relatedness of the dependent variables. On the other hand, if we write differential equations governing the current $i(t)$ in a circuit and the price of tea in China, $p(t)$, we would hardly expect those equations to be coupled and, indeed, it would hardly make sense to group them as part of the same system.

EXAMPLE 2. *LC Circuit.* For the circuit shown in Fig. 3, the same reasoning as above gives the integro-differential equations

$$\begin{aligned} \frac{1}{C_1}\int i_1\,dt + L\frac{d}{dt}(i_1 - i_2) &= E(t), \\ \frac{1}{C_2}\int i_2\,dt + L\frac{d}{dt}(i_2 - i_1) &= 0 \end{aligned} \tag{5}$$

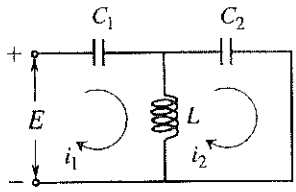


Figure 3. *LC* circuit.

on the currents $i_1(t)$ and $i_2(t)$ or, differentiating to eliminate the integral signs,

$$\begin{aligned} Li_1'' - Li_2'' + \frac{1}{C_1}i_1 &= E'(t), \\ Li_2'' - Li_1'' + \frac{1}{C_2}i_2 &= 0. \end{aligned} \tag{6}$$

Whereas (4) was a first-order system, (6) is of second order. ∎

EXAMPLE 3. *Mass-Spring System.* This time consider a mechanical system, shown in Fig. 4 and comprised of masses and springs. The masses rest on a frictionless table and are subjected to applied forces $F_1(t)$, $F_2(t)$, respectively. When the displacements x_1 and x_2 are zero, the springs are neither stretched nor compressed, and we seek the equations of motion of the system, that is, the differential equations governing $x_1(t)$ and $x_2(t)$.

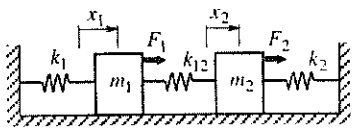


Figure 4. Mass-spring system.

The relevant physics is Newton's second law of motion, and Hooke's law for each of the three springs, as were discussed in Section 1.3. To proceed, it is useful to make a concrete assumption on x_1 and x_2. Specifically, suppose that at the instant t we have $x_1 > x_2 > 0$, as assumed in Fig. 5 (which figure, in the study of mechanics, is called a *free-body diagram*). Then the left spring is stretched by x_1 so it exerts a force to the left, on m_1, equal (according to Hooke's law) to k_1x_1. The middle spring is compressed by $x_1 - x_2$ so it exerts a force $k_{12}(x_1 - x_2)$ to the left on m_1 and to the right on m_2, and the right spring is compressed by x_2 and exerts a force k_2x_2 to the left on m_2, as shown in the figure. With the help of the information given in Fig. 5, Newton's second law for each of the two masses gives

$$\begin{aligned} m_1x_1'' &= -k_1x_1 - k_{12}(x_1 - x_2) + F_1(t), \\ m_2x_2'' &= -k_2x_2 + k_{12}(x_1 - x_2) + F_2(t) \end{aligned} \tag{7}$$

as the desired equations of motion – or, rearranging terms,

$$\begin{aligned} m_1x_1'' + (k_1 + k_{12})x_1 - k_{12}x_2 &= F_1(t), \\ m_2x_2'' - k_{12}x_1 + (k_2 + k_{12})x_2 &= F_2(t). \end{aligned} \tag{8}$$

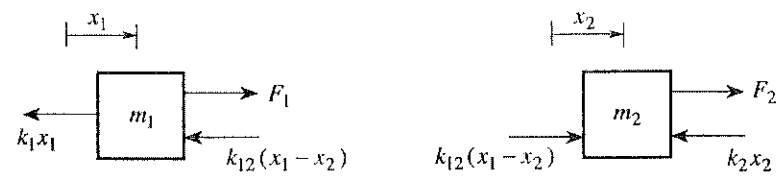


Figure 5. Free-body diagram of the masses.

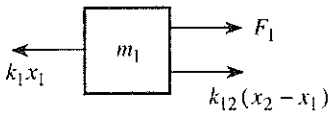


Figure 6. Revised free-body diagram for m_1.

COMMENT. Our assumption that $x_1 > x_2 > 0$ was only for definiteness; the resulting equations (8) are insensitive to whatever such assumption is made. For instance, suppose that we assume, instead, that $x_2 > x_1 > 0$. Then the middle spring is stretched by $x_2 - x_1$, so the free-body diagram of m_1 changes to that shown in Fig. 6, and Newton's law for m_1 gives $m_1x_1'' = -k_1x_1 + k_{12}(x_2 - x_1) + F_1(t)$, which is seen to be equivalent to the first of equations (7). Similarly for m_2. ∎

3.9.2. Existence and uniqueness. The fundamental theorem regarding existence and uniqueness is as follows.*

THEOREM 3.9.1 *Existence and Uniqueness for Linear First-Order Systems*
Let the functions $a_{11}(t), a_{12}(t), \ldots, a_{nn}(t)$ and $f_1(t), \ldots, f_n(t)$ be continuous on a closed interval I. And let numbers $b_1, \ldots, b_n$ be given such that

$$x_1(a) = b_1, \quad x_2(a) = b_2, \ \cdots, \ x_n(a) = b_n, \tag{9}$$

where a is a given point in I. Then the system

$$\begin{aligned} x_1' &= a_{11}(t)x_1 + a_{12}(t)x_2 + \cdots + a_{1n}(t)x_n + f_1(t), \\ &\vdots \\ x_n' &= a_{n1}(t)x_1 + a_{n2}(t)x_2 + \cdots + a_{nn}(t)x_n + f_n(t), \end{aligned} \tag{10}$$

*There is a subtle point that is worth noting, namely, that (10) is not quite of the same form as the general first-order linear system (2) in that its left-hand sides are simply $x_1', \ldots, x_n'$ rather than linear combinations of those terms. (What follows presumes that you have already studied the sections on matrices, rank, and Gauss–Jordan reduction.) The idea is that (2) can be reduced to the form (10) by elementary row operations, as in the Gauss–Jordan reduction of linear algebraic equations – unless the rank of the $\{a_{jk}(t)\}$ matrix is less than n. In that case, such operations would yield at least one equation, at the bottom of the system, which has no derivatives in it. If not all of the coefficients of the undifferentiated x_j terms in that equation are zero, then one could use that equation to solve for one of the x_j's in terms of the others and use that result to reduce the system by one unknown and one equation; if all of the coefficients of the undifferentiated x_j terms in that equation are zero, then that equation would either be $0 = 0$, which could be discarded, or zero equal to some nonzero prescribed function of t, which would cause the system to have no solution. To avoid these singular cases, it is conventional to use the form (10), rather than (2), in the existence and uniqueness theorem.

subject to the initial conditions (9), has a unique solution on the entire interval I.

Observe that we have added the word "entire" for emphasis, for recall from Section 2.4 that the Existence and Uniqueness Theorem 2.4.1 for the *non*linear initial-value problem $y'(x) = f(x, y)$ with initial condition $y(a) = b$ is a local one; it guarantees the existence of a unique solution over *some* interval $|x - a| < h$, but it does not tell us how big h can be. In contrast, Theorem 3.9.1 tells us that the solution exists and is unique over the entire interval I over which the specified conditions are met.

EXAMPLE 4. *Example 1, Revisited.* Suppose that we add initial conditions, say $i_1(0) = b_1, i_2(0) = b_2, i_3(0) = b_3$ to the system (4) governing the RL circuit of Example 1. If the $E_j(t)$'s are continuous on $0 \le t \le T$ and the L_j's are nonzero [so we can divide through by them in reducing (4) to the form of (10)] then, according to Theorem 3.9.1, the initial-value problem has a solution on $0 \le t \le T$, and it is unique. ∎

It would appear that Theorem 3.9.1 does not apply to the system (8) of Example 3 because the latter is of second order rather than first. However, and this is important, higher-order systems can be reduced to first-order ones by introducing artificial, or auxiliary, dependent variables.

EXAMPLE 5. Reduce the second-order system (8) to a first-order system. The idea is to introduce artificial dependent variables u and v according to $x_1' = u$ and $x_2' = v$ because then the second-order derivatives x_1'' and x_2'' become first-order derivatives u' and v', respectively. Thus, (8) can be re-expressed, equivalently, as the first-order system

$$\begin{aligned} x_1'(t) &= u, \\ u'(t) &= -\frac{k_1 + k_{12}}{m_1} x_1 + \frac{k_{12}}{m_1} x_2 + \frac{1}{m_1} F_1(t), \\ x_2'(t) &= v, \\ v'(t) &= \frac{k_{12}}{m_2} x_1 - \frac{k_2 + k_{12}}{m_2} x_2 + \frac{1}{m_2} F_2(t). \end{aligned} \tag{11}$$

To see that this system is of the form (10), let "x_1" $= x_1$, "x_2" $= u$, "x_3" $= x_2$, and "x_4" $= v$. Then $a_{11} = a_{13} = a_{14} = 0$, $a_{12} = 1$, $f_1(t) = 0$, $a_{21} = -(k_1 + k_{12})/m_1$, $a_{22} = a_{44} = 0$, $a_{23} = k_{12}/m_1$, $f_2(t) = F_1(t)/m_1$, and so on. All of the $a_{jk}(t)$ coefficients are constants and hence continuous for all t. Let the forcing functions $F_1(t)$ and $F_2(t)$ be continuous on $0 \le t < \infty$.

Thus, according to Theorem 3.9.1, if we prescribe initial conditions $x_1(0)$, $u(0)$, $x_2(0)$, $v(0)$, then the initial-value problem consisting of (11), together with those initial conditions, will have a unique solution for $x_1(t), u(t), x_2(t), v(t)$. Equivalently, the initial-value problem consisting of (8), together with prescribed initial values $x_1(0)$, $x_1'(0)$, $x_2(0)$, $x_2'(0)$, will have a unique solution for $x_1(t), x_2(t)$. ∎

Consider one more example on auxiliary variables.

EXAMPLE 6. Consider the third-order equation

$$x''' + 2tx'' - x' + (\sin t)x = \cos t, \tag{12}$$

which *is* a system, a system of n equations in n unknowns, where $n = 1$. To reduce it to a system of first-order equations, introduce auxiliary variables u, v according to $x' = u$ and $x'' = u' = v$. Then

$$\begin{aligned} x' &= u, \\ u' &= v, \\ v' &= -(\sin t)x + u - 2tv + \cos t \end{aligned} \tag{13}$$

is the desired equivalent first-order system, where the last of the three equations follows from the first two together with (12). ∎

3.9.3. Solution by elimination. We now give a method of solution of systems of linear differential equations for the special case of constant coefficients, a method of elimination that is well suited to systems that are small enough for us to carry out the steps by hand.

We introduce the method with an example after first recalling (from Section 3.3) the idea of a linear differential operator,

$$\begin{aligned} L &= a_0(t)\frac{d^n}{dt^n} + a_1(t)\frac{d^{n-1}}{dt^{n-1}} + \cdots + a_n(t) \\ &= a_0(t)D^n + a_1(t)D^{n-1} + \cdots + a_n(t), \end{aligned}$$

where D denotes $\dfrac{d}{dt}$, D^2 denotes $\dfrac{d^2}{dt^2}$, and so on. By $L[x]$ we mean the function $a_0\dfrac{d^n x}{dt^n} + a_1\dfrac{d^{n-1}x}{dt^{n-1}} + \cdots + a_n x$. We say that L is of order n (if a_0 is not identically zero) and that it "acts" on x, or "operates" on x. Further, by $L_1L_2[x]$ we mean $L_1[L_2[x]]$; that is, first the operator immediately to the left of x acts on x, then the operator to the left of that acts on the result. Two operators, say L_1 and L_2, are said to be equal if $L_1[x] = L_2[x]$ for all functions $x(t)$ (that are sufficiently differentiable for L_1 and L_2 to act on them). Finally, *in general, differential operators do not commute*: $L_1L_2 \neq L_2L_1$. For instance, if $L_1 = D$ and $L_2 = tD$, then $L_1L_2[x] = D(tDx) = D(tx') = tx'' + x'$, whereas $L_2L_1[x] = tD(Dx) = tDx' = tx''$. However, they *do commute if their a_j coefficients are constants*. For instance,

$$(2D - 1)(D + 3)x = (2D - 1)(x' + 3x) = 2x'' + 6x' - x' - 3x,$$

and

$$(D + 3)(2D - 1)x = (D + 3)(2x' - x) = 2x'' - x' + 6x' - 3x$$

are identical for all functions $x(t)$, so $(2D-1)(D+3) = (D+3)(2D-1)$.

EXAMPLE 7. To solve the system

$$x' - x - y = 3t, \tag{14a}$$
$$x' + y' - 5x - 2y = 5, \tag{14b}$$

it is convenient to begin by re-expressing it as

$$(D-1)x - y = 3t, \tag{15a}$$
$$(D-5)x + (D-2)y = 5, \tag{15b}$$

or

$$L_1[x] + L_2[y] = 3t, \tag{16a}$$
$$L_3[x] + L_4[y] = 5, \tag{16b}$$

where $L_1 = D-1, L_2 = -1$, and so on. To solve by the method of elimination, let us operate on (16a) with L_3 and on (16b) with L_1, giving

$$L_3L_1[x] + L_3L_2[y] = L_3[3t], \tag{17a}$$
$$L_1L_3[x] + L_1L_4[y] = L_1[5], \tag{17b}$$

where we have used the linearity of L_3 in writing $L_3\,[L_1[x] + L_2[y]]$ as $L_3L_1[x] + L_3L_2[y]$ in obtaining (17a) and, similarly, the linearity of L_1 in obtaining (17b). Subtracting one equation from the other, and cancelling the x terms because $L_3L_1 = L_1L_3$, enables us to eliminate x and to obtain the equation

$$(L_1L_4 - L_3L_2)\,[y] = L_1[5] - L_3[3t] \tag{18}$$

on y alone. At this point we can return to the non-operator form, with $L_1L_4 - L_3L_2 = (D-1)(D-2) - (D-5)(-1) = D^2 - 2D - 3$ and $L_1[5] - L_3[3t] = (D-1)(5) - (D-5)(3t) = 15t - 8$. Thus,

$$y'' - 2y' - 3y = 15t - 8, \tag{19}$$

which admits the general solution

$$y(t) = Ae^{3t} + Be^{-t} - 5t + 6. \tag{20}$$

To find $x(t)$, we can proceed in the same manner. This time, operate on (16a) with L_4 and on (16b) with L_2:

$$L_4L_1[x] + L_4L_2[y] = L_4[3t] \tag{21a}$$
$$L_2L_3[x] + L_2L_4[y] = L_2[5], \tag{21b}$$

and subtraction gives

$$(L_4L_1 - L_2L_3)\,[x] = L_4[3t] - L_2[5], \tag{22}$$

or

$$x'' - 2x' - 3x = 8 - 6t, \tag{23}$$

with general solution

$$x(t) = Ce^{3t} + Ee^{-t} + 2t - 4. \tag{24}$$

(We avoid using D as an integration constant because $D = d/dt$ here.)

It might appear that A, B, C, E are all arbitrary, but don't forget that x and y are related through (14), so these constants might be related as well. In fact, putting (20) and (24) into (14a) gives, after cancellation of terms,

$$(2C - A)e^{3t} - (2E + B)e^{-t} = 0, \tag{25}$$

and the linear independence of e^{3t} and e^{-t} requires that $A = 2C$ and $B = -2E$. Putting (20) and (24) into (14b) gives this same result.

Thus, the general solution of (14) is

$$x(t) = Ce^{3t} + Ee^{-t} + 2t - 4, \tag{26a}$$
$$y(t) = 2Ce^{3t} - 2Ee^{-t} - 5t + 6. \tag{26b}$$

COMMENT 1. With hindsight, it would have been easier to eliminate y first and solve for x since we could have put that x [namely, as given by (26a)] into (14a) and solved that equation for y. That step would have produced (26b) directly.

COMMENT 2. Notice that (14) is not of the "standard" form (10) because (14b) has both x' and y' in it. While we need it to be in that form to apply Theorem 3.9.1, we do not need the system to be of that form to apply the method of elimination. ∎

A review of the steps in the elimination process reveals that the operators $L_1, \ldots, L_4$ might just as well have been constants, by the way we have manipulated them. In fact, a useful way to organize the procedure is to use Cramer's rule (Section 10.6). For instance, if we have two differential equations

$$L_1[x] + L_2[y] = f_1(t), \tag{27a}$$
$$L_3[x] + L_4[y] = f_2(t), \tag{27b}$$

we can, heuristically, use Cramer's rule to write

$$x = \frac{\begin{vmatrix} f_1 & L_2 \\ f_2 & L_4 \end{vmatrix}}{\begin{vmatrix} L_1 & L_2 \\ L_3 & L_4 \end{vmatrix}} = \frac{L_4[f_1] - L_2[f_2]}{L_1L_4 - L_2L_3}, \tag{28a}$$

$$y = \frac{\begin{vmatrix} L_1 & f_1 \\ L_3 & f_2 \end{vmatrix}}{\begin{vmatrix} L_1 & L_2 \\ L_3 & L_4 \end{vmatrix}} = \frac{L_1[f_2] - L_3[f_1]}{L_1L_4 - L_2L_3}. \tag{28b}$$

Of course, the division by an operator on the right-hand sides of (28a,b) is not defined, so we need to put the $L_1L_4 - L_2L_3$ back up on the left-hand side, where it came from. That step gives

$$(L_1L_4 - L_2L_3)[x] = L_4[f_1] - L_2[f_2] \tag{29a}$$
$$(L_1L_4 - L_2L_3)[y] = L_1[f_2] - L_3[f_1], \tag{29b}$$

which equations correspond to (22) and (18), respectively, in Example 7. Again, this approach is heuristic, but it does give the correct result and is readily applied – and extended to systems of three equations in three unknowns, four in four unknowns, and so on.

What might possibly go wrong with our foregoing solution of (27)? In the application of Cramer's rule to linear algebraic equations, the case where the determinant in the denominator vanishes is singular, and either there are no solutions (the system is "inconsistent") or there is an infinite number of them (the system is "redundant"). Likewise, the system (27) is singular if $L_1L_4 - L_2L_3$ is zero and is either inconsistent (with no solution) or redundant (with infinitely many linearly independent solutions). For instance, the system

$$Dx + 2Dy = 1, \tag{30a}$$
$$2Dx + 4Dy = 3 \tag{30b}$$

has $L_1L_4 - L_2L_3 = 4D^2 - 4D^2 = 0$ and has no solution since the left-hand sides are in the ratio 1:2, whereas the right-hand sides are in the ratio 1:3. However, if we change the 3 to a 2, then the new system still has $L_1L_4 - L_2L_3 = 0$ but is now consistent. Indeed, then the second equation is merely twice the first and can be discarded, leaving the single equation $Dx + 2Dy = 1$ in the two unknowns $x(t)$ and $y(t)$. We can choose one of these arbitrarily and use $Dx + 2Dy = 1$ to solve for the other, so there are infinitely many linearly independent solutions. Understand that the singular nature of (30), and the modified system, is intrinsic to those systems and is not a fault of the method of elimination.

In the generic case, however, $L_1L_4 - L_2L_3 \neq 0$ and we can solve (29a) and (29b) for $x(t)$ and $y(t)$, respectively. It can be shown* that the number of independent arbitrary integration constants is the same as the degree of the determinantal polynomial $L_1L_4 - L_2L_3$. In Example 7, for instance, $L_1L_4 - L_2L_3 = D^2 - 2D - 3$ is of second degree, so we could have known in advance that there would be two independent arbitrary constants.

EXAMPLE 8. *Mass-Spring System in Fig. 4.* Let us study the two-mass system shown in Fig. 4, and let $m_1 = m_2 = k_1 = k_{12} = k_2 = 1$ and $F_1(t) = F_2(t) = 0$, for definiteness. Then equations (8) become

*See pages 144–150 in the classic treatise by E. L. Ince, *Ordinary Differential Equations* (New York: Dover, 1956).

$$\left(D^2+2\right)x_1 - x_2 = 0, \tag{31a}$$
$$-x_1 + \left(D^2+2\right)x_2 = 0. \tag{31b}$$

With $L_1 = L_4 = D^2+2$ and $L_2 = L_3 = -1$, and $f_1(t) = f_2(t) = 0$, (29a,b) become

$$\left(D^4+4D^2+3\right)x_1 = 0, \tag{32a}$$
$$\left(D^4+4D^2+3\right)x_2 = 0, \tag{32b}$$

so (Exercise 2)

$$x_1(t) = A\cos t + B\sin t + C\cos\sqrt{3}\,t + E\sin\sqrt{3}\,t, \tag{33a}$$
$$x_2(t) = F\cos t + G\sin t + H\cos\sqrt{3}\,t + I\sin\sqrt{3}\,t. \tag{33b}$$

To determine any relationships among the constants $A, B, \ldots, I$, we put (33) into (31a) [or (32b), the result would be the same] and find that

$$(A-F)\cos t + (B-G)\sin t - (C+H)\cos\sqrt{3}\,t - (E+I)\sin\sqrt{3}\,t = 0,$$

from which we learn that $F = A$, $G = B$, $H = -C$, and $I = -E$, so the general solution of (31) is

$$x_1(t) = A\cos t + B\sin t + C\cos\sqrt{3}\,t + E\sin\sqrt{3}\,t, \tag{34a}$$
$$x_2(t) = A\cos t + B\sin t - C\cos\sqrt{3}\,t - E\sin\sqrt{3}\,t. \tag{34b}$$

The determinantal polynomial was of fourth degree and, as asserted above, there are four independent arbitrary integration constants. There are important things to say about the result expressed in (34):

COMMENT 1. It will be more illuminating to re-express (34) in the form

$$x_1(t) = G\sin(t+\phi) + H\sin\left(\sqrt{3}\,t+\psi\right), \tag{35a}$$
$$x_2(t) = G\sin(t+\phi) - H\sin\left(\sqrt{3}\,t+\psi\right), \tag{35b}$$

where the four constants G, H, ϕ, ψ are determined from the initial conditions $x_1(0)$, $x_1'(0)$, $x_2(0)$, and $x_2'(0)$. While neither $x_1(t)$ nor $x_2(t)$ is a pure sinusoid, each is a superposition of two pure sinusoids, the frequencies of which are characteristics of the system (i.e., independent of the initial conditions). Those frequencies, $\omega = 1$ rad/sec and $\omega = \sqrt{3}$ rad/sec, are the **natural frequencies** of the system. If the initial conditions are such that $H = 0$, then the motion is of the form

$$x_1(t) = G\sin(t+\phi), \quad x_2(t) = G\sin(t+\phi); \tag{36}$$

that is, the two masses swing in unison at the lower frequency $\omega = 1$. Such a motion is called a **low mode** motion because it is at the lower of the two natural frequencies. If instead the initial conditions are such that $G = 0$, then

$$x_1(t) = H\sin(\sqrt{3}\,t+\psi), \quad x_2(t) = -H\sin(\sqrt{3}\,t+\psi); \tag{37}$$

the masses swing in opposition, at the higher frequency $\omega = \sqrt{3}$ rad/sec, so the latter is called a **high mode** motion. For instance, the initial conditions $x_1(0) = x_2(0) = 1$ and $x_1'(0) = x_2'(0) = 0$ give (Exercise 7) the purely low mode motion

$$\begin{aligned} x_1(t) &= \sin(t + \pi/2) = \cos t, \\ x_2(t) &= \sin(t + \pi/2) = \cos t, \end{aligned} \tag{38}$$

and the conditions $x_1(0) = 1$, $x_2(0) = -1$, and $x_1'(0) = x_2'(0) = 0$ give the purely high mode motion

$$\begin{aligned} x_1(t) &= \sin(\sqrt{3}\,t + \pi/2) = \cos\sqrt{3}\,t, \\ x_2(t) &= -\sin(\sqrt{3}\,t + \pi/2) = -\cos\sqrt{3}\,t. \end{aligned} \tag{39}$$

If, instead, $x_1(0) = 1$ and $x_2(0) = x_1'(0) = x_2'(0) = 0$, say, then both G and H will be nonzero and the motion will be a linear combination of the low and high modes.

COMMENT 2. Why is the frequency corresponding to the masses swinging in opposition higher than that corresponding to the masses swinging in unison? Remember from the single-mass case studied in Section 3.5 that the natural frequency in that case is $\sqrt{k/m}$; that is, the stiffer the system (the larger the value of k), the higher the frequency. For the two-mass system, observe that in the low mode the middle spring is completely inactive, whereas in the high mode it is being stretched and compressed. Thus, there is more stiffness encountered in the high mode, so the high mode frequency is higher.

COMMENT 3. Just as the free vibration of a single mass is governed by one differential equation, $mx'' + kx = 0$, and has a single mode of vibration with natural frequency $\omega = \sqrt{k/m}$, a two-mass system is governed by two differential equations and its general vibration is a linear combination of two modes (unison and opposition in this example), each with its own natural frequency. Similarly, the free vibration of an n-mass system will be governed by n differential equations, and its general vibration will be a linear combination of n distinct modes, each with its own pattern and natural frequency. In the limit, we can think of a continuous system, such as a beam, as an infinite-mass system, an infinite number of tiny masses connected together. In that limiting case, in place of an infinite number of ordinary differential equations we obtain a *partial* differential equation on the deflection $y(x, t)$, solution of which yields the general solution as a linear combination of an infinite number of discrete modes of vibration. In applications it is important to know the natural frequencies of a given system because if it is driven by a harmonic forcing function, then it will have a large, perhaps catastrophic, response if the driving frequency is close to one of the natural frequencies.

COMMENT 4. Finally, we note that molecules and atoms can be modeled as mass-spring systems, and the spectrum of the natural frequencies are of great importance in determining their allowable energy levels. ∎

We will have more to say about the foregoing example later, when we study matrix theory and the eigenvalue problem.

Observe that once a system of linear constant coefficient equations is converted by the process of elimination to a set of uncoupled equations such as (32a,b), the

homogeneous solutions of those equations can be sought in the usual exponential form. In fact, one can do that even at the outset, without first going through the process of elimination. For instance, to solve (31a,b) one can start out by seeking a solution in the form $x_1(t) = \xi_1 e^{rt}$ and $x_2(t) = \xi_2 e^{rt}$. Putting those forms into (31a,b) gives what is known as an *eigenvalue problem* on the unknown constants ξ_1, ξ_2 and r. That discussion is best reserved for the chapters on matrix theory and linear algebra, as an important application of the eigenvalue problem, so we will not pursue it in the present section.

Closure. Systems of ordinary differential equations arise in the modeling of physical systems that involve more than one dependent variable. For instance, in modeling an ecological system such as the fish populations in a given lake, the dependent variables might be the populations of each fish species, as a function of the independent variable t. Surely these populations are interrelated (for instance, one species might be the primary food supply for another species), so the governing differential equations will be coupled. It is precisely the coupling that produces the interest in this section because if they are not coupled then we can solve them, individually, by the methods developed in preceding sections.

Our first step was to give the basic existence and uniqueness theorem. That theorem guaranteed both existence and uniqueness, under rather mild conditions of continuity, over an interval that is known in advance. The theorem applied to first-order systems, but we showed that systems of higher order can be converted to first-order systems by suitable introduction of auxiliary dependent variables.

Then we outlined a method of elimination for systems with constant coefficients. Elimination is similar to the steps in the solution of linear algebraic equations by Gauss elimination, where the coefficients of the unknowns are operators rather than numbers. The correct result can even be obtained by using Cramer's rule, provided that the determinantal operator in the denominator does not vanish, and provided that we move that operator back "upstairs" – as we did in converting (28) to (29). If the operator does vanish, then the problem is singular and there will be no solution or infinitely many linearly independent solutions.

In subsequent chapters on matrix theory we shall return to systems of linear differential equations with constant coefficients and develop additional solution techniques that are based upon the so-called eigenvalue problem.

Computer software. Often, one can solve systems of differential equations using computer-algebra systems. For instance, to find the general solution of the system

$$\begin{aligned}(D+1)x + 2y &= 0,\\ 3x + (D+2)y &= 0\end{aligned}$$

using *Maple*, enter

$$\text{dsolve}(\{\text{diff}(x(t),t) + x(t) + 2*y(t) = 0,\ 3*x(t) + \text{diff}(y(t),t)$$
$$+2*y(t) = 0\}, \{x(t), y(t)\});$$

and return. The result is the general solution

$$\{\, y(t) = _C2 \exp(t) + 3/2\, _C1 \exp(-4t),$$
$$x(t) = -_C2 \exp(t) + _C1 \exp(-4t)\,\}$$

If we wish to include initial conditions $x(0) = 3$, $y(0) = 2$, use instead the command

$$\text{dsolve}(\{\text{diff}(x(t), t) + x(t) + 2 * y(t) = 0,\ 3 * x(t) + \text{diff}(y(t), t) + 2 * y(t) = 0,$$
$$x(0) = 3,\ y(0) = 2\},\ \{x(t), y(t)\});$$

The result is the particular solution

$$\{y(t) = -\exp(t) + 3\exp(-4t),\ \ x(t) = \exp(t) + 2\exp(-4t)\}$$

Alternatively, one can first define the two equations and then call them in the dsolve command. That is, enter

$$\text{deq1} := \text{diff}(x(t), t) + x(t) + 2 * y(t) = 0:$$
$$\text{deq2} := 3 * x(t) + \text{diff}(y(t), t) + 2 * y(t) = 0:$$

The colon at the end of each line indicates that it is a definition, not a command. Commands are followed by semicolons. Now, enter the dsolve command:

$$\text{dsolve}(\{\text{deq1}, \text{deq2}, x(0) = 3,\ y(0) = 2\},\ \{x(t), y(t)\});$$

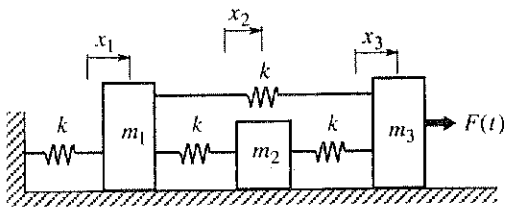
and return. The result is

$$\{y(t) = -\exp(t) + 3\exp(-4t),\ \ x(t) = \exp(t) + 2\exp(-4t)\}$$

EXERCISES 3.9

1. Derive the solution (20) of (19).

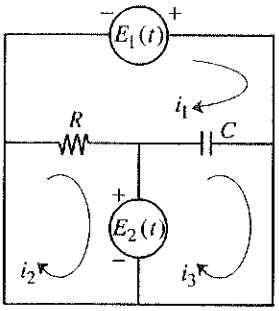
2. Derive the solutions (33a,b) of (32a,b).



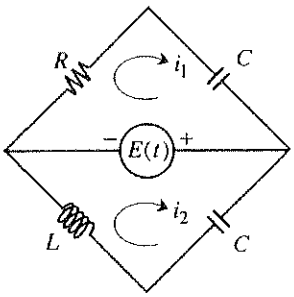
3. Derive the system of differential equations governing the displacements $x_j(t)$, using the assumption that $x_1 > x_2 > x_3 > 0$. Repeat the derivation assuming instead that $x_3 > x_2 > x_1 > 0$ and again, assuming that $x_1 > x_3 > x_2 > 0$, and show that the resulting equations are the same, independent of these different assumptions.

4.(a),(b),(c) Derive the system of differential equations governing the currents $i_j(t)$, but you need not solve them. State any physical laws that you use.

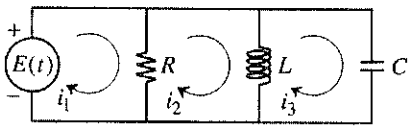
(a)



(b)



(c)



5. Obtain the general solution by the method of elimination, either step-by-step or using the Cramer's rule shortcut.

(a) $(D-1)x + Dy = 0$
$(D+1)x + (2D+2)y = 0$

(b) $(D-1)x + 2Dy = 0$
$(D+1)x + 4Dy = 0$

(c) $Dx + (D-1)y = 5$
$2(D+1)x + (D+1)y = 0$

(d) $x' + y' = y + t$
$x' - 3y' = -x + 2$

(e) $x' = \sin t - y$
$y' = -9x + 4$

(f) $x' = x - 8y$
$y' = -x - y - 3t^2$

(g) $x' = 2x + 6y - t + 7$
$y' = 2x - 2y$

(h) $2x' + y' + x + y = t^2 - 1$
$x' + y' + x + y = 0$

(i) $x' + y' + x - y = e^t$
$x' + 2y' + 2x - 2y = 1 - t$

(j) $x' - 3x + y = 4\sin 2t$
$3x + y' - y = 6$

(k) $x'' = x - 4y$
$y'' = -2x - y$

(l) $x'' = x - 2y$
$y'' = 2x - 4y$

(m) $x'' - x + 2y = 0$
$-2x + y'' + 4y = 1 - t^2$

(n) $x'' - x + 3y = 0$
$y'' + x + y = 4$

(o) $x'' + x + y = 24$
$y'' + 3x - y = -8$

(p) $x'' + y'' = x$
$3x'' - y'' = y + 6$

(q) $(2D^2 + 3)x + (2D + 1)y = 4e^{3t} - 7$
$Dx + (D - 2)y = 2$

6. (a)–(q) Find the general solution of the corresponding problem in Exercise 5 using computer software. Separately, make up any set of initial conditions, and use the computer to find the particular solution corresponding to those initial conditions.

7. (*Mass-spring system of Examples 3 and 8*) (a) Derive the particular solutions (38) and (39) from the general solution (35) by applying the given sets of initial conditions.
(b) Evaluate G, H, ϕ, ψ for the initial conditions $x_1(0) = 1, x_2(0) = x_1'(0) = x_2'(0) = 0$, and show that both modes are present in the solution. Obtain a computer plot of $x_1(t)$ and $x_2(t)$, over $0 \le t \le 20$ (so as to show several cycles).
(c) Same as (b), for $x_1'(0) = 1, x_1(0) = x_2(0) = x_2'(0) = 0$.
(d) Same as (b), for $x_1(0) = x_2(0) = 0, x_1'(0) = 2, x_2'(0) = 3$.
(e) Same as (b), for $x_1(0) = x_2(0) = 0, x_1'(0) = 2$, $x_2'(0) = -1$.

8. (*Chemical kinetics*) Two substances, with concentrations $x(t)$ and $y(t)$, react to form a third substance, with concentration $z(t)$. The reaction is governed by the system $x' + \alpha x = 0, z' = \beta y$ and $x + y + z = \gamma$, where α, β, γ are known positive constants. Solve for $x(t)$, $y(t)$, $z(t)$, subject to the initial conditions $z(0) = z'(0) = 0$ for these cases:

(a) $\alpha \neq \beta$
(b) $\alpha = \beta$ HINT: Apply l'Hôpital's rule to your answer to part (a).

9. (*Motion of a charged mass*) Consider a particle of mass m, carrying an electrical charge q, and moving in a uniform magnetic field of strength B. The field is in the positive z direction. The equations of motion of the particle are

$$\begin{aligned} mx'' &= qBy', \\ my'' &= -qBx', \\ mz'' &= 0, \end{aligned} \tag{9.1}$$

where $x(t), y(t), z(t)$ are the x, y, z displacements as a function of the time t.

(a) Find the general solution of (9.1) for $x(t), y(t), z(t)$. How many independent arbitrary constants of integration are there?

(b) Show that by a suitable choice of initial conditions the motion can be a circle in the x, y plane, of any desired radius R and centered at any desired point x_0, y_0. Propose such a set of initial conditions.
(c) Besides a circular motion in a constant z plane, are any other types of motion possible? Explain.

10. Show that the given system is singular (i.e., either inconsistent or redundant). If it has no solutions show that; if it has solutions find them.

(a) $x' - x + y = t$
$x'' + y' - x + y = t^2$

(b) $(D-1)x + y = t$
$(D^2-1)x + (D+1)y = t+1$

(c) $(D+1)x - Dy = e^t$
$(D^2-1)x - (D^2-D)y = 0$

(d) $(D+1)x + Dy = e^t$
$(D^2-1)x + (D^2-D)y = 3t$

Chapter 3 Review

A differential equation is far more tractable, insofar as analytical solution is concerned, if it is linear than if it is nonlinear. We could see a hint of that even in Chapter 2, where we were able to derive the general solution of the general first-order linear equation $y' + p(x)y = q(x)$ but had success with nonlinear equations only in special cases. In fact, for linear equations of any order (first, second, or higher) a number of important results follow.

The most important is that for an nth-order *linear* differential equation $L[y] = f(x)$, with constant coefficients or not, a general solution is expressible as the sum of a general solution $y_h(x)$ to the homogeneous equation $L[y] = 0$, and any particular solution $y_p(x)$ to the full equation $L[y] = f$:

$$y(x) = y_h(x) + y_p(x).$$

In turn, $y_h(x)$ is expressible as an arbitrary linear combination of any n LI (linearly independent) solutions of $L[y] = 0$:

$$y_h(x) = C_1 y_1(x) + \cdots + C_n y_n(x).$$

Thus, linear independence is introduced early, in Section 3.2, and theorems are provided for testing a given set of functions to see if they are LI or not. We then show how to find the solutions $y_1(x), \ldots, y_n(x)$ for the following two extremely important cases: for constant-coefficient equations and for Cauchy–Euler equations.

For *constant-coefficient equations* the idea is to seek $y_h(x)$ in the exponential form $e^{\lambda x}$. Putting that form into $L[y] = 0$ gives an nth-degree polynomial equation on λ, called the *characteristic equation*. Each nonrepeated root λ_j contributes a solution $e^{\lambda_j x}$, and each repeated root λ_j of order k contributes k solutions $e^{\lambda_j x}, xe^{\lambda_j x}, \ldots, x^{k-1}e^{\lambda_j x}$.

For *Cauchy–Euler equations* the form $e^{\lambda x}$ does *not* work. Rather, the idea is to seek $y_h(x)$ in the power form x^λ. Each nonrepeated root λ_j contributes a solution x^{λ_j}, and each repeated root λ_j of order k contributes k solutions $x^{\lambda_j}, (\ln x)x^{\lambda_j}, \ldots, (\ln x)^k x^{\lambda_j}$.

Two different methods are put forward for finding particular solutions, the method of *undetermined coefficients* and Lagrange's method of *variation of parameters*. Undetermined coefficients is easier to apply but is subject to the conditions that

(i) besides being linear, L must be of constant-coefficient type, and

(ii) repeated differentiation of each term in f must produce only a finite number of LI terms.

Variation of parameters, on the other hand, merely requires L to be linear. According to the method, we vary the parameters (i.e., the constants of integration in y_h) $C_1, \ldots, C_n$, and seek $y_p(x) = C_1(x)y_1(x) + \cdots + C_n(x)y_n(x)$. Putting that form into the given differential equation gives one condition on the $C_j(x)$'s. That condition is augmented by $n-1$ additional conditions that are designed to preclude the presence of derivatives of the $C_j(x)$'s that are of order higher than first.

In Section 3.8 we study the *harmonic oscillator*, both damped and undamped, both free and driven. Of special interest are the concepts of *natural frequency* for the undamped case, *critical damping, amplitude- and frequency-response curves, resonance*, and *beats*. This application is of great importance in engineering and science and should be understood thoroughly.

Finally, Section 3.9 is devoted to systems of linear differential equations. We give an existence/uniqueness theorem and show how to solve systems by elimination.

Chapter 4

Power Series Solutions

PREREQUISITES: This chapter presumes a familiarity with the complex plane and the algebra of complex numbers, material which is covered in Section 21.2.

4.1 Introduction

In Chapter 2 we presented a number of methods for obtaining analytical closed form solutions of first-order differential equations, some of which methods could be applied even to nonlinear equations. In Chapter 3 we studied equations of second order and higher, and found them to be more difficult. Restricting our discussion to linear equations, even then we were successful in developing solutions only for the (important) cases of equations with constant coefficients and Cauchy-Euler equations. We also found that we can solve nonconstant-coefficient equations if we can factor the differential operator, but such factorization can be accomplished only in exceptional cases.

In Chapter 4 we continue to restrict our discussion to linear equations, but we now study nonconstant-coefficient equations. That case is so much more difficult than the constant-coefficient case that we do two things: we consider only second-order equations, and we give up on the prospect of finding solutions in closed form and seek solutions in the form of infinite series.

To illustrate the idea that is developed in the subsequent sections, consider the simple example

$$\frac{dy}{dx} + y = 0. \tag{1}$$

To solve by the series method, we seek a solution in the form of a power series expansion about any desired point $x = x_0$, $y(x) = \sum_{n=0}^{\infty} a_n(x - x_0)^n$, where the a_n coefficients are to be determined so that the assumed form satisfies the given differential equation (1), if possible. If we choose $x_0 = 0$ for simplicity, then

$$y(x) = \sum_{0}^{\infty} a_n x^n = a_0 + a_1 x + a_2 x^2 + \cdots, \tag{2a}$$

and

$$\frac{dy}{dx} = \frac{d}{dx}\left(a_0 + a_1x + a_2x^2 + \cdots\right) = a_1 + 2a_2x + 3a_3x^2 + \cdots. \tag{2b}$$

Putting (2a,b) into (1) gives

$$\left(a_1 + 2a_2x + 3a_3x^2 + \cdots\right) + \left(a_0 + a_1x + a_2x^2 + \cdots\right) = 0, \tag{3}$$

or, rearranging terms,

$$(a_1 + a_0) + (2a_2 + a_1)\,x + (3a_3 + a_2)\,x^2 + \cdots = 0. \tag{4}$$

If we realize that the right side of (4) is really $0 + 0x + 0x^2 + \cdots$, then, by equating coefficients of like powers of x on both sides of (4), we obtain $a_1 + a_0 = 0$, $2a_2 + a_1 = 0$, $3a_3 + a_2 = 0$, and so on. Thus,

$$\begin{aligned} a_1 &= -a_0, \\ a_2 &= -a_1/2 = -(-a_0)/2 = a_0/2, \\ a_3 &= -a_2/3 = -(a_0/2)/3 = -a_0/6, \end{aligned} \tag{5}$$

and so on, where a_0 remains arbitrary. Thus, we have

$$y(x) = a_0\left(1 - x + \frac{1}{2}x^2 - \frac{1}{6}x^3 + \cdots\right), \tag{6}$$

as the general solution to (1). Here, a_0 is the constant of integration; we could rename it C, for example, if we wish. Thus, we have the solution – not in closed form but as a power series. In this simple example we are able to "sum the series" into closed form, that is, to identify it as the Taylor series of e^{-x}, so that our general solution is really $y(x) = Ce^{-x}$. However, for nonconstant-coefficient differential equations we are generally not so fortunate, and must leave the solution in series form.

As simple as the above steps appear, there are several questions that need to be addressed before we can have confidence in the result given by (6):

(i) In (2b) we differentiated an infinite series term by term. That is, we interchanged the order of the differentiation and the summation and wrote

$$\frac{d}{dx}\sum a_nx^n = \sum \frac{d}{dx}\left(a_nx^n\right). \tag{7}$$

That step looks reasonable, but observe that it amounts to an interchange in the order of two operations, the summation and the differentiation, and it is possible that reversing their order might give different results. For instance, do we get the same results if we put toothpaste on our toothbrush and then brush, or if we brush and then put toothpaste on the brush?

(ii) Re-expressing (3) in the form of (4) is based on a supposition that we can add series term by term:

$$\sum A_n + \sum B_n = \sum (A_n + B_n) . \tag{8}$$

Again, that step looks reasonable, but is it necessarily correct?

(iii) Finally, inferring (5) from (4) is based on a supposition that if

$$\sum A_n x^n = \sum B_n x^n \tag{9}$$

for all x in some interval of interest, then it must be true that $A_n = B_n$ for each n. Though reasonable, does it really follow that for the sums to be the same the corresponding individual terms need to be the same?

Thus, there are some technical questions that we need to address, and we do that in the next section. Our approach, in deriving (6), was heuristic, not rigorous, since we did not attend to the issues mentioned above. We can sidestep the several questions of rigor that arose in deriving the series (6) if, instead, we verify, a posteriori, that (6) does satisfy the given differential equation (1). However, that procedure begs exactly the same questions: termwise differentiation of the series, termwise addition of series, and equating the coefficients of like powers of x on both sides of the equation.

Here is a brief outline of this chapter:

4.2 *Power Series Solutions.* In Section 4.2, we review infinite series, power series, and Taylor series, then we show how to find solutions to the equation $y'' + p(x)y' + q(x)y = 0$ in the form of a power series about a chosen point x_0 if $p(x)$ and $q(x)$ are sufficiently well-behaved at x_0.

4.3 *The Method of Frobenius.* If $p(x)$ and $q(x)$ are not sufficiently well-behaved at x_0, then the singular behavior of p and/or q gets passed on in some form to the solutions of the differential equation; hence those solutions cannot be found in power series form. Yet, if $p(x)$ and $q(x)$ are not too singular at x_0, then solutions can still be found, but in a more general form, a so-called Frobenius series. Section 4.3 puts forward the theoretical base for such solutions and the procedure whereby to obtain them.

4.4 *Legendre Functions.* This section focuses on a specific important example, the Legendre equation $(1 - x^2)y'' - 2xy' + \lambda y = 0$, where λ is a constant.

4.5 *Singular Integrals; Gamma Function.* Singular integrals are defined and their convergence is discussed. An important singular integral, the gamma function, is introduced and studied.

4.6 *Bessel Functions.* Besides the Legendre equation, we need to study the extremely important Bessel equation, $x^2y'' + xy' + (x^2 - \nu^2)y = 0$, where ν is a constant, but preparatory to that study we first need to introduce singular integrals and the gamma function, which will be needed again in Chapter 5 in any case.

4.2 Power Series Solutions

4.2.1. Review of power series. Whereas a finite sum,

$$\sum_{k=1}^{N} a_k = a_1 + a_2 + \cdots + a_N, \tag{1}$$

is well-defined thanks to the commutative and associative laws of addition, an infinite sum, or **infinite series**,

$$\sum_{k=1}^{\infty} a_k = a_1 + a_2 + a_3 + \cdots, \tag{2}$$

is not. For example, is the series $\sum_1^\infty (-1)^{k-1} = 1 - 1 + 1 - 1 + \cdots$ equal to $(1-1) + (1-1) + \cdots = 0 + 0 + \cdots = 0$? Is it (by grouping differently) $1 - (1-1) - (1-1) - \cdots = 1 - 0 - 0 - \cdots = 1$? In fact, besides grouping the numbers in different ways we could rearrange their order as well. The point, then, is that (2) is not self-explanatory, it needs to be defined; we need to decide, or be told, how to do the calculation. To give the traditional definition of (2), we first define the sequence of **partial sums** of the series (2) as

$$s_1 = a_1, \qquad s_2 = a_1 + a_2, \qquad s_3 = a_1 + a_2 + a_3, \tag{3}$$

and so on:

$$s_n = \sum_{k=1}^{n} a_k, \tag{4}$$

where a_k is called the kth **term** of the series. If the limit of the sequence s_n exists, as $n \to \infty$, and equals some number s, then we say that the series (2) is **convergent**, and that it **converges to** s; otherwise it is **divergent**. That is, an infinite series is defined as the limit (if that limit exists) of its sequence of partial sums:

$$\sum_{k=1}^{\infty} a_k \equiv \lim_{n\to\infty} \sum_{k=1}^{n} a_k = \lim_{n\to\infty} s_n = s. \tag{5}$$

That definition, known as **ordinary convergence**, is not the only one possible. For instance, another definition, due to Cesàro, is discussed in the exercises. However, ordinary convergence is the traditional definition and is the one that is understood unless specifically stated otherwise.

Recall from the calculus that by $\lim_{n\to\infty} s_n = s$, in (5), we mean that to each number $\epsilon > 0$, no matter how small, there exists an integer N such that $|s - s_n| < \epsilon$ for all $n > N$. (Logically, the words "no matter how small" are unnecessary, but we include them for emphasis.) In general, the smaller the chosen ϵ, the larger the N that is needed, so that N is a function of ϵ.

The significance of the limit concept cannot be overstated, for in mathematics it is often as limits of "old things" that we introduce "new things." For instance,

the derivative is introduced as the limit of a difference quotient, the Riemann integral is introduced as the limit of a sequence of Riemann sums, infinite series are introduced as limits of sequences of partial sums, and so on.

To illustrate the definition of convergence given above, consider two simple examples. The series $1 + 1 + 1 + \cdots$ diverges because $s_n = n$ fails to approach a limit as $n \to \infty$. However, for a series to diverge its partial sums need not grow unboundedly. For instance, the series $1-1+1-1+\cdots$, mentioned above, diverges because its sequence of partial sums (namely, $1, 0, 1, 0, 1, \ldots$) fails to approach a limit. Of course, determining whether a series is convergent or divergent is usually much harder than for these examples. Ideally, one would like a theorem that gives necessary and sufficient conditions for convergence. Here is such a theorem.

THEOREM 4.2.1 *Cauchy Convergence Theorem*
An infinite series is convergent if and only if its sequence of partial sums s_n is a *Cauchy sequence* – that is, if to each $\epsilon > 0$ (no matter how small) there corresponds an integer $N(\epsilon)$ such that $|s_m - s_n| < \epsilon$ for all m and n greater than N.

Unfortunately, this theorem is difficult to apply, so one develops (in the calculus) an array of theorems (i.e., tests for convergence/divergence), each of which is more specialized (and hence less powerful) than the Cauchy convergence theorem, but easier to apply. For instance, if in Theorem 4.2.1 we set $m = n - 1$, then the stated condition becomes: to each $\epsilon > 0$ (no matter how small) there corresponds an integer $N(\epsilon)$ such that $|s_m - s_n| = |a_n| < \epsilon$ for all $n > N$. The latter is equivalent to saying that $a_n \to 0$ as $n \to \infty$. Thus, we have the specialized, but readily applied, theorem that *for the series $\sum^{\infty} a_n$ to converge, it is necessary (but not sufficient) that $a_n \to 0$ as $n \to \infty$.* From this theorem it follows immediately that the series $1 + 1 + 1 + \cdots$ and $1 - 1 + 1 - 1 + \cdots$, cited above, both diverge because in each case the terms do not tend to zero.

Let us now focus on the specific needs of this chapter, **power series** – that is, series of the form

$$\sum_{0}^{n} a_n(x - x_0)^n = a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + \cdots, \tag{6}$$

where the a_n's are numbers called the **coefficients** of the series, x is a variable, and x_0 is a fixed point called the **center** of the series. We say that the expansion is "about the point x_0." In a later chapter we study complex series, but in this chapter we restrict all variables and constants to be real. Notice that the quantity $(x - x_0)^n$ on the left side of (6) is the indeterminate form 0^0 when $n = 0$ and $x = x_0$; that form must be interpreted as 1 if the leading term of the series is to be a_0, as desired.

The terms in (6) are now functions of x rather than numbers, so that the series may converge at some points on the x axis and diverge at others. At the very least (6) converges at $x = x_0$ since then it reduces to the single term a_0.

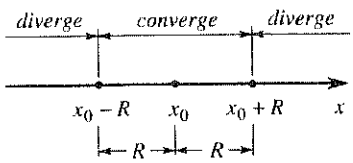


Figure 1. Interval of convergence of power series.

THEOREM 4.2.2 *Interval of Convergence of Power Series*
The power series (6) converges at $x = x_0$. If it converges at other points as well, then those points necessarily comprise an interval $|x - x_0| < R$ centered at x_0 and, possibly, one or both endpoints of that interval (Fig. 1), where R can be determined from either of the formulas

$$R = \frac{1}{\lim\limits_{n\to\infty} \left|\dfrac{a_{n+1}}{a_n}\right|} \qquad \text{or,} \qquad R = \frac{1}{\lim\limits_{n\to\infty} \sqrt[n]{|a_n|}}, \tag{7a,b}$$

if the limits in the denominators exist and are nonzero. If the limits in (7a,b) are zero, then (6) converges for all x (i.e., for every finite x, no matter how large), and we say that "$R = \infty$." If the limits fail to exist by virtue of being infinite, then $R = 0$ and (6) converges only at x_0.

We call $|x - x_0| < R$ the **interval of convergence**, and R the **radius of convergence**. If a power series converges to a function f on some interval, we say that it **represents** f on that interval, and we call f its **sum function**.

EXAMPLE 1. Consider $\sum_0^\infty n!\,x^n$, so $a_n = n!$ and $x_0 = 0$. Then (7a) is easier to apply than (7b), and gives $R = 1/\lim\limits_{n\to\infty} \dfrac{(n+1)!}{n!} = 1/\lim\limits_{n\to\infty}(n+1) = 1/\infty = 0$, so the series converges only at $x = x_0 = 0$. ∎

EXAMPLE 2. Consider $\sum_0^\infty (-1)^n\,[(x+5)/2]^n$. Then $a_n = (-1)^n/2^n$, $x_0 = -5$, and (7a) gives $R = 1/\lim\limits_{n\to\infty}\left|\dfrac{(-1)^{n+1}}{2^{n+1}}\dfrac{2^n}{(-1)^n}\right| = 1/\lim\limits_{n\to\infty}\dfrac{1}{2} = 1/\left(\dfrac{1}{2}\right) = 2$, so the series converges in $|x+5| < 2$ and diverges in $|x+5| > 2$. For $|x+5| = 2$ ($x = -7, -3$) the theorem gives no information. However, we see that for $x = -7$ and -3 the terms do not tend to zero as $n \to \infty$, so the series diverges for $x = -7$ and -3. ∎

EXAMPLE 3. Consider $\displaystyle\sum_4^\infty \frac{(x-1)^n}{(n+1)^n}$. Then $a_n = (n+1)^{-n}$, $x_0 = 1$, and (7b) gives $R = 1/\lim\limits_{n\to\infty}\sqrt[n]{(n+1)^{-n}} = 1/\lim\limits_{n\to\infty}\dfrac{1}{n+1} = 1/0 = \infty$, so the series converges for all x; that is, the interval of convergence is $|x-1| < \infty$. ∎

EXAMPLE 4. Consider the series

$$1 + \frac{(x-3)^2}{5} + \frac{(x-3)^4}{5^2} + \cdots = \sum_0^\infty \frac{1}{5^n}(x-3)^{2n}. \tag{8}$$

This series is not of the form (6) because the powers of $x-3$ proceed in steps of 2. However, if we set $X = (x-3)^2$, then we have the standard form $\sum_0^\infty \frac{1}{5^n} X^n$, with $a_n = 1/5^n$ and $\lim_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right| = \lim_{n\to\infty} \left|\frac{5^n}{5^{n+1}}\right| = \frac{1}{5}$. Thus, $R = 5$, and the series converges in $|X| < 5$ (i.e., $|x-3| < \sqrt{5}$), and diverges in $|X| > 5$ (i.e., $|x-3| > \sqrt{5}$). ∎

Recall from our introductory example, in Section 4.1, that several questions arose regarding the manipulation of power series. The following theorem answers those questions and, therefore, will be needed when we apply the power series method of solution.

THEOREM 4.2.3 *Manipulation of Power Series*
(a) *Termwise differentiation (or integration) permissible.* A power series may be differentiated (or integrated) termwise (i.e., term by term) within its interval of convergence I. The series that results has the same interval of convergence I and represents the derivative (or integral) of the sum function of the original series.
(b) *Termwise addition (or subtraction or multiplication) permissible.* Two power series (about the same point x_0) may be added (or subtracted or multiplied) termwise within their common interval of convergence I. The series that results has the same interval of convergence I and represents the sum (or difference or product) of their two sum functions.
(c) *If two power series are equal, then their corresponding coefficients must be equal.* That is, for

$$\sum_0^\infty a_n(x-x_0)^n = \sum_0^\infty b_n(x-x_0)^n \tag{9}$$

to hold in some common interval of convergence, it must be true that $a_n = b_n$ for each n. In particular, if

$$\sum_0^\infty a_n(x-x_0)^n = 0 \tag{10}$$

in some interval, then each a_n must be zero.

Part (a) means that if $f(x) = \sum_0^\infty a_n(x-x_0)^n$ within I, then

$$f'(x) = \frac{d}{dx}\sum_0^\infty a_n(x-x_0)^n = \sum_0^\infty \frac{d}{dx}\left[a_n(x-x_0)^n\right] = \sum_1^\infty n a_n(x-x_0)^{n-1} \tag{11}$$

and

$$\begin{aligned}\int_a^b f(x)\,dx &= \int_a^b \sum_0^\infty a_n(x-x_0)^n\,dx \\ &= \sum_0^\infty a_n \int_a^b (x-x_0)^n\,dx \\ &= \sum_0^\infty a_n \frac{(b-x_0)^{n+1}-(a-x_0)^{n+1}}{n+1} \qquad (12)\end{aligned}$$

within I, where a, b are any two points within I.

Part (b) means that if $f(x) = \sum_0^\infty a_n(x-x_0)^n$ and $g(x) = \sum_0^\infty b_n(x-x_0)^n$ on I, then

$$f(x) \pm g(x) = \sum_0^\infty (a_n \pm b_n)(x-x_0)^n, \qquad (13)$$

and, with $z = x - x_0$ for brevity,

$$\begin{aligned}f(x)g(x) &= \left(\sum_0^\infty a_n z^n\right)\left(\sum_0^\infty b_n z^n\right) \\ &= (a_0 + a_1 z + \cdots)(b_0 + b_1 z + \cdots) \\ &= a_0\left(b_0 + b_1 z + b_2 z^2 + \cdots\right) + a_1 z\left(b_0 + b_1 z + b_2 z^2 + \cdots\right) \\ &\quad + a_2 z^2\left(b_0 + b_1 z + b_2 z^2 + \cdots\right) + \cdots \\ &= a_0 b_0 + (a_0 b_1 + a_1 b_0)\,z + \cdots \\ &= \sum_0^\infty (a_0 b_n + a_1 b_{n-1} + \cdots + a_n b_0)\,z^n \qquad (14)\end{aligned}$$

within I. The series on the right-hand side of (14) is known as the **Cauchy product** of the two series. Of course, if the two convergence intervals have different radii, then the common interval means the smaller of the two.

In summary, we see that convergent power series can be manipulated in essentially the same way as if they were finite-degree polynomials.

The last items to address, before coming to the power series method of solution of differential equations, are Taylor series and analyticity. Recall from the calculus that the **Taylor series** of a given function $f(x)$ about a chosen point x_0, which we denote here as TS $f|_{x_0}$, is defined as the infinite series

$$\begin{aligned}\text{TS } f|_{x_0} &= f(x_0) + \frac{f'(x_0)}{1!}(x-x_0) + \frac{f''(x_0)}{2!}(x-x_0)^2 + \cdots \\ &= \sum_0^\infty \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n, \qquad (15)\end{aligned}$$

where $0! = 1$. The purpose of Taylor series is to represent the given function, so the fundamental question is: does it? Does the Taylor series really converge to $f(x)$ on some x interval, in which case we can write, in place of (15),

$$f(x) = \sum_{0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n. \tag{16}$$

For that to be the case we need three conditions to be met:

(i) First, we need f to *have* a Taylor series (15) about that point. Namely, f must be infinitely differentiable at x_0 so that all of the coefficients $f^{(n)}(x_0)/n!$ in (15) exist.

(ii) Second, we need the resulting series in (15) to *converge* in some interval $|x - x_0| < R$, for $R > 0$.

(iii) Third, we need the sum of the Taylor series to equal f in the interval, so that the Taylor series *represents* f over that interval – which is, after all, our objective.

The third condition might seem strange, for how could the Taylor series of $f(x)$ converge, but to something other than $f(x)$? Such cases can indeed be put forward, but they are somewhat pathological and not likely to be encountered in applications.

If a function is represented in some nonzero interval $|x - x_0| < R$ by its Taylor series [i.e., TS $f|_{x_0}$ exists, and converges to $f(x)$ there], then f is said to be **analytic** at x_0. If a function is not analytic at x_0, then it is **singular** there.

Most functions encountered in applications are analytic for all x, or for all x with the exception of one or more points called **singular points** of f. (Of course, the points are not singular, the function is.) For instance, polynomial functions, $\sin x$, $\cos x$, e^x, and e^{-x} are analytic for all x. On the other hand, $f(x) = 1/(x-1)$ is analytic for all x except $x = 1$, where f and all of its derivatives are undefined, fail to exist. The function $f(x) = \tan x = \sin x/\cos x$ is analytic for all x except $x = n\pi/2$ ($n = \pm 1, \pm 3, \ldots$), where it is undefined because $\cos x$ vanishes in the denominator.

The function $f(x) = x^{4/3}$ is analytic for all x except $x = 0$, for even though $f(0)$ and $f'(0)$ exist, the subsequent derivatives $f''(0)$, $f'''(0)$,... do not (Fig. 2). In fact, $f(x) = x^\alpha$ is singular at $x = 0$ for *any* noninteger value of α.

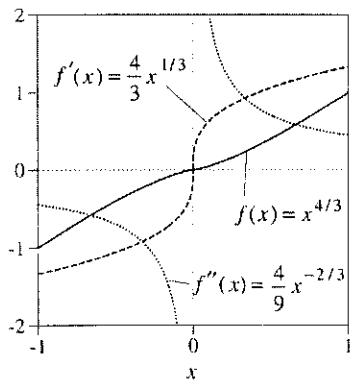


Figure 2. $f(x) = x^{4/3}$ and its first two derivatives.

Observe that there is a subtle difficulty here. We know how to test a given Taylor series for convergence since a Taylor series is a power series, and Theorem 4.2.2 on power series convergence even gives formulas for determining the radius of convergence R. But how can we determine if the sum function (i.e., the function to which the series converges) is the same as the original function f? We won't be able to answer this question until we study complex variable theory, in later chapters. However, we repeat that the cases where the Taylor series of f converges, but not to f, are exceptional and will not occur in the present chapter, so it will suffice

to understand analyticity at x_0 to correspond to the convergence of the Taylor series in some nonzero interval about x_0. In fact, it is also exceptional for f to have a Taylor series about a point (i.e., be infinitely differentiable at that point) and to have that Taylor series fail to converge in some nonzero interval about x_0. Thus, *as a rule of thumb that will suffice until we study complex variable theory, we will test a function for analyticity at a given point simply by seeing if it is infinitely differentiable at that point.*

4.2.2. Power series solution of differential equations. We can now state the following basic theorem.

THEOREM 4.2.4 *Power series solution*
If p and q are analytic at x_0, then every solution of

$$y'' + p(x)y' + q(x)y = 0 \tag{17}$$

is too, and can therefore be found in the form

$$y(x) = \sum_{0}^{\infty} a_n (x - x_0)^n. \tag{18}$$

Further, the radius of convergence of every solution (18) is at least as large as the smaller of the radii of convergence of TS $p|_{x_0}$ and TS $q|_{x_0}$.

Although we will not prove this theorem, we shall explain why one can expect it to be true. Since p and q are analytic at the chosen point x_0, they admit convergent Taylor series about x_0, so that we can write (17) as

$$y'' + \left[p(x_0) + p'(x_0)(x - x_0) + \cdots\right] y' + \left[q(x_0) + q'(x_0)(x - x_0) + \cdots\right] y = 0. \tag{19}$$

Locally, near x_0, we can approximate (19) as

$$y'' + p(x_0)y' + q(x_0)y = 0,$$

all solutions of which are either exponential or x times an exponential, and are therefore analytic and representable in the form (18), as claimed.

In many applications, $p(x)$ and $q(x)$ are **rational functions**, that is, one polynomial in x divided by another. Let $F(x) = N(x)/D(x)$ be any rational function, where the numerator and denominator polynomials are $N(x)$ and $D(x)$, respectively, and where any common factors have been canceled. It will be shown, when we study complex variable theory, that $F(x)$ is singular only at those points in the complex plane where $D = 0$, at the *zeros* of D, so that a Taylor expansion of F about a point x_0 on the x axis will have a radius of convergence which, we know in advance, will be equal to the distance from x_0 on the x axis to the nearest zero

of D in the complex plane. For instance, if $F(x) = (2+3x)/[(4+x)(9+x^2)]$, then D has zeros at -4 and $\pm 3i$. Thus, if we expand F about $x = 2$, say, then the radius of convergence will be the distance from 2 to the nearest zero, which is $+3i$ (or, equally, $-3i$), namely, $\sqrt{13}$ (Fig. 3). If, instead, we expand about $x = -6$, say, then the radius of convergence will be 2, the distance from -6 to the zero of D at -4.

EXAMPLE 5. Solve

$$y'' + y = 0 \tag{20}$$

by the power series method. Of course, this equation is elementary. We know the solution and do not need the power series method to find it. Let us use it nevertheless, as a first example, to illustrate the method.

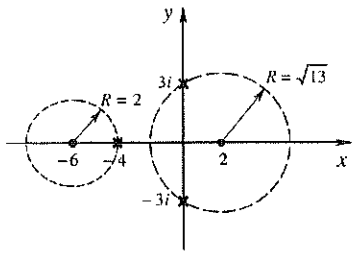


Figure 3. Disks of convergence in z plane ($z = x + iy$).

We can choose the point of expansion x_0 in (18) as any point at which both $p(x)$ and $q(x)$ in (17) are analytic. In the present example, $p(x) = 0$ and $q(x) = 1$ are analytic for all x, so we can choose the point of expansion x_0 to be whatever we like. Let $x_0 = 0$, for simplicity. Then Theorem 4.2.4 assures us that all solutions can be found in the form

$$y(x) = \sum_0^\infty a_n x^n, \tag{21a}$$

and their series will have infinite radii of convergence. Within that (infinite) interval of convergence we have, from Theorem 4.2.3(a),

$$y'(x) = \sum_1^\infty n a_n x^{n-1}, \tag{21b}$$

$$y''(x) = \sum_2^\infty n(n-1) a_n x^{n-2}, \tag{21c}$$

so (20) becomes

$$\sum_{n=2}^\infty n(n-1) a_n x^{n-2} + \sum_{n=0}^\infty a_n x^n = 0. \tag{22}$$

We would like to combine the two sums, but to do that we need the exponents of x to be the same, whereas they are $n-2$ and n. To have the same exponents, let us set $n-2 = m$ in the first sum, just as one might make a change of variables in an integral. In the integral, one would then need to change the integration limits, if necessary, consistent with the change of variables; here, we need to do likewise with the summation limits. With $n - 2 = m$, $n = \infty$ corresponds to $m = \infty$, and $n = 2$ corresponds to $m = 0$, so (22) becomes

$$\sum_{m=0}^\infty (m+2)(m+1) a_{m+2} x^m + \sum_{n=0}^\infty a_n x^n = 0. \tag{23}$$

Next, and we shall explain this step in a moment, let $m = n$ in the first sum in (23):

$$\sum_{n=0}^\infty (n+2)(n+1) a_{n+2} x^n + \sum_{n=0}^\infty a_n x^n = 0 \tag{24}$$

or, with the help of Theorem 4.2.3(b),

$$\sum_{n=0}^{\infty} \left[(n+2)(n+1)a_{n+2} + a_n\right] x^n = 0. \tag{25}$$

Finally, it follows from Theorem 4.2.3(c) that each coefficient in (25) must be zero:

$$(n+2)(n+1)a_{n+2} + a_n = 0. \qquad (n = 0, 1, 2, \ldots) \tag{26}$$

Before using (26), let us explain our setting $m = n$ in (23) since that step might seem to contradict the preceding change of variables $n - 2 = m$. The point to appreciate is that m in the first sum in (23) is a *dummy index* just as t is a *dummy variable* in $\int_0^1 t^2\,dt$. (We shall use the word *index* for a discrete variable; m takes on only integer values, not a continuous range of values.) Just as $\int_0^1 t^2\,dt = \int_0^1 r^2\,dr = \int_0^1 x^2\,dx = \cdots = \frac{1}{3}$, the sums in (23) are insensitive to whether the dummy index is m or n:

$$\sum_{m=0}^{\infty} (m+2)(m+1)a_{m+2}x^m = 2a_2 + 6a_3x + 12a_4x^2 + \cdots,$$

and

$$\sum_{n=0}^{\infty} (n+2)(n+1)a_{n+2}x^n = 2a_2 + 6a_3x + 12a_4x^2 + \cdots$$

are identical, even though the summation indices are different.

Equation (26) is known as a **recursion** (or **recurrence**) **formula** on the unknown coefficients since it gives us the nth coefficient in terms of preceding ones. Specifically,

$$a_{n+2} = -\frac{1}{(n+2)(n+1)}a_n, \qquad (n = 0, 1, 2, \ldots) \tag{27}$$

so that

$$\begin{aligned}
n = 0: &\quad a_2 = -\frac{1}{(2)(1)}a_0,\\
n = 1: &\quad a_3 = -\frac{1}{(3)(2)}a_1,\\
n = 2: &\quad a_4 = -\frac{1}{(4)(3)}a_2 = \frac{1}{(4)(3)(2)(1)}a_0 = \frac{1}{4!}a_0,\\
n = 3: &\quad a_5 = -\frac{1}{(5)(4)}a_3 = \frac{1}{5!}a_1,
\end{aligned} \tag{28}$$

and so on, where a_0 and a_1 remain arbitrary and are, in effect, the integration constants. Putting these results into (21a) gives

$$\begin{aligned}
y(x) &= a_0 + a_1x - \frac{1}{2!}a_0x^2 - \frac{1}{3!}a_1x^3 + \frac{1}{4!}a_0x^4 + \frac{1}{5!}a_1x^5 + \cdots\\
&= a_0\left(1 - \frac{1}{2!}x^2 + \frac{1}{4!}x^4 - \cdots\right) + a_1\left(x - \frac{1}{3!}x^3 + \frac{1}{5!}x^4 - \cdots\right)
\end{aligned} \tag{29}$$

or

$$y(x) = a_0y_1(x) + a_1y_2(x), \tag{30}$$

where $y_1(x)$ and $y_2(x)$ are the series within the first and second pairs of parentheses, respectively. From their series, we recognize $y_1(x)$ as $\cos x$ and $y_2(x)$ as $\sin x$ but, in general, we can't expect to identify the power series in terms of elementary functions because normally we reserve the power series method for nonelementary equations (except for pedagogical examples such as this). Thus, let us continue to call the series "$y_1(x)$" and "$y_2(x)$."

We don't need to check the series for convergence because Theorem 4.2.4 guarantees that they will converge for all x. We should, however, check to see if y_1, y_2 are LI (linearly independent), so that (30) is a general solution of (20). To do so, it suffices to evaluate the Wronskian $W[y_1, y_2](x)$ at a single point, say $x = 0$:

$$W[y_1, y_2](x) = \begin{vmatrix} y_1(0) & y_2(0) \\ y_1'(0) & y_2'(0) \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} = 1, \tag{31}$$

which is nonzero. It follows from that result and Theorem 3.2.3 that y_1, y_2 are LI on the entire x axis, so (30) is indeed a general solution of (20) for all x. Actually, since there are only two functions it would have been easier to apply Theorem 3.2.4: y_1, y_2 are LI because neither one is a constant multiple of the other.

COMMENT 1. To evaluate $y_1(x)$ or $y_2(x)$ at a given x, we need to add enough terms of the series to achieve the desired accuracy. For small values of x (i.e., for x's that are close to the point of expansion x_0, which in this case is 0) just a few terms may suffice. For example, the first four terms of $y_1(x)$ give $y_1(0.5) = 0.877582$, whereas the exact value is 0.877583 (to six decimal places). As x increases, more and more terms are needed for comparable accuracy. The situation is depicted graphically in Fig. 4, where we plot the partial sums s_3 and s_6, along with the sum function $y_1(x)$ (i.e., $\cos x$). Observe that the larger n is, the broader is the x interval over which the n-term approximation s_n stays close to the sum function. However, whereas $y_1(x)$ is oscillatory and remains between -1 and $+1$, $s_n(x)$ is a polynomial, and therefore it eventually tends to $+\infty$ or $-\infty$ as x increases ($-\infty$ if n is even and $+\infty$ if n is odd). Observe that if we do need to add a great many terms, then it is useful to have an expression for the general term in the series. In this example it is not hard to establish that

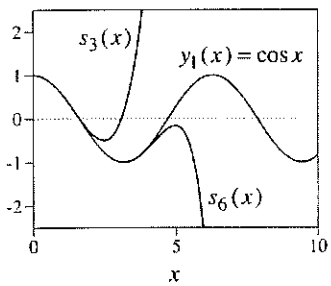


Figure 4. Partial sums of $y_1(x)$, compared with $y_1(x)$.

$$y_1(x) = \sum_0^\infty (-1)^n \frac{x^{2n}}{(2n)!}, \qquad y_2(x) = \sum_0^\infty (-1)^n \frac{x^{2n+1}}{(2n+1)!}. \tag{32}$$

COMMENT 2. Besides obtaining the values of the solutions $y_1(x)$ and $y_2(x)$, one is usually interested in determining some of their properties. Some properties can be obtained directly from the series. For instance, in this case we can see that $y_1(-x) = y_1(x)$ and $y_2(-x) = -y_2(x)$ (so that the graphs of y_1 and y_2 are symmetric and antisymmetric, respectively, about $x = 0$), and that $y_1'(x) = -y_2(x)$ and $y_2'(x) = y_1(x)$. The differential equation is also a source of information.

COMMENT 3. We posed (20) without initial conditions. Suppose that we wish to impose the initial conditions $y(0) = 4$ and $y'(0) = -1$. Then, from (30),

$$\begin{aligned} y(0) &= 4 = a_0 y_1(0) + a_1 y_2(0), \\ y'(0) &= -1 = a_0 y_1'(0) + a_1 y_2'(0). \end{aligned} \tag{33}$$

From the series representations of y_1 and y_2 in (30), we see that $y_1(0) = 1$, $y_2(0) = 0$, $y_1'(0) = 0$, and $y_2'(0) = 1$, so we can solve (33) for a_0 and a_1: $a_0 = 4$ and $a_1 = -1$, hence the desired particular solution is $y(x) = 4y_1(x) - y_2(x)$, on $-\infty < x < \infty$. ∎

We can now see more clearly how to select the point of expansion x_0, besides selecting it to be a point at which p and q in (17) are analytic. We have emphasized that the series solutions are especially convenient when the calculation point x is close to x_0, for then only a few terms of the series may suffice for the desired accuracy. Thus, if our interest is limited to the interval $6 < x < 10$, say, then it would make sense to seek series solutions about a point somewhere in that interval, such as a midpoint or endpoint, rather than about some distant point such as $x = 0$.

In the event that initial conditions are supplied at some point x_i, then it is especially helpful to let x_0 be x_i because when we apply the initial conditions we will need to know the values of $y_1(x_i)$, $y_2(x_i)$, $y_1'(x_i)$, and $y_2'(x_i)$, as we did in (33). If x_0 is other than x_i, then each of these evaluations requires the summing of an infinite series, whereas if it is chosen as x_i then these evaluations are trivial (as in Comment 3, above).

EXAMPLE 6. Solve the initial-value problem

$$(x-1)y'' + y' + 2(x-1)y = 0, \qquad y(4) = 5, \quad y'(4) = 0 \tag{34}$$

on the interval $4 \le x < \infty$. To get (34) into the standard form $y'' + p(x)y' + q(x)y = 0$, we divide by $x - 1$ (which is permissible since $x - 1 \neq 0$ on the interval of interest):

$$y'' + \frac{1}{x-1}y' + 2y = 0, \tag{35}$$

so $p(x) = 1/(x-1)$ and $q(x) = 2$. These are analytic for all x except $x = 1$, where $p(x)$ is undefined. In particular, they are analytic at the initial point $x = 4$, so let us choose $x_0 = 4$ and seek

$$y(x) = \sum_0^\infty a_n(x-4)^n. \tag{36}$$

To proceed we can use either form (34) or (35). Since we are expanding each term in the differential equation about $x = 4$, we need to expand $x - 1$ and $2(x-1)$ if we use (34), or the $1/(x-1)$ factor if we use (35). The former is easier since

$$x - 1 = 3 + (x - 4) \tag{37}$$

is merely a two-term Taylor series, whereas (Exercise 6)

$$\frac{1}{x-1} = \frac{1}{3}\sum_0^\infty \frac{(-1)^n}{3^n}(x-4)^n \tag{38}$$

is an infinite series. Thus, let us use (34). Putting (36) and its derivatives and (37) into (34) gives

$$[3 + (x-4)]\sum_2^\infty n(n-1)a_n(x-4)^{n-2} + \sum_1^\infty na_n(x-4)^{n-1}$$

$$+2\left[3+(x-4)\right]\sum_{0}^{\infty} a_n(x-4)^n = 0 \tag{39}$$

or, absorbing the $3+(x-4)$ terms into the series that they multiply and setting $z = x-4$ for compactness,

$$\sum_{2}^{\infty} 3n(n-1)a_n z^{n-2} + \sum_{2}^{\infty} n(n-1)a_n z^{n-1}$$
$$+\sum_{1}^{\infty} na_n z^{n-1} + \sum_{0}^{\infty} 6a_n z^n + \sum_{0}^{\infty} 2a_n z^{n+1} = 0. \tag{40}$$

To adjust all z exponents to n, let $n-2=m$ in the first sum, $n-1=m$ in the second and third, and $n+1=m$ in the last:

$$\sum_{0}^{\infty} 3(m+2)(m+1)a_{m+2}z^m + \sum_{1}^{\infty}(m+1)ma_{m+1}z^m$$
$$+\sum_{0}^{\infty}(m+1)a_{m+1}z^m + \sum_{0}^{\infty} 6a_n z^n + \sum_{1}^{\infty} 2a_{m-1}z^m = 0. \tag{41}$$

Next, we change all of the m indices to n. Then we have z^n in each sum, but we cannot yet combine the five sums because the lower summation limits are not all the same; three are 0 and two are 1. We can handle that problem as follows. The lower limit in the second sum can simply be changed from 1 to 0 because the zeroth term is zero anyhow (due to the m factor). And the lower limit in the last sum can be changed to 0 if we simply agree that the a_{-1}, that occurs in the zeroth term, be zero by definition. Then (41) becomes

$$\sum_{0}^{\infty} 3(n+2)(n+1)a_{n+2}z^n + \sum_{0}^{\infty}(n+1)na_{n+1}z^n$$
$$+\sum_{0}^{\infty}(n+1)a_{n+1}z^n + \sum_{0}^{\infty} 6a_n z^n + \sum_{0}^{\infty} 2a_{n-1}z^n = 0 \tag{42}$$

or

$$\sum_{0}^{\infty}\left[3(n+2)(n+1)a_{n+2} + (n+1)^2 a_{n+1} + 6a_n + 2a_{n-1}\right] z^n = 0, \tag{43}$$

with $a_{-1} \equiv 0$.

Setting the square-bracketed coefficients to zero for each n [Theorem 4.2.3(c)] then gives the recursion formula

$$3(n+2)(n+1)a_{n+2} + (n+1)^2 a_{n+1} + 6a_n + 2a_{n-1} = 0$$

or

$$a_{n+2} = -\frac{n+1}{3(n+2)}a_{n+1} - \frac{2}{(n+2)(n+1)}a_n - \frac{2}{3(n+2)(n+1)}a_{n-1} \tag{44}$$

for $n = 0, 1, 2, \ldots$. Thus,

$$\begin{aligned}
n = 0: \quad & a_2 = -\frac{1}{6}a_1 - a_0 - \frac{1}{3}a_{-1} = -\frac{1}{6}a_1 - a_0 \\
n = 1: \quad & a_3 = -\frac{2}{9}a_2 - \frac{1}{3}a_1 - \frac{1}{9}a_0 \\
& \quad = -\frac{2}{9}\left(-\frac{1}{6}a_1 - a_0\right) - \frac{1}{3}a_1 - \frac{1}{9}a_0 = -\frac{8}{27}a_1 + \frac{1}{9}a_0 \\
n = 2: \quad & a_4 = -\frac{1}{4}a_3 - \frac{1}{6}a_2 - \frac{1}{18}a_1 \\
& \quad = -\frac{1}{4}\left(-\frac{8}{27}a_1 + \frac{1}{9}a_0\right) - \frac{1}{6}\left(-\frac{1}{6}a_1 - a_0\right) - \frac{1}{18}a_1 \\
& \quad = \frac{5}{108}a_1 + \frac{5}{36}a_0,
\end{aligned} \tag{45}$$

and so on, where a_0 and a_1 remain arbitrary. Putting these expressions for the a_n's back into (36) then gives

$$\begin{aligned}
y(x) &= a_0 + a_1(x-4) + \left(-\frac{1}{6}a_1 - a_0\right)(x-4)^2 + \left(-\frac{8}{27}a_1 + \frac{1}{9}a_0\right)(x-4)^3 \\
&\quad + \left(\frac{5}{108}a_1 + \frac{5}{36}a_0\right)(x-4)^4 + \cdots \\
&= a_0\left[1 - (x-4)^2 + \frac{1}{9}(x-4)^3 + \frac{5}{36}(x-4)^4 + \cdots\right] \\
&\quad + a_1\left[(x-4) - \frac{1}{6}(x-4)^2 - \frac{8}{27}(x-4)^3 + \frac{5}{108}(x-4)^4 + \cdots\right] \\
&= a_0 y_1(x) + a_1 y_2(x),
\end{aligned} \tag{46}$$

where $y_1(x), y_2(x)$ are the functions represented by the bracketed series. To test y_1, y_2 for linear independence it is simplest to use Theorem 3.2.4: y_1, y_2 are LI because neither one is a constant multiple of the other. Thus, $y(x) = a_0 y_1(x) + a_1 y_2(x)$ is a general solution of $(x-1)y'' + y' + 2(x-1)y = 0$.

Imposing the initial conditions is easy because the expansions are about the initial point $x = 4$:

$$\begin{aligned}
y(4) &= 5 = a_0 y_1(4) + a_1 y_2(4) = a_0(1) + a_1(0), \\
y'(4) &= 0 = a_0 y_1'(4) + a_1 y_2'(4) = a_0(0) + a_1(1),
\end{aligned} \tag{47}$$

so $a_0 = 5$ and $a_1 = 0$, and

$$y(x) = 5y_1(x) = 5\left[1 - (x-4)^2 + \frac{1}{9}(x-4)^3 + \frac{5}{36}(x-4)^4 + \cdots\right] \tag{48}$$

is the desired particular solution.

COMMENT. Recall that Theorem 4.2.4 guaranteed that the power series solution would have a radius of convergence R at least as large as 3 – namely, the distance from the center of the expansion ($x_0 = 4$) to the singularity in $1/(x-1)$ at $x = 1$. For comparison, let us

determine R from our results. In this example it is difficult to obtain a general expression for a_n. (Indeed, we didn't attempt to; we were content to develop the first several terms of the series, knowing that we could obtain as many more as we wish, from the recursion formula.) Can we obtain R without an explicit expression for a_n? Yes, we can use the recursion formula (44), which tells us that $a_{n+2} \sim -\frac{1}{3}a_{n+1}$ as $n \to \infty$ or, equivalently, $a_{n+1} \sim -\frac{1}{3}a_n$. Then, from (7a),

$$R = \frac{1}{\lim_{n\to\infty}\left|\frac{a_{n+1}}{a_n}\right|} = \frac{1}{\lim_{n\to\infty}\left|-\frac{1}{3}\right|} = \frac{1}{\frac{1}{3}} = 3.$$

Thus, if we were hoping to obtain the solution over the entire interval $4 \le x < \infty$ we are disappointed to find that the power series converges only over $1 < x < 7$, and hence only over the $4 \le x < 7$ part of the problem domain. Does this result mean that the solution is singular at $x = 7$ and can't be continued beyond, or that it doesn't exist beyond $x = 7$? No, the convergence is simply being limited by the singularity at $x = 1$, which lies outside of the problem domain $4 \le x < \infty$. For further discussion of this point, see Exercise 12. ∎

Closure. In Section 4.2.1 we reviewed the basic concepts of series and power series and, in Theorem 4.2.3, we listed the properties of power series that are needed to solve differential equations. In Section 4.2.2 we provided a basis for the power series solution method of Theorem 4.2.4 and then showed, by two examples, how to implement it.

It is best to use summation notation, as we did in Examples 5 and 6, because it is more concise and leads to the recursion relation. (But that notation is not essential to the method; for example, we did not use it in our introductory example in Section 4.1.) The recursion relation is important because it permits the calulation of as many coefficients of the series as we desire, and because it can be used in determining the radius of convergence of the resulting series solutions.

The method may be outlined as follows:

(1) Write the differential equation in the standard form $y'' + p(x)y' + q(x)y = 0$ to identify $p(x)$ and $q(x)$ and their singularities (if any).

(2) Choose an expansion point x_0 at which p and q are analytic. If initial conditions are given at some point, it is suggested that that point be used as x_0.

(3) The least possible radius of convergence can be predicted as the distance (in the complex plane) from x_0 to the nearest singular point of p and q.

(4) Seeking $y(x)$ in the form of a power series about x_0, put that form into the differential equation, and also expand all coefficients of y'', y', y about x_0 as well.

(5) By changing dummy indices of summation and lower summation limits, as necessary, obtain a form where each summation has the same exponent on $x - x_0$ and the same summation limits.

(6) Combine the sums into a single sum.

(7) Set the coefficient of $(x - x_0)^n$, within that sum, to zero; that step gives the recursion formula.

(8) From the recursion formula, obtain as many of the coefficients as desired and hence the solution form $y(x) = Ay_1(x) + By_2(x)$, where A, B are arbitrary constants and $y_1(x), y_2(x)$ are power series. If possible, obtain expressions for the general term in each of those series.

(9) Verify that y_1, y_2 are LI.

Computer software. One can use computer software to generate Taylor series and also to obtain power series solutions to differential equations. Using *Maple*, for instance, the relevant commands are **taylor** and **dsolve**.

For example, to obtain the Taylor series of $1/(a - x)$ about $x = 0$, up to terms of third order, where a is a constant, enter

$$\text{taylor}(1/(a - x),\ x = 0,\ 4);$$

and return. The result is

$$\frac{1}{a} + \frac{1}{a^2}x + \frac{1}{a^3}x^2 + \frac{1}{a^4}x^3 + O(x^4)$$

where the $O(x^4)$ denotes that there are more terms, of order 4 and higher.

To obtain a power series solution to $y'' + y = 0$ about the point $x = 0$, enter

$$\text{dsolve}(\text{diff}(y(x),\ x, x) + y(x) = 0,\ y(x),\ \text{type} = \text{series});$$

and return. The result is

$$y(x) = y(0) + D(y)(0)x - \frac{1}{2}y(0)x^2 - \frac{1}{6}D(y)(0)x^3 + \frac{1}{24}y(0)x^4 + \frac{1}{120}D(y)(0)x^5 + O(x^6)$$

where $D(y)(0)$ means $y'(0)$. The default is to expand about $x = 0$ and to go as far as the fifth-order term. If we want an expansion about $x = 4$, say, and through the seventh-order term, enter

$$\text{Order} := 8;$$

to set the order then return and enter

$$\text{dsolve}(\{\text{diff}(y(x),\ x, x) + y(x) = 0,\ y(4) = a,\ D(y)(4) = b\},\ y(x),\ \text{type} = \text{series});$$

and return. Here, the initial conditions merely serve to establish the point about which the expansion is desired. The result is

$$y(x) = a + b(x-4) - \frac{1}{2}a(x-4)^2 - \frac{1}{6}b(x-4)^3 + \frac{1}{24}a(x-4)^4 + \frac{1}{120}b(x-4)^5 - \frac{1}{720}a(x-4)^6 - \frac{1}{5040}b(x-4)^7 + O\left((x-4)^8\right)$$

EXERCISES 4.2

1. Use (7a) or (7b) to determine the radius of convergence R of the given power series.

(a) $\sum_0^\infty nx^n$ (b) $\sum_0^\infty (-1)^n n^{1000} x^n$

(c) $\sum_5^\infty e^n x^n$ (d) $\sum_0^{1000} n!\, x^n$

(e) $\sum_0^\infty \left(\frac{x+3}{2}\right)^n$ (f) $\sum_2^\infty (n-1)^3 (x-5)^n$

(g) $\sum_1^\infty \frac{n^{50}}{n!}(x+7)^n$ (h) $\sum_2^\infty (\ln n)^{n+1}(x-2)^n$

(i) $\sum_3^\infty \frac{(-1)^n}{4^n}(x+2)^{3n}$ (j) $\sum_0^\infty \frac{n}{2^n}(x-5)^{2n}$

(k) $\sum_0^\infty \frac{n^6}{3^n+n}(x+4)^{8n+1}$ (l) $\sum_0^\infty \frac{(-1)^{n+1}}{4^n+1}(x-3)^{2n+1}$

2. Determine the radius of convergence R of the Taylor series expansion of the given rational function, about the specified point x_0, using the ideas given in the paragraph preceding Example 5. Also, prepare a sketch analogous to those in Fig. 3.

(a) $\frac{1}{x^2+1}$, $x_0 = 0$

(b) $\frac{1}{x^2+9}$, $x_0 = 2$

(c) $\frac{x^3-2x+1}{x+7}$, $x_0 = 5$

(d) $\frac{x+1}{x+2}$, $x_0 = -26$

(e) $\frac{(x+1)^5}{x^2+3x+2}$, $x_0 = -4$

(f) $\frac{x^2-3x+1}{x^2+2x+4}$, $x_0 = 3$

(g) $\frac{x^2+x-2}{x^3-x^2+4x-4}$, $x_0 = 2$

(h) $\frac{x^2-3x+2}{x-1}$, $x_0 = 0$

3. Work out the Taylor series of the given function, about the given point x_0, and use (7a) or (7b) to determine its radius of convergence.

(a) e^x, $x_0 = 1$ (b) e^{-x}, $x_0 = -2$

(c) $\sin x$, $x_0 = \pi$ (d) $\sin x$, $x_0 = \pi/2$

(e) $\cos x$, $x_0 = \pi/2$ (f) $\cos x$, $x_0 = \pi$

(g) $\cos x$, $x_0 = 5$ (h) $\ln x$, $x_0 = 1$

(i) x^2, $x_0 = 3$ (j) $2x^3 - 4$, $x_0 = 0$

(k) $\cos(x-2)$, $x_0 = 2$ (l) $\frac{1}{1-x^{10}}$, $x_0 = 0$

(m) $\frac{x^2}{1+x^{18}}$, $x_0 = 0$ (n) $\sin(3x^{10})$, $x_0 = 0$

4. Use computer software to obtain the first 12 nonzero terms in the Taylor series expansion of the given function f, about the given point x_0, and obtain a computer plot of f and the partial sums $s_3(x)$, $s_6(x)$, $s_9(x)$, and $s_{12}(x)$ over the given interval I.

(a) $f(x) = e^{-x}$, $x_0 = 0$, $I: 0 < x < 4$
(b) $f(x) = \sin x$, $x_0 = 0$, $I: 0 < x < 10$
(c) $f(x) = \ln x$, $x_0 = 1$, $I: 0 < x < 2$
(d) $f(x) = 1/(1-x)$, $x_0 = 0$, $I: -1 < x < 1$
(e) $f(x) = 1/x$, $x_0 = 2$, $I: 0 < x < 4$
(f) $f(x) = 1/(1+x^2)$, $x_0 = 0$, $I: -1 < x < 1$
(g) $f(x) = 4/(4+x+x^2)$, $x_0 = 0$, $I: -1.3 < x < 0.36$

5. (*Geometric series*) (a) Show that

$$\boxed{\frac{1}{1-x} = 1 + x + x^2 + \cdots + x^{n-1} + \frac{x^n}{1-x}} \qquad (5.1)$$

is an *identity* for all $x \neq 1$ and any positive integer n, by multiplying through by $1-x$ (which is nonzero since $x \neq 1$) and simplifying.

(b) The identity (5.1) can be used to study the Taylor series

known as the **geometric series** $\sum_{k=0}^{\infty} x^k$ since, according to (5.1), its partial sum $s_n(x)$ is

$$s_n(x) = \sum_{k=0}^{n-1} x^k = \frac{1-x^n}{1-x}. \qquad (x \neq 1) \tag{5.2}$$

Show, from (5.2), that the sequence $s_n(x)$ converges, as $n \to \infty$, for $|x| < 1$, and diverges for $|x| > 1$.
(c) Determine, by any means, the convergence or divergence of the geometric series for the points at the ends of the interval of convergence, $x = \pm 1$. NOTE: The formula (5.2) is quite striking because it reduces $s_n(x)$ to the *closed form* $(1-x^n)/(1-x)$, direct examination of which gives not only the interval of convergence but also the sum function $1/(1-x)$. It is rare that one can reduce $s_n(x)$ to closed form.

6. (a) Derive the Taylor series of $1/(x-1)$ about $x = 4$ using the Taylor series formula (16), and show that your result agrees with (38).
(b) Show that the same result is obtained (more readily) by writing

$$\frac{1}{x-1} = \frac{1}{3+(x-4)} = \frac{1}{3}\frac{1}{1+\frac{x-4}{3}} \tag{6.1}$$

and using the geometric series formula

$$\boxed{\frac{1}{1-t} = \sum_{0}^{\infty} t^n \qquad (|t| < 1)} \tag{6.2}$$

from Exercise 5, with $t = -(x-4)/3$. Further, deduce the x interval of convergence of the result from the convergence condition $|t| < 1$ in (6.2).

7. For each of the following differential equations do the following: Identify $p(x)$ and $q(x)$ and, from them, determine the least possible guaranteed radius of convergence of power series solutions about the specified point x_0; seeking a power series solution form, about that point, obtain the recursion formula and the first four nonvanishing terms in the power series for $y_1(x)$ and $y_2(x)$; verify that y_1, y_2 are LI.

(a) $y'' + 2y' + y = 0, \quad x_0 = 0$
(b) $y'' + 2y' = 0, \quad x_0 = 0$
(c) $y'' + 2y' = 0, \quad x_0 = 3$
(d) $xy'' + y' + y = 0, \quad x_0 = -5$
(e) $xy'' - 2y' + xy = 0, \quad x_0 = 1$
(f) $x^2y'' - y = 0, \quad x_0 = 2$
(g) $xy'' + (3+x)y' + xy = 0, \quad x_0 = -3$
(h) $y'' + y' + (1+x+x^2)y = 0, \quad x_0 = 0$
(i) $y'' - (1+x^2)y = 0, \quad x_0 = 0$
(j) $y'' - x^3y = 0, \quad x_0 = 0$
(k) $y'' + x^3y' + y = 0, \quad x_0 = 0$
(l) $y'' + xy' + x^2y = 0, \quad x_0 = 0$
(m) $y'' + (x-1)^2y = 0, \quad x_0 = 2$

8. (a)–(m) Use computer software to obtain the general solution, in power series form, for the corresponding problem given in Exercise 7, about the given expansion point.

9. (*Airy equation*) For the **Airy equation**,

$$y'' - xy = 0, \qquad (-\infty < x < \infty) \tag{9.1}$$

derive the power series solution

$$\begin{aligned} y(x) &= a_0y_1(x) + a_1y_2(x) \\ &= a_0\left(1 + \sum_{1}^{\infty} \frac{x^{3n}}{2\cdot 3\cdots(3n-1)(3n)}\right) \\ &\quad + a_1\left(x + \sum_{1}^{\infty} \frac{x^{3n+1}}{3\cdot 4\cdots(3n)(3n+1)}\right) \end{aligned} \tag{9.2}$$

and verify that it is a general solution. NOTE: These series are not summable in closed form in terms of elementary functions thus, certain linear combinations of y_1 and y_2 are adopted as a usable pair of LI solutions. In particular, it turns out to be convenient (for reasons that are not obvious) to use the **Airy functions** $Ai(x)$ and $Bi(x)$, which satisfy these initial conditions: $Ai(0) = 0.35502$, $Ai'(0) = -0.25881$ and $Bi(0) = 0.61493$, $Bi'(0) = 0.44829$.

10. Use computer software to obtain power series solutions of the following initial-value problems, each defined on $0 \le x < \infty$, through terms of eighth order, and obtain a computer plot of $s_2(x)$, $s_4(x)$, $s_6(x)$, and $s_8(x)$.

(a) $y'' + 4y' + y = 0, \quad y(0) = 1, \quad y'(0) = 0$
(b) $y'' + x^2y = 0, \quad y(0) = 2, \quad y'(0) = 0$
(c) $y'' - xy' + y = 0, \quad y(0) = 0, \quad y'(0) = 1$
(d) $(1+x)y'' + y = 0, \quad y(0) = 2, \quad y'(0) = 0$
(e) $(3+x)y'' + y' + y = 0, \quad y(0) = 0, \quad y'(0) = 1$
(f) $(1+x^2)y'' + y = 0, \quad y(0) = 1, \quad y'(0) = 1$

11. From the given recursion formula alone, determine the radius of convergence of the corresponding power series solutions.

(a) $(n+3)(n+2)a_{n+2} - (n+1)^2a_{n+1} + na_n = 0$
(b) $(n+1)a_{n+2} + 5na_{n+1} + a_n - a_{n-1} = 0$
(c) $(n+1)^2a_{n+2} + (2n^2+1)a_{n+1} - 4a_n = 0$
(d) $(n+1)a_{n+2} - 3(n+2)a_n = 0$
(e) $na_{n+2} + 4na_{n+1} + 3a_n = 0$
(f) $n^2a_{n+2} - 3(n+2)^2a_{n+1} + 3a_{n-1} = 0$

12. In the Comment at the end of Example 6 we wondered what the divergence of the series solution over $7 < x < \infty$

implied about the nature of the solution over that part of the domain. To gain insight, we propose studying a simple problem with similar features. Specifically, consider the problem

$$(x-1)y' + y = 0, \qquad y(4) = 5 \tag{12.1}$$

on the interval $4 \le x < \infty$.
(a) Solve (12.1) analytically, and show that the solution is

$$y(x) = \frac{15}{x-1} \tag{12.2}$$

over $4 \le x < \infty$. Sketch the graph of (12.2), showing it as a solid curve over the domain $4 \le x < \infty$, and dotted over $-\infty < x < 4$.
(b) Solve (12.1), instead, by seeking $y(x) = \sum_0^\infty a_n(x-4)^n$.
(c) Show that the solution obtained in (b) is, in fact, the Taylor expansion of (12.2) about $x = 4$ and that it converges only in $|x-4| < 3$ so that it represents the solution (12.2) only over the $4 \le x < 7$ part of the domain, even though the solution (12.2) exists and is perfectly well-behaved over $7 < x < \infty$.

13. Rework Example 5 without using the $\sum$ summation notation. That is, just write out the series, as we did in the introductory example of Section 4.1. Keep powers of x up to and including fifth order, x^5, and show that your result agrees (up to terms of fifth order) with that given in (29).

14. Rework Example 6 without using the $\sum$ summation notation. That is, just write out the series as we did in the introductory example of Section 4.1. Keep powers of $x-4$ up to and including fourth order $(x-4)^4$, and show that your result agrees (up to terms of fourth order) with that given in (46).

15. (*Cesàro summability*) Although (5) gives the usual definition of infinite series, it is not the only possible one nor the only one used. For example, according to **Cesàro summability**, which is especially useful in the theory of Fourier series, one defines

$$\sum_1^\infty a_n \equiv \lim_{N\to\infty} \frac{s_1 + s_2 + \cdots + s_N}{N}, \tag{15.1}$$

that is, the limit of the arithmetic means of the partial sums. It can be shown that if a series converges to s according to "ordinary convergence" [equation (5)], then it will also converge to the same value in the Cesàro sense. Yet, there are series that diverge in the ordinary sense but that converge in the Cesàro sense. Show that for the geometric series (see Exercise 5),

$$\frac{s_1 + s_2 + \cdots + s_N}{N} = \frac{1}{1-x} - \frac{x}{N}\frac{1-x^N}{(1-x)^2} \tag{15.2}$$

for all $x \ne 1$, and use that result to show that the Cesáro definition gives divergence for all $|x| > 1$ and for $x = 1$, and convergence for $|x| < 1$, as does ordinary convergence, but that for $x = -1$ it gives convergence to $1/2$, whereas according to ordinary convergence the series diverges for $x = -1$.

4.3 The Method of Frobenius

4.3.1. Singular points. In this section we continue to consider series solutions of the equation

$$y'' + p(x)y' + q(x)y = 0. \tag{1}$$

From Section 4.2, we know that we can find two LI solutions as power series expansions about any point x_0 at which both p and q are analytic. We call such a point x_0 an **ordinary point** of the equation (1). Typically, p and q are analytic everywhere on the x axis except perhaps at one or more singular points, so that all points of the x axis, except perhaps a few, are ordinary points. In that case one can readily select such an x_0 and develop two LI power series solutions about that point.

Nevertheless, in the present section we examine singular points more closely, and show that one can often obtain modified series solutions about singular points.

Why should we want to develop a method of finding series solutions about a singular point when we can stay away from the singular point and expand about an ordinary point? There are at least two reasons, which are explained later in this section.

Proceeding, we begin by classifying singular points as follows:

DEFINITION 4.3.1 *Regular and Irregular Singular Points of (1)*
Let x_0 be a singular point of p and/or q. We classify it as a regular or irregular singular point of equation (1) as follows: x_0 is
(a) a **regular singular point of** (1) if $(x - x_0)p(x)$ and $(x - x_0)^2 q(x)$ are analytic at x_0,
(b) an **irregular singular point** of (1) if it is not a regular singular point.

EXAMPLE 1. Consider $x(x-1)^2 y'' - 3y' + 5y = 0$ or, dividing by $x(x-1)^2$ to put the equation in the standard form $y'' + p(x)y' + q(x)y = 0$,

$$y'' - \frac{3}{x(x-1)^2}y' + \frac{5}{x(x-1)^2}y = 0. \tag{2}$$

Thus, $p(x) = -3/[x(x-1)^2]$ and $q(x) = 5/[x(x-1)^2]$. These are analytic for all x except for $x = 0$ and $x = 1$, so every x is an ordinary point except for those points. Let us classify those two singular points:

$$\begin{aligned}
x_0 = 0: \quad & (x - x_0)p(x) = (x-0)\left(-\frac{3}{x(x-1)^2}\right) = -\frac{3}{(x-1)^2} \\
& (x - x_0)^2 q(x) = (x-0)^2\left(\frac{5}{x(x-1)^2}\right) = \frac{5x}{(x-1)^2} \\
x_0 = 1: \quad & (x - x_0)p(x) = (x-1)\left(-\frac{3}{x(x-1)^2}\right) = -\frac{3}{x(x-1)} \\
& (x - x_0)^2 q(x) = (x-1)^2\left(\frac{5}{x(x-1)^2}\right) = \frac{5}{x}.
\end{aligned} \tag{3a,b,c,d}$$

To classify the singular point at $x = 0$, consider (3a) and (3b). Since the right-hand sides of (3a) and (3b) are analytic* at 0, we classify $x = 0$ as a regular singular point of (2). (The fact that those right-hand sides are singular elsewhere, at $x = 1$, is irrelevant.) To classify the singular point at $x = 1$, we turn to (3c) and (3d). Whereas the right-hand side of (3d) is analytic at $x = 1$, the right-hand side of (3c) is not, so we classify the singular point at $x = 1$ as an irregular singular point of (2). ∎

EXAMPLE 2. Consider the case

$$y'' + \sqrt{x}\,y = 0. \qquad (0 \le x < \infty) \tag{4}$$

---

*Recall the rule of thumb given in the last sentence of Section 4.2.1, that we will classify a function as analytic at a given point if it is infinitely differentiable at that point.

Then $p(x) = 0$ and $q(x) = \sqrt{x}$, and these are analytic (infinitely differentiable) for all $x > 0$, but not at $x = 0$ because $q(x)$ is not even once differentiable there, let alone infinitely differentiable. To classify the singular point at $x = 0$, observe that $(x-x_0)p(x) = (x)(0) = 0$ is analytic at $x = 0$, but $(x - x_0)^2 q(x) = x^2\sqrt{x} = x^{5/2}$ is not; it is twice differentiable there (those derivatives being zero), but all higher derivatives are undefined at $x = 0$. Thus, $x = 0$ is an irregular singular point of (4). (See Exercise 2.) ∎

4.3.2. Method of Frobenius. To develop the method of Frobenius, we require that the singular point about which we expand be a *regular* singular point. Before stating theorems and working examples, let us motivate the idea behind the method. We consider the equation

$$y'' + p(x)y' + q(x)y = 0 \tag{5}$$

to have a regular singular point at the origin (and perhaps other singular points as well). There is no loss of generality in assuming it to be at the origin since, if it is at $x = x_0 \neq 0$, we can always make a change of variable $\xi = x - x_0$ to move it to the origin in terms of the new variable ξ (Exercise 3). Until stated otherwise, let us assume that the interval of interest is $x > 0$.

We begin by multiplying equation (5) by x^2 and rearranging terms as

$$x^2y'' + x\left[xp(x)\right]y' + \left[x^2q(x)\right]y = 0. \tag{6}$$

Since $x = 0$ is a regular singular point, it follows that xp and x^2q can be expanded about the origin in convergent Taylor series, so we can write

$$x^2y'' + x\left(p_0 + p_1x + \cdots\right)y' + \left(q_0 + q_1x + \cdots\right)y = 0. \tag{7}$$

Locally, in the neighborhood of $x = 0$, we can approximate (7) as

$$x^2y'' + p_0xy' + q_0y = 0, \tag{8}$$

which is a Cauchy-Euler equation. As such, (8) has at least one solution in the form x^r, for some constant r. Returning to (7), it is reasonable to expect that equation, likewise, to have at least one solution that behaves like x^r (for the same value of r) in the neighborhood of $x = 0$. More completely, we expect it to have at least one solution of the form

$$y(x) = x^r\left(a_0 + a_1x + a_2x^2 + \cdots\right), \tag{9}$$

where the power series factor is needed to account for the deviation of $y(x)$, away from $x = 0$, from its asymptotic behavior $y(x) \sim a_0x^r$ as $x \to 0$. That is, in place of the power series expansion

$$y(x) = \sum_0^\infty a_nx^n \tag{10}$$

that is guaranteed to work when $x = 0$ is an ordinary point of (5), it appears that we should seek $y(x)$ in the more general form

$$\boxed{y(x) = x^r \sum_0^\infty a_n x^n = \sum_0^\infty a_n x^{n+r}} \tag{11}$$

if $x = 0$ is a regular singular point. Is (11) really different from (10)? Yes, because whereas (10) necessarily represents an analytic function, (11) represents a nonanalytic function because of the x^r factor (unless r is a nonnegative integer).

Let us try the solution form (11) in an example.

EXAMPLE 3. The equation

$$6x^2y'' + 7xy' - (1 + x^2)y = 0, \qquad (0 < x < \infty) \tag{12}$$

has a regular singular point at $x = 0$ because whereas $p(x) = 7x/(6x^2) = 7/(6x)$ and $q(x) = -(1 + x^2)/(6x^2)$ are singular at $x = 0$, $xp(x) = 7/6$ and $x^2q(x) = -(1 + x^2)/6$ are analytic there. Let us seek $y(x)$ in the form (11). Putting that form into (12) gives

$$6x^2 \sum_0^\infty (n + r)(n + r - 1)a_n x^{n+r-2} + 7x \sum_0^\infty (n + r)a_n x^{n+r-1} - (1 + x^2) \sum_0^\infty a_n x^{n+r} = 0 \tag{13}$$

or

$$\sum_0^\infty 6(n + r)(n + r - 1)a_n x^{n+r} + \sum_0^\infty 7(n + r)a_n x^{n+r} - \sum_0^\infty a_n x^{n+r} - \sum_0^\infty a_n x^{n+r+2} = 0. \tag{14}$$

Letting $n + r + 2 = m + r$ in the last sum, (14) becomes

$$\sum_0^\infty [6(n + r)(n + r - 1) + 7(n + r) - 1]\, a_n x^{n+r} - \sum_2^\infty a_{m-2} x^{m+r} = 0. \tag{15}$$

Changing the lower limit in the last sum to 0, with the understanding that $a_{-2} = a_{-1} \equiv 0$, and changing the m's to n's, we can finally combine the sums as

$$\sum_0^\infty \left\{ \left[6(n + r)^2 + n + r - 1\right] a_n - a_{n-2} \right\} x^{n+r} = 0, \tag{16}$$

where we have also simplified the square bracket in (15) to $6(n + r)^2 + n + r - 1$. From (16), we infer the recursion formula

$$\left[6(n + r)^2 + n + r - 1\right] a_n - a_{n-2} = 0 \tag{17}$$

for each $n = 0, 1, 2, \ldots$.

In particular, $n = 0$ gives

$$\left(6r^2 + r - 1\right) a_0 - a_{-2} = 0 \tag{18}$$

or, since $a_{-2} \equiv 0$,

$$\left(6r^2 + r - 1\right) a_0 = 0. \tag{19}$$

Observe carefully that we may assume, with no loss of generality, that $a_0 \neq 0$ for, if $a_0 = 0$ in (9) and $a_1 \neq 0$, then we can factor an x out of the series and absorb it into the x^r. Rather, let us assume, in writing (9) that all such factors have already been absorbed into the x^r, and that a_0 is the first nonvanishing coefficient.

Proceeding, with $a_0 \neq 0$, it follows from (19) that

$$6r^2 + r - 1 = 0 \tag{20}$$

so $r = -1/2$ and $1/3$. We expect each of the choices, $r = -1/2$ and $r = 1/3$, to lead to a Frobenius-type solution (11), and the two solutions thus obtained to be LI. Let us see. First, set $r = -1/2$. The corresponding recursion formula (17) is then

$$a_n = \frac{1}{6\left(n - \frac{1}{2}\right)^2 + n - \frac{3}{2}} a_{n-2} \tag{21}$$

for $n = 1, 2, \ldots$, since the $n = 0$ case has already been carried out:

$$\begin{aligned}
n = 1: &\quad a_1 = a_{-1} = 0, \\
n = 2: &\quad a_2 = \frac{1}{14} a_0, \\
n = 3: &\quad a_3 = \frac{1}{39} a_1 = 0, \\
n = 4: &\quad a_4 = \frac{1}{76} a_2 = \frac{1}{(76)(14)} a_0, \\
n = 5: &\quad a_5 = \frac{1}{125} a_3 = 0, \\
n = 6: &\quad a_6 = \frac{1}{186} a_4 = \frac{1}{(186)(76)(14)} a_0,
\end{aligned} \tag{22}$$

and so on. From these results we have the solution

$$\begin{aligned}
y(x) &= a_0 x^{-1/2} \left[1 + \frac{1}{14} x^2 + \frac{1}{(76)(14)} x^4 + \frac{1}{(186)(76)(14)} x^6 + \cdots\right] \\
&= a_0 y_1(x),
\end{aligned} \tag{23}$$

where a_0 remains arbitrary.

Next, set $r = 1/3$. The corresponding recursion formula (17) is then

$$a_n = \frac{1}{6\left(n + \frac{1}{3}\right)^2 + n - \frac{2}{3}} a_{n-2}, \tag{24}$$

and proceeding as we did for $r = -1/2$, we obtain the solution (Exercise 4)

$$\begin{aligned} y(x) &= a_0 x^{1/3}\left[1 + \frac{1}{34}x^2 + \frac{1}{(116)(34)}x^4 + \frac{1}{(246)(116)(34)}x^6 + \cdots\right] \\ &= a_0 y_2(x), \end{aligned} \tag{25}$$

where a_0 remains arbitrary. [Of course, the a_0's in (23) and (25) have nothing to do with each other; they are independent arbitrary constants.] According to Theorem 3.2.4, the solutions y_1 and y_2 are LI because neither is a scalar multiple of the other, so a general solution to (12) is $y(x) = C_1 y_1(x) + C_2 y_2(x)$, where y_1 and y_2 are given in (23) and (25).

What are the regions of convergence of the series in (23) and (25)? Though we don't have the general terms for those series, we can use their recursion formulas, (21) and (24), respectively, to study their convergence. Consider the series in (23) first. Its recursion formula is (21) or, equivalently,

$$a_{n+2} = \frac{1}{6\left(n+2-\frac{1}{2}\right)^2 + n + 2 - \frac{3}{2}} a_n. \tag{26}$$

We need to realize that the a_{n+2} on the left side is really the next coefficient after a_n, the "a_{n+1}" in Theorem 4.2.2, since every other term in the series is missing (because $a_1 = a_3 = a_5 = \cdots = 0$). Thus, (26) gives

$$\lim_{n\to\infty}\left|\frac{\text{"}a_{n+1}\text{"}}{a_n}\right| = \lim_{n\to\infty}\frac{1}{6\left(n+\frac{3}{2}\right)^2 + n + \frac{1}{2}} = 0, \tag{27}$$

and it follows from Theorem 4.2.2 that $R = \infty$; the series converges for all x. Of course, the $x^{-1/2}$ factor in (23) "blows up" at $x = 0$, so (23) is valid in $0 < x < \infty$, which is the full interval specified in (12).

Similarly, we can show that the series in (24) converges for all x, so (25) is valid over the full interval $0 < x < \infty$. ∎

With Example 3 completed, we can come back to the important question that we posed near the beginning of Section 4.3.1: "Why should we want to develop a method of finding series solutions about a singular point when we can stay away from the singular point and expand about an ordinary point?". Observe that our Frobenius-type solution $y(x) = C_1 y_1(x) + C_2 y_2(x)$, with $y_1(x)$ and $y_2(x)$ given by (23) and (25), was valid on the full interval $0 < x < \infty$. Furthermore, it even showed us the singular behavior at the origin explicitly:

$$\begin{aligned} y(x) &= C_1 x^{-1/2}\left(1 + \frac{1}{14}x^2 + \cdots\right) + C_2 x^{1/3}\left(1 + \frac{1}{34}x^2 + \cdots\right) \\ &\sim C_1 \frac{1}{\sqrt{x}} \end{aligned} \tag{28}$$

as $x \to 0$. In contrast, if we had avoided the singular point $x = 0$ and pursued power series expansions about an ordinary point, say $x = 2$, then the resulting

solution would have been valid only in $0 < x < 4$, and it would not have explicitly displayed the $1/\sqrt{x}$ singular behavior at $x = 0$.

Let us review the ideas presented above, and get ready to state the main theorem. If $x = 0$ is a regular singular point of the equation $y'' + p(x)y' + q(x) = 0$, which we rewrite as

$$x^2y'' + x\left[xp(x)\right]y' + \left[x^2q(x)\right]y = 0, \tag{29}$$

then $xp(x)$ and $x^2q(x)$ admit Taylor series representations $xp(x) = p_0 + p_1x + \cdots$ and $x^2q(x) = q_0 + q_1x + \cdots$. Locally then, near $x = 0$, (29) can be approximated as

$$x^2y'' + p_0xy' + q_0y = 0, \tag{30}$$

which is a Cauchy-Euler equation. Cauchy-Euler equations, we recall, always admit at least one solution in the form x^r, and this fact led us to seek solutions $y(x)$ to (29) that behave like $y(x) \sim x^r$ as $x \to 0$, or

$$y(x) = x^r\sum_0^\infty a_nx^n = \sum_0^\infty a_nx^{n+r}, \tag{31}$$

where the $\sum_0^\infty a_nx^n$ factor is to account for the deviation of y from the local behavior $y(x) \sim x^r$ away from $x = 0$. Putting (31) into

$$x^2y'' + x\left(p_0 + p_1x + \cdots\right)y' + \left(q_0 + q_1x + \cdots\right)y = 0 \tag{32}$$

gives

$$\begin{aligned}\sum_0^\infty(n+r)(n+r-1)a_nx^{n+r} + (p_0 + p_1x + \cdots)\sum_0^\infty(n+r)a_nx^{n+r}\\ + (q_0 + q_1x + \cdots)\sum_0^\infty a_nx^{n+r} \;=\; 0,\end{aligned} \tag{33}$$

and equating coefficients of the various powers of x to zero gives

$$x^r : \left[r(r-1) + p_0r + q_0\right]a_0 = 0, \tag{34a}$$

$$x^{r+1} : \left[(r+1)r + p_0(r+1) + q_0\right]a_1 + (p_1r + q_1)a_0 = 0, \tag{34b}$$

$$\begin{aligned}x^{r+2} : \;&\left[(r+2)(r+1) + p_0(r+2) + q_0\right]a_2 + (\text{etc})a_1\\ &+(\text{etc})a_0 = 0,\end{aligned} \tag{34c}$$

$$\begin{aligned}x^{r+3} : \;&\left[(r+3)(r+2) + p_0(r+3) + q_0\right]a_3 + (\text{etc})a_2 + (\text{etc})a_1\\ &+(\text{etc})a_0 = 0,\end{aligned} \tag{34d}$$

and so on, where we've used "etc's" for brevity since we're most interested, here, in showing the *form* of the equations. Assuming, without loss of generality, that $a_0 \neq 0$, (34a) gives

$$\boxed{r^2 + (p_0 - 1)r + q_0 = 0,} \tag{35}$$

which quadratic equation for r is called the **indicial equation**; in Example 3 the indicial equation was equation (20). Let the roots be r_1 and r_2. Setting $r = r_1$ in (34b,c,d,...) gives a system of linear algebraic equations to find $a_1, a_2, \ldots$ in terms of a_0, with a_0 remaining arbitrary. Next, we set $r = r_2$ in (34b,c,d,...) and again try to solve for $a_1, a_2, \ldots$ in terms of a_0. If all goes well, those steps should produce two LI solutions of the differential equation $y'' + p(x)y' + q(x) = 0$.

The process is straightforward and was carried out successfully in Example 3. Can anything go wrong? Yes. One potential difficulty is that *the indicial equation might have repeated roots* ($r_1 = r_2$), in which case the procedure gives only one solution. To seek guidance as to how to find a second LI solution, realize that the same situation occurred for the simplified problem, the Cauchy-Euler equation (30): if, seeking $y(x) = x^r$ in (30), we obtain a repeated root for r, then a second solution can be found (by the method of reduction of order) to be of the form x^r times $\ln x$. Similarly, for $y'' + p(x)y' + q(x) = 0$, as we shall see in the theorem below [specifically, (41b)].

The other possible difficulty, which is more subtle, occurs *if the roots differ by a nonzero integer.* For observe that if we denote the bracketed coefficient of a_0 in (34a) as $F(r)$, then the coefficient of a_1 in (34b) is $F(r+1)$, that of a_2 in (34c) is $F(r+2)$, and so on. To illustrate, suppose that $r_1 = r_2 + 1$, so that the roots differ by 1. Then not only will $F(r)$ vanish in (34a) when we are using $r = r_2$, but so will $F(r+1)$ in (34b) [though not $F(r+2)$ in (34c), nor $F(r+3)$ in (34d), etc.], in which case (34b) becomes $0a_1 + (p_1 r_2 + q_1)a_0 = 0$. If $p_1 r_2 + q_1$ happens not to be zero then the equation (34b) cannot be satisfied, and there is no set of a_n's that satisfy the system (34). Thus, for the algebraically smaller root r_2 (e.g., -6 is algebraically smaller than 2), no solution is found. But if $p_1 r_2 + q_1$ does equal zero, then (34b) becomes $0a_1 = 0$ and a_1 (in addition to a_0) remains arbitrary. Then (34c,d,...) give $a_2, a_3, \ldots$ as linear combinations of a_0 and a_1, and one obtains a general solution

$$\begin{aligned} y(x) &= a_0 x^{r_2} \text{ (a power series) } + a_1 x^{r_2} \text{ (a different power series)} \\ &= a_0 y_1(x) + a_1 y_2(x), \end{aligned} \tag{36}$$

where a_0, a_1 are arbitrary and y_1, y_2 are LI.

If, however, r_2 gives no solution, then we can turn to r_1. For r_1 the difficulty cited in the preceding paragraph does not occur, and the method produces a single solution "$y_2(x)$."

If, instead, $r_1 = r_2 + 2$, say, then the same sort of difficulty shows up, but not until (34c). Similarly, if $r_1 = r_2 + 3$, $r_1 = r_2 + 4$, and so on.

The upshot is that if r_1, r_2 differ by a nonzero integer, then the algebraically smaller root r_2 leads either to no solution or to a general solution. In either case, the larger root r_1 leads to one solution.

The theorem is as follows.

THEOREM 4.3.1 *Regular Singular Point; Frobenius Solution*
Let $x = 0$ be a regular singular point of the differential equation

$$y'' + p(x)y' + q(x)y = 0, \qquad (x > 0) \tag{37}$$

with $xp(x) = p_0 + p_1 x + \cdots$ and $x^2 q(x) = q_0 + q_1 x + \cdots$ having radii of convergence R_1, R_2 respectively. Let r_1, r_2 be the roots of the indicial equation

$$r^2 + (p_0 - 1)r + q_0 = 0, \tag{38}$$

where $r_1 \geq r_2$ if the roots are real. (Otherwise they are complex conjugates.) Seeking $y(x)$ in the form

$$y(x) = x^r \sum_0^\infty a_n x^n = \sum_0^\infty a_n x^{n+r}, \qquad (a_0 \neq 0) \tag{39}$$

with $r = r_1$ inevitably leads to a solution

$$y_1(x) = x^{r_1} \sum_0^\infty a_n x^n, \qquad (a_0 \neq 0) \tag{40}$$

where $a_1, a_2, \ldots$ are known multiples of a_0, which remains arbitrary. For definiteness, we choose $a_0 = 1$ in (40). The form of the second LI solution, $y_2(x)$, depends on r_1 and r_2 as follows:

(i) r_1 and r_2 *distinct and not differing by an integer.* (Complex conjugate roots belong to this case.) Then with $r = r_2$, (39) yields

$$y_2(x) = x^{r_2} \sum_0^\infty b_n x^n, \qquad (b_0 \neq 0) \tag{41a}$$

where the b_n's are generated by the same recursion relation as the a_n's, but with $r = r_2$ instead of $r = r_1$; $b_1, b_2, \ldots$ are known multiples of b_0, which is arbitrary. For definiteness, we choose $b_0 = 1$ in (41a).
(ii) *Repeated roots*, $r_1 = r_2 \equiv r$. Then $y_2(x)$ can be found in the form

$$y_2(x) = y_1(x) \ln x + x^r \sum_1^\infty c_n x^n. \tag{41b}$$

(iii) $r_1 - r_2$ *equal to an integer.* Then the smaller root r_2 leads to both solutions, $y_1(x)$ and $y_2(x)$, or to neither. In either case, the larger root r_1 gives the single solution (40). In the latter case, $y_2(x)$ can be found in the form

$$y_2(x) = \kappa y_1(x) \ln x + x^{r_2} \sum_0^\infty d_n x^n, \tag{41c}$$

where the constant κ may turn out to be zero, in which case there is no logarithmic term in (41c).

The radius of convergence of each of the series in (40) and (41) is at least as large as the smaller of R_1, R_2.

If (37) is on $x < 0$ rather than $x > 0$, then the foregoing is valid, provided that each x^r, x^{r_1}, x^{r_2} and $\ln x$ is changed to $|x|^r$, $|x|^{r_1}$, $|x|^{r_2}$ and $\ln|x|$, respectively.

Outline of Proof of (ii): Our discussion preceding this theorem contained an outline of the proof of case (i), and also some discussion of case (iii). Here, let us focus on case (ii) and, again, outline the main ideas behind a proof. We consider the case of repeated roots, where $r_1 = r_2 \equiv r$. Since $y_1(x)$, given by (40), is a solution, then so is $y(x) = Ay_1(x)$, where A is arbitrary. To find $y_2(x)$, let us use reduction of order; that is, seek $y_2(x) = A(x)y_1(x)$, where $y_1(x)$ is known and $A(x)$ is to be found. Putting that form into (37) gives

$$A''y_1 + A'\left(2y_1' + py_1\right) + A\left(y_1'' + py_1' + qy_1\right) = 0. \tag{42}$$

Since y_1 satisfies (37), the last term in (42) is zero, so (42) becomes

$$A''y_1 + A'\left(2y_1' + py_1\right) = 0. \tag{43}$$

Replacing A'' by dA'/dx, multiplying through by dx, and dividing by y_1 and A' gives

$$\frac{dA'}{A'} + 2\frac{dy_1}{y_1} + p\,dx = 0. \tag{44}$$

Integrating,

$$\ln|A'| + 2\ln|y_1| + \int p(x)\,dx = \text{constant, say } \ln C, \text{ for } C > 0,$$

so

$$\ln\frac{|A'|\,|y_1|^2}{C} = -\int p(x)\,dx,$$

and

$$\begin{aligned}\left|A'(x)\right| &= C\,\frac{e^{-\int p(x)\,dx}}{y_1^2(x)} = C\,\frac{e^{-\int\left(\frac{p_0}{x}+p_1+p_2x+\cdots\right)dx}}{\left[x^r\left(1+a_1x+\cdots\right)\right]^2}\\ &= C\,\frac{e^{-p_0\ln x}e^{(-p_1x-\cdots)}}{x^{2r}\left(1+2a_1x+\cdots\right)},\end{aligned} \tag{45}$$

where we write $\ln x$ rather than $\ln|x|$ since $x > 0$ here, by assumption.

Since $\exp\left(-\int p(x)\,dx\right) > 0$, we see from the first equality in (45) that $A'(x)$ is either everywhere positive or everywhere negative on the interval. Thus, we can drop the absolute value signs around $A'(x)$ if we now allow C to be positive or

negative. Further, $e^{-p_0 \ln x} = e^{\ln x^{-p_0}} = x^{-p_0}$, and $e^{(-p_1 x - \cdots)}/(1 + 2a_1 x + \cdots)$ is analytic at $x = 0$ and can be expressed in Taylor series form as $1 + \kappa_1 x + \kappa_2 x^2 + \cdots$, so

$$A'(x) = C \frac{1}{x^{2r+p_0}} (1 + \kappa_1 x + \cdots). \tag{46}$$

For r to be a double root of the indicial equation (38), it is necessary that $2r + p_0 = 1$, in which case integration of (46) gives

$$A(x) = C (\ln x + \kappa_1 x + \cdots). \tag{47}$$

Finally, setting $C = 1$ with no loss of generality, we have the form

$$\begin{aligned} y_2(x) &= A(x) y_1(x) = (\ln x + \kappa_1 x + \cdots) y_1(x) \\ &= y_1(x) \ln x + (\kappa_1 x + \cdots) x^r (1 + a_1 x + \cdots) \\ &= y_1(x) \ln x + x^r \sum_1^\infty c_n x^n, \end{aligned} \tag{48}$$

as given in (41b). ■

In short, the Frobenius method is guaranteed to provide two LI solutions to (37) if $x = 0$ is a regular singular point of that equation. If $x = 0$ is an irregular singular point, the theorem does not apply. That case is harder, and we have no analogous theory to guide us.

EXAMPLE 4. *Case (ii).* Solve the equation

$$x^2 y'' - (x + x^2) y' + y = 0; \qquad (0 < x < \infty) \tag{49}$$

that is, find two LI solutions. The only singular point of (49) is $x = 0$, and it is a regular singular point. Seeking

$$y(x) = \sum_0^\infty a_n x^{n+r}, \qquad (a_0 \neq 0) \tag{50}$$

substitution of that form into (49) gives

$$\begin{aligned} &\sum_0^\infty (n+r)(n+r-1) a_n x^{n+r} - \sum_0^\infty (n+r) a_n x^{n+r} \\ &\qquad - \sum_0^\infty (n+r) a_n x^{n+r+1} + \sum_0^\infty a_n x^{n+r} = 0. \end{aligned} \tag{51}$$

Set $n + 1 = m$ in the third sum, change the lower limit from $n = 0$ to $m = 1$, extend that limit back to 0 by defining $a_{-1} \equiv 0$, change the m's back to n's, and combine the four sums. Those steps give

$$\sum_0^\infty \{[(n+r)(n+r-1) - (n+r) + 1] a_n - (n+r-1) a_{n-1}\} x^{n+r} = 0$$

and hence the recursion formula

$$[(n+r)(n+r-1)-(n+r)+1]\,a_n-(n+r-1)a_{n-1}=0, \tag{52}$$

for $n=0,1,2,\ldots$. For $n=0$, (52) becomes $(r^2-2r+1)a_0-(r-1)a_{-1}=0$. Since $a_{-1}\equiv 0$ and $a_0\neq 0$, the latter gives the indicial equation

$$r^2-2r+1=0, \tag{53}$$

with repeated roots $r=1,1$. Thus, this example illustrates case (ii). Putting $r=1$ into (52), we obtain the recursion formula

$$a_n=\frac{1}{n}a_{n-1} \tag{54}$$

for $n=1,2,\ldots$. Thus, $a_1=a_0$, $a_2=\frac{1}{2}a_1=\frac{1}{2}a_0$, $a_3=\frac{1}{3}a_2=\frac{1}{3!}a_0$ and we can see that $a_n=\frac{1}{n!}a_0$, so

$$\begin{aligned} y(x)&=x\left(a_0+\frac{a_0}{1!}x+\frac{a_0}{2!}x^2+\cdots\right)\\ &=a_0\sum_0^\infty\frac{x^{n+1}}{n!}=a_0y_1(x). \end{aligned} \tag{55}$$

In this case we can identify the series as xe^x, but we are not always this fortunate, so let us keep working with the series form in this example.

Theorem 4.3.1 tells us that y_2 can be found in the form (41b), where $r=1$ and the c_n's are to be determined. Putting that form into (49) gives

$$\begin{aligned} x^2y_2''-(x+x^2)y_2'+y_2=&\left[x^2y_1''-(x+x^2)y_1'+y_1\right]\ln x+2xy_1'-(2+x)y_1\\ &+\sum_1^\infty n(n+1)c_nx^{n+1}+\sum_1^\infty c_nx^{n+1}\\ &-\sum_1^\infty(n+1)c_nx^{n+1}-\sum_1^\infty(n+1)c_nx^{n+2}=0, \end{aligned} \tag{56}$$

where $y_1(x)=\sum_0^\infty x^{n+1}/n!$. The square-bracketed terms in (56) cancel to zero because y_1 is a solution of (49). If we move $2xy_1'-(2+x)y_1$ to the right-hand side, and write out the various series, then (56) becomes

$$c_1x^2+4c_2x^3+9c_3x^4+\cdots-2c_1x^3-3c_2x^4-\cdots=-x^2-x^3-\frac{1}{2}x^4-\cdots, \tag{57}$$

and equating coefficients of like powers of x, on the left- and right-hand sides, gives

$$\begin{aligned} x^2:&\quad c_1=-1,\\ x^3:&\quad 4c_2-2c_1=-1,\\ x^4:&\quad 9c_3-3c_2=-\tfrac{1}{2}, \end{aligned} \tag{58}$$

and so on. Thus, $c_1 = -1$, $c_2 = -\frac{3}{4}$, $c_3 = -\frac{11}{36}, \ldots$ and

$$\begin{aligned} y_2(x) &= y_1(x)\ln x + \sum_1^\infty c_n x^{n+1} \\ &= \left(x + x^2 + \frac{1}{2}x^3 + \frac{1}{6}x^4 + \cdots\right)\ln x - x^2 - \frac{3}{4}x^3 - \frac{11}{36}x^4 - \cdots. \end{aligned} \tag{59}$$

If, instead, we retain the summation notation, then in place of (57) we obtain, after manipulation and simplification,

$$\sum_1^\infty \left(n^2 c_n - n c_{n-1}\right) x^{n+1} = -\sum_1^\infty \frac{1}{(n-1)!} x^{n+1} \tag{60}$$

[where $c_0 \equiv 0$ because there is no c_0 term in (41b)] and hence the recursion formula

$$n^2 c_n - n c_{n-1} = -\frac{1}{(n-1)!}$$

or, more conveniently,

$$c_n = \frac{1}{n} c_{n-1} - \frac{1}{n\,n!}. \qquad (c_0 \equiv 0) \tag{61}$$

Solving (61) gives

$$\begin{aligned} c_1 &= -1, \\ c_2 &= -\frac{1}{2!}\left(1 + \frac{1}{2}\right) = -\frac{3}{4}, \\ c_3 &= -\frac{1}{3!}\left(1 + \frac{1}{3}\right) = -\frac{11}{36}, \end{aligned} \tag{62}$$

and so on. These results agree with those obtained from (58), but the gain, here, is that (61) can give us as many c_n's as we wish. In fact, by carrying (62) further we can see that

$$c_n = -\frac{1}{n!}\left(1 + \frac{1}{2} + \cdots + \frac{1}{n}\right) \tag{63}$$

for any $n = 1, 2, \ldots$. [The price that we paid for that gain was that we needed to manipulate the series by shifting some of the summation indices and summation limits in order to obtain (60).]

COMMENT 1. In this example we were able to sum the series in (55) and obtain $y_1(x)$ in the closed form

$$y_1(x) = xe^x. \tag{64}$$

In such a case it is more convenient to seek y_2 by actually carrying out the reduction-of-order process that lay behind the general form (41b). Thus, seek $y_2(x) = A(x)y_1(x)$. The steps were already carried out in the preceding outline of the proof of Theorem 4.3.1, so we can immediately use (45). With $C = 1$,

$$A'(x) = \frac{e^{-\int p(x)\,dx}}{y_1^2(x)} = \frac{e^{-\int(-\frac{1}{x}-1)dx}}{x^2 e^{2x}} = \frac{e^{\ln x} e^x}{x^2 e^{2x}} = \frac{e^{-x}}{x}, \tag{65}$$

so

$$A(x) = \int \frac{e^{-x}}{x}\,dx = \int \left(\frac{1}{x} - 1 + \frac{x}{2!} - \frac{x^2}{3!} + \cdots\right) dx$$
$$= \ln x + \int \sum_1^\infty (-1)^n \frac{x^{n-1}}{n!}\,dx = \ln x + \sum_1^\infty (-1)^n \frac{x^n}{n\,n!} \tag{66}$$

and

$$y_2(x) = A(x)y_1(x) = \left[\ln x + \sum_1^\infty (-1)^n \frac{x^n}{n\,n!}\right] xe^x, \tag{67}$$

which expression is found, upon expanding the e^x, to agree with (59).

COMMENT 2. As a matter of fact, we can leave the integral in (66) intact because it is a tabulated function. Specifically, the integral of e^{-x}/x, though nonelementary, comes up often enough for it to have been defined, in the literature, as a special function:

$$\boxed{E_1(x) = \int_x^\infty \frac{e^{-t}}{t}\,dt} \qquad (x > 0) \tag{68}$$

is known as the **exponential integral**. Among its various properties is the expansion

$$E_1(x) = -\gamma - \ln x - \sum_1^\infty (-1)^n \frac{x^n}{n\,n!}, \qquad (x > 0) \tag{69}$$

where $\gamma = 0.5772157$ is **Euler's constant**. Using the $E_1(x)$ function, we can express

$$y_2(x) = A(x)y_1(x) = \left(\int_a^x \frac{e^{-t}}{t}\,dt\right) xe^x$$
$$= \left(\int_a^\infty \frac{e^{-t}}{t}\,dt - \int_x^\infty \frac{e^{-t}}{t}\,dt\right) xe^x = [E_1(a) - E_1(x)]\,xe^x, \tag{70}$$

for any $a > 0$. The $E_1(a)xe^x$ term is merely $E_1(a)$ times $y_1(x)$, so it can be dropped with no loss. Further, the factor -1 in front of the $E_1(x)xe^x$ can likewise be dropped. Thus, in this example we were able to obtain both solutions in closed form, $y_1(x) = xe^x$ and $y_2(x) = E_1(x)xe^x$.

COMMENT 3. Observe that the Taylor series

$$xp(x) = x\left(\frac{-x - x^2}{x^2}\right) = -1 - x, \qquad x^2q(x) = x^2\left(\frac{1}{x^2}\right) = 1 \tag{71}$$

both terminate, hence they surely converge with $R_1 = \infty$ and $R_2 = \infty$, respectively. Thus, Theorem 4.3.1 guarantees that $\sum_0^\infty a_n x^n$ in (40) and $\sum_0^\infty c_n x^n$ in (41b) will likewise have infinite radii of convergence. Of course the $\ln x$ in (41b) tends to $-\infty$ as $x \to 0$, but nevertheless our solutions y_1 and y_2 are indeed valid on the full interval $0 < x < \infty$. ∎

EXAMPLE 5. *Case (iii).* Solve the equation

$$xy'' + y = 0; \qquad (0 < x < \infty) \tag{72}$$

that is, find two LI solutions. The only singular point of (72) is $x = 0$, and it is a regular singular point. Seeking

$$y(x) = \sum_0^\infty a_n x^{n+r}, \qquad (a_0 \neq 0)$$

(72) becomes

$$\sum_0^\infty (n+r)(n+r-1)a_n x^{n+r-1} + \sum_0^\infty a_n x^{n+r} = 0. \tag{73}$$

Set $n - 1 = m$ in the first sum, in which case the lower summation limit becomes -1, then change the m's back to n's. In the second sum change the lower limit to -1, with the understanding that $a_{-1} \equiv 0$. Then (73) becomes

$$\sum_{n=-1}^\infty \left[(n+r+1)(n+r)a_{n+1} + a_n\right] x^{n+r} = 0,$$

so we have the recursion formula

$$(n+r+1)(n+r)a_{n+1} + a_n = 0, \qquad (a_{-1} \equiv 0,\ a_0 \neq 0) \tag{74}$$

for $n = -1, 0, 1, 2, \ldots$. Setting $n = -1$, and using $a_{-1} = 0$ and $a_0 \neq 0$, gives the indicial equation

$$r(r-1) = 0, \tag{75}$$

with roots $r_1 = 1$ and $r_2 = 0$. These differ by an integer, so that the problem is of case (iii) type. Let us try the smaller root first. That root will, according to Theorem 4.3.1, lead to both solutions or to neither. With $r = r_2 = 0$, (74) becomes $(n+1)na_{n+1} + a_n = 0$. Having already used $n = -1$, to obtain (74), we next set $n = 0$. That step gives $0 + a_0 = 0$, so that $a_0 = 0$, which contradicts the assumption that $a_0 \neq 0$. Thus, $r = r_2 = 0$ yields *no* solutions. Thus, we will use the larger root, $r = r_1 = 1$, to obtain one solution, and then (41c) to obtain the other.

With $r = r_1 = 1$, (74) gives

$$a_{n+1} = -\frac{1}{(n+2)(n+1)} a_n. \tag{76}$$

Working out the first several a_n's from (76), we find that

$$a_n = -\frac{(-1)^n}{(n+1)(n!)^2} a_0,$$

so

$$y(x) = \sum_0^\infty \frac{(-1)^n a_0}{(n+1)(n!)^2} x^{n+1} = a_0 y_1(x),$$

where

$$y_1(x) = \sum_0^\infty \frac{(-1)^n}{(n+1)(n!)^2} x^{n+1}. \tag{77}$$

[Remember, throughout, that $0! = 1$ and $(-1)^0 = 1$.]

To find y_2, we use (41c) and seek

$$y_2(x) = \kappa y_1(x) \ln x + \sum_0^\infty d_n x^n. \tag{78}$$

Putting (78) into (72) gives

$$\kappa x^2 y_1'' \ln x + 2\kappa x y_1' - \kappa y_1 + \sum_0^\infty n(n-1) d_n x^n$$
$$+\kappa x y_1 \ln x + \sum_0^\infty d_n x^{n+1} = 0. \tag{79}$$

Cancelling the $\ln x$ terms [because y_1 satisfies (72)], re-expressing the last sum in (79) as $\sum_0^\infty d_n x^{n+1} = \sum_1^\infty d_{m-1} x^m = \sum_0^\infty d_{n-1} x^n$, where $d_{-1} \equiv 0$, and putting (77) in for the y_1 terms, (79) becomes

$$\begin{aligned}\sum_0^\infty [n(n-1)d_n + d_{n-1}]\, x^n &= \kappa y_1 - 2\kappa x y_1' \\ &= -\kappa \sum_0^\infty \frac{(-1)^n (2n+1)}{(n+1)(n!)^2} x^{n+1} \\ &= -\kappa \sum_1^\infty \frac{(-1)^{n-1}(2n-1)}{n[(n-1)!]^2} x^n,\end{aligned} \tag{80}$$

where, to obtain the last equality, we let $n+1 = m$ and then set $m = n$. Equating coefficients of like powers of x gives the recursion formula

$$n(n-1)d_n + d_{n-1} = -\kappa \frac{(-1)^{n-1}(2n-1)}{n[(n-1)!]^2} \tag{81}$$

for $n = 1, 2, \ldots$. [We can begin with $n = 1$ because equating the constant terms on both sides of (81) merely gives $0 = 0$.] Letting $n = 1, 2, \ldots$ gives

$$\begin{aligned} n = 1: &\quad d_0 = -\kappa, \\ n = 2: &\quad d_2 = \frac{3}{4}\kappa - \frac{1}{2} d_1, \\ n = 3: &\quad d_3 = -\frac{7}{36}\kappa + \frac{1}{12} d_1, \\ n = 4: &\quad d_4 = \frac{35}{1728}\kappa - \frac{1}{144} d_1, \end{aligned} \tag{82}$$

and so on, where d_1 remains arbitrary. Thus, the series in (78) is

$$\begin{aligned}\sum_0^\infty d_n x^n = \kappa &\left(-1 + \frac{3}{4}x^2 - \frac{7}{36}x^4 + \frac{35}{1728}x^5 + \cdots\right) \\ &+ d_1 \left(x - \frac{1}{2}x^2 + \frac{1}{12}x^3 - \frac{1}{144}x^4 + \cdots\right).\end{aligned} \tag{83}$$

The series multiplying d_1, on the right side of (83), is identical to $y_1(x)$, given by (77), so we can set $d_1 = 0$ without loss. With $d_1 = 0$, we see that the entire right side of (78) is scaled by κ, which has remained arbitrary, so there is no loss in setting $\kappa = 1$.

Thus, $y_2(x)$ is given by (78), wherein $y_1(x)$ is given by (77) and the d_n's by (81), with d_1 taken to be zero and $\kappa = 1$. ∎

EXAMPLE 6. *Case (iii).* Solve

$$4x^2y'' + 4xy' - y = 0 \tag{84}$$

by the method of Frobenius. This has been a long and arduous section so we will only outline the solution to (84). Seeking a Frobenius expansion $y(x) = \sum_0^\infty a_n x^{n+r}$ about the regular singular point $x = 0$, we obtain the indicial equation $4r^2 - 1 = 0$, so $r = \pm 1/2$, which corresponds to case (iii) of Theorem 4.3.1. We find that the larger root $r_1 = 1/2$ leads to the one-term solution $y(x) = a_0 x^{1/2}$ (i.e., $a_1 = a_2 = \cdots = 0$), and that the smaller root $r_2 = -1/2$ leads to $y(x) = a_0 x^{-1/2} + a_1 x^{1/2}$ (i.e., $a_2 = a_3 = \cdots = 0$), which is the general solution. We did not, in (84), specify the x interval of interest. Suppose that it is $x < 0$. Then a general solution of (84) is $y(x) = a_0|x|^{-1/2} + a_1|x|^{1/2}$, and that solution is valid on the entire interval $x < 0$.

In fact, (84) is an elementary equation, a Cauchy-Euler equation, so we could have solved it more easily. But we wanted to show that it can nonetheless be solved by the Frobenius method, and that that method does indeed give the correct one-term solutions. ∎

One final point: what if the indicial equation gives complex roots $r = \alpha \pm i\beta$? This issue came up in Section 3.6.1 as well, for the Cauchy-Euler equation. Our treatment here is virtually the same as in Section 3.6.1 and is left for Exercise 10.

Closure. The Frobenius theory, embodied in Theorem 3.4.1, enables us to find a general solution to any second-order linear ordinary differential equaton with a regular singular point at $x = 0$, in the form of generalized power series expansions about that point, possibly with $\ln x$ included. There are exactly three possible cases: if the roots of the indicial equation (38) are r_1, r_2, where $r_1 \geq r_2$ if they are real, then if the roots are distinct and not differing by an integer (which includes the case where the roots are complex) then LI solutions are given by (40) and (41a); if the roots are repeated then LI solutions are given by (40) and (41b); and if $r_1 - r_2$ is an integer then LI solutions are given by (40) and (41c). Theorem 3.4.1 is by no means of theoretical interest alone, since applications, especially the solution by separation of variables of the classical partial differential equations of mathematical physics and engineering (such as the diffusion, Laplace, and wave equations), often lead to nonconstant-coefficient second-order linear differential equations with regular singular points, such as the well known Legendre and Bessel equations. We devote Sections 4.4 and 4.6 to those two important cases.

Computer software. It is fortunate that computer-algebra systems can even generate Frobenius-type solutions, fortunate because the hand calculations can be quite tedious, as our examples have shown. Thus, we urge you to study the theory in this

section on the one hand and to learn how to use such software as well. To illustrate, let us use the *Maple* dsolve command (discussed at the end of Section 4.2) to obtain a Frobenius-type solution of the differential equation $xy'' + y = 0$ about the regular singular point $x = 0$; this was our Example 6. Enter

$$\text{dsolve}(x * \text{ diff}(y(x),\ x,\ x) + y(x) = 0,\ y(x),\ \text{type} = \text{series});$$

and return. The resulting output

$$\begin{aligned} y(x) = {} & _C1\, x \left(1 - \frac{1}{2}x + \frac{1}{12}x^2 - \frac{1}{144}x^3 + \frac{1}{2880}x^4 - \frac{1}{86400}x^5 + O(x^6)\right) \\ & + _C2 \left[\ln(x)\left(-x + \frac{1}{2}x^2 - \frac{1}{12}x^3 + \frac{1}{144}x^4 - \frac{1}{2880}x^5 + O(x^6)\right)\right. \\ & \left. + \left(1 - \frac{3}{4}x^2 + \frac{7}{36}x^3 - \frac{35}{1728}x^4 + \frac{101}{86400}x^5 + O(x^6)\right)\right] \end{aligned}$$

is found to agree with the general solution that we generated in Example 5.

EXERCISES 4.3

1. For each equation, identify all singular points (if any), and classify each as regular or irregular. For each regular singular point use Theorem 4.3.1 to determine the minimum possible radii of convergence of the series that will result in (40) and (41) (but you need not work out those series).

(a) $y'' - x^3y' + xy = 0$
(b) $xy'' - (\cos x)y' + 5y = 0$
(c) $(x^2 - 3)y'' - y = 0$
(d) $x(x^2 + 3)y'' + y = 0$
(e) $(x + 1)^2y'' - 4y' + (x + 1)y = 0$
(f) $y'' + (\ln x)y' + 2y = 0$
(g) $(x - 1)(x + 3)^2y'' + y' + y = 0$
(h) $xy'' + (\sin x)y' - (\cos x)y = 0$
(i) $x(x^4 + 2)y'' + y = 0$
(j) $(x^4 - 1)y'' + xy' - x^2y = 0$
(k) $(x^4 - 1)^3y'' + (x^2 - 1)^2y' - y = 0$
(l) $(x^4 - 1)^3y'' - 3(x + 1)^2y' + x(x + 1)y = 0$
(m) $(xy')' - 5y = 0$
(n) $\left[x^3(x - 1)y'\right]' + 2y = 0$
(o) $2x^2y'' - xy' + 7y = 0$
(p) $xy'' + 4y' = 0$
(q) $x^2y'' - 3y = 0$
(r) $2x^2y'' + \sqrt{\pi}\, y = 0$

2. Sometimes one can change an irregular singular point to a regular singular point, by suitable change of variables, so that the Frobenius theory can be applied. The purpose of this exercise is to present such a case. We noted, in Example 3, that $y'' + \sqrt{x}\, y = 0$ $(x > 0)$ has an irregular singular point at $x = 0$, because of the $\sqrt{x}$.

(a) Show that if we change the independent variable from x to t, say, according to $\sqrt{x} = t$, then the equation on $y(x(t)) = Y(t)$ is

$$Y''(t) - \frac{1}{t}Y'(t) + 4t^3Y(t) = 0. \qquad (t > 0) \qquad (2.1)$$

(b) Show that (2.1) has a regular singular point at $t = 0$ (which point corresponds to $x = 0$).
(c) Obtain a general solution of (2.1) by the Frobenius method. (If possible, give the general term of any series obtained.) Putting $t = \sqrt{x}$ in that result, obtain the corresponding general solution of $y'' + \sqrt{x}\, y = 0$. Is that general solution for $y(x)$ of Frobenius form? Explain.
(d) Use computer software to find a general solution.

3. In each case, there is a regular singular point at the left end of the stated x interval; call that point x_0. Merely introduce a change of independent variable, from x to t, according to $x - x_0 = t$, and obtain the new differential equation on $y(x(t)) = Y(t)$. You need not solve that equation.

(a) $(x-1)y'' + y' - y = 0, \quad (1 < x < \infty)$
(b) $(x^2-1)y'' + y = 0, \quad (1 < x < \infty)$
(c) $(x+3)y'' - 2(x+3)y' - 4y = 0, \quad (-3 < x < \infty)$
(d) $(x-5)^2y'' + 2(x-5)y' - y = 0, \quad (5 < x < \infty)$

4. Derive the series solution (25).

5. Make up a differential equation that will have as the roots of its indicial equation

(a) $1, 4$ (b) $3, 3$ (c) $1/2, 2$ (d) $-1/2, 1/2$
(e) $2 \pm 3i$ (f) $-1, -1$ (g) $-2/3, 5$ (h) $-1 \pm i$
(i) $(1 \pm 2i)/3$ (j) $5/4, 8/3$

6. In each case verify that $x = 0$ is a regular singular point, and use the method of Frobenius to obtain a general solution $y(x) = Ay_1(x) + By_2(x)$ of the given differential equation, on the interval $0 < x < \infty$. That is, determine $y_1(x)$ and $y_2(x)$. On what interval can you be certain that your solution is valid? HINT: See Theorem 4.3.1.

(a) $2xy'' + y' + x^3y = 0$
(b) $xy'' + y' - xy = 0$
(c) $xy'' + y' + x^8y = 0$
(d) $xy'' + y' + xy = 0$
(e) $x^2y'' + xy' - y = 0$
(f) $x^2y'' - x^2y' - 2y = 0$
(g) $x^2y'' + xy' - (1+2x)y = 0$
(h) $x^2y'' + xy' - y = 0$
(i) $xy'' + xy' + (1+x)y = 0$
(j) $3xy'' + y' + y = 0$
(k) $x(1+x)y'' + y = 0$
(l) $x^2(2+x)y'' - y = 0$
(m) $x^2y'' - (2+3x)y = 0$
(n) $5xy'' + y' + 8x^2y = 0$
(o) $xy'' + e^xy = 0$
(p) $2xy'' + e^xy' + y = 0$
(q) $16x^2y'' + 8xy' - 3y = 0$
(r) $16x^2y'' + 8xy' - (3+x)y = 0$
(s) $x^2y'' + xy' + (\sin x)y = 0$
(t) $5(xy)'' - 9y' + xy = 0$
(u) $(xy')' - y = 0$
(v) $(xy')' - 2y' - y = 0$

7. (a)–(x) Use computer software to obtain a general solution of the corresponding differential equation in Exercise 6.

8. Use the method of Frobenius to obtain a general solution to the equation $xy'' + cy' = 0$ on $x > 0$, where c is a real constant. You may need to treat different cases, depending upon c.

9. (a) The equation

$$(x^2 - x)y'' + (4x-2)y' + 2y = 0, \quad (0 < x < 1) \quad (9.1)$$

has been "rigged" to have, as solutions, $1/x$ and $1/(1-x)$. Solve (9.1) by the method of Frobenius, and show that you do indeed obtain those two solutions.
(b) You may have wondered how we made up the equation (9.1) so as to have the two desired solutions. Here, we ask you to make up a linear homogenous second-order differential equation that has two prescribed LI solutions $F(x)$ and $G(x)$.

10. (*Complex roots*) Since $p(x)$ and $q(x)$ are real-valued functions, p_0 and q_0 are real. Thus, if the indicial equation (38) has complex roots they will be complex conjugates, $r = \alpha \pm i\beta$, so case (i) of Theorem 3.4.1 applies, and the method of Frobenius will give a general solution of the form

$$\begin{aligned} y(x) &= Ay_1(x) + By_2(x) \\ &= Ax^{\alpha+i\beta}\textstyle\sum_0^\infty a_nx^n + Bx^{\alpha-i\beta}\sum_0^\infty b_nx^n. \end{aligned} \quad (10.1)$$

(a) Show that the b_n's will be the complex conjugates of the a_n's: $b_n = \overline{a}_n$.
(b) Recalling, from Section 3.6.1, that

$$x^{\alpha\pm i\beta} = x^\alpha\left[\cos(\beta\ln x) \pm i\sin(\beta\ln x)\right], \quad (10.2)$$

show that (10.1) [with b_n replaced by $\overline{a}_n$, according to the result found in part (a) above] can be re-expressed in terms of real functions as

$$\begin{aligned} y(x) = Cx^\alpha\big[&\cos(\beta\ln x)\textstyle\sum_0^\infty c_nx^n \\ &- \sin(\beta\ln x)\textstyle\sum_0^\infty d_nx^n\big] \\ +Dx^\alpha\big[&\cos(\beta\ln x)\textstyle\sum_0^\infty d_nx^n \\ &+ \sin(\beta\ln x)\textstyle\sum_0^\infty c_nx^n\big], \end{aligned} \quad (10.3)$$

where c_n, d_n are the real and imaginary parts of a_n, respectively: $a_n = c_n + id_n$.
(c) Find a general solution of the form (10.3) for the equation

$$x^2y'' + x(1+x)y' + y = 0.$$

That is, determine α, β and c_n, d_n in (10.3), through $n = 3$, say.
(d) The same as (c), for $x^2y'' + xy' + (1-x)y = 0$.

4.4 Legendre Functions

4.4.1. Legendre polynomials. The differential equation

$$\boxed{\left(1-x^2\right)y''-2xy'+\lambda y=0,} \tag{1}$$

where λ is a constant, is known as **Legendre's equation**, after the French mathematician *Adrien-Marie Legendre* (1752–1833). The x interval of interest is $-1 < x < 1$, and (1) has regular singular points at each endpoint, $x = \pm 1$. In this section we study aspects of the Legendre equation and its solutions that will be needed in applications in later chapters. There, we will be interested in power series solutions about the ordinary point $x = 0$,

$$y(x)=\sum_{k=0}^{\infty} a_k x^k. \tag{2}$$

Putting (2) into (1) leads to the recursion formula (Exercise 1)

$$a_{k+2}=\frac{k(k+1)-\lambda}{(k+1)(k+2)}a_k. \qquad (k=0,1,2,\cdots) \tag{3}$$

Setting $k = 0, 1, 2, \ldots$, in turn, shows that a_0 and a_1 are arbitrary, and that subsequent a_k's can be expressed, alternately, in terms of a_0 and a_1:

$$a_2=-\frac{\lambda}{2}a_0, \quad a_3=\frac{2-\lambda}{6}a_1, \quad a_4=-\frac{(6-\lambda)\lambda}{24}a_0,$$

and so on, and we have the general solution

$$\begin{aligned} y(x) &= a_0\left[1-\frac{\lambda}{2}x^2-\frac{(6-\lambda)\lambda}{24}x^4-\frac{(20-\lambda)(6-\lambda)\lambda}{720}x^6-\cdots\right] \\ &\quad +a_1\left[x+\frac{2-\lambda}{6}x^3+\frac{(12-\lambda)(2-\lambda)}{120}x^5+\cdots\right] \\ &= a_0y_1(x)+a_1y_2(x) \end{aligned} \tag{4}$$

of (1). To determine the radii of convergence of the two series in (4) we can use the recursion formula (3) and Theorem 4.2.2, provided that we realize that the a_{k+2} on the left side of (3) is really the next coefficient after a_k, the "a_{k+1}" in Theorem 4.2.2 since every other term in each series is missing. Thus, (3) gives

$$\lim_{k\to\infty}\left|\frac{\text{"}a_{k+1}\text{"}}{a_k}\right|=\lim_{k\to\infty}\left|\frac{k(k+1)-\lambda}{(k+1)(k+2)}\right|=1, \tag{5}$$

and it follows from Theorem 4.2.2 that $R = 1$, so each series converges in $-1 < x < 1$.

In physical applications of the Legendre equation, such as finding the steady-state temperature distribution within a sphere subjected to a known temperature distribution on its surface, one needs solutions that are *bounded* on $-1 < x < 1$. [$F(x)$ being **bounded** on an interval I means that there exists a finite constant M such that $|F(x)| \leq M$ for all x in I. If $F(x)$ is not bounded then it is **unbounded**.] However, for arbitrary values of the parameter λ the functions $y_1(x)$ and $y_2(x)$ given in (4) grow unboundedly as $x \to \pm 1$, as illustrated in Fig. 1 for $\lambda = 1$. If you studied Section 4.3, then you could investigate why that is so by developing a Frobenius-type solution about the right endpoint $x = 1$, which is a regular singular point of (1). Doing so, you would find a $\ln(1-x)$ term, within the solutions, which is infinite at $x = 1$. Similarly, a Frobenius solution about $x = -1$ would reveal a $\ln(1+x)$ term, which is infinite at $x = -1$. Evidently, $y_1(x)$ and $y_2(x)$, above, contain linear combinations of $\ln(1-x)$ and $\ln(1+x)$ [of course, one cannot see them explicitly in (4) because (4) is an expansion about $x = 0$, not $x = 1$ or $x = -1$] so they grow unboundedly as $x \to \pm 1$.

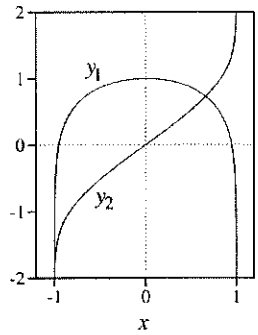


Figure 1. $y_1(x)$ and $y_2(x)$ in (4), for $\lambda = 1$.

Nonetheless, for certain specific values of λ one series or the other, in (4), will *terminate* and thereby be bounded on the interval since it is then a finite degree polynomial! Specifically, if λ happens to be such that

$$\lambda = n(n+1) \tag{6}$$

for any integer $n = 0, 1, 2, \ldots$, then we can see from (3) that one of the two series terminates at $k = n$: if $\lambda = n(n+1)$, where n is even, then the even-powered series terminates at $k = n$ (because $a_{n+2} = a_{n+4} = \cdots = 0$). For example, if $n = 2$ and $\lambda = 2(2+1) = 6$, then the $6 - \lambda$ factor in every term after the second, in the even-powered series, causes all those terms to vanish, so the series terminates as $1 - 3x^2$. Similarly, if $\lambda = n(n+1)$, where n is odd, then the odd-powered series terminates at $k = n$. The first five such λ's, and their corresponding polynomial solutions of (1), are shown in the second and third columns of Table 1. These

Table 1. The first five Legendre polynomials.

n	$\lambda = n(n+1)$	Polynomial Solution	Legendre Polynomial $P_n(x)$
0	0	1	$P_0(x) = 1$
1	2	x	$P_1(x) = x$
2	6	$1 - 3x^2$	$P_2(x) = \frac{1}{2}(3x^2 - 1)$
3	12	$x - \frac{5}{3}x^3$	$P_3(x) = \frac{1}{2}(5x^3 - 3x)$
4	20	$1 - 10x^2 + \frac{35}{3}x^4$	$P_4(x) = \frac{1}{8}(35x^4 - 30x^2 + 3)$

polynomial solutions can, of course, be scaled by any desired numerical factor. Scaling them to equal unity at $x = 1$, by convention, they become the so-called **Legendre polynomials**. Thus, the Legendre polynomial $P_n(x)$ is a polynomial

solution of the Legendre equation

$$(1-x^2)y'' - 2xy' + n(n+1)y = 0, \tag{7}$$

scaled so that $P_n(1) = 1$. In fact, it can be shown that they are given explicitly by the formula

$$P_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}\left[(x^2-1)^n\right], \tag{8}$$

which result is known as **Rodrigues's formula**.

4.4.2. Orthogonality of the P_n's. For reasons that will not be evident until we study vector spaces (Section 9.6), the integral formula

$$\boxed{\int_{-1}^{1} P_j(x)P_k(x)\,dx = 0, \qquad (j \neq k)} \tag{9}$$

is known as the **orthogonality relation**. By virtue of (9), we say that $P_j(x)$ and $P_k(x)$ are orthogonal to each other – provided that they are different Legendre polynomials (i.e., $j \neq k$).

Proof of (9) is not difficult. Noting that the $(1-x^2)y'' - 2xy'$ terms in (7) can be combined as $[(1-x^2)y']'$, we begin by considering $\int_{-1}^{1}\left[(1-x^2)P_j'\right]' P_k\,dx$ and integrating by parts until all the derivatives have been transferred from P_j to P_k:

$$\begin{aligned}\int_{-1}^{1}\left[(1-x^2)P_j'\right]' P_k\,dx &= (1-x^2)\,P_j'P_k\Big|_{-1}^{1} - \int_{-1}^{1}(1-x^2)P_j'P_k'\,dx \\ &= 0 - (1-x^2)\,P_k'P_j\Big|_{-1}^{1} + \int_{-1}^{1} P_j\left[(1-x^2)P_k'\right]'dx. \end{aligned} \tag{10}$$

The next to last term is zero because of the $1-x^2$ factor, just as the boundary term following the first equal sign is zero. Since P_j and P_k are solutions of the Legendre equation

$$\left[(1-x^2)y'\right]' + n(n+1)y = 0 \tag{11}$$

for $n = j$ and k, respectively, we can use (11) to re-express (10) as

$$-j(j+1)\int_{-1}^{1} P_jP_k\,dx = -k(k+1)\int_{-1}^{1} P_jP_k\,dx \tag{12}$$

or

$$[k(k+1) - j(j+1)]\int_{-1}^{1} P_j(x)P_k(x)\,dx = 0. \tag{13}$$

Since $j \neq k$, it follows from (13) that $\int_{-1}^{1} P_jP_k\,dx = 0$, as was to be proved.

We will see later that (9) is but a special case of the more general orthogonality relation found in the Sturm-Liouville theory, which theory will be essential to us when we solve partial differential equations.

4.4.3. Generating function and properties. Besides (9), another important property of Legendre polynomials is expressed by the formula

$$\boxed{\left(1-2xr+r^2\right)^{-1/2}=\sum_0^\infty P_n(x)r^n. \qquad (|x|\le 1,\ |r|<1)} \tag{14}$$

That is, if we regard the left side of (14) as a function of r and expand it in a Taylor series about $r=0$, then the coefficient of r^n turns out to be $P_n(x)$. Thus, $\left(1-2xr+r^2\right)^{-1/2}$ is called the **generating function** for the P_n's (Exercise 4).

Equation (14) is the source of considerable additional information about the P_n's. For instance, by changing x to $-x$ in (14) it can be seen that

$$P_n(-x)=(-1)^nP_n(x). \tag{15}$$

Now, if $f(-x)=f(x)$, then the graph of f is symmetric about $x=0$ and we say that f is an **even function** of x. If, instead, $f(-x)=-f(x)$, then the graph of f is antisymmetric about $x=0$ and we say that f is an **odd function** of x. Noting that the $(-1)^n$ is $+1$ if n is an even integer and -1 if n is an odd integer, then we see from (15) that $P_n(x)$ is an even function of x if n is an even integer, and an odd function of x if n is an odd integer, as is seen to be true for the P_n's that are shown in Fig. 2.

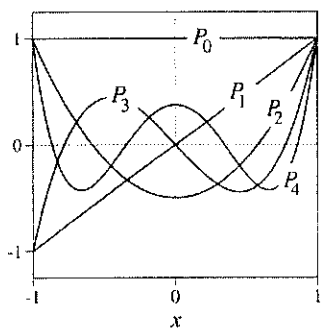


Figure 2. Graphs of the first five Legendre polynomials.

Also, by taking $\partial/\partial r$ of (14) one can show (Exercise 6) that

$$nP_n(x)=(2n-1)xP_{n-1}(x)-(n-1)P_{n-2}(x), \qquad (n=2,3,\ldots) \tag{16}$$

which is a recursion relation giving P_n in terms of P_{n-1} and P_{n-2}. Or by taking $\partial/\partial x$ of (14) instead, one can show (Exercise 7) that

$$P_n'(x)-2xP_{n-1}'(x)+P_{n-2}'(x)=P_{n-1}(x). \qquad (n=2,3,\ldots) \tag{17}$$

Finally, squaring both sides of (14) and integrating on x from -1 to $+1$, and using the orthogonality relation (9), one can show that

$$\boxed{\int_{-1}^{1}\left[P_n(x)\right]^2dx=\frac{2}{2n+1}, \qquad (n=0,1,2,\ldots)} \tag{18}$$

which result is a companion to (9); it covers the case where $j=k$ ($=n$, say). We will need (9) and (18) in later chapters.

Closure. Our principal application of Legendre's equation and Legendre polynomials, in this text, is in connection with the solution of the Laplace equation in spherical coordinates. There, we need to know how to expand a given function in terms of the Legendre polynomials $P_0(x), P_1(x), P_2(x), \ldots$, and the theory behind such expansions is covered in Section 17.6 on the Sturm-Liouville theory.

To help put that idea into perspective, recall from a first course in physics or mechanics or calculus that one can expand any given vector in 3-space in terms of

orthogonal (perpendicular) vectors "$\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}$." That fact is of great importance and was probably used extensively in those courses. Remarkably, we will be able to generalize the idea of vectors so as to regard *functions* as vectors. It will turn out that the set of Legendre polynomials $P_0, P_1, \ldots$ constitute an infinite orthogonal set of vectors such that virtually any given function defined on $-1 \le x \le 1$ can be expanded in terms of them, just as any given "arrow vector" in 3-space can be expanded in terms of $\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}$. In the present section we have not gotten that far, but the results obtained here will be used later, when we finish the story.

For a more extensive treatment of Legendre functions, Bessel functions, and the various other important special functions of mathematical physics, see, for instance, D. E. Johnson and J. R. Johnson, *Mathematical Methods in Engineering and Physics* (Englewood Cliffs, NJ: Prentice Hall, 1982). Even short of a careful study of the other special functions – such as those associated with the names Bessel, Hermite, Laguerre, Chebyshev, and Mathieu – we recommend browsing through a book like Johnson and Johnson so as to at least become aware of these functions and the circumstances under which they arise.

Computer software. In *Maple*, $P_n(x)$ is denoted as $P(n, x)$. To obtain $P_7(x)$, say, enter

with(orthopoly):

and return, then

$$P(7, x);$$

and return. The result is $\frac{429}{16}x^7 - \frac{693}{16}x^5 + \frac{315}{16}x^3 - \frac{35}{16}x$.

EXERCISES 4.4

1. Putting (2) into (1), derive the recursion formula (3).

2. Obtain (4) using computer software.

3. Use Rodrigues's formula, (8), to reproduce the first five Legendre polynomials, cited in Table 1.

4. Expanding the left-hand side of (14) in a Taylor series in r, about $r = 0$, through r^3, say, verify that the coefficients of $r^0, \ldots, r^3$ are indeed $P_0(x), \ldots, P_3(x)$, respectively.

5. We stated that by changing x to $-x$ in (14) it can be seen that $P_n(-x) = (-1)^n P_n(x)$. Show those steps and explain your reasoning.

6. (a) We stated that by taking $\partial/\partial r$ of (14) one obtains (16). Show those steps and explain your reasoning.
(b) Verify (16) for $n = 2$.
(c) Verify (16) for $n = 3$.

7. (a) We stated that by taking $\partial/\partial x$ of (14) one obtains (17). Show those steps and explain your reasoning.
(b) Verify (17) for $n = 2$.
(c) Verify (17) for $n = 3$.

8. (a) Derive (18) as follows. Squaring (14) and integrating from -1 to 1, obtain

$$\int_{-1}^{1} \frac{dx}{1 - 2rx + r^2} = \int_{-1}^{1} \sum_{m=0}^{\infty} r^m P_m(x) \sum_{n=0}^{\infty} r^n P_n(x)\, dx. \tag{8.1}$$

Integrating the left side, and using the orthogonality relation (9) to simplify the right side, obtain

$$\frac{1}{r} \ln\left(\frac{1+r}{1-r}\right) = \sum_{n=0}^{\infty} \left\{ \int_{-1}^{1} [P_n(x)]^2\, dx \right\} r^{2n}. \tag{8.2}$$

Finally, expanding the left-hand side in a Taylor series in r, show that (18) follows.
(b) Verify (18), by working out the integral, for the cases $n = 0, 1$, and 2.

9. (*Integral representation of* P_n) It can be shown that

$$P_n(x) = \frac{1}{\pi}\int_0^{\pi}\left(x+\sqrt{x^2-1}\,\cos t\right)^n dt, \quad (n=0,1,2,\ldots) \tag{9.1}$$

which is called **Laplace's integral form** for $P_n(x)$. Here, we ask you to verify (9.1) for the cases $n = 0, 1$, and 2, by working out the integral for those cases.

10. We sought power series expansions of (1) about the ordinary point $x = 0$ and, for the case where $\lambda = n(n+1)$, we obtained a bounded solution [namely, the Legendre polynomial $P_n(x)$] and an unbounded solution. Instead, seek a Frobenius-type solution about the regular singular points $x = 1$ and $x = -1$, for the case where

(a) $n = 0 \quad (\lambda = 0)$
(b) $n = 1 \quad (\lambda = 1)$
(c) $n = 2 \quad (\lambda = 2)$

11. (*Legendre functions of second kind*) For the Legendre equation (7) on the interval $-1 < x < 1$, we obtained the bounded solution $y(x) = P_n(x)$. In this exercise we seek a second LI solution, denoted as $Q_n(x)$ and called the **Legendre function of the second kind**. Then the general solution of (7) can be expressed as

$$y(x) = AP_n(x) + BQ_n(x). \tag{11.1}$$

(a) For the special case $n = 0$, solve (7) and show that a second LI solution is $\ln[(1+x)/(1-x)]$. Scaling this solution by $1/2$, we define

$$Q_0(x) = \frac{1}{2}\ln\left(\frac{1+x}{1-x}\right). \tag{11.2}$$

Sketch the graph of $Q_0(x)$ on $-1 < x < 1$, and notice that $|Q_0(x)| \to \infty$ as $x \to \pm 1$.
(b) More generally, consider any nonnegative integer n. With only $P_n(x)$ in hand, seek a second solution (by reduction of order) in the form $y(x) = A(x)P_n(x)$, and show that $Q_n(x)$ is given by

$$Q_n(x) = C_nP_n(x)\int^x \frac{dt}{(1-t^2)\,[P_n(t)]^2} + D_nP_n(x). \tag{11.3}$$

(c) Evaluating the integral in (11.3), show that the first two Q_n's are

$$Q_0(x) = \frac{1}{2}C_0\ln\left(\frac{1+x}{1-x}\right) + D_0, \tag{11.4a}$$

$$Q_1(x) = C_1\left[\frac{x}{2}\ln\left(\frac{1+x}{1-x}\right) - 1\right] + D_1x. \tag{11.4b}$$

By convention, choose $C_0 = 1$, $D_0 = 0$, $C_1 = 1$ and $D_1 = 0$, so that

$$Q_0(x) = \frac{1}{2}\ln\left(\frac{1+x}{1-x}\right), \tag{11.5a}$$

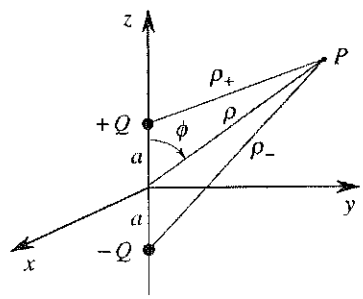
$$Q_1(x) = \frac{x}{2}\ln\left(\frac{1+x}{1-x}\right) - 1. \tag{11.5b}$$

(d) The recursion formula (16) holds for the Q_n's as well as the P_n's. Thus, with Q_0 and Q_1 in hand we can use (16) to obtain Q_2, Q_3, and so on. Do that: show that

$$Q_2(x) = \frac{3x^2-1}{4}\ln\left(\frac{1+x}{1-x}\right) - \frac{3}{2}x, \tag{11.6}$$

and obtain $Q_3(x)$ as well.

12. (*Electric field induced by two charges*) Given a positive charge Q and a negative charge $-Q$, a distance $2a$ apart, let us introduce a coordinate system as in the figure below. Since the charges lie on the z axis, it follows that the electric field that they induce will be symmetric about the z axis.



(a) Specifically, the *electric potential* (i.e., the *voltage*) Φ induced by a charge q is $\Phi = (1/4\pi\epsilon_0)(q/r)$, where the physical constant ϵ_0 is the permittivity of free space and r is the distance from the charge to the field point. Thus the potential induced at the field point P shown in the figure is

$$\Phi = \frac{1}{4\pi\epsilon_0}\left(\frac{Q}{\rho_+} - \frac{Q}{\rho_-}\right). \tag{12.1}$$

Show that (12.1) gives

$$\Phi(\rho,\phi) = \frac{Q}{4\pi\epsilon_0}\left(\frac{1}{\sqrt{a^2+\rho^2-2a\rho\cos\phi}} - \frac{1}{\sqrt{a^2+\rho^2+2a\rho\cos\phi}}\right),$$

$$= \begin{cases} \dfrac{1}{4\pi\epsilon_0}\dfrac{2Q}{\rho}\displaystyle\sum_{n=1,3,\ldots}^{\infty}\left(\frac{a}{\rho}\right)^n P_n(\cos\phi) & (\rho > a) \\ \dfrac{1}{4\pi\epsilon_0}\dfrac{2Q}{a}\displaystyle\sum_{n=1,3,\ldots}^{\infty}\left(\frac{\rho}{a}\right)^n P_n(\cos\phi). & (\rho < a) \end{cases} \tag{12.2}$$

(b) With the point P fixed, imagine letting a become arbitrarily small. Show, from (12.2), that we obtain

$$\Phi(\rho,\phi) \sim \frac{1}{4\pi\epsilon_0}2Qa\frac{\cos\phi}{\rho^2} \tag{12.3}$$

as $a \to 0$. Thus, $\Phi(\rho,\phi) \to 0$ as $a \to 0$, as makes sense, because the positive and negative charges cancel each other as they are moved together. However, observe that if, as a is decreased, Q is increased such that the product Qa is held constant, then (12.3) becomes

$$\Phi(\rho,\phi) \sim \frac{\mu}{4\pi\epsilon_0}\frac{\cos\phi}{\rho^2}, \tag{12.4}$$

where $\mu = 2Qa$ is called the **dipole moment**, and the charge configuration is said to constitute an **electric dipole**. If, for instance, a molecule is comprised of equal and opposite charges, $+Q$ and $-Q$, displaced by a very small distance $2a$, then, even though a is not tending to zero and Q to infinity, the field induced by that molecule is approximately equal to that of an idealized dipole of strength $\mu = 2Qa$, at points sufficiently far away (i.e., for $\rho/a >> 1$).

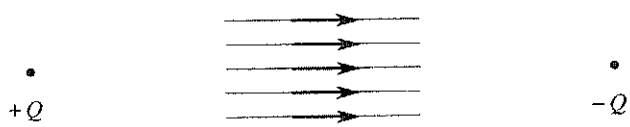
(c) As a different limit of interest, imagine the point P as fixed, and this time let a become arbitrarily *large*. Show, from (12.2), that we obtain

$$\Phi(\rho,\phi) \sim \frac{1}{4\pi\epsilon_0}\frac{2Q}{a^2}\rho\cos\phi = \frac{1}{4\pi\epsilon_0}\frac{2Q}{a^2}z \tag{12.5}$$

as $a \to \infty$. Notice that if, as a is increased, Q is increased such that Q/a^2 is held constant, then the *electric field intensity* E (which, we recall from a course in physics, is the negative of the derivative of the potential) is a constant:

$$E = -\frac{d\Phi}{dz} = -\frac{1}{4\pi\epsilon_0}\frac{Q}{2a^2}; \tag{12.6}$$

that is, we have a uniform field. Thus, a uniform electric field can be thought of as resulting from moving apart two charges, $+Q$ and $-Q$, and at the same time increasing their strength Q such that Q/a^2 is held constant as $a \to \infty$. Similarly, in fluid mechanics, a uniform fluid velocity field can be thought of as resulting from moving a fluid "source" of strength $+Q$ and a fluid "sink" of strength $-Q$ apart in such a way that Q/a^2 is held constant as $a \to \infty$, where $2a$ is their separation distance, as sketched schematically in the figure.



4.5 Singular Integrals; Gamma Function

This section amounts to a diversion from the main stream of this chapter because when we come to Bessel functions in the next section we need to know about the gamma function, and to study the gamma function we need to know about singular integrals. Furthermore, both of these topics are needed in Chapter 5 on the Laplace transform. So let us take time out to introduce them here.

4.5.1. Singular integrals. An integral is said to be **singular** (or **improper**) if one or both integration limits are infinite and/or if the integrand is unbounded on the interval; otherwise, it is **regular** (or **proper**).

For example, if

$$\begin{aligned} I_1 &= \textstyle\int_3^\infty xe^{-2x}\,dx, & I_2 &= \textstyle\int_0^5 e^{-x}\,dx/\sqrt{x}, \\ I_3 &= \textstyle\int_{-1}^2 dx/(x-1), & I_4 &= \textstyle\int_0^{100} \sqrt{x}\,e^x\,dx, \end{aligned} \tag{1}$$

then I_1 is singular due to the infinite limit, I_2 is singular because the integrand tends to ∞ as $x \to 0$, I_3 is singular because the integrand is unbounded (tends to $-\infty$ as $x \to 1$ from the left, and tends to $+\infty$ as $x \to 1$ from the right), and I_4 is regular.

Most of our interest will be in integrals that are singular by virtue of an infinite upper limit (illustrated by I_1) and/or a singularity in the integrand at the left endpoint (illustrated by I_2), so we limit this brief discussion to those cases. Other cases are considered in the exercises.

Consider the first type, $\int_a^\infty f(x)\,dx$. Analogous to our definition

$$\sum_{n=0}^{\infty} a_n \equiv \lim_{N\to\infty} \sum_{n=0}^{N} a_n \tag{2}$$

of an infinite series, we define

$$\boxed{I = \int_a^\infty f(x)\,dx \equiv \lim_{X\to\infty} \int_a^X f(x)\,dx.} \tag{3}$$

If the limit exists, we say that I is **convergent**; if not, it is **divergent**.

Recall, from our review of infinite series in Section 4.2, that the necessary and sufficient condition for the convergence of an infinite series is given by the Cauchy convergence theorem, but that theorem is difficult to apply. Thus, in the calculus, we studied a wide variety of specialized but more easily applied methods and theorems. For instance, one proves, in the calculus, that the p-**series**,

$$\sum_{1}^{\infty} \frac{1}{n^p}, \tag{4}$$

converges if $p > 1$ and diverges if $p \le 1$, the case $p = 1$ giving the well known (and divergent) **harmonic series** $\sum_1^\infty \frac{1}{n}$. That is, the terms need to die out fast enough, as n increases, for the series to converge. As p is increased, they die out faster and faster, and the borderline case is $p = 1$, with convergence requiring $p > 1$.

Then one establishes one or more *comparison tests*. For instance: If $S_1 = \sum_0^\infty a_n$ and $S_2 = \sum_0^\infty b_n$ are series of finite positive terms, and $a_n \sim Kb_n$ as $n \to \infty$ for some finite constant K, then S_1 and S_2 both converge or both diverge. (The lower limits are inconsequential insofar as convergence/divergence is concerned and have been taken to be 0 merely for definiteness.)

For instance, to determine the convergence or divergence of the series $S = \sum_1^\infty \frac{2n+3}{n^4+5}$, we observe that $\frac{2n+3}{n^4+5} \sim \frac{2}{n^3}$ as $n \to \infty$. Now, $\sum_1^\infty \frac{1}{n^3}$ is convergent

because it is a p-series with $p = 3 > 1$, and by the comparison test stated above it follows that S is convergent too.

Our development for determining the convergence/divergence of singular integrals is analogous to the development described above for infinite series. Analogous to the p-series, we study the **horizontal p-integral**,

$$I = \int_a^\infty \frac{1}{x^p}\,dx, \qquad (a > 0) \tag{5}$$

where p is a constant. (The name "horizontal p-integral" is not standard, and is explained below.) The latter integral is simple enough so that we can determine its convergence/divergence by direct evaluation. Then we can use that result, in conjunction with comparison tests, to determine the convergence/divergence of more complicated integrals. Proceeding,

$$I = \int_a^\infty \frac{1}{x^p}\,dx = \lim_{X\to\infty}\int_a^X \frac{1}{x^p}\,dx = \begin{cases} \lim_{X\to\infty} \frac{1}{1-p}\, x^{1-p}\big|_a^X & (p \neq 1) \\ \lim_{X\to\infty} \ln x\big|_a^X\,. & (p = 1) \end{cases} \tag{6}$$

Now, $\lim_{X\to\infty} \ln X$ is infinite and hence does not exist, and similarly for $\lim_{X\to\infty} X^{1-p}$ if $p < 1$, whereas the latter does exist if $p > 1$. Thus,

---

THEOREM 4.5.1 *Horizontal p-Integral*
The horizontal p-integral, (5), converges if $p > 1$ and diverges if $p \leq 1$.

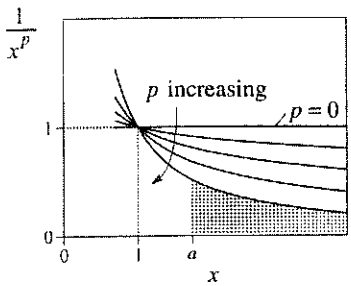


Figure 1. The effect, on $1/x^p$, of varying p.

That result is easy to remember since the p-series, likewise, converges if $p > 1$ and diverges if $p \leq 1$. Graphically, the idea is that p needs to be positive enough (namely, $p > 1$) so that the infinitely long sliver of area (shaded in Fig. 1) is squeezed thin enough to have a finite area.

We state the following comparison tests without proof.

THEOREM 4.5.2 *Comparison Tests*
Let $I_1 = \int_a^\infty f(x)\,dx$ and $I_2 = \int_a^\infty g(x)\,dx$, where $f(x)$ and $g(x)$ are positive (and bounded) on $a \leq x < \infty$.
(a) If there exist constants K and X such that $f(x) \leq Kg(x)$ for all $x \geq X$, then the convergence of I_2 implies the convergence of I_1, and the divergence of I_1 implies the divergence of I_2.
(b) If $f(x) \sim Cg(x)$ as $x \to \infty$, for some finite constant C, then I_1 and I_2 both converge or both diverge.

Of course, K must be finite. Actually, (b) is implied by (a), but we have included it explicitly since it is a simpler statement and is easier to use. Note

that C cannot be zero because the notation $f(x) \sim 0$ makes no sense. That is, $f(x) \sim g(x)$ as $x \to x_0$ means that $f(x)/g(x) \to 1$ as $x \to x_0$, and $f(x)/0$ cannot possibly tend to 1.

EXAMPLE 1. Consider $I = \displaystyle\int_9^\infty \frac{2x+3}{x^4+5}\,dx$. Since $\dfrac{2x+3}{x^4+5} \sim \dfrac{2}{x^3}$ as $x \to \infty$, and $\int_9^\infty dx/x^3$ is a convergent p-integral ($p = 4 > 1$), it follows from Theorem 4.5.2(b) that I is convergent.

COMMENT. If, instead, the integrand were $(2x+3)/(x^2+5)$, then the integral would be divergent because $(2x+3)/(x^2+5) \sim 2/x$, and $\int_9^\infty dx/x$ is a divergent p-integral ($p = 1$). It would be incorrect to argue that the integral converges because $(2x+3)/(x^2+5) \to 0$ as $x \to \infty$. Tending to zero is not enough; the integrand must tend to zero fast enough. ∎

Since the integrand of the integral in question might not be positive, as was assumed in Theorem 4.5.2, the following theorem is useful.

THEOREM 4.5.3 *Absolute Convergence*
If $\displaystyle\int_a^\infty |f(x)|\,dx$ converges, then so does $\displaystyle\int_a^\infty f(x)\,dx$, and we say that the latter converges **absolutely**.

EXAMPLE 2. Consider $I = \displaystyle\int_2^\infty \frac{\sin x}{3x^2+1}\,dx$, the integrand of which is not everywhere positive. We have

$$\left|\frac{\sin x}{3x^2+1}\right| \le \frac{1}{3x^2+1} \sim \frac{1}{3x^2} \quad \text{as } x \to \infty. \tag{7}$$

Now, $\int_2^\infty dx/x^2$ is a convergent p-integral ($p = 2 > 1$). Thus, by the asymptotic relation in (7) and Theorem 4.5.2(b), $\int_2^\infty dx/(3x^2+1)$ converges. Next, by the inequality in (7) and Theorem 4.5.2(a), $\int_2^\infty \left|\sin x/(3x^2+1)\right|\,dx$ converges. Finally, by Theorem 4.5.3, I converges. ∎

EXAMPLE 3. Consider $I = \int_0^\infty x^{100}e^{-0.01x}dx$. It might appear that this integral diverges because of the dramatic growth of the x^{100}, in spite of the $e^{-0.01x}$ decay. Let us see. Writing

$$\begin{aligned} x^{100}e^{-0.01x} &= \frac{x^{100}}{e^{0.01x}} = \frac{x^{100}}{1+(0.01x)+\frac{(0.01x)^2}{2!}+\cdots} \\ &< \frac{x^{100}}{\frac{(0.01x)^{102}}{102!}} = \frac{(102!)10^{200}}{x^2}, \end{aligned} \tag{8}$$

we see, by comparison with the p-integral, with $p = 2$, that I converges. ∎

EXAMPLE 4. Observe that

$$I = \int_3^\infty \frac{dx}{x \ln x} = \int_{x=3}^{x=\infty} \frac{d(\ln x)}{\ln x} = \lim_{X\to\infty} \ln(\ln x)\Big|_3^X = \infty, \tag{9}$$

so I is divergent. This example illustrates just how weakly $\ln x \to \infty$ as $x \to \infty$, for the integral of $1/x$ is borderline divergent ($p = 1$), and the $\ln x$ in the denominator does not even provide enough help, as $x \to \infty$, to produce borderline convergence! ∎

So much for the case where the upper limit is ∞. The other case that we consider is that in which the integrand "blows up" (i.e., tends to $+\infty$ or $-\infty$) at a finite endpoint, say the left endpoint $x = a$. If the integrand $f(x)$ blows up as $x \to a$, then in the same spirit as (3) we define

$$\boxed{I = \int_a^b f(x)\,dx \equiv \lim_{\epsilon\to 0} \int_{a+\epsilon}^b f(x)\,dx,} \tag{10}$$

where $\epsilon \to 0$ through positive values.

We first consider the so-called **vertical p-integral**

$$I = \int_0^b \frac{1}{x^p}\,dx. \qquad (b > 0) \tag{11}$$

According to (10),

$$I = \int_0^b \frac{1}{x^p}\,dx = \lim_{\epsilon\to 0} \int_\epsilon^b \frac{1}{x^p}\,dx = \begin{cases} \lim_{\epsilon\to 0} \frac{1}{1-p}\, x^{1-p}\Big|_\epsilon^b & (p \neq 1) \\ \lim_{\epsilon\to 0} \ln x\Big|_\epsilon^b. & (p = 1) \end{cases} \tag{12}$$

Now, $\lim_{\epsilon\to 0} \ln \epsilon$ is infinite ($-\infty$) and hence does not exist and, similarly for $\lim_{\epsilon\to 0} \epsilon^{1-p}$ if $p > 1$, whereas the latter limit does exist if $p < 1$. Thus,

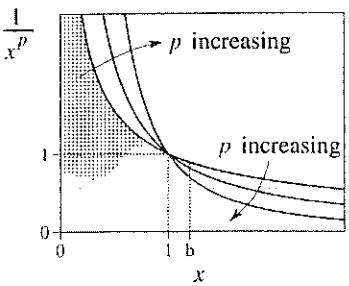


Figure 2. The effect, on $1/x^p$, of varying p.

THEOREM 4.5.4 *Vertical p-Integral*
The vertical p-integral, (11), converges if $p < 1$ and diverges if $p > 1$.

Recall that as p is increased, the *horizontal* sliver of area (shaded in Fig. 1) is squeezed thinner and thinner. For $p > 1$ it is thin enough to have finite area. However, the effect near $x = 0$ is the opposite: increasing p causes the singularity at $x = 0$ to become stronger, and the *vertical* column of area (shaded in Fig. 2) to become thicker. Thus, to squeeze the vertical column thin enough for it to have finite area, we need to make p small enough; namely, we need $p < 1$.

The motivation behind the terms "horizontal" and "vertical" p-integrals should now be apparent; the former involves the horizontal sliver shown (shaded) in Fig. 1,

and the vertical p-integral involves the vertical sliver shown (shaded) in Fig. 2.

Next, we add the following comparison test:

THEOREM 4.5.5 *Comparison Test*
Let $I = \int_0^b f(x)\,dx$, where $0 < b < \infty$. If $f(x) \sim K/x^p$ as $x \to 0$ for some constants K and p, and $f(x)$ is continuous on $0 < x \le b$, then I converges if $p < 1$ and diverges if $p \ge 1$.

EXAMPLE 5. Test the integral $\int_0^4 \left(\sin 2x/x^{3/2}\right) dx$ for convergence/divergence. Evidently, the integrand blows up as $x \to 0$ and needs to be examined there more closely. Recalling the Taylor series $\sin 2x = (2x) - (2x)^3/3! + (2x)^5/5! - \cdots$, we see that $\sin 2x \sim 2x$ as $x \to 0$ [as can be verified, if you wish, by applying l'Hôpital's rule to show that $\sin(2x)/2x \to 1$ as $x \to 0$], so

$$\frac{\sin 2x}{x^{3/2}} \sim \frac{2x}{x^{3/2}} = \frac{2}{x^{1/2}}. \tag{13}$$

Thus, according to Theorem 4.5.5, with $p = 1/2$, the integral is convergent. ∎

Example 5 concludes our introduction to singular integrals, and we are now prepared to study the gamma function.

4.5.2. Gamma function. The integral

$$\boxed{\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt \qquad (x > 0)} \tag{14}$$

is nonelementary; that is, it cannot be evaluated in closed form in terms of the so-called elementary functions. Since it arises frequently, it has been given a name, the **gamma function**, and has been studied extensively.

Observe that the integral is singular for two reasons: first, the upper limit is ∞ and, second, the integrand blows up as $t \to 0$ if the exponent $x - 1$ is negative. To determine its convergence or divergence, we can separate the two singularities by breaking the integral into the sum of an integral from $t = 0$ to $t = \tau$, say, for any $\tau > 0$, plus another from τ to ∞.* In the first, we have $t^{x-1}e^{-t} \sim t^{x-1} = 1/t^{1-x}$

*That is, if $f(t)$ is unbounded as $t \to 0$, then the integral

$$\int_0^\infty f(t)\,dt \equiv \lim_{\substack{\epsilon \to 0 \\ T \to \infty}} \int_\epsilon^T f(t)\,dt = \lim_{\substack{\epsilon \to 0 \\ T \to \infty}} \left[\int_\epsilon^\tau f(t)\,dt + \int_\tau^T f(t)\,dt\right]$$

$$= \lim_{\epsilon \to 0} \int_\epsilon^\tau f(t)\,dt + \lim_{T \to \infty} \int_\tau^T f(t)\,dt = \int_0^\tau f(t)\,dt + \int_\tau^\infty f(t)\,dt$$

exists if and only if each of the last two integrals exist.

as $t \to 0$, and by Theorem 4.5.5 we see that we have convergence if $1 - x < 1$ (i.e., $x > 0$), and divergence if $x \leq 0$. In the part from 0 to ∞, we have convergence no matter how large x is, due to the e^{-t}. Thus, the integral in (14) is convergent only if $x > 0$; hence the parenthetic stipulation in (14).

An important property of the gamma function can be derived from the definition (14) by integration by parts. With "u"$= t^{x-1}$ and "dv"$= e^{-t}dt$,

$$\Gamma(x) = -t^{x-1}e^{-t}\Big|_0^\infty + (x-1)\int_0^\infty t^{x-2}e^{-t}\,dt. \tag{15}$$

The integral in (15) converges [and is $\Gamma(x-1)$] only if $x > 1$ (rather than $x > 0$, because the exponent on t is now $x - 2$), in which case the boundary term vanishes. Thus, (15) becomes

$$\boxed{\Gamma(x) = (x-1)\Gamma(x-1). \qquad (x > 1)} \tag{16}$$

The latter is a *recursion formula* because it gives Γ at one point in terms of Γ at another point. In fact, if we compute $\Gamma(x)$, by numerical integration, over a unit interval such as $0 < x \leq 1$, then (16) enables us to compute $\Gamma(x)$ for all $x > 1$. For example,

$$\Gamma(3.2) = 2.2\,\Gamma(2.2) = (2.2)(1.2)\Gamma(1.2) = (2.2)(1.2)(0.2)\Gamma(0.2), \tag{17}$$

and one can find $\Gamma(0.2)$ in a table. (Actually, tabulations are normally given over the interval $1 < x \leq 2$ because accurate integration is difficult if x is close to 0. In fact, tables are no longer essential since the gamma function is available within most computer libraries.) Note, in particular, that if n is a positive integer, then

$$\begin{aligned}\Gamma(n+1) &= n\Gamma(n) = n(n-1)\Gamma(n-1)\\ &= \cdots = n(n-1)(n-2)\cdots(1)\Gamma(1),\end{aligned} \tag{18}$$

and since

$$\Gamma(1) = \int_0^\infty e^{-t}\,dt = 1, \tag{19}$$

(18) becomes

$$\boxed{\Gamma(n+1) = n!.} \tag{20}$$

Thus, the gamma function can be evaluated analytically at positive integer values of its argument. Another x at which the integration can be carried out is $x = 1/2$, and the result is

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}. \tag{21}$$

Derivation of (21) is interesting and is left for the exercises.

Recall that (14) defines $\Gamma(x)$ only for $x > 0$; for $x < 0$ the integral diverges and (14) is meaningless. What is a reasonable way to extend the definition of $\Gamma(x)$

to negative x? Recall that if we know $\Gamma(x)$, then we can compute $\Gamma(x+1)$ from the recursion formula $\Gamma(x+1) = x\Gamma(x)$. Instead of using this formula to step to the right [for instance, recall that knowing $\Gamma(0.2)$ we were able, in (17), to compute $\Gamma(3.2)$], we can turn it around and use it to step to the left. Thus, let us *define*

$$\Gamma(x) \equiv \frac{\Gamma(x+1)}{x} \qquad \text{for } x < 0. \tag{22}$$

For example,

$$\Gamma(-2.6) = \frac{\Gamma(-1.6)}{-2.6} = \frac{\Gamma(-0.6)}{(-2.6)(-1.6)} = \frac{\Gamma(0.4)}{(-2.6)(-1.6)(-0.6)}, \tag{23}$$

where $\Gamma(0.4)$ is known because its argument is positive. The resulting graph of $\Gamma(x)$ is shown in Fig. 3.

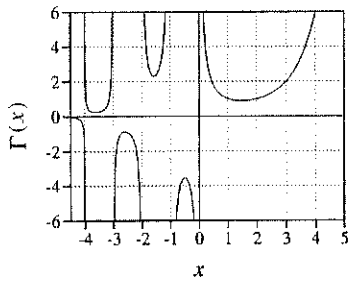


Figure 3. Gamma function, $\Gamma(x)$.

In summary then, $\Gamma(x)$ is defined for all $x \neq 0, -1, -2, \ldots$ by the integral (14) together with the leftward-stepping recursion formula (16). The singularity of $\Gamma(x)$ at $x = 0$ propagates to $x = -1, -2, \ldots$ by virtue of that formula. Especially notable is the fact that $\Gamma(x) = (x-1)!$ at $x = 1, 2, 3, \ldots$, and for this reason $\Gamma(x)$ is often referred to as the **generalized factorial function**.

A great many integrals are not themselves gamma function integrals but can be evaluated by making suitable changes of variables so as to reduce them to gamma function integrals.

EXAMPLE 6. Evaluate $I = \int_0^\infty t^{2/3} e^{-\sqrt{t}}\, dt$. Setting $\sqrt{t} = u$, we obtain

$$I = \int_0^\infty \left(u^2\right)^{2/3} e^{-u} 2u\, du = 2\int_0^\infty u^{7/3} e^{-u}\, du = 2\Gamma\left(\frac{10}{3}\right), \tag{24}$$

where $\Gamma(10/3)$ can be obtained from tables or a computer. ∎

4.5.3. Order of magnitude. In some of the foregoing examples it was important to assess the relative magnitudes of two given functions. In Example 3, for instance, the x^{100} grows as $x \to \infty$ while the $e^{-0.01x}$ decays. Which one "wins," and by what margin determines whether the integral converges or diverges.

Of particular interest are the relative growth and decay of the exponential, algebraic, and logarithmic functions as $x \to \infty$ and $x \to 0$, and we state the following elementary results as a theorem, both for emphasis and for reference.

THEOREM 4.5.6 *Relative Growth and Decay*
For any choice of positive real numbers α and β,

$$x^\alpha e^{-\beta x} \to 0 \quad \text{as } x \to \infty, \tag{25a}$$

$$(\ln x)/x^\alpha \to 0 \quad \text{as } x \to \infty, \tag{25b}$$

$$x^\alpha \ln x \to 0 \quad \text{as } x \to 0. \tag{25c}$$

Proof of (25a) can proceed as a generalization of (8) or by using l'Hôpital's rule, and (25b,c) can be proved by l'Hôpital's rule. To prove (25c), for example, observe that $x^\alpha \ln x \to (0)(-\infty)$, which result is indeterminate. To use l'Hôpital's rule we need to have $0/0$ or ∞/∞. Thus, express $x^\alpha \ln x$ as $(\ln x)/x^{-\alpha}$, which tends to $-\infty/\infty$ as $x \to 0$. Then l'Hôpital's rule gives

$$\lim_{x\to 0} \frac{\ln x}{x^{-\alpha}} = \lim_{x\to 0} \frac{x^{-1}}{-\alpha x^{-\alpha-1}} = \lim_{x\to 0} \left(-\frac{x^\alpha}{\alpha}\right) = 0.$$

We say that x^α exhibits *algebraic growth* as $x \to \infty$, and we see from (25a) that algebraic growth x^α is no match for exponential decay $e^{-\beta x}$, no matter how large α is and no matter how small β is! Of course, it follows from (25a) that $x^{-\alpha}e^{\beta x} \to \infty$ as $x \to \infty$: *algebraic decay* is no match for exponential growth. Just as exponential growth is extremely strong, logarithmic growth is extremely weak for (25b) shows that x^α dominates $\ln x$ as $x \to \infty$, no matter how small α is. Similarly as $x \to 0$: $x^{-\alpha} \to \infty$ and $\ln x \to -\infty$ (recall that $\ln x$ is zero at $x = 1$, increases without bound as $x \to \infty$, and decreases without bound as $x \to 0$; sketch it), and (25c), rewritten as $(\ln x)/x^{-\alpha} \to 0$, shows that $x^{-\alpha} \to \infty$ faster than $\ln x \to -\infty$, no matter how small α is.

Crudely then, we can think of $\ln x$ as being of the order of x to an infinitesimal positive power as $x \to \infty$, and of the order of x to an infinitesimal negative power as $x \to 0$. In contrast, one can think, crudely, of e^x as being of the order of x to an arbitrarily large positive power as $x \to \infty$, and e^{-x} as being of the order of x to an arbitrarily large negative power as $x \to \infty$.

When considering the relative strength of functions as x tends to some value x_0, constant scale factors are of no consequence no matter how large or small they may be. For instance, $(87 \ln x)/x^{0.01} \to 0$ as $x \to \infty$ just as $(\ln x)/x^{0.01}$ does. Thus, in place of the asymptotic notation $f(x) \sim g(x)$ as $x \to x_0$, which means that $f(x)/g(x) \to 1$ as $x \to x_0$, we will sometimes use the "**big oh**" notation

$$\boxed{f(x) = O(g(x)) \quad \text{as } x \to x_0} \tag{26a}$$

to mean that*

$$f(x) \sim Cg(x) \quad \text{as } x \to x_0, \tag{26b}$$

for some finite nonzero constant C. For instance, whereas

$$f(x) = \frac{\sqrt{1+\sqrt{3}}}{\Gamma(1+\sqrt{5})}\, x^{-1/2} + 673 \ln x \;\sim\; \frac{\sqrt{1+\sqrt{3}}}{\Gamma(1+\sqrt{5})}\, x^{-1/2} \tag{27}$$

as $x \to 0$, it is simpler to write $f(x) = O(x^{-1/2})$ as $x \to 0$. That is, the scale factor $C = \sqrt{1+\sqrt{3}}/\Gamma(1+\sqrt{5})$ can be omitted insofar as the order of magnitude

*Actually, the notation (26a) means that $f(x)/g(x)$ is bounded as $x \to x_0$. Though our usage is more restricted, it is consistent with the definition just given, for if (26b) holds, then surely $f(x)/g(x)$ is bounded as $x \to x_0$. Though more restricted, our definition (26b) of (26a) is sufficient for our purposes and is easier to understand and use.

of f is concerned. In words, we say that f is big oh of $x^{-1/2}$ as $x \to 0$. Of course, x_0 can be any point in (26); often, but not always, x_0 is 0 or ∞.

As one more illustration of the big oh notation, observe from the Taylor series

$$\sin x = x - \frac{x^3}{6} + \frac{x^5}{120} - \frac{x^7}{5040} + \cdots \tag{28}$$

that each of the following is true:

$$\sin x = O(x), \tag{29a}$$

$$\sin x = x + O(x^3), \tag{29b}$$

$$\sin x = x - \frac{x^3}{6} + O(x^5), \tag{29c}$$

and so on, as $x \to 0$. For instance, l'Hôpital's rule shows that $(\sin x)/x \to 1$ as $x \to 0$, so $\sin x \sim x$; hence (26b) holds, with $C = 1$, so (29a) is correct. Similarly, l'Hôpital's rule shows that $(\sin x - x)/x^3 \to -1/6$, so $\sin x - x \sim -x^3/6$; hence $\sin x - x = O(x^3)$ or $\sin x = x + O(x^3)$ so (29b) is correct.

The big oh notation is especially useful in working with series. For instance, (29b) states that if we retain only the leading term of the Taylor series (28), then the error thereby incurred is of order $O(x^3)$. Put differently, the portion omitted, $-\frac{x^3}{6} + \frac{x^5}{120} - \frac{x^7}{5040} + \cdots$, is simply $O(x^3)$.

Closure. In Section 4.5.1, we define singular integrals as integrals in which something is infinite: one or both integration limits and/or the integrand. We make such integrals meaningful by defining them as limits of regular integrals. Just as the convergence and divergence of infinite series is a long story, so is the convergence and divergence of singular integrals, but our aim here is to consider only types that will arise in this text. Though the convergence of singular integrals of the type $\int_0^\infty f(x)\,dx$ and of infinite series $\sum_1^\infty a_n$ bear a strong resemblance (e.g., the p-series and horizontal p-integral both converge for $p > 1$ and diverge for $p \le 1$), one should by no means expect all results about infinite series to merely carry over. For instance, for series convergence it is necessary (but not sufficient) that $a_n \to 0$ as $n \to \infty$, but it is *not* necessary for the convergence of $\int_0^\infty f(x)\,dx$ that $f(x) \to 0$ as $x \to \infty$. For instance, we state without proof that $\int_0^\infty \sin(x^2)\,dx$ converges, even though $\sin(x^2)$ does not tend to zero as $x \to \infty$.

In Section 4.5.2, we introduce a specific and useful singular integral, the gamma function, and obtain its recursion formula and some of its values. The exercises indicate some of its many applications.

In the final section, 4.5.3, our aim is to clarify the relative orders of magnitude of exponential, algebraic, and logarithmic functions. It is important for you to be familiar with the results listed in Theorem 4.5.6, just as you are familiar with the relative weights of cannonballs and feathers. We also introduce a simple big oh notation which is especially useful in Chapter 6 on the numerical integration of differential equations.

Computer software. Many integrals can be evaluated by symbolic computer software. With *Maple*, for instance, the relevant command is **int**. To evaluate the integral I in Example 6, for instance, enter

$$\text{int}(t\hat{}(2/3) * \exp(-t\hat{}(1/2)),\ t = 0..\text{infinity});$$

and return. The result is

$$\frac{112}{81}\frac{\pi\sqrt{3}}{\Gamma(2/3)}$$

which looks different from the result obtained in Example 6 but is actually equivalent to that result (Exercise 12). To evaluate the latter, enter

evalf(");

and return. The result is 5.556316963.

EXERCISES 4.5

1. If $\alpha > 0$ and $\beta > 0$ show that, no matter how large α is and no matter how small β is,

(a) $x^{\alpha}e^{-\beta x} \to 0$ as $x \to \infty$
(b) $x^{-\alpha}e^{\beta x} \to \infty$ as $x \to \infty$

2. If $\alpha > 0$, show that no matter how small α is,

(a) $(\ln x)/x^{\alpha} \to 0$ as $x \to \infty$
(b) $x^{\alpha}/\ln x \to 0$ as $x \to 0$

3. Show whether the given integral converges or diverges. As usual, be sure to explain your reasoning.

(a) $\displaystyle\int_0^{\infty} \frac{dx}{x^4+2}$

(b) $\displaystyle\int_0^{\infty} \frac{x^3\,dx}{x^5+2}$

(c) $\displaystyle\int_0^{\infty} \frac{x^3\,dx}{x^4+2}$

(d) $\displaystyle\int_0^{\infty} \frac{x^{3.2}\,dx}{x^4+100}$

(e) $\displaystyle\int_0^{\infty} \frac{dx}{x^2}$

(f) $\displaystyle\int_0^{\infty} \frac{dx}{\sqrt{x}}$

(g) $\displaystyle\int_4^{\infty} \frac{\sin^2 x\,dx}{\sqrt{x}(x-1)}$

(h) $\displaystyle\int_5^{\infty} \frac{\cos x\,dx}{x(x-1)}$

(i) $\displaystyle\int_1^{2} \frac{e^x\,dx}{\sqrt{x-1}}$ HINT: Let $\xi = x - 1$.

(j) $\displaystyle\int_0^{1} \frac{dx}{x^2\cos x}$

4. Show whether the given integral converges or diverges.

(a) $\displaystyle\int_1^{\infty} \frac{\ln x\,dx}{x^2}$ HINT: Show that $\ln x < x^{1/4}$ for all sufficiently large x, by showing that $(\ln x)/x^{1/4} \to 0$ as $x \to \infty$.

(b) $\displaystyle\int_0^{2} \frac{\ln x\,dx}{\sqrt{x}}$ HINT: Make the change of variables $x = 1/\xi$ and use the hint given in part (a).

5. For what p's, if any, does $\displaystyle\int_0^{\infty} dx/x^p$ converge? Explain.

6. Enter the indicated change of variables in the given singular integral, and state whether the resulting integral is singular or not.

(a) $\displaystyle\int_2^{\infty} \frac{dx}{x^4+2}, \quad x = \frac{1}{\xi}$

(b) $\displaystyle\int_0^{3} \frac{dx}{\sqrt{x}}, \quad x = \frac{1}{\xi}$

(c) $\displaystyle\int_2^{\infty} \frac{\cos x\,dx}{\sqrt{x}}, \quad x = \xi^2$

(d) $\displaystyle\int_2^\infty \frac{\cos x\,dx}{\sqrt{x}}, \quad x = \xi^4$

7. For what range of α's (such as $0 < \alpha < 2$, $\alpha > 4$, no α's, etc.) does the given integral converge? Explain.

(a) $\displaystyle\int_0^\infty \frac{dx}{x^\alpha + 3}$ (b) $\displaystyle\int_0^4 \frac{x^\alpha\,dx}{x^4+1}$

(c) $\displaystyle\int_0^\infty \frac{x^\alpha\,dx}{x+1}$ (d) $\displaystyle\int_0^2 x^\alpha \sin x\,dx$

(e) $\displaystyle\int_1^2 \left(x^2-1\right)^\alpha dx$ (f) $\displaystyle\int_1^\infty \frac{x^\alpha\,dx}{\sqrt{x^2-1}}$

8. Evaluate, using a suitable recursion formula and the known value $\Gamma(1/2) = \sqrt{\pi}$. Repeat the evaluation using computer software.

(a) $\Gamma(3.5)$ (b) $\Gamma(-3.5)$ (c) $\Gamma(6.5)$ (d) $\Gamma(-0.5)$

9. Derive (21), that $\Gamma(1/2) = \sqrt{\pi}$. HINT: Show that

$$\Gamma(1/2) = 2\int_0^\infty e^{-u^2}du,$$

so that

$$\begin{aligned}[\Gamma(1/2)]^2 &= 4\int_0^\infty e^{-u^2}du\int_0^\infty e^{-v^2}dv\\ &= 4\int_0^\infty\int_0^\infty e^{-(u^2+v^2)}du\,dv.\end{aligned}$$

Regarding the latter as a double integral in a Cartesian u, v plane, change from u, v to polar coordinates r, θ. The resulting double integral should be easier to evaluate.

10. Show by suitable change of variables that

(a) $\displaystyle\int_0^\infty e^{-x^p}dx = \frac{\Gamma(1/p)}{p} \quad (p > 0)$

(b) $\displaystyle\int_0^1 x^m(\ln x)^n\,dx = (-1)^n\frac{n!}{(m+1)^{n+1}}$
(m, n nonnegative integers)

(c) $\displaystyle\int_0^\infty x^2e^{-x^2}dx = \frac{\sqrt{\pi}}{4}$

(d) $\displaystyle\int_0^\infty xe^{-\sqrt{x}}dx = 12$

(e) $\displaystyle\int_0^\infty (x-1)^2e^{-x^3}dx = \frac{1}{3}\left[1 - 2\Gamma\left(\frac{2}{3}\right) + \Gamma\left(\frac{1}{3}\right)\right]$

11. Evaluate as many of the integrals in Exercise 10 as possible using computer software.

12. In Example 6 we obtained the value $2\Gamma(10/3)$. Using *Maple*, instead, show that the result is $(112/81)\pi\sqrt{3}/\Gamma(2/3)$. Then, use any formulas given in this section or in these exercises to show that the two results are equivalent.

13. Deduce, from the formula given in Exercise 10(a), that $\Gamma(x) \sim 1/x$ as x tends to zero through positive values.

14. (*Beta function*) Derive the result

$$\boxed{B(p,q) = \int_0^1 x^{p-1}(1-x)^{q-1}dx = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)},} \tag{14.1}$$

for $p > 0$, $q > 0$; $B(p,q)$ is known as the **beta function**. HINT: Putting $x = u^2$ in

$$\Gamma(p) = \int_0^\infty x^{p-1}e^{-x}dx, \tag{14.2}$$

show that

$$\Gamma(q)\Gamma(p) = 4\int_0^\infty u^{2p-1}e^{-u^2}du\int_0^\infty v^{2q-1}e^{-v^2}dv. \tag{14.3}$$

Regarding the latter as a double integral in a Cartesian u, v plane, change from u, v to polar coordinates r, θ. Making one more change of variables in each integral, the r integral gives $\Gamma(p+q)$ and the θ integral gives $B(p,q)$.

15. Derive, from (14.1) above, the alternative forms:

(a) $$B(p,q) = \int_0^\infty \frac{t^{p-1}}{(1+t)^{p+q}}\,dt \tag{15.1}$$
$(p > 0,\ q > 0)$

HINT: Set $x = t/(1+t)$ in (14.1).

(b) $$B(p,q) = 2\int_0^{\pi/2} \cos^{2q-1}\theta\,\sin^{2p-1}\theta\,d\theta \tag{15.2}$$
$(p > 0,\ q > 0)$

16. Using any results from the preceding two exercises, show that

(a) $$\int_0^{\pi/2} \cos^q\theta\,\sin^p\theta\,d\theta = \frac{1}{2}B\left(\frac{p+1}{2}, \frac{q+1}{2}\right) \tag{16.1}$$
$(p > -1,\ q > -1)$

(b) $$\int_0^{\pi/2} \tan^p \theta \, d\theta = \int_0^{\pi/2} \cot^p \theta \, d\theta = \frac{1}{2} B\left(\frac{1+p}{2}, \frac{1-p}{2}\right) = \frac{\pi}{2\cos\frac{p\pi}{2}} \tag{16.2}$$

for $-1 < p < 1$. HINT: You may use (17.1), below.

(c) $$\int_0^{\infty} \frac{x^a \, dx}{(1+x^b)^c} = \frac{1}{b} B\left(\frac{a+1}{b}, \frac{cb-a-1}{b}\right) = \frac{1}{b} \frac{\Gamma\left(\frac{a+1}{b}\right)\Gamma\left(\frac{cb-a-1}{2}\right)}{\Gamma(c)} \tag{16.3}$$

for $a > -1$, $b > 0$, $bc - a > 1$.

17. It can be shown, from the residue theorem of the complex integral calculus, that

$$\int_0^{\infty} \frac{x^{\alpha-1}}{1+x} \, dx = \frac{\pi}{\sin \alpha\pi} \quad (0 < \alpha < 1) \tag{17.1}$$

Using this result, (15.1), and (14.1), show that

$$\Gamma(\alpha)\Gamma(1-\alpha) = \frac{\pi}{\sin \alpha\pi}. \quad (0 < \alpha < 1) \tag{17.2}$$

18. (*Period of oscillation of pendulum*) Conservation of energy dictates that the angular displacement $\theta(t)$ of a pendulum, of length l and mass m, satisfies the differential equation

$$\frac{1}{2} m \left(l\dot{\theta}\right)^2 + mg\left(l - l\cos\theta\right) = mg\left(l - l\cos\theta_0\right). \tag{18.1}$$

(a) From (18.1), show that

$$\sqrt{\frac{l}{2g}} \int_0^{\theta_0} \frac{d\theta}{\sqrt{\cos\theta - \cos\theta_0}} = \int_0^{T/4} dt, \tag{18.2}$$

where T is the period and θ_0 is the maximum swing. We expect T to depend on θ_0, so we denote it as $T(\theta_0)$. For the case $\theta_0 = \pi/2$, show that

$$T(\pi/2) = \sqrt{\frac{2\pi l}{g}} \frac{\Gamma(1/4)}{\Gamma(3/4)}.$$

NOTE: You may use results from the preceding exercises.
(b) At first glance, it appears from (18.2) that $T(\theta_0) \to 0$ as $\theta_0 \to 0$. Is that assessment correct? Explain.

19. Let $F(x) = 4/(1+x^2) = 4 - 4x^2 + 4x^4 - \cdots$, $G(x) = \dfrac{2+3x}{5-4x}$, $H(x) = \dfrac{7x^3 - x + 1}{x^2+4}$, $I(x) = \dfrac{2x - 3\ln x}{1+x^2}$, $J(x) = 2e^{-x} + 3x$, and $K(x) = \sin 3x$. Verify the truth of each:

(a) $F(x) = O(1)$ as $x \to 0$
(b) $F(x) = 4 + O(x^2)$ as $x \to 0$
(c) $F(x) = 4 - 4x^2 + O(x^4)$ as $x \to 0$
(d) $G(x) = O(1)$ as $x \to \infty$
(e) $H(x) = O(x)$ as $x \to 0$
(f) $H(x) = O(x)$ as $x \to \infty$
(g) $H(x) = O(1)$ as $x \to 0$
(h) $I(x) = O(x^{-1})$ as $x \to \infty$
(i) $I(x) = O(\ln x)$ as $x \to 0$
(j) $J(x) = O(x)$ as $x \to \infty$
(k) $J(x) = O(1)$ as $x \to 0$
(l) $K(x) = O(1)$ as $x \to 0$

4.6 Bessel Functions

The differential equation

$$\boxed{x^2 y'' + xy' + \left(x^2 - \nu^2\right) y = 0,} \tag{1}$$

where ν is a nonnegative real number, is known as **Bessel's equation of order** ν. The equation was studied by *Friedrich Wilhelm Bessel* (1784–1846), director of the astronomical observatory at Königsberg, in connection with his work on planetary motion. Outside of planetary motion, the equation appears prominently in a wide range of applications such as steady and unsteady diffusion in cylindrical regions, and one-dimensional wave propagation and diffusion in variable

cross-section media, and it is one of the most important differential equations in mathematical physics. Dividing through by the leading coefficient x^2, we see from $p(x) = 1/x$ and $q(x) = \left(x^2 - \nu^2\right)/x^2$ that there is one singular point, $x = 0$, and that it is a regular singular point because $xp(x) = 1$ and $x^2 q(x) = x^2 - \nu^2$ are analytic at $x = 0$.

4.6.1. $\nu \neq$ integer. Consider the case where the parameter ν is not an integer. Seeking a Frobenius solution about the regular singular point $x = 0$,

$$y(x) = \sum_{k=0}^{\infty} a_k x^{k+r}, \qquad (a_0 \neq 0) \tag{2}$$

gives (Exercise 1)

$$\sum_{k=0}^{\infty} \left\{ \left[(k+r)^2 - \nu^2\right] a_k + a_{k-2} \right\} x^{k+r} = 0, \tag{3}$$

where $a_0 \neq 0$ and $a_{-2} = a_{-1} \equiv 0$. Equating to zero the coefficient of each power of x in (3) gives

$$k = 0: \quad (r^2 - \nu^2) a_0 = 0, \tag{4a}$$

$$k = 1: \quad \left[(r+1)^2 - \nu^2\right] a_1 = 0, \tag{4b}$$

$$k \geq 2: \quad \left[(r+k)^2 - \nu^2\right] a_k + a_{k-2} = 0. \tag{4c}$$

Since $a_0 \neq 0$, (4a) gives the indicial equation $r^2 - \nu^2 = 0$, with the distinct roots $r = \pm\nu$. First, let $r = +\nu$. Then (4b) gives $a_1 = 0$ and (4c) gives the recursion relation

$$a_k = -\frac{1}{k(k+2\nu)} a_{k-2}. \qquad (k \geq 2) \tag{5}$$

From (5), together with the fact that $a_1 = 0$, it follows that $a_1 = a_3 = a_5 = \cdots = 0$ and that

$$a_{2k} = \frac{(-1)^k}{2^{2k} k!\,(\nu+k)(\nu+k-1)\cdots(\nu+1)} a_0. \tag{6}$$

If ν were an integer, then the ever-growing product $(\nu+k)(\nu+k-1)\cdots(\nu+1)$ could be simplified into closed form as $\nu!/(\nu+k)!$. But since ν is not an integer, we seek to accomplish such simplification by means of the generalized factorial function, the gamma function, which is studied in Section 4.5. Specifically, if we recall the gamma function recursion formula $\Gamma(x) = (x-1)\Gamma(x-1)$, then $\Gamma(\nu+k+1) = (\nu+k)\Gamma(\nu+k) = (\nu+k)(\nu+k-1)\Gamma(\nu+k-1) = \cdots = (\nu+k)(\nu+k-1)\cdots(\nu+1)\Gamma(\nu+1)$, which gives $(\nu+k)(\nu+k-1)\cdots(\nu+1) = \Gamma(\nu+k+1)/\Gamma(\nu+1)$. With this replacement, (6) becomes

$$a_{2k} = \frac{(-1)^k \Gamma(\nu+1)}{2^{2k} k!\,\Gamma(\nu+k+1)} a_0, \tag{7}$$

so we have the solution

$$y(x) = a_0 2^\nu \Gamma(\nu+1) \sum_{k=0}^{\infty} \frac{(-1)^k}{k!\,\Gamma(\nu+k+1)} \left(\frac{x}{2}\right)^{2k+\nu}. \tag{8}$$

Dropping the $a_0 2^\nu \Gamma(\nu+1)$ scale factor, we call the resulting solution the **Bessel function of the first kind, of order** ν:

$$\boxed{J_\nu(x) = \left(\frac{x}{2}\right)^\nu \sum_{k=0}^{\infty} \frac{(-1)^k}{k!\,\Gamma(\nu+k+1)} \left(\frac{x}{2}\right)^{2k}.} \tag{9}$$

To obtain a second linearly independent solution, we turn to the other indicial root, $r = -\nu$. There is no need to retrace all of our steps; all we need to do is to change ν to $-\nu$ everywhere on the right side of (9). Denoting the result as $J_{-\nu}(x)$, the **Bessel function of the first kind, of order** $-\nu$, we have

$$\boxed{J_{-\nu}(x) = \left(\frac{x}{2}\right)^{-\nu} \sum_{k=0}^{\infty} \frac{(-1)^k}{k!\,\Gamma(k-\nu+1)} \left(\frac{x}{2}\right)^{2k}.} \tag{10}$$

Both series, (9) and (10), converge for all x, as follows from Theorem 4.3.1 with $R_1 = R_2 = \infty$ or from Theorem 4.2.2 and the recursion formula (5). The leading terms of the series in (9) and (10) are constants times x^ν and $x^{-\nu}$, respectively, so neither of the solutions $J_\nu(x)$ and $J_{-\nu}(x)$ is a scalar multiple of the other. Thus, they are LI, and we conclude that

$$y(x) = AJ_\nu(x) + BJ_{-\nu}(x) \tag{11}$$

is a general solution of (1).

Writing (9) and (10),

$$J_\nu(x) = x^\nu \left[\frac{1}{\Gamma(\nu+1)2^\nu} - \frac{1}{\Gamma(\nu+2)2^{\nu+2}}x^2 + \cdots\right], \tag{12}$$

$$J_{-\nu}(x) = x^{-\nu} \left[\frac{1}{\Gamma(1-\nu)2^{-\nu}} - \frac{1}{\Gamma(2-\nu)2^{-\nu+2}}x^2 + \cdots\right]. \tag{13}$$

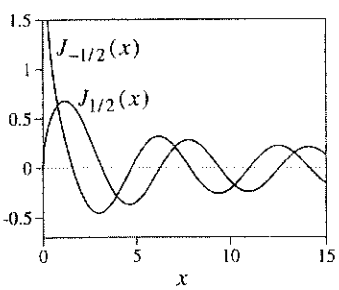


Figure 1. $J_{1/2}(x)$ and $J_{-1/2}(x)$.

Since the power series within the square brackets tend to $1/\Gamma(\nu+1)2^\nu$ and $1/\Gamma(1-\nu)2^{-\nu}$, respectively, as $x \to 0$, we see that $J_\nu(x) \sim [1/\Gamma(\nu+1)2^\nu]x^\nu$ and $J_{-\nu}(x) \sim [1/\Gamma(1-\nu)2^{-\nu}]x^{-\nu}$ as $x \to 0$. It is simpler and more concise to use the big oh notation introduced in Section 4.5.3, and say that $J_\nu(x) = O(x^\nu)$ and $J_{-\nu}(x) = O(x^{-\nu})$ as $x \to 0$. Thus, the $J_\nu(x)$'s tend to zero and the $J_{-\nu}(x)$'s tend to infinity as $x \to 0$. As representative, we have plotted $J_{1/2}(x)$ and $J_{-1/2}(x)$ in Fig. 1. In fact, for the half-integer values $\nu = \pm 1/2, \pm 3/2, \pm 5/2, \ldots$ the series in (9) and (10) can be shown to represent elementary functions. For instance (Exercise 5),

$$J_{1/2}(x) = \sqrt{\frac{2}{\pi x}}\,\sin x, \qquad J_{-1/2}(x) = \sqrt{\frac{2}{\pi x}}\,\cos x. \tag{14a,b}$$

4.6.2. ν = integer. If ν is a positive integer n, then $\Gamma(\nu+k+1) = \Gamma(n+k+1) = (n+k)!$ in (9), so we have from (9) the solution

$$\boxed{J_n(x) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!\,(k+n)!} \left(\frac{x}{2}\right)^{2k+n}} \tag{15}$$

of (1). For instance,

$$J_0(x) = 1 - \frac{x^2}{2^2} + \frac{x^4}{2^4(2!)^2} - \frac{x^6}{2^6(3!)^2} + \cdots, \tag{16a}$$

$$J_1(x) = \frac{x}{2} - \frac{x^3}{2^3 2!} + \frac{x^5}{2^5 2!\,3!} - \frac{x^7}{2^7 3!\,4!} + \cdots. \tag{16b}$$

We need to be careful with (10) because if $\nu = n$, then the $\Gamma(k-n+1)$ in (10) is, we recall from Section 4.5.2, undefined when its argument is zero or a negative integer – namely, for $k = 0, 1, \ldots, n-1$. One could say that $\Gamma(k-n+1)$ is infinite for those k's, so $1/\Gamma(k-n+1)$ equals zero for $k = 0, 1, \ldots, n-1$, and it equals $1/(k-n)!$ for $k = n, n+1, \ldots$, in which case (10) becomes

$$J_{-n}(x) = \sum_{k=n}^{\infty} \frac{(-1)^k}{k!\,(k-n)!} \left(\frac{x}{2}\right)^{2k-n}. \tag{17}$$

[The resulting equation (17) is correct, but our reasoning was not rigorous since $\Gamma(k-n+1)$ is *undefined* at $k = 0, 1, \ldots, n-1$, rather than "∞." A rigorous line of approach is suggested in Exercise 10.] Replacing the dummy summation index k by m according to $k - n = m$,

$$J_{-n}(x) = \sum_{m=0}^{\infty} \frac{(-1)^{m+n}}{(m+n)!\,m!} \left(\frac{x}{2}\right)^{2m+n}.$$

If $(-1)^n$ is factored out, the series that remains is the same as that given in (15), so that

$$J_{-n}(x) = (-1)^n J_n(x). \tag{18}$$

The result is that $J_n(x)$ and $J_{-n}(x)$ are linearly dependent, since (18) tells us that one is a scalar multiple of the other. Thus, whereas $J_\nu(x)$ and $J_{-\nu}(x)$ are LI and give the general solution (11) if ν is not an integer, we have only one linearly independent solution thus far for the case where $\nu = n$, namely, $y_1(x) = J_n(x)$ given by (15). To obtain a second LI solution $y_2(x)$ we rely on Theorem 4.3.1.

Let us begin with $n = 0$. Then we have the case of repeated indicial roots, $r = \pm n = \pm 0$, which corresponds to case (ii) of that theorem. Accordingly, we seek $y_2(x)$ in the form

$$y_2(x) = J_0(x) \ln x + \sum_{1}^{\infty} c_k x^k. \tag{19}$$

Doing so, we can evaluate the c_k's, and we obtain

$$y_2(x) = J_0(x)\ln x + \left(\frac{x}{2}\right)^2 - \left(1+\frac{1}{2}\right)\frac{1}{(2!)^2}\left(\frac{x}{2}\right)^4 + \cdots, \tag{20}$$

which is called $\mathbf{Y}_0(x)$, the Neumann function of order zero. Thus, Theorem 4.3.1 leads us to the two LI solutions $y_1(x) = J_0(x)$ and $y_2(x) = \mathbf{Y}_0(x)$, so we can use them to form a general solution of (1). However, following Weber, it proves to be convenient and standard to use, in place of $\mathbf{Y}_0(x)$, a linear combination of $J_0(x)$ and $\mathbf{Y}_0(x)$, namely,

$$y_2(x) = \frac{2}{\pi}\left[\mathbf{Y}_0(x) + (\gamma - \ln 2)J_0(x)\right] \equiv Y_0(x), \tag{21}$$

where

$$Y_0(x) = \frac{2}{\pi}\left[\left(\ln\frac{x}{2}+\gamma\right)J_0(x) + \frac{x^2}{2^2} - \left(1+\frac{1}{2}\right)\frac{x^4}{2^4(2!)^2} + \left(1+\frac{1}{2}+\frac{1}{3}\right)\frac{x^6}{2^6(3!)^2} - \cdots\right] \tag{22}$$

is Weber's **Bessel function of the second kind, of order zero**; $\gamma = 0.5772157$ is known as **Euler's constant** and is sometimes written as C, and $Y_0(x)$ is sometimes written as $N_0(x)$. The graphs of $J_0(x)$ and $Y_0(x)$ are shown in Fig. 2. Important features are that $J_0(x)$ and $Y_0(x)$ look a bit like damped cosine and sine functions, except that $Y_0(x)$ tends to $-\infty$ as $x \to 0$. Specifically, we see from (16a) and (22) that

$$J_0(x) \sim 1, \qquad Y_0(x) \sim \frac{2}{\pi}\ln x \tag{23a,b}$$

as $x \to 0$, and it can be shown (Exercise 6) that

$$J_0(x) \sim \sqrt{\frac{2}{\pi x}}\cos\left(x - \frac{\pi}{4}\right), \qquad Y_0(x) \sim \sqrt{\frac{2}{\pi x}}\sin\left(x - \frac{\pi}{4}\right) \tag{24a,b}$$

as $x \to \infty$. Indeed, we can see from (24) why the Weber Bessel function Y_0 is a nicer companion for J_0 than the Neumann Bessel function $\mathbf{Y}_0$, for

$$\mathbf{Y}_0(x) \sim \frac{1}{\sqrt{\pi x}}\left[\left(\frac{\pi}{2} - \gamma + \ln x\right)\sin x - \left(\frac{\pi}{2} + \gamma - \ln x\right)\cos x\right] \tag{24c}$$

as $x \to \infty$; surely (24b) makes a nicer companion for (24a) than does (24c).

It might appear, from Fig. 2 and (24), that the *zeros* of J_0 and Y_0 [i.e., the roots of $J_0(x) = 0$ and $Y_0(x) = 0$] are equally spaced, but they are not; they approach an equal spacing only as $x \to \infty$. For instance, the first several zeros of J_0 are 2.405, 5.520, 8.654, 11.792, 14.931. Their differences are 3.115, 3.134, 3.138, 3.139, and these are seen to rapidly approach a constant [namely, π, the spacing between the zeros of $\cos(x - \pi/4)$ in (22a)]. The zeros of the various Bessel functions turn out to be important, and they are tabulated to many significant figures.

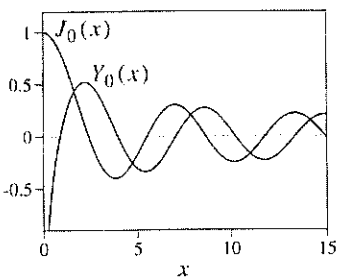


Figure 2. J_0 and Y_0.

For $n = 1, 2, \ldots$ the indicial roots $r = \pm n$ differ by an integer, which corresponds to case (iii) of Theorem 4.3.1. Using that theorem, and the ideas given above for Y_0 we obtain Weber's **Bessel function of the second kind, of order** n,

$$\boxed{\begin{aligned} Y_n(x) = \frac{2}{\pi}\Bigg[&\left(\ln\frac{x}{2}+\gamma\right)J_n(x) - \frac{1}{2}\sum_{k=0}^{n-1}\frac{(n-k-1)!}{k!}\left(\frac{x}{2}\right)^{2k-n} \\ &-\frac{1}{2}\sum_{k=0}^{\infty}\frac{(-1)^k\left[\phi(k)+\phi(k+n)\right]}{k!\,(k+n)!}\left(\frac{x}{2}\right)^{2k+n}\Bigg], \end{aligned}} \tag{25}$$

which formula holds for $n = 0, 1, 2, \ldots$; $\phi(0) = 0$ and $\phi(k) = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{k}$ for $k \geq 1$.

4.6.3. General solution of Bessel equation. Thus, we have two different general solution forms for (1), depending on whether ν is an integer or not: $y(x) = AJ_\nu(x) + BJ_{-\nu}(x)$ if ν is not an integer, and $y(x) = AJ_n(x) + BY_n(x)$ if $\nu = n = 0, 1, 2, \ldots$. It turns out that if we define

$$\boxed{Y_\nu(x) \equiv \frac{(\cos\nu\pi)\,J_\nu(x) - J_{-\nu}(x)}{\sin\nu\pi}} \tag{26}$$

for noninteger ν, then the limit of $Y_\nu(x)$ as $\nu \to n$ $(n = 0, 1, 2, \ldots)$ gives the same result as (25). Furthermore, $J_\nu(x)$ and $Y_\nu(x)$ are LI (Exercise 11) so the upshot is that we can express the general solution of (1) as

$$\boxed{y(x) = AJ_\nu(x) + BY_\nu(x)} \tag{27}$$

for *all* values of ν, with Y_ν defined by (25) and (26) for integer and noninteger values of ν, respectively. The graphs of several J_n's and Y_n's are shown in Fig. 3.

For reference, we cite the following asymptotic behavior:

$$J_n(x) \sim \frac{1}{2^n n!}x^n, \qquad (n = 0, 1, 2, \ldots) \tag{28a}$$

$$Y_n(x) \sim \begin{cases} \dfrac{2}{\pi}\ln x, & (n = 0) \\ -\dfrac{2^n(n-1)!}{\pi}\dfrac{1}{x^n}, & (n = 1, 2, \ldots) \end{cases} \tag{28b}$$

as $x \to 0$, and

$$J_n(x) \sim \sqrt{\frac{2}{\pi x}}\cos(x - \psi_n), \qquad (n = 0, 1, 2, \ldots) \tag{29a}$$

$$Y_n(x) \sim \sqrt{\frac{2}{\pi x}}\sin(x - \psi_n), \qquad (n = 0, 1, 2, \ldots) \tag{29b}$$

as $x \to \infty$, where $\psi_n = (2n+1)\pi/4$. Observe the sort of conservation expressed in (28a,b): as n increases, the Y_n's develop stronger singularities ($\ln x$, x^{-1}, x^{-2}, ...)

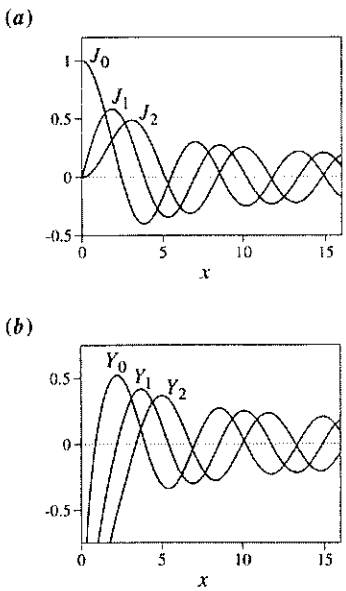


Figure 3. J_n's and Y_n's.

while the J_n's develop stronger zeros $(1, x, x^2, \ldots)$. (We say that x^5 has a stronger zero at the origin than x^3, for example, because $x^5/x^3 \to 0$ as $x \to 0$.) Finally, we call attention to the interlacing of the zeros of J_n and Y_n. All of these features can be seen in Fig. 3.

In summary, our key result is the general solution (27) with J_ν given by (9), Y_ν by (25) for integer ν and by (26) for noninteger ν, and with $J_{-\nu}$ in (26) given by (10).

4.6.4. Hankel functions. (Optional) Recall that the harmonic oscillator equation $y'' + y = 0$ has two preferred bases: $\cos x$, $\sin x$ and e^{ix}, e^{-ix}. Usually, the former is used because those functions are real valued, but sometimes the complex exponentials are more convenient. The connection between them is given by Euler's formulas:

$$e^{ix} = \cos x + i \sin x,$$
$$e^{-ix} = \cos x - i \sin x.$$

Analogous to the complex basis e^{ix}, e^{-ix} for the equation $y'' + y = 0$, a complex-valued basis is defined for the Bessel equation (1):

$$H_\nu^{(1)}(x) \equiv J_\nu(x) + i\, Y_\nu(x) \tag{30a}$$
$$H_\nu^{(2)}(x) \equiv J_\nu(x) - i\, Y_\nu(x). \tag{30b}$$

These are called the **Hankel functions** of the first and second kind, respectively, of order ν. Thus, alternatively to (27), we have the general solution

$$\boxed{y(x) = AH_\nu^{(1)}(x) + BH_\nu^{(2)}(x)} \tag{31}$$

of (1).

As a result of (29a,b), the Hankel functions $H_n^{(1)}(x)$, $H_n^{(2)}(x)$ have the pure complex exponential behavior

$$H_n^{(1)}(x) \sim \sqrt{\frac{2}{\pi x}}\, e^{i(x-\psi_n)}, \tag{32a}$$
$$H_n^{(2)}(x) \sim \sqrt{\frac{2}{\pi x}}\, e^{-i(x-\psi_n)} \tag{32b}$$

as $x \to \infty$.

The Hankel functions are particularly useful in the study of wave propagation.

4.6.5. Modified Bessel equation. Besides the Bessel equation of order ν, one also encounters the **modified Bessel equation of order** ν, $x^2y'' + xy' + \left(-x^2 - \nu^2\right) y = 0$, where the only difference is the minus sign in front of the second x^2 term. Let us limit our attention, for brevity, to the case where ν is an integer n, so we have

$$x^2y'' + xy' + \left(-x^2 - n^2\right) y = 0. \tag{33}$$

The change of variables $t = ix$ (or $x = -it$) converts (33) to the Bessel equation

$$t^2Y'' + tY' + \left(t^2 - n^2\right) Y = 0, \tag{34}$$

where $y(x) = y(-it) \equiv Y(t)$ and the primes on Y denote d/dt. Since a general solution to (34) is $Y(t) = AJ_n(t) + BY_n(t)$ we have, immediately, the general solution

$$y(x) = AJ_n(ix) + BY_n(ix) \tag{35}$$

of (33). From (15),

$$J_n(ix) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!\,(k+n)!} \left(\frac{ix}{2}\right)^{2k+n} = i^n \sum_{k=0}^{\infty} \frac{1}{k!\,(k+n)!} \left(\frac{x}{2}\right)^{2k+n} \tag{36}$$

so we can absorb the i^n into A and be left with the real-valued solution

$$\boxed{I_n(x) \equiv i^{-n} J_n(ix) = \sum_{k=0}^{\infty} \frac{1}{k!\,(k+n)!} \left(\frac{x}{2}\right)^{2k+n},} \tag{37}$$

known as the **modified Bessel function of the first kind, and order** n. In place of $Y_n(ix)$ it is standard to introduce, as a second real-valued solution, the **modified Bessel function of the second kind, and order** n,

$$K_n(x) = \frac{\pi}{2} i^{n+1} \left[J_n(ix) + i\,Y_n(ix)\right]. \tag{38}$$

For instance,

$$I_0(x) = 1 + \frac{x^2}{2^2} + \frac{x^4}{2^4(2!)^2} + \frac{x^6}{2^6(3!)^2} + \cdots \tag{39a}$$

$$K_0(x) = -\left(\ln\frac{x}{2} + \gamma\right) I_0(x) + (1)\frac{x^2}{2^2(1!)^2} + \left(1 + \frac{1}{2}\right)\frac{x^4}{2^4(2!)^2} + \left(1 + \frac{1}{2} + \frac{1}{3}\right)\frac{x^6}{2^6(3!)^2} + \cdots, \tag{39b}$$

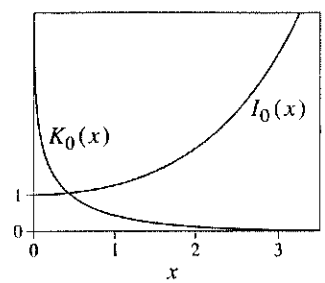


Figure 4. I_0 and K_0.

and the graphs of these functions are plotted in Fig. 4.

As a general solution of (33) we have

$$\boxed{y(x) = AI_n(x) + BK_n(x).} \tag{40}$$

Whereas the Bessel functions are oscillatory, the modified Bessel functions are not.

To put the various Bessel functions in perspective, observe that the relationship between the J, Y solutions of the Bessel equation and the I, K solutions of the modified Bessel equation is similar to that between the solutions $\cos x$, $\sin x$ of the harmonic oscillator equation $y'' + y = 0$ and the $\cosh x$, $\sinh x$ solutions of the "modified harmonic oscillator" equation $y'' - y = 0$. For instance, just as

$\cos(ix) = \cosh(x)$ and $\sin(ix) = i\sinh(x)$, $I_n(x)$ and $K_n(x)$ are linear combinations of $J_n(ix)$ and $Y_n(ix)$.

Finally, the asymptotic behavior of I_n and K_n is as follows:

$$I_n(x) \sim \frac{1}{n!}\left(\frac{x}{2}\right)^n, \qquad (n = 0, 1, 2, \ldots) \tag{41a}$$

$$K_n(x) \sim \begin{cases} -\ln x & (n = 0) \\ \frac{(n-1)!}{2}\left(\frac{2}{x}\right)^n, & (n = 1, 2, \ldots) \end{cases} \tag{41b}$$

as $x \to 0$, and

$$I_n(x) \sim \frac{1}{\sqrt{2\pi x}}e^x, \qquad (n = 0, 1, 2, \ldots) \tag{42a}$$

$$K_n(x) \sim \sqrt{\frac{\pi}{2x}}e^{-x}, \qquad (n = 0, 1, 2, \ldots) \tag{42b}$$

as $x \to \infty$. As n increases, the I_n's develop stronger zeros at $x = 0$ $(1, x, x^2, \ldots$ while the K_n's develop stronger singularities there $(\ln x, x^{-1}, x^{-2}, \ldots)$.

4.6.6. Equations reducible to Bessel equations. The results discussed in Sections 4.6.1–4.6.5 are all the more important because there are many equations which, although they are not Bessel equations or modified Bessel equations, can be reduced to Bessel or modified Bessel equations by changes of variables and then solved in closed form in terms of Bessel or modified Bessel functions.

EXAMPLE 1. Solve

$$xy'' + y' + \kappa^2 xy = 0. \tag{43}$$

Equation (43) is not quite a Bessel equation of order zero because of the κ^2. Let us try to absorb the κ^2 by a change of variable. Specifically, scale x as $t = \alpha x$, where α is to be determined. Then $\frac{d}{dx} = \left(\frac{d}{dt}\right)\left(\frac{dt}{dx}\right) = \alpha\frac{d}{dt}$, so $\frac{d^2}{dx^2} = \alpha^2\frac{d^2}{dt^2}$ and (43) becomes, after division by α,

$$tY'' + Y' + \frac{\kappa^2}{\alpha^2}tY = 0, \tag{44}$$

where $y(x) = y(t/\alpha) \equiv Y(t)$. Thus, we can absorb the κ^2 in (44) by choosing $\alpha = \kappa$. Then (44) is a Bessel equation of order zero with general solution $Y(t) = AJ_0(t) + BY_0(t)$ so

$$y(x) = AJ_0(t) + BY_0(t) = AJ_0(\kappa x) + BY_0(\kappa x) \tag{45}$$

is a general solution of (43). ∎

More generally, the equation

$$\boxed{\frac{d}{dx}\left(x^a\frac{dy}{dx}\right) + bx^c y = 0,} \tag{46}$$

where a, b, c are real numbers, can be transformed to a Bessel equation by transforming both independent and dependent variables. Because of the powers of x in (46), it seems promising to change variables from $x, y(x)$ to $t, u(t)$ according to the forms $t = Ax^B$, $u = x^C y$, and to try to find A, B, C so that the new equation, on $u(t)$, is a Bessel equation. It turns out that that plan works and one finds that under the change of variables

$$t = \alpha\sqrt{b}\,x^{1/\alpha} \qquad \text{and} \qquad u = x^{-\nu/\alpha}y \tag{47}$$

equation (46) becomes the Bessel equation of order ν,

$$t^2\frac{d^2u}{dt^2} + t\frac{du}{dt} + \left(t^2 - \nu^2\right)u = 0, \tag{48}$$

if we choose

$$\alpha = \frac{2}{c-a+2} \qquad \text{and} \qquad \nu = \frac{1-a}{c-a+2}. \tag{49}$$

[The latter is meaningless if $c-a+2=0$, but in that case (46) is merely a Cauchy-Euler equation.] Thus, if Z_ν denotes any Bessel function solution of (48), then putting (47) into $u(t) = Z_\nu$ and solving for y gives the solution

$$\boxed{y(x) = x^{\nu/\alpha}Z_\nu\left(\alpha\sqrt{|b|}\,x^{1/\alpha}\right)} \tag{50}$$

of (46). If $b > 0$, then Z_ν denotes J_ν and Y_ν, and if $b < 0$, then Z_ν denotes I_ν and K_ν (though we gave formulas for I_ν and K_ν only for the case where ν is an integer).

EXAMPLE 2. Solve the equation

$$y'' + 3\sqrt{x}\,y = 0. \qquad (0 < x < \infty) \tag{51}$$

Comparing (51) with (46), we see that $a = 0$, $b = 3$, and $c = 1/2$, so $\alpha = 2/(1/2-0+2) = 4/5$ and $\nu = 1/(1/2-0+2) = 2/5$. Thus, (50) gives

$$y(x) = x^{1/2}Z_{2/5}\left(\frac{4}{5}\sqrt{3}\,x^{5/4}\right), \tag{52}$$

and

$$y(x) = \sqrt{x}\left[AJ_{2/5}\left(\frac{4}{5}\sqrt{3}\,x^{5/4}\right) + BY_{2/5}\left(\frac{4}{5}\sqrt{3}\,x^{5/4}\right)\right] \tag{53}$$

is a general solution of (51). ∎

EXAMPLE 3. Solve

$$xy'' + 3y' + y = 0, \qquad (0 < x < \infty) \tag{54}$$

or

$$y'' + \frac{3}{x}y' + \frac{1}{x}y = 0. \tag{55}$$

Writing out (46) as $x^a y'' + ax^{a-1}y' + bx^c y = 0$ or

$$y'' + \frac{a}{x}y' + bx^{c-a}y = 0, \tag{56}$$

and comparing (56) and (55) term by term gives $a = 3$, $b = 1$, and $c - a = -1$, so $c = 2$. Hence, $\alpha = 2/(2-3+2) = 2$ and $\nu = (1-3)/(2-3+2) = -2$, so (50) becomes

$$y(x) = x^{-2/2} Z_{-2}\left(2\sqrt{1}\, x^{1/2}\right) = x^{-1} Z_2\left(2\sqrt{x}\right) \tag{57}$$

and

$$y(x) = \frac{1}{x}\left[AJ_2\left(2\sqrt{x}\right) + BY_2\left(2\sqrt{x}\right)\right] \tag{58}$$

is a general solution of (54). ∎

NOTE: In the second equality of (57) we changed the Z_{-2} to Z_2. More generally, if the ν that we compute in (49) turns out to be negative we can always change the Z_ν in (50) to $Z_{|\nu|}$, for if ν is a negative integer $-n$, then the Z_ν in (50) gives J_{-n} and Y_{-n}; but (18) told us that J_{-n} is identical to J_n, to within a constant scale factor, and it can likewise be shown that Y_{-n} is identical to Y_n, to within a constant scale factor [namely, $Y_{-n}(x) = (-1)^n Y_n(x)$]. And if ν is negative but not an integer, then the Z_ν in (50) gives J_ν and $J_{-\nu}$, and that is equivalent to $Z_{-\nu}$ giving $J_{-\nu}$ and J_ν.

EXAMPLE 4. Solve

$$(xy')' - 5x^3 y = 0. \tag{59}$$

We see from (46) that $a = 1$, $b = -5$, $c = 3$, so $\alpha = 1/2$, $\nu = 0$ and

$$y(x) = x^0 Z_0\left(\frac{1}{2}\sqrt{|-5|}\, x^2\right),$$

so

$$y(x) = AI_0\left(\frac{\sqrt{5}}{2}x^2\right) + BK_0\left(\frac{\sqrt{5}}{2}x^2\right) \tag{60}$$

is a general solution of (59). ∎

Closure. In this section we studied the Bessel equation

$$x^2 y'' + xy' + \left(x^2 - \nu^2\right) y = 0 \tag{61}$$

and the modified Bessel equation

$$x^2 y'' + xy' + \left(-x^2 - \nu^2\right) y = 0. \tag{62}$$

For heuristic purposes, it is useful to keep in mind the similarity between the Bessel equation and the harmonic oscillator equation

$$y'' + y = 0, \tag{63}$$

and between the modified Bessel equation and the "modified harmonic oscillator" equation

$$y'' - y = 0. \tag{64}$$

For large x, the left side of (61) becomes

$$y'' + \frac{1}{x}y' + \left(1 - \frac{\nu^2}{x^2}\right) y \approx y'' + y \tag{65}$$

so we expect qualitative similarity between the solutions of (61) and those of (63). In fact, the solutions $J_\nu(x)$ and $Y_\nu(x)$ of (61) do tend to harmonic functions as $x \to \infty$, like the cosine and sine solutions of (63), and the y'/x term in (65) causes some damping of those harmonic functions, by a factor of $1/\sqrt{x}$. Thus, the general solution

$$y(x) = AJ_\nu(x) + BY_\nu(x) \tag{66}$$

of (61) is similar, qualitatively, to that of (63). Further, just as one can use pure complex exponential solutions of (63) according to the Euler definitions, one can introduce the Hankel functions in essentially the same way, and write the general solution of (61), alternatively, as

$$y(x) = AH_\nu^{(1)}(x) + BH_\nu^{(1)}(x). \tag{67}$$

Likewise, for the modified Bessel equation (62), the left side of which becomes

$$y'' + \frac{1}{x}y' + \left(1 - \frac{\nu^2}{x^2}\right) y \approx y'' + y \tag{68}$$

for large x, we find nonoscillatory solutions analogous to the hyperbolic cosine and sine solutions of (64).

So much for large x. As $x \to 0$, the Y_ν solutions of (61) are unbounded as are the K_ν solutions of (62).

Computer software. As a general rule of thumb, if we can derive a solution to a given differential equation by hand, we can probably obtain it using computer software. For instance, if, using *Maple*, we attempt to solve the nonconstant-coefficient differential equation (54) by the command

$$\text{dsolve}(x * \text{diff}(y(x), x, x) + 3 * \text{diff}(y(x), x) + y(x) = 0,\ y(x));$$

we do obtain the same general solution as was obtained here in Example 3.

EXERCISES 4.6

1. Putting the solution form (2) into the Bessel equation (1), derive the recursion relation (3).

2. Solve (1) for the case where $\nu = 1/2$, by the method of Frobenius. Show that your two LI solutions can be expressed in closed form as given in (14a,b).

3. Show that with $Y_0(x)$ defined by (21), the asymptotic behavior given in (24b) follows from (24a) and (24c).

4. (*Recursion formulas*) It can be shown that

$$\frac{d}{dx}\left[x^{\nu} Z_{\nu}(x)\right] = \begin{cases} x^{\nu} Z_{\nu-1}(x), & \left(Z = J, Y, I, H^{(1)}, H^{(2)}\right) \\ -x^{\nu} Z_{\nu-1}(x), & (Z = K) \end{cases} \tag{4.1}$$

$$\frac{d}{dx}\left[x^{-\nu} Z_{\nu}(x)\right] = \begin{cases} -x^{-\nu} Z_{\nu+1}(x), & \left(Z = J, Y, K, H^{(1)}, H^{(2)}\right) \\ x^{-\nu} Z_{\nu+1}(x), & (Z = I) \end{cases} \tag{4.2}$$

where the "$Z = \cdots$" means that the formula holds with Z equal to each of the itemized functions. In particular, the formula

$$\frac{d}{dx} Z_0(x) = \begin{cases} -Z_1(x), & \left(Z = J, Y, K, H^{(1)}, H^{(2)}\right) \\ Z_1(x), & (Z = I) \end{cases} \tag{4.3}$$

corresponding to (4.2) for the case $\nu = 0$, is useful in evaluating certain Bessel function integrals by integration by parts.

(a) Verify (4.1) and (4.2) for the case where Z is J (i.e., J_ν, $J_{-\nu}$ and J_n) using the series given in the text for those functions.

(b) From the formulas given above, show that

$$Z_{\nu+1}(x) = \frac{2\nu}{x} Z_{\nu}(x) - Z_{\nu-1}(x), \quad \left(Z = J, Y, H^{(1)}, H^{(2)}\right) \tag{4.4}$$

and

$$I_{\nu+1}(x) = -\frac{2\nu}{x} I_{\nu}(x) - I_{\nu-1}(x), \tag{4.5}$$

$$K_{\nu+1}(x) = \frac{2\nu}{x} K_{\nu}(x) + K_{\nu-1}(x). \tag{4.6}$$

(c) Use computer software to differentiate $x^3 J_3(x)$, $x^2 Y_2(x)$, $x^5 I_5(x)$, $x^{-2} K_2(x)$, $J_0(x)$, $Y_0(x)$, $I_0(x)$, and $K_0(x)$, and show that the results are in accord with the formulas (4.1)–(4.3).

5. (*Half-integer formulas*) (a) Putting $\nu = 1/2$ in (9) and (10), show that they give

$$J_{1/2}(x) = \sqrt{\frac{2}{\pi x}} \sin x \tag{5.1}$$

and

$$J_{-1/2}(x) = \sqrt{\frac{2}{\pi x}} \cos x. \tag{5.2}$$

(b) Derive, from (4.4), the recursion formula

$$J_{n+1/2}(x) = \frac{2n-1}{x} J_{n-1/2}(x) - J_{n-3/2}(x). \tag{5.3}$$

(c) It follows from (5.1)–(5.3) that *all* $J_{n+1/2}$'s are expressible in closed form in terms of $\sin x$, $\cos x$, and powers of x. Derive those expressions for $J_{3/2}(x)$ and $J_{-3/2}(x)$.

6. (a) (*Normal form*) By making the change of variables $y(x) = \sigma(x) v(x)$, from y to v, in $y'' + p(x) y' + q(x) y = 0$, show that the first derivative term can be eliminated by choosing σ such that $2\sigma' + p\sigma = 0$. Show that the result is the *normal form* (i.e., canonical or simplest form)

$$v'' + \left(\frac{\sigma'' + p\sigma' + q\sigma}{\sigma}\right) v = 0, \tag{6.1}$$

where

$$\sigma(x) = e^{-\frac{1}{2}\int^x p(t)\,dt}. \tag{6.2}$$

(b) (*Large-x behavior of Bessel functions*) For the Bessel equation (1), show that $\sigma(x) = 1/\sqrt{x}$ and that (6.1) is

$$v'' + \left(1 - \frac{\nu^2 - 1/4}{x^2}\right) v = 0. \tag{6.3}$$

NOTE: If we write $1 - (\nu^2 - 1/4)/x^2 \approx 1$ for large x, then (6.3) becomes $v'' + v \approx 0$, so we expect that $v(x) \approx A\cos(x + \phi)$ or, equivalently, $A\sin(x + \phi)$, where

A and ϕ are arbitrary constants. Thus,

$$y(x) = \sigma(x)v(x) \approx \frac{A}{\sqrt{x}}\cos(x+\phi) \quad \text{or} \quad \frac{A}{\sqrt{x}}\sin(x+\phi), \tag{6.4}$$

which forms are the same as those given by (24a) and (24b). Thus, we expect every solution of (1) to behave according to (6.4) as $x \to \infty$. Evaluating the constants A and ϕ corresponding to a particular solution, such as $J_0(x)$ or $Y_0(x)$, is complicated and will not be discussed here.

7. Recall from Example 1 that $J_n(\kappa x)$ satisfies the differential equation $x^2y'' + xy' + (\kappa^2x^2 - n^2)y = 0$ or, equivalently,

$$(xy')' + \left(\kappa^2 x - \frac{n^2}{x}\right)y = 0. \quad (n = 0, 1, 2, \ldots) \tag{7.1}$$

Let the x interval be $0 < x < c$, and suppose that κ is chosen so that $J_n(\kappa c) = 0$; i.e., κc is any of the zeros of $J_n(x) = 0$. The purpose of this exercise is to derive the formula

$$\boxed{\int_0^c [J_n(\kappa x)]^2\, x\, dx = \frac{c^2}{2}[J_{n+1}(\kappa c)]^2,} \tag{7.2}$$

which will be be needed when we show how to use the Sturm-Liouville theory to expand functions on $0 < x < c$ in terms of Bessel functions. In turn, that concept will be needed later in our study of partial differential equations. To derive (7.2), we suggest the following steps.

(a) Multiplying (7.1) by $2xy'$ and integrating on x from 0 to c, obtain

$$(xy')^2\big|_0^c + 2\int_{x=0}^{x=c}(\kappa^2x^2 - n^2)\, y\, dy = 0. \tag{7.3}$$

(b) Show that with $y = J_n(\kappa x)$, the $(xy')^2$ term is zero at $x = 0$ for $n = 0, 1, 2, \ldots$, and that at $x = c$ it is $c^2\kappa^2[J_{n+1}(\kappa c)]^2$. HINT: It follows from (4.2) that $J_n'(x) = -J_{n+1}(x) + \frac{n}{x}J_n(x)$.
(c) Thus, show that (7.3) reduces to

$$c^2\kappa^2[J_{n+1}(\kappa c)]^2 + 2\kappa^2\int_{x=0}^{x=c} x^2y\, dy - n^2y^2\big|_{x=0}^{x=c} = 0. \tag{7.4}$$

(d) Show that the $n^2y^2\big|_{x=0}^{x=c}$ term is zero for any $n = 0, 1, 2, \ldots$, integrate the remaining integral by parts and show that the resulting boundary term is zero, and thus obtain the desired result (7.2).

8. (*Generating function for J_n*) Just as there is a "generating function" for the Legendre polynomials [see (9) in Section 4.4], there is one for the Bessel functions. Specifically, it can be shown that

$$\boxed{e^{\frac{x}{2}\left(t - \frac{1}{t}\right)} = \sum_{n=-\infty}^{\infty} J_n(x)t^n,} \tag{8.1}$$

and the left-hand side is called the **generating function** for the J_n's.

(a) We do not ask you to derive (8.1) but only to verify the $n = 0$ term. That is, expanding $e^{xt/2}$ and $e^{-x/2t}$ in Maclaurin series (one in ascending powers of t and one in powers of $1/t$) and multiplying these series together, show that the coefficient of t^0 is $J_0(x)$.
(b) Equation (8.1) is useful for deriving various properties of the J_n's. For example, taking $\partial/\partial x$ of both sides, show that

$$\frac{d}{dx}J_n(x) = \frac{1}{2}[J_{n-1}(x) - J_{n+1}(x)]. \quad (n = 1, 2, \ldots) \tag{8.2}$$

(c) Similarly, taking $\partial/\partial t$ of both sides, show that

$$J_{n+1}(x) = \frac{x}{2(n+1)}[J_n(x) + J_{n+2}(x)]. \tag{8.3}$$

(d) Using computer software, differentiate $J_0(x)$ and $J_1(x)$ and show that the results agree with (8.2).

9. (*Integral representation of J_n*) Besides the generating function (preceding exercise), another source of information about the J_n's is the integral representation

$$\boxed{J_n(x) = \frac{1}{\pi}\int_0^{\pi}\cos(n\theta - x\sin\theta)d\theta.} \tag{9.1}$$

Verify (9.1) for the case $n = 0$ by using the Taylor series of $\cos t$, where $t = n\theta - x\sin\theta$ and integrating term by term. HINT: You may use any of the formulas given in the exercises to Section 4.5.

10. To derive (17) from (10), we argued that $1/\Gamma(k-n+1) = 0$ for $k = 0, 1, \ldots, n-1$, on the grounds that for those k's $\Gamma(k-n+1)$ is infinite. However, while it is true that the gamma function becomes infinite as its argument approaches $0, -1, -2, \ldots$, it is not rigorous to say that it is infinite *at* those points; it is simply undefined there. Here, we ask you to verify that the $k = 0, 1, \ldots, n-1$ terms are zero so that the correct lower limit in (17) is, indeed, $k = n$. For definiteness, let $\nu = 3$ and $r = -\nu = -3$. (You should then be able to generalize the result for the case of any positive integer ν, but we do not ask you to do that; $\nu = 3$ will suffice.) HINT: Rather than work from (17), go back to the formulas (4a,b,c).

11. It was stated, below (26), that J_ν and Y_ν are LI. Prove that claim. HINT: Use (25) for $\nu = n$ and (26) for $\nu \neq n$.

12. Each differential equation is given on $0 < x < \infty$. Use (50) to obtain a general solution. Gamma functions that appear need not be evaluated.

(a) $y'' + 4x^2 y = 0$ (b) $xy'' - 2y' - x^2 y = 0$
(c) $xy'' - 2y' + xy = 0$ (d) $4y'' + 9xy = 0$
(e) $y'' + \sqrt[3]{x}\, y = 0$ (f) $y'' - xy = 0$
(g) $xy'' + 3y' - xy = 0$ (h) $4xy'' + y = 0$
(i) $4xy'' + 2y' + xy = 0$ (j) $xy'' + 3y' - 4xy = 0$
(k) $xy'' + y' - 9x^2 y = 0$ (l) $y'' + xy = 0$

13. (a)–(l) Solve the corresponding problems in Exercise 12, this time using computer software.

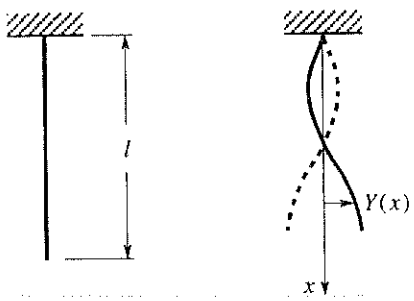
14. (a) Use (50) to find a general solution of

$$xy'' + 3y' + 9xy = 0. \quad (0 < x < \infty)$$

(b) Find a particular solution satisfying the boundary conditions $y(0) = 6,\ y'(0) = 0$.
(c) Show that there is *no* particular solution satisfying the initial conditions $y(0) = 6,\ y'(0) = 2$. Does that result contradict the result stated in Theorem 3.3.1? Explain.

15. Use (50) to solve $y'' + 4y = 0$, and show that your solution agrees with the known elementary solution. You may use any results given in these exercises.

16. (*Lateral vibration of hanging rope*) Consider a flexible rope or chain that hangs from the ceiling under the sole action of gravity (see the accompanying sketch). If we pull



the rope to one side and let go, it will oscillate from side to side in a complicated pattern which amounts to a superposition of many different modes, each having a specific shape $Y(x)$ and temporal frequency ω. It can be shown (from Newton's second law of motion) that each shape $Y(x)$ is governed by the differential equation

$$[\rho g(l - x)Y']' + \rho\omega^2 Y = 0, \quad (0 < x < l) \qquad (16.1)$$

where ρ is the mass per unit length and g is the acceleration of gravity.

(a) Derive the general solution

$$Y(x) = AJ_0\left(\frac{2\omega}{\sqrt{g}}\sqrt{l - x}\right) + BY_0\left(\frac{2\omega}{\sqrt{g}}\sqrt{l - x}\right) \qquad (16.2)$$

of (16.1). HINT: It may help to first make the change of variables $l - x = \xi$. NOTE: Observe from (16.2) that the displacement Y will be unbounded at the free end $x = l$ because of the logarithmic singularity in Y_0 when its argument is zero (namely, when $x = l$). Mathematically, that singularity can be traced to the vanishing of the coefficient $\rho g(l - x)$ in (16.1) at $x = l$, which vanishing introduces a regular singular point of (16.1) at $x = l$ and results in the logarithmic singularity in the solution (16.2). Physically, observe that the coefficient $\rho g(l - x)$ in (16.1) represents the tension in the rope. The greater the tension the smaller the displacement (as anyone who has strung a clothesline knows). Hence the vanishing of the tension $\rho g(l - x)$ at the free end leads to the mathematical possibility of unbounded displacements there. In posing suitable boundary conditions, it is appropriate to preclude such unbounded displacements there by prescribing the boundary condition that $Y(l)$ be bounded; that is, a "boundedness condition." Imposing that condition implies that $B = 0$, so that the solution (16.2) reduces to $Y(x) = AJ_0\left(\frac{2\omega}{\sqrt{g}}\sqrt{l - x}\right)$.

(b) As a second boundary condition, set $Y(0) = 0$. That condition does not lead to the evaluation of A (which remains arbitrary); rather, it permits us to find the allowable temporal frequencies ω. If the first three zeros of $J_0(x)$ are $x = 2.405$, 5.520, and 8.654, evaluate the first three frequencies ω (in terms of g and l) and the corresponding mode shapes $Y(x)$ (to within the arbitrary scale factor A). Sketch those mode shapes by hand over $0 < x < l$.
(c) Use computer software to obtain the zeros quoted above (2.405, 5.520, 8.654), and to obtain computer plots of the three mode shapes. (Set $A = 1$, say.)

Chapter 4 Review

In this chapter we present methods for the solution of second-order homogeneous differential equations with nonconstant coefficients.

The most important general results are Theorems 4.2.4 and 4.3.1, which guarantee specific forms of series solutions about ordinary and regular singular points, respectively. About an ordinary point one can find two LI power series solutions and hence the general solution. About a regular singular point, say $x = 0$, one can find two LI solutions, by the method of Frobenius, in terms of power series and power series modified by the multiplicative factors $|x|^r$ and/or $\ln|x|$, where r is found by solving a quadratic equation known as the *indicial equation*. The combination of these forms is dictated by whether the roots r are repeated or distinct and, if distinct, whether they differ by an integer or not. Note that the $|x|^r$ and $\ln|x|$ factors introduce singularities in the solutions (unless r is a nonnegative integer).

Besides these general results, we meet these special functions:

Exponential integral (Section 4.3): $E_1(x) = \displaystyle\int_x^\infty \frac{e^{-t}}{t}\,dt, \quad (x > 0)$

Gamma function (Section 4.5): $\Gamma(x) = \displaystyle\int_0^\infty t^{x-1}e^{-t}\,dt, \quad (x > 0)$

and study these important differential equations and find solutions for them:

Legendre equation (Section 4.4): $(1 - x^2)y'' - 2xy' + \lambda y = 0$

Solutions that are bounded on $-1 \le x \le 1$ exist only if $\lambda = n(n+1)$ for $n = 0, 1, 2, \ldots$, and these are the Legendre polynomials $P_n(x)$:

$$P_0(x) = 1, \quad P_1(x) = x, \quad P_2(x) = \frac{1}{2}(3x^2 - 1), \ldots.$$

Bessel equation (Section 4.6): $x^2y'' + xy' + \left(x^2 - \nu^2\right)y = 0$

$$\text{General solution:} \qquad y(x) = \begin{cases} AJ_\nu(x) + BY_\nu(x) \\ CH_\nu^{(1)}(x) + DH_\nu^{(2)}(x) \end{cases}$$

where J_ν, Y_ν are Bessel functions of the first and second kind, respectively, of order ν, and $H_\nu^{(1)}, H_\nu^{(2)}$ are the Hankel functions of the first and second kind, respectively, of order ν.

Modified Bessel equation (Section 4.6): $x^2y'' + xy' + \left(-x^2 - \nu^2\right)y = 0$
For brevity, we consider only the case where $\nu = n$ is an integer.

$$\text{General solution:} \qquad y(x) = AI_n(x) + BK_n(x),$$

where I_n, K_n are modified Bessel functions of the first and second kinds, respectively, of order n.

NOTE: We suggest that to place the many Bessel function results in perspective it is helpful to see the Bessel equation $x^2y'' + xy' + (x^2 - \nu^2)y = 0$ and the modified Bessel equation $x^2y'' + xy' + (-x^2 - \nu^2)y = 0$ as analogous to the harmonic oscillator equation $y'' + y = 0$ and the "modified harmonic oscillator equation" $y'' - y = 0$. For instance, the oscillatory $J_\nu(x)$ and $Y_\nu(x)$ solutions of the Bessel equation are analogous to the oscillatory $\cos x$ and $\sin x$ solutions of $y'' + y = 0$, and the complex Hankel function solutions $H_\nu^{(1)}(x)$ and $H_\nu^{(2)}(x)$ are analogous to the complex e^{ix} and e^{-ix} solutions. Similarly, the nonoscillatory $I_\nu(x)$ and $K_\nu(x)$ solutions of the modified Bessel equation are analogous to the nonoscillatory e^x and e^{-x} solutions of the equation $y'' - y = 0$.

Equations reducible to Bessel equations (Section 4.6): The equation

$$\frac{d}{dx}\left(x^a \frac{dy}{dx}\right) + bx^c y = 0,$$

where a, b, c are real numbers, has solutions

$$y(x) = x^{\nu/\alpha} Z_{|\nu|}\left(\alpha\sqrt{|b|}\, x^{1/\alpha}\right),$$

where

$$\alpha = \frac{2}{c - a + 2}, \qquad \nu = \frac{1 - a}{c - a + 2}.$$

$Z_{|\nu|}$ denotes $J_{|\nu|}$ and $Y_{|\nu|}$ if $b > 0$, and $I_{|\nu|}$ and $K_{|\nu|}$ if $b < 0$.

Chapter 5

Laplace Transform

5.1 Introduction

The Laplace transform is an example of an **integral transform**, namely, a relation of the form

$$F(s) = \int_a^b K(t,s)\, f(t)\, dt, \tag{1}$$

which transforms a given function $f(t)$ into another function $F(s)$; $K(t,s)$ is called the **kernel** of the transform, and $F(s)$ is known as the **transform** of $f(t)$. Thus, whereas a function sends one number into another [for example, the function $f(x) = x^2$ sends the point $x = 3$ on an x axis into the point $f = 9$ on an f axis], (1) sends one function into another, namely, it sends $f(t)$ into $F(s)$. Probably the most well known integral transform is the **Laplace transform**, where $a = 0$, $b = \infty$, and $K(t,s) = e^{-st}$. In that case (1) takes the form

$$\boxed{F(s) = \int_0^\infty f(t)\, e^{-st}\, dt.} \tag{2}$$

The parameter s can be complex, but we limit it to real values in this chapter. Besides the notation $F(s)$ used in (2), the Laplace transform of $f(t)$ is also denoted as $L\{f(t)\}$ or as $\overline{f}(s)$, and in a given application we will use whichever of these three notations seems best.

The basic idea behind any transform is that the given problem can be solved more readily in the "transform domain." To illustrate, consider the use of the natural logarithm in numerical calculation. While the addition of two numbers is arithmetically simple, their multiplication can be quite laborious; for example, try working out 2.761359×8.247504 by hand. Thus, given two positive numbers u and v, suppose we wish to compute their product $y = uv$. Taking the logarithm of both sides gives $\ln y = \ln uv$. But $\ln uv = \ln u + \ln v$, so we have $\ln y = \ln u + \ln v$. Thus, whereas the original problem was one of multiplication, the problem in the "transform domain" is merely one of addition. The idea, then, is to look up $\ln u$ and $\ln v$

in a table and to add these two values. With the sum in hand, we again enter the table, this time using it in the reverse direction to find the antilog, y. (Of course, with pocket calculators and computers available logarithm tables are no longer needed, as they were fifty years ago, but the transform nature of the logarithm remains the same, whether we use tables or not.)

Similarly, the logarithm reduces exponentiation to multiplication since if $y = u^v$, then $\ln y = \ln(u^v) = v \ln u$, and it reduces division to subraction.

Analogously, given a linear ordinary differential equation with constant coefficients, we see that if we take a Laplace transform of all terms in the equation then we obtain a linear algebraic equation on the transform $X(s)$ of the unknown function $x(t)$. That equation can be solved for $X(s)$ by simple algebra and the solution $x(t)$ obtained from a Laplace transform table. The method is especially attractive for nonhomogeneous differential equations with forcing functions which are step functions or impulse functions; we study those cases in Section 5.5.

Observe that we have departed from our earlier usage of x as the independent variable. Here we use t and consider the interval $0 \leq t < \infty$ because in most (though not all) applications of the Laplace transform the independent variable is the time t, with $0 \leq t < \infty$.

A brief outline of this chapter follows:

5.2 *Calculation of the Transform.* In this section we study the existence of the transform, and its calculation.

5.3 *Properties of the Transform.* Three properties of the Laplace transform are discussed: linearity of the transform and its inverse, the transform of derivatives, and the convolution theorem. These are crucial in the application of the method to the solution of ordinary differential equations, homogeneous or not.

5.4 *Application to the Solution of Differential Equations.* Here, we demonstrate the principal application of the Laplace transform, namely, to the solution of linear ordinary differential equations.

5.5 *Discontinuous Forcing Functions; Heaviside Step Function.* Discontinuous forcing functions are common in engineering and science. In this section we introduce the Heaviside step function and demonstrate its use.

5.6 *Impulsive Forcing Function; Dirac Impulse Function.* Likewise common are impulsive forcing functions such as the force imparted to a mass by a hammer blow. In this section we introduce the Dirac delta function to model such impulsive actions.

5.7 *Additional Properties.* There are numerous useful properties of the transform beyond the three discussed in Section 5.3. A number of these are given here, as a sequence of theorems.

5.2 Calculation of the Transform

The first question to address is whether the transform $F(s)$ of a given function $f(t)$ exists – that is, whether the integral

$$F(s) = \int_0^\infty f(t)\, e^{-st}\, dt \tag{1}$$

converges. Before giving an existence theorem, we define two terms.

First, we say that $f(t)$ is of **exponential order** as $t \to \infty$ if there exist real constants K, c, and T such that

$$|f(t)| \le K e^{ct} \tag{2}$$

for all $t \ge T$. That is, the set of functions of exponential order is the set of functions that do not grow faster than exponentially, which includes most functions of engineering interest.

EXAMPLE 1. Is $f(t) = \sin t$ of exponential order? Yes: $|\sin t| \le 1$ for all t, so (2) holds with $K = 1$, $c = 0$, and $T = 0$. Of course, these values are not uniquely chosen for (2) holds also with $K = 7$, $c = 12$, and $T = 100$, for instance. ∎

EXAMPLE 2. Is $f(t) = t^2$ of exponential order? l'Hôpital's rule gives

$$\lim_{t\to\infty} \frac{t^2}{e^{ct}} = \lim_{t\to\infty} \frac{2t}{c\, e^{ct}} = \lim_{t\to\infty} \frac{2}{c^2\, e^{ct}} = 0$$

if $c > 0$. Choose $c = 1$, say. Then, from the definition of limit, there must be a T such that $t^2/e^{ct} \le 0.06$, say, for all $t \ge T$. Thus, $|f(t)| = t^2 \le 0.06\, e^t$ for all $t \ge T$, hence $f(t)$ is of exponential order. ∎

On the other hand, the function $f(t) = e^{t^2}$ is not of exponential order because

$$\lim_{t\to\infty} \frac{e^{t^2}}{e^{ct}} = \lim_{t\to\infty} e^{t^2 - ct} = \infty, \tag{3}$$

no matter how large c is.

We say that $f(t)$ is **piecewise continuous** on $a \le t \le b$ if there exist a finite number of points $t_1, t_2, \ldots, t_N$ such that $f(t)$ is continuous on each open subinterval $a < t < t_1$, $t_1 < t < t_2, \ldots, t_N < t < b$, and has a finite limit as t approaches each endpoint from the interior of that subinterval. For instance, the function $f(t)$ shown in Fig. 1 is piecewise continuous on the interval $0 \le t \le 4$. The values of f *at* the endpoints $a, t_1, t_2, \ldots, b$ are not relevant to whether or not f is piecewise continuous; hence we have not even indicated those values in Fig. 1. For instance, the limit of f as t tends to 2 from the left exists and is 5, and the limit of f as t tends to 2 from the right exists and is 10, so the value of f at $t = 2$ does not matter. Thus, piecewise continuity allows for the presence of jump discontinuities.

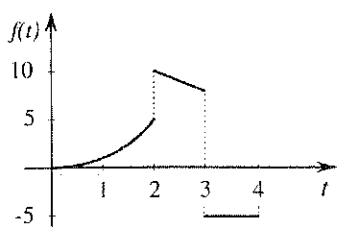


Figure 1. Piecewise continuity.

We can now provide a theorem that gives sufficient conditions on $f(t)$ for the existence of its Laplace transform $F(s)$.

THEOREM 5.2.1 *Existence of the Laplace Transform*
Let $f(t)$ satisfy these conditions: **(i)** $f(t)$ is piecewise continuous on $0 \le t \le A$, for every $A > 0$, and **(ii)** $f(t)$ is of exponential order as $t \to \infty$, so that there exist real constants K, c, and T such that $|f(t)| \le Ke^{ct}$ for all $t \geq T$. Then the Laplace transform of $f(t)$, namely, $F(s)$ given by (1) exists for all $s > c$.

Proof: We need to show only that the singular integral in (1) is convergent. Breaking it up as

$$\int_0^\infty f(t)\, e^{-st}\, dt = \int_0^T f(t)\, e^{-st}\, dt + \int_T^\infty f(t)\, e^{-st}\, dt, \tag{4}$$

the first integral on the right exists since the integrand is piecewise continuous on the finite interval $0 < t < T$. In the second integral, $|f(t)e^{-st}| = |f(t)|e^{-st} \le Ke^{-(s-c)t}$. Now, $\int_T^\infty Ke^{-(s-c)t}\, dt$ is convergent for $s > c$, so $\int_T^\infty f(t)\, e^{-st}\, dt$ is absolutely convergent – hence, by Theorem 4.5.3, convergent. ∎

Being thus assured by Theorem 5.2.1 that the transform $F(s)$ exists for a large and useful class of functions, we proceed to illustrate the evaluation of $F(s)$ for several elementary functions, say $f(t) = 1$, e^{at}, $\sin at$, where a is a real number, and $1/\sqrt{t}$.

EXAMPLE 3. If $f(t) = 1$, then the conditions of Theorem 5.2.1 are met for any $c \ge 0$, so according to Theorem 5.2.1, $F(s)$ should exist for all $s > 0$. Let us see.

$$F(s) = \int_0^\infty e^{-st}\, dt = \lim_{B\to\infty} \left.\frac{e^{-st}}{-s}\right|_0^B = \frac{1}{s}, \tag{5}$$

where the limit does indeed exist for all $s > 0$. Such restriction on s will cause no difficulty in applications. ∎

EXAMPLE 4. If $f(t) = e^{at}$, the conditions of Theorem 5.2.1 are met for any $c \ge a$ so according to the theorem, $F(s)$ should exist for all $s > a$. In fact,

$$F(s) = \int_0^\infty e^{at}\, e^{-st}\, dt = \lim_{B\to\infty} \left.\frac{e^{-(s-a)t}}{-(s-a)}\right|_0^B = \frac{1}{s-a}, \tag{6}$$

where the limit does indeed exist for all $s > a$. ∎

EXAMPLE 5. If $f(t) = \sin at$, then the conditions of Theorem 5.2.1 are met for any $c \ge 0$ so $F(s)$ should exist for all $s > 0$. In fact, integrating by parts twice gives

$$F(s) = \int_0^\infty \sin at\, e^{-st}\, dt$$

$$= \lim_{B\to\infty}\left[\left(\sin at\frac{e^{-st}}{-s} + \frac{a}{s}\cos at\frac{e^{-st}}{-s}\right)\Bigg|_0^B - \frac{a^2}{s^2}\int_0^B \sin at\, e^{-st}\,dt\right]$$

$$= (0-0) + \left(0 - \frac{a}{-s^2}\right) - \frac{a^2}{s^2}F(s), \tag{7}$$

where the limit exists if $s > 0$. The latter can be solved for $F(s)$, and gives

$$F(s) = \frac{a}{s^2+a^2} \qquad (s > 0) \tag{8}$$

as the transform of $\sin at$.

COMMENT. An alternative approach, which requires a knowledge of the algebra of complex numbers (Section 21.2), is as follows:

$$\int_0^\infty \sin at\, e^{-st}\,dt = \int_0^\infty \left(\operatorname{Im} e^{iat}\right) e^{-st}\,dt$$

$$= \operatorname{Im}\int_0^\infty e^{-(s-ia)t}\,dt = \operatorname{Im}\left[\lim_{B\to\infty}\frac{e^{-(s-ia)t}}{-(s-ia)}\Bigg|_0^B\right]$$

$$= \operatorname{Im}\frac{1}{s-ia} = \operatorname{Im}\frac{1}{s-ia}\frac{s+ia}{s+ia} = \frac{a}{s^2+a^2}, \tag{9}$$

as before, where the fourth equality follows because

$$\left|e^{-(s-ia)B}\right| = \left|e^{-sB}\right|\left|e^{iaB}\right| = e^{-sB} \to 0 \tag{10}$$

as $B \to \infty$, if $s > 0$. In (10) we have used the fact that $|e^{iaB}| = |\cos aB + i\sin aB| = \sqrt{\cos^2 aB + \sin^2 aB} = 1$. ∎

EXAMPLE 6. If $f(t) = 1/\sqrt{t}$, then

$$F(s) = \int_0^\infty t^{-1/2}e^{-st}\,dt = \int_0^\infty \sqrt{\frac{s}{\tau}}\,e^{-\tau}\,\frac{d\tau}{s} = \frac{1}{\sqrt{s}}\int_0^\infty \tau^{-1/2}e^{-\tau}\,d\tau, \tag{11}$$

where we have used the substitution $st = \tau$. Having studied the gamma function in Section 4.5, we see that the final integral is $\Gamma(1/2) = \sqrt{\pi}$ so

$$F(s) = \sqrt{\frac{\pi}{s}}, \tag{12}$$

for $s > 0$ (Exercise 4). Observe that $f(t) = 1/\sqrt{t}$ does not satisfy the conditions of Theorem 5.2.1 because it is not piecewise continuous on $0 \le t < \infty$ since it does not have a limit as $t \to 0$. Nonetheless, the singularity at $t = 0$ is not strong enough to cause divergence of the integral in (11), and hence the transform exists. Thus, remember that all we need is convergence of the integral in (1); the conditions in Theorem 5.2.1 are sufficient, not necessary. ∎

From these examples we could begin to construct a Laplace transform table, with $f(t)$ in one column and its transform $F(s)$ in another. Such a table is sup-

plied in Appendix C. More extensive ones are available,* and one can also obtain transforms and their inverses directly using computer software.

Tables can be used in either direction. For example, just as the transform of e^{at} is $1/(s-a)$, it is also true that the unique function whose transform is $1/(s-a)$ is e^{at}. Operationally, we say that

$$L\{e^{at}\} = \frac{1}{s-a} \quad \text{and} \quad L^{-1}\left\{\frac{1}{s-a}\right\} = e^{at}, \tag{13}$$

where L is the **Laplace transform operator** defined by

$$\boxed{L\{f(t)\} = \int_0^\infty f(t)\, e^{-st}\, dt,} \tag{14}$$

and L^{-1} is the **inverse Laplace transform operator**. It turns out that L^{-1} is, like L, an integral operator, namely

$$\boxed{L^{-1}\{F(s)\} = \frac{1}{2\pi i}\int_{\gamma-i\infty}^{\gamma+i\infty} F(s)\, e^{st}\, ds,} \tag{15}$$

where γ is a sufficiently positive real number. The latter is an integration in a complex s plane, and to carry out such integrations one needs to study the complex integral calculus. If, for instance, we would put $1/(s-a)$ into the integrand in (15), for $F(s)$ and carry out the integration, we would obtain e^{at}. We will return to (15) near the end of this text, when we study the complex integral calculus, but we will not use it in our present discussion; instead, we will rely on tables (and computer software) to obtain inverses. In fact, there are entire books on the Laplace transform that do not even contain the inversion formula (15). Our purpose in presenting it here is to show that the inverse operator is, like L, an integral operator, and to close the operational "loop:" $L\{f(t)\} = F(s)$, and then $L^{-1}\{F(s)\} = f(t)$.

What can we say about the existence and uniqueness of the inverse transform? Although we do not need to go into them here, there are conditions that $F(s)$ must satisfy if the inversion integral, in (15), is to exist, to converge. Thus, if one writes a function $F(s)$ at random, it may not have an inverse; there may be no function $f(t)$ whose transform is that particular $F(s)$. For instance, there is no function $f(t)$ whose transform is s^2 because its inversion integral is divergent. But suppose that we can, indeed, find an inverse for a given $F(s)$. Is that inverse necessarily unique; might there be more than one function $f(t)$ having the same transform? Strictly speaking, the answer is always yes. For instance, not only does the function $f(t) = 1$ have the transform $F(s) = 1/s$ (as found in Example 3), but so does the function

$$g(t) = \begin{cases} 1, & 0 \le t < 3, \quad 3 < t < \infty \\ 500, & t = 3 \end{cases}$$

*See, for example, A. Erdèlyi (ed.), *Tables of Integral Transforms*, Vol. 1 (New York: McGraw-Hill, 1954).

have the transform $G(s) = 1$ because the integrands in $\int_0^\infty g(t)\, e^{-st}\, dt$ and $\int_0^\infty f(t)\, e^{-st}\, dt$ differ only at the single point $t = 3$. Since there is no area under a single point (of finite height), $G(s)$ and $F(s)$ are identical: $G(s) = F(s) = 1/s$.

Clearly, one can construct an *infinite* number of functions, each having $1/s$ as its transform, but in a practical sense these functions differ only superficially. In fact, it is known from **Lerch's theorem*** that the inverse transform is unique to within an additive *null function*, a function $N(t)$ such that $\int_0^T N(t)\, dt = 0$ for every $T > 0$, so we can be content that the inverse transform is essentially unique.

Closure. Theorem 5.2.1 guarantees the existence of the Laplace transform of a given function $f(t)$, subject to the (sufficient but not necessary) conditions that f be piecewise continuous on $0 \le t < \infty$ and of exponential order as $t \to \infty$. We proceed to demonstrate the evaluation of the transforms of several simple functions, and discuss the building up of a transform table. Regarding the use of such a table, one needs to know whether the inverse transform found in the table is necessarily unique, and we use Lerch's theorem to show that for practical purposes we can, indeed, regard inverses as unique.

Computer software. On *Maple*, the **laplace** and **invlaplace** commands give transforms and inverses, respectively, provided that we enter readlib(laplace) first. To illustrate, the commands

```
readlib(laplace) :
laplace (1 + t^(-1/2) = x(t), t, s);
```

give the transform of $1 + t^{-1/2}$ as

$$\frac{1}{s} + \frac{\sqrt{\pi}}{\sqrt{s}} \tag{16}$$

and the command

```
invlaplace (a/(s^2 + a^2), s, t);
```

(17)

gives the inverse transform of $a/(s^2 + a^2)$ as $\sin at$.

EXERCISES 5.2

1. Show whether or not the given function is of exponential order. If it is, determine a suitable set of values for K, c, and T in (2).

(a) $5e^{4t}$ (b) $-10e^{-5t}$ (c) $\sinh 2t$
(d) $\cosh 3t$ (e) $\sinh t^2$ (f) $e^{4t} \sin t$
(g) $\cos t^3$ (h) t^{100} (i) $1/(t+2)$
(j) $\cosh t^4$ (k) $6t + e^t \cos t$ (l) t^{1000}

2. If $f(t)$ is of exponential order, does it follow that df/dt is too? HINT: Consider $f(t) = \sin e^{t^2}$.

*See, for example, D. V. Widder, *The Laplace Transform* (Princeton, NJ: Princeton University Press, 1941).

3. If $f(t)$ and $g(t)$ are each of exponential order, does it follow that $f(g(t))$ is too? HINT: Consider the case where $f(t) = e^t$ and $g(t) = t^2$.

4. In Example 6 we state that the result (12) holds if $s > 0$. Show why that condition is needed.

5. Does $L\{t^{-3/2}\}$ exist? Explain.

6. Does $L\{t^{-2/3}\}$ exist? Explain.

7. Derive $L\{\cos at\}$ two ways: using integration by parts and using the fact that $\cos at = \operatorname{Re} e^{iat}$. (See Example 5.)

8. Derive $L\{e^{at}\sin bt\}$ two ways: using integration by parts and using $\sin bt = \operatorname{Im} e^{ibt}$.

9. Derive $L\{e^{at}\cos bt\}$ two ways: using integration by parts and using $\cos bt = \operatorname{Re} e^{ibt}$.

10. Derive by integration the Laplace transform for each of the following entries in Appendix C:

(a) entry 5 (b) entry 6 (c) entry 7 (d) entry 8

11. Derive entry 11 in Appendix C two ways:

(a) by writing the transform as $\frac{1}{2i}\int_0^\infty t\,e^{-(s-ia)t}\,dt - \frac{1}{2i}\int_0^\infty t\,e^{-(s+ia)t}\,dt$ and using integration by parts or, if you prefer, by writing the transform as the single integral $\operatorname{Im}\int_0^\infty t\,e^{-(s-ia)t}\,dt$ and using integration by parts;

(b) by differentiating both sides of the known transform

$$\int_0^\infty \sin at\, e^{-st}\,dt = \frac{a}{s^2+a^2}$$

(which we derived in Example 5) with respect to s, assuming the validity of the interchange

$$\frac{d}{ds}\int_0^\infty \sin at\, e^{-st}\,dt = \int_0^\infty \frac{d}{ds}\left(\sin at\, e^{-st}\right)\,dt$$

in the order of integration and differentiation.

12. Use the idea in Exercise 11(b) to derive

(a) entry (12) from entry (4) (b) entry (7) from entry (1)
(c) entry (13) from entry (5) (d) entry (14) from entry (6)
(e) entry (15) from entry (2)

13. Show that $L\{e^{at}\} = 1/(s-a)$ holds even if $a = \operatorname{Re} a + i \operatorname{Im} a$ is complex, provided that $s > \operatorname{Re} a$.

14. Use computer software to verify the given entry in Appendix C in both directions. That is, show that the transform of $f(t)$ is $F(s)$, and also show that the inverse of $F(s)$ is $f(t)$.

(a) 1–3 (b) 4–7 (c) 8–10 (d) 11–13
(e) 14–16 (f) 17–19 (g) 20–22

5.3 Properties of the Transform

When we studied the integral calculus we might have evaluated a few simple integrals directly from the definition of the Riemann integral, but for the most part we learned how to evaluate integrals using a number of properties. For instance, we used linearity (whereby $\int_a^b [\alpha u(x) + \beta v(x)]\,dx = \alpha\int_a^b u(x)\,dx + \beta\int_a^b v(x)\,dx$ for any constants α, β, and any functions u, v, if the two integrals on the right exist), integration by parts, and the fundamental theorem of the calculus (which enabled us to generate a long list of integrals from an already known long list of derivatives). Our plan for the Laplace transform is not much different; we work out a handful of transforms by direct integration, and then rely on a variety of properties of the transform and its inverse to extend that list substantially. There are many such properties, but in this section we present only the handful that will be essential when we apply the Laplace transform method to the solution of differential equations, in the next section. Additional properties are discussed in the final section of this chapter.

We begin with the linearity property of the transform and its inverse.

THEOREM 5.3.1 *Linearity of the Transform*
If $u(t)$ and $v(t)$ are any two functions such that the transforms $L\{u(t)\}$ and $L\{v(t)\}$ both exist, then

$$\boxed{L\{\alpha u(t) + \beta v(t)\} = \alpha L\{u(t)\} + \beta L\{v(t)\}} \tag{1}$$

for any constants α, β.

Proof: We have

$$\begin{aligned}
L\{\alpha u(t) + \beta v(t)\} &= \int_0^\infty [\alpha u(t) + \beta v(t)]\, e^{-st}\, dt \\
&= \lim_{B\to\infty} \int_0^B [\alpha u(t) + \beta v(t)]\, e^{-st}\, dt \\
&= \lim_{B\to\infty} \left[\alpha \int_0^B u(t)\, e^{-st}\, dt + \beta \int_0^B v(t)\, e^{-st}\, dt\right] \\
&= \alpha \lim_{B\to\infty} \int_0^B u(t)\, e^{-st}\, dt + \beta \lim_{B\to\infty} \int_0^B v(t)\, e^{-st}\, dt \\
&= \alpha \int_0^\infty u(t)\, e^{-st}\, dt + \beta \int_0^\infty v(t)\, e^{-st}\, dt \\
&= \alpha L\{u(t)\} + \beta L\{v(t)\}\,, \qquad (2)
\end{aligned}$$

where the third equality follows from the linearity property of Riemann integration, and the fourth equality amounts to the result, from the calculus, that $\lim[\alpha f(B) + \beta g(B)] = \alpha \lim f(B) + \beta \lim g(B)$ as $B \to B_0$, if the latter two limits exist. ■

EXAMPLE 1. To evaluate the transform of $6 - 5e^{4t}$, for example, we need merely know the transforms of the simpler functions 1 and e^{4t} for $L\{6 - 5e^{4t}\} = 6L\{1\} - 5L\{e^{4t}\}$. Now, $L\{1\} = 1/s$ for $s > 0$, and $L\{e^{4t}\} = 1/(s-4)$ for $s > 4$ so

$$L\{6 - 5e^{4t}\} = 6\frac{1}{s} - 5\frac{1}{s-4} = \frac{s-24}{s(s-4)}$$

for $s > 4$. ■

THEOREM 5.3.2 *Linearity of the Inverse Transform*
For any $U(s)$ and $V(s)$ such that the inverse transforms $L^{-1}\{U(s)\} = u(t)$ and $L^{-1}\{V(s)\} = v(t)$ exist,

$$\boxed{L^{-1}\{\alpha U(s)+\beta V(s)\}=\alpha L^{-1}\{U(s)\}+\beta L^{-1}\{V(s)\}} \tag{3}$$

for any constants α, β.

Proof: Equation (3) follows either upon taking L^{-1} of both sides of (1) or from the linearity property of the integral in the inversion formula [equation (15) in Section 5.2]. ■

EXAMPLE 2. Asked to evaluate the inverse of $F(s) = 3/(s^2+3s-10)$, we turn to Appendix C but do not find this $F(s)$ in the column of transforms. However, we can simplify $F(s)$ by using partial fractions. Accordingly, we express

$$\begin{aligned}\frac{3}{s^2+3s-10} &= \frac{A}{s+5}+\frac{B}{s-2}\\ &= \frac{(A+B)s+(-2A+5B)}{s^2+3s-10}.\end{aligned} \tag{4}$$

To make the latter an identity we equate the coefficients of s^1 and s^0 in the numerators of the left- and right-hand sides: s^1 gives $0 = A+B$, and s^0 gives $3 = -2A+5B$ so $A = 3/7$ and $B = -3/7$. Then

$$\begin{aligned}L^{-1}\left\{\frac{3}{s^2+3s-10}\right\} &= L^{-1}\left\{\frac{-3/7}{s+5}+\frac{3/7}{s-2}\right\}\\ &= -\frac{3}{7}L^{-1}\left\{\frac{1}{s+5}\right\}+\frac{3}{7}L^{-1}\left\{\frac{1}{s-2}\right\}\\ &= -\frac{3}{7}e^{-5t}+\frac{3}{7}e^{2t},\end{aligned} \tag{5}$$

where the second equality follows from Theorem 5.3.2, and the last equality follows from entry 2 in Appendix C.

COMMENT. Actually, we could have used entry 9 in Appendix C, which says that

$$L^{-1} = \left\{\frac{b}{(s-a)^2+b^2}\right\} = e^{at}\sin bt, \tag{6}$$

for if we equate $(s-a)^2+b^2 = s^2-2as+a^2+b^2$ to $s^2+3s-10$ we see that $a = -3/2$ and $b = \pm 7i/2$. Choose $b = +7i/2$, say ($b = -7i/2$ will give the same result). Then

$$\begin{aligned}L^{-1}\left\{\frac{3}{s^2+3s-10}\right\} &= \frac{3}{7i/2}L^{-1}\left\{\frac{7i/2}{(s+3/2)^2+(7i/2)^2}\right\}\\ &= \frac{6}{7i}e^{-3t/2}\sin(7it/2) = \frac{6}{7i}e^{-3t/2}\,\frac{e^{i(7it/2)}-e^{-i(7it/2)}}{2i}\\ &= \frac{3}{7}\left(-e^{-5t}+e^{2t}\right),\end{aligned} \tag{7}$$

where the first equality is true by linearity and the second follows from (6). This result is the same as the one found above by partial fractions. This example illustrates the fact that we can often invert a given transform in more than one way. ∎

If we are going to apply the Laplace transform method to differential equations, we need to know how to take transforms of derivatives.

THEOREM 5.3.3 *Transform of the Derivative*
Let $f(t)$ be continuous and $f'(t)$ be piecewise continuous on $0 \le t \le t_0$ for every finite t_0, and let $f(t)$ be of exponential order as $t \to \infty$ so that there are constants K, c, T such that $|f(t)| \le Ke^{ct}$ for all $t > T$. Then $L\{f'(t)\}$ exists for all $s > c$, and

$$\boxed{L\{f'(t)\} = s\,L\{f(t)\} - f(0).} \tag{8}$$

Proof: Since $L\{f'(t)\} = \lim_{B\to\infty} \int_0^B f'(t)\,e^{-st}\,dt$, consider the integral

$$\mathrm{I} = \int_0^B f'(t)\,e^{-st}\,dt = \int_0^{t_1} f'(t)\,e^{-st}\,dt + \cdots + \int_{t_n}^B f'(t)\,e^{-st}\,dt, \tag{9}$$

where $t_1, \ldots, t_n$ are the points, in $0 < t < B$, at which f' is discontinuous. Integrating by parts gives

$$\begin{aligned}\mathrm{I} &= f(t)\,e^{-st}\Big|_0^{t_1} + \cdots + f(t)\,e^{-st}\Big|_{t_n}^B \\ &\quad + s\int_0^{t_1} f(t)\,e^{-st}\,dt + \cdots + s\int_{t_n}^B f(t)\,e^{-st}\,dt.\end{aligned} \tag{10}$$

By virtue of the continuity of f, the boundary terms at $t_1, \ldots, t_n$ cancel in pairs so that, after recombining the integrals in (10), we have

$$\mathrm{I} = f(B)\,e^{-sB} - f(0) + s\int_0^B f(t)\,e^{-st}\,dt. \tag{11}$$

Since f is of exponential order as $t \to \infty$ it follows that $f(B)\,e^{-sB} \to 0$ as $B \to \infty$. Thus,

$$\begin{aligned}L\{f'(t)\} &= \lim_{B\to\infty}\left[f(B)\,e^{-sB} - f(0) + s\int_0^B f(t)\,e^{-st}\,dt\right] \\ &= 0 - f(0) + s\,L\{f(t)\},\end{aligned} \tag{12}$$

as was to be proved. ∎

The foregoing result can be used to obtain the transforms of higher derivatives as well. For example, if $f'(t)$ satisfies the conditions imposed on f in Theorem 5.2.3, then replacement of f by f' in (8) gives

$$L\{f''\} = s\,L\{f'\} - f'(0) = s\,[s\,L\{f\} - f(0)] - f'(0).$$

If, besides f', f also satisfies the conditions of Theorem 5.3.3, so that the $L\{f\}$ term on the right side exists, then

$$L\{f''\} = s^2\,L\{f\} - s\,f(0) - f'(0). \tag{13}$$

Similarly,

$$L\{f'''\} = s^3\,L\{f\} - s^2\,f(0) - s\,f'(0) - f''(0), \tag{14}$$

if f'', f', and f satisfy the conditions of Theorem 5.3.3, and so on for the transforms of higher-order derivatives.

The last of the major properties of the Laplace tansform that we discuss in this section is the Laplace convolution theorem.

THEOREM 5.3.4 *Laplace Convolution Theorem*
If $L\{f(t)\} = F(s)$ and $L\{g(t)\} = G(s)$ both exist for $s > c$, then

$$\boxed{L^{-1}\{F(s)G(s)\} = \int_0^t f(\tau)\,g(t-\tau)\,d\tau} \tag{15}$$

or, equivalently,

$$\boxed{L\left\{\int_0^t f(\tau)\,g(t-\tau)\,d\tau\right\} = F(s)G(s)} \tag{16}$$

for $s > c$.

Proof: Since (15) and (16) are equivalent statements, it suffices to prove just one, say (16). By definition,

$$L\left\{\int_0^t f(\tau)\,g(t-\tau)\,d\tau\right\} = \int_0^\infty \left\{\int_0^t f(\tau)\,g(t-\tau)\,d\tau\right\} e^{-st}\,dt. \tag{17}$$

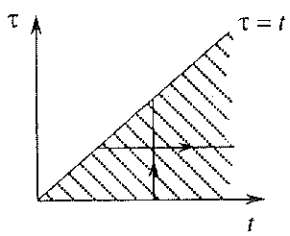


Figure 1. Region of integration.

Regarding the latter as an iterated integral over a 45° wedge in a τ, t plane as shown in Fig. 1, let us invert the order of integration in (17). Recalling from the calculus the equivalent notations

$$\int_c^d \left\{\int_a^b f(x,y)\,dx\right\} dy = \int_c^d \int_a^b f(x,y)\,dx\,dy = \int_c^d dy \int_a^b f(x,y)\,dx \tag{18}$$

for iterated integrals (where a, b, c, d are constants, say), inverting the order of integration in (17) gives

$$\begin{aligned} L\left\{\int_0^t f(\tau)\,g(t-\tau)\,d\tau\right\} &= \int_0^\infty \int_\tau^\infty f(\tau)\,g(t-\tau)\,e^{-st}dt\,d\tau \\ &= \int_0^\infty f(\tau)\,d\tau \int_\tau^\infty g(t-\tau)\,e^{-st}\,dt \\ &= \int_0^\infty f(\tau)\,d\tau \int_0^\infty g(\mu)\,e^{-s(\mu+\tau)}\,d\mu \\ &= \int_0^\infty f(\tau)\,e^{-s\tau}\,d\tau \int_0^\infty g(\mu)\,e^{-s\mu}\,d\mu. \end{aligned} \tag{19}$$

The last product is simply $F(s)$ times $G(s)$, so the theorem is proved. ■

The integral on the right side of (15) is called the **Laplace convolution** of f and g and is denoted as $f * g$. It too is a function of t:

$$\boxed{(f*g)(t) \equiv \int_0^t f(\tau)\,g(t-\tau)\,d\tau.} \tag{20}$$

CAUTION: Be sure to see that the inverse of the product, $L^{-1}\{F(s)G(s)\}$, is not simply the algebraic product of the inverses, $f(t)g(t)$; rather, it is (according to Theorem 5.3.4) their convolution, $(f*g)(t)$.

EXAMPLE 3. In Example 2 we inverted $F(s) = 3/(s^2+3s-10)$ in two different ways. Let us now obtain the inverse by still another method, the convolution theorem:

$$\begin{aligned} L^{-1}\left\{\frac{3}{s^2+3s-10}\right\} &= 3L^{-1}\left\{\frac{1}{s-2}\,\frac{1}{s+5}\right\} = 3L^{-1}\left\{\frac{1}{s-2}\right\} * L^{-1}\left\{\frac{1}{s+5}\right\} \\ &= 3e^{2t} * e^{-5t} = 3\int_0^t e^{2\tau}e^{-5(t-\tau)}\,d\tau \\ &= \frac{3}{7}\left(e^{2t}-e^{-5t}\right), \end{aligned} \tag{21}$$

which is the same result as obtained in Example 2. ■

Observe that in equation (15) it surely doesn't matter if we write $F(s)G(s)$ or $G(s)F(s)$ because ordinary multiplication is commutative. Yet it is not clear that the results are the same, $\int_0^t f(\tau)\,g(t-\tau)\,d\tau$ in one case and $\int_0^t g(\tau)\,f(t-\tau)\,d\tau$ in the other. Nonetheless, these results are indeed the same, proof of which claim is left as an exercise. In fact, although the convolution is not an ordinary product it does share several of the properties of ordinary multiplication:

$$f*g = g*f, \qquad \text{(commutative)} \tag{22a}$$

$$f * (g * h) = (f * g) * h, \qquad \text{(associative)} \tag{22b}$$
$$f * (g + h) = f * g + f * h, \qquad \text{(distributive)} \tag{22c}$$
$$f * 0 = 0. \tag{22d}$$

Closure. The properties studied in this section – linearity, the transform of a derivative, and the convolution theorem, should be thoroughly understood. All are used in the next section, where we use the Laplace transform method to solve differential equations. The convolution property, in particular, should be studied carefully.

The convolution theorem is useful in both directions. If we have a transform $H(s)$ that is difficult to invert, it may be possible to factor H as $F(s)G(s)$, where F and G are more easily inverted. If so, then $h(t)$ is given, according to (15), as the convolution of $f(t)$ and $g(t)$. Furthermore, we may need to find the transform of an integral that is in convolution form. If so, then the transform is given easily by (16).

Finally, we mention that convolution integrals arise in *hereditary* systems, systems whose behavior at time t depends not only on the state of the system at that instant but also on its past history. Examples occur in the study of viscoelasticity and population dynamics.

EXERCISES 5.3

1. Find the inverse of the given transform two different ways: using partial fractions and using the convolution theorem. Cite any entries used from Appendix C.

(a) $3/[s(s+8)]$ (b) $1/(3s^2+5s-2)$
(c) $1/(s^2-a^2)$ (d) $5/[(s+1)(3s+2)]$
(e) $1/(s^2+s)$ (f) $2/(2s^2-s-1)$

2.(a)–(f) Find the inverse of the corresponding transform in Exercise 1 using computer software.

3. Use entry 9 in Appendix C to evaluate the inverse of each. If necessary, use entry 10 as well. NOTE: See the Comment in Example 2.

(a) $1/(s^2+8s)$ (b) $1/(s^2-3s+3)$
(c) $1/(s^2-s)$ (d) $1/(s^2-s-2)$
(e) $s/(s^2-2s+2)$ (f) $(s+1)/(s^2+4s+6)$
(g) $(s+1)/(s^2-s)$ (h) $(2s-1)/(s^2-6s+5)$

4. Use (8) together with mathematical induction to verify the general formula $L\{f^{(n)}\} = s^n L\{f\} - s^{n-1}f(0) - s^{n-2}f'(0) - \cdots - f^{(n-1)}(0)$, which is valid if $f^{(n-1)}, f^{(n-2)}, \ldots, f'$, and f satisfy the conditions of Theorem 5.3.3.

5. Prove

(a) equation (22a) (b) equation (22b)
(c) equation (22c) (d) equation (22d)

6. Prove that $L\{f * g * h\} = F(s)G(s)H(s)$ or, equivalently, that $L^{-1}\{F(s)G(s)H(s)\} = f * g * h$. NOTE: Does $f * g * h$ mean $(f * g) * h$ or $f * (g * h)$? According to the associative property (22b) it doesn't matter; they are equal.

7. To illustrate the result stated in Exercise 6, find the inverse of $1/s^3$ as $L^{-1}\left\{\frac{1}{s^3}\right\} = L^{-1}\left\{\frac{1}{s}\frac{1}{s}\frac{1}{s}\right\} = 1 * 1 * 1$, and show that the result agrees with that given directly in Appendix C.

8. Factoring $\frac{s}{(s^2+a^2)^2} = \frac{s}{s^2+a^2}\,\frac{1}{s^2+a^2}$, it follows from the convolution theorem and entries 3 and 4 of Appendix C that

$$L^{-1}\left\{\frac{s}{(s^2+a^2)^2}\right\} = \cos at * \frac{\sin at}{a}.$$

Evaluate this convolution and show that the result agrees with that given directly by entry 11.

9. Verify (8) and (13) directly, for each given $f(t)$, by working out the left- and right-hand sides and showing that they are equal. You may use the table in Appendix C to evaluate $L\{f''(t)\}$, $L\{f'(t)\}$, and $L\{f(t)\}$.

(a) e^{3t} (b) $e^{-4t} + 2$ (c) $t^2 + 5t - 1$
(d) $\sinh 4t$ (e) $\cosh 3t + 5t^6$ (f) $4t^{3/2} - \cos 2t$

10. Evaluate the transform of each:

(a) $\int_0^t e^{t-\tau} \sin 2\tau \, d\tau$ (b) $\int_0^t \cos 3(t-\tau) \, d\tau$
(c) $\int_0^t (t-\tau)^8 e^{-3\tau} \, d\tau$ (d) $\int_0^t \cosh 3(t-\tau) \, d\tau$
(e) $\int_0^t \frac{\sin \tau}{\sqrt{t-\tau}} \, d\tau$ (f) $\int_0^t \tau^{5.2} \sinh 4(t-\tau) \, d\tau$

11. We emphasized that $L^{-1}\{F(s)G(s)\}$ equals the convolution of f and g; in general, it does not merely equal the product $f(t)g(t)$. Show that $L^{-1}\{F(s)G(s)\} \neq f(t)g(t)$ for each given pair of functions.

(a) $f(t) = t, \quad g(t) = e^t$ (b) $f(t) = \sin t, \quad g(t) = 4$
(c) $f(t) = t, \quad g(t) = t^2$ (d) $f(t) = \cos t, \quad g(t) = t + 6$

5.4 Application to the Solution of Differential Equations

Our object, in this section, is to explain the use of the Laplace transform in solving linear constant-coefficient differential equations on the interval $0 < t < \infty$, with initial conditions at $t = 0$.

EXAMPLE 1. We've already studied the important case of the harmonic oscillator – both free and forced, both damped and undamped. Consider the undamped mechanical oscillator shown in Fig. 1, with a forcing function that is a constant: $f(t) = F_0$. Recall that the displacement $x(t)$ then satisfies the equation

$$mx'' + kx = f(t) = F_0. \tag{1}$$

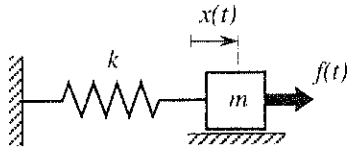


Figure 1. Mechanical oscillator.

Further, we assume the initial conditions $x(0)$ and $x'(0)$ are known.

To apply the Laplace transform, we transform equation (1). That is, we multiply each term in (1) by the Laplace kernel e^{-st} and integrate on t from 0 to ∞. Operationally, we use L to denote that step:

$$L\{mx'' + kx\} = L\{F_0\}. \tag{2}$$

By the linearity of L (Theorem 5.3.1), we can rewrite (2) as

$$mL\{x''(t)\} + kL\{x(t)\} = F_0 \, L\{1\}. \tag{3}$$

Recalling from Theorem 5.3.2 that $L\{x''(t)\} = s^2X(s) - sx(0) - x'(0)$, noting that $L\{x(t)\} \equiv X(s)$, and obtaining $L\{1\} = 1/s$ from Appendix C, (3) becomes

$$m\left[s^2X(s) - sx(0) - x'(0)\right] + kX(s) = F_0 \frac{1}{s}. \tag{4}$$

The point to appreciate is that whereas in the t domain we had the linear *differential* equation (1) on $x(t)$, in the transform domain, or s domain, we now have the linear *algebraic* equation (4) on $X(s)$. The solution now amounts to solving (4) by simple algebra for $X(s)$. Doing so gives

$$X(s) = \frac{sx(0) + x'(0)}{s^2 + \omega^2} + \frac{F_0}{ms\,(s^2 + \omega^2)}, \tag{5}$$

where $\omega = \sqrt{k/m}$ is the natural frequency.

With the solving for $X(s)$ completed, we now invert (5) to obtain $x(t)$:

$$\begin{aligned} x(t) &= L^{-1}\left\{\frac{sx(0)+x'(0)}{s^2+\omega^2} + \frac{F_0}{ms\,(s^2+\omega^2)}\right\} \\ &= x(0)L^{-1}\left\{\frac{s}{s^2+\omega^2}\right\} + x'(0)L^{-1}\left\{\frac{1}{s^2+\omega^2}\right\} + \frac{F_0}{m}L^{-1}\left\{\frac{1}{s\,(s^2+\omega^2)}\right\}, \end{aligned} \tag{6}$$

where the second equality follows from the linearity of the L^{-1} operator (Theorem 5.3.2). Appendix C gives

$$L^{-1}\left\{\frac{s}{s^2+\omega^2}\right\} = \cos\omega t \qquad \text{and} \qquad L^{-1}\left\{\frac{1}{s^2+\omega^2}\right\} = \frac{\sin\omega t}{\omega}, \tag{7}$$

but the third inverse in (6) is not found in the table. We could evaluate it with the help of partial fractions, but it is easier to use the convolution theorem:

$$\begin{aligned} L^{-1}\left\{\frac{1}{s\,(s^2+\omega^2)}\right\} &= L^{-1}\left\{\frac{1}{s}\,\frac{1}{s^2+\omega^2}\right\} = L^{-1}\left\{\frac{1}{s}\right\} * L^{-1}\left\{\frac{1}{s^2+\omega^2}\right\} \\ &= 1 * \frac{\sin\omega t}{\omega} = \int_0^t (1)\left(\frac{\sin\omega(t-\tau)}{\omega}\right)d\tau = \frac{1-\cos\omega t}{\omega^2}, \end{aligned} \tag{8}$$

so (6), (7), and (8) give the desired particular solution as

$$x(t) = x(0)\,\cos\omega t + \frac{x'(0)}{\omega}\,\sin\omega t + \frac{F_0}{k}\,(1-\cos\omega t)\,. \tag{9}$$

For instance, if $x(0) = x'(0) = 0$, then $x(t) = (F_0/k)(1-\cos\omega t)$ as depicted in Fig. 2. Does it seem correct that the constant force F_0 should cause an oscillation? Yes, for imagine rotating the apparatus 90° so that the mass hangs down. Then we can think of F_0 as the downward gravitational force on m. In static equilibrium, the mass will hang down an amount $x = F_0/k$. If we release it from $x = 0$, it will fall and then oscillate about the equilibrium position $x = F_0/k$, as shown in Fig. 2.

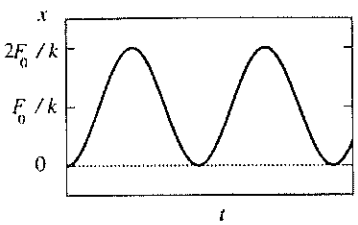


Figure 2. Release from rest.

COMMENT 1. Recall that $f * g = g * f$, so we can write the convolution integral either as $\int_0^t f(\tau)\,g(t-\tau)\,d\tau$ or as $\int_0^t g(\tau)\,f(t-\tau)\,d\tau$; that is, we can let the argument of f be τ and the argument of g be $t-\tau$, or vice versa, whichever we choose. In (8) we chose the τ argument for 1 and the $t-\tau$ argument for $(\sin\omega t)/\omega$. (Of course, if we change all the t's in 1 to τ's we still have 1 because there are no t's in 1.) Alternatively, we could have expressed the inverse in (8) as

$$\frac{\sin\omega t}{\omega} * 1 = \int_0^t \left(\frac{\sin\omega\tau}{\omega}\right)(1)\,d\tau = \frac{1-\cos\omega\tau}{\omega^2},$$

as obtained in (8).

COMMENT 2. Observe that (9) is the particular solution satisfying the initial conditions $x = x(0)$ and $x' = x'(0)$ at $t = 0$. If those quantities are not prescribed, we can replace them by arbitrary constants, and then (9) amounts to the general solution of $mx'' + kx =$

F_0. Thus, the method gives either a particular solution or a general solution, whichever is desired.

COMMENT 3. If, instead of the specific forcing function $f(t) = F_0$ we allow $f(t)$ to be an unspecified function, then we have, in place of (5),

$$X(s) = \frac{sx(0) + x'(0)}{s^2 + \omega^2} + \frac{F(s)}{m(s^2 + \omega^2)}, \tag{10}$$

and, in place of (9),

$$x(t) = x(0)\cos\omega t + \frac{x'(0)}{\omega}\sin\omega t + \frac{1}{m}L^{-1}\left\{\frac{F(s)}{s^2 + \omega^2}\right\}. \tag{11}$$

Using the convolution theorem to write

$$L^{-1}\left\{\frac{F(s)}{s^2 + \omega^2}\right\} = L^{-1}\left\{\frac{1}{s^2 + \omega^2}\right\} * L^{-1}\{F(s)\} = \frac{\sin\omega t}{\omega} * f(t), \tag{12}$$

gives

$$x(t) = x(0)\cos\omega t + \frac{x'(0)}{\omega}\sin\omega t + \frac{1}{m\omega}\int_0^t \sin\omega\tau\, f(t-\tau)\, d\tau \tag{13}$$

as the solution. ∎

With Example 1 completed, there are several observations that can be made about the method. First, consider the general second-order equation

$$x'' + ax' + bx = f(t), \tag{14}$$

where a, b are constants, although the following discussion applies to higher-order equations as well. If we solve (14) by the methods of Chapter 3, then we need both homogeneous and particular solutions. To find the homogeneous solution we need to factor the characteristic polynomial $\lambda^2 + a\lambda + b$ or, equivalently, to find the roots of the characteristic equation $\lambda^2 + a\lambda + b = 0$. Solving (14) by the Laplace transform instead, we obtain, and need to invert,

$$X(s) = \frac{(s+a)x(0) + x'(0)}{s^2 + as + b} + \frac{F(s)}{s^2 + as + b}. \tag{15}$$

Whether we invert these terms by partial fractions or by some other method, their inversion depends, essentially, on our being able to factor the $s^2 + as + b$ denominator. That polynomial is none other than the characteristic polynomial corresponding to (14). Thus, whether we solve (14) by seeking exponential solutions for the homogeneous equation and then seeking a particular solution, or by using the Laplace transform, we need to face up to the same task, finding the roots of the characteristic equation.

Second, observe that if we invert the $F(s)/(s^2 + as + b)$ term in (15) by the convolution theorem, then we convolve the inverse of $F(s)$, namely, $f(t)$, with

the inverse of $1/(s^2 + as + b)$. Therefore, if we use the convolution theorem, then there is no need to evaluate the transform $F(s)$ of $f(t)$ when transforming the given differential equation.

Third, observe how the initial conditions become "built in," when we take the transform of the differential equation. Thus, there is no need to apply them at the end.

Fourth, recall that Laplace transforms come with restrictions on s. For instance, $L\{1\} = 1/s$ for $s > 0$. However, such restrictions in no way impede the solution steps in using the Laplace transform method, and once we invert $X(s)$ to obtain $x(t)$ they are no longer relevant.

Fifth, we need to realize that when we apply the Laplace transform method to a differential equation, we take the transform of the unknown and one or more of its derivatives, but since we don't yet know the solution we don't yet know whether or not these functions are transformable. The procedure, then, is to assume that they are transformable in order to proceed, and to verify that they are once the solution is in hand.

Finally, and most important, understand that the power of the Laplace transform, in solving linear constant-coefficient differential equations, is in its ability to convert such an equation to a linear algebraic equation on $X(s)$, which ability flows from the fact that the transform of $f'(t)$ is merely a multiple of $F(s)$ plus a constant (and therefore similarly for $f'', f''', \ldots$). Indeed, the transform $L\{f(t)\} = \int_0^\infty f(t)\, e^{-st}\, dt$ was *designed* so as to have this property. That is, the "kernel" e^{-st} was designed so as to imbue the transform with that property.

EXAMPLE 2. Solve the initial-value problem

$$y^{(iv)} - y = 0; \qquad y(0) = 1,\ y'(0) = y''(0) = y'''(0) = 0 \tag{16}$$

for $y(x)$. That the independent and dependent variables are x, y, rather than x, t, is immaterial to the application of the Laplace transform; the transform of $y(x)$ is now $Y(s) = \int_0^\infty y(x)\, e^{-sx}\, dx$. Taking the transform of (16) gives

$$\left[s^4 Y(s) - s^3 y(0) - s^2 y'(0) - s y''(0) - y'''(0)\right] - Y(s) = 0. \tag{17}$$

Putting the initial conditions into (17), and solving for $Y(s)$, gives

$$Y(s) = \frac{s^3}{s^4 - 1}. \tag{18}$$

To invert the latter, we can use partial fractions:

$$\begin{aligned} \frac{s^3}{s^4 - 1} &= \frac{A}{s+1} + \frac{B}{s-1} + \frac{C}{s+i} + \frac{D}{s-i} \\ &= \frac{(s-1)(s^2+1)A + (s+1)(s^2+1)B + (s-i)(s^2-1)C + (s+i)(s^2-1)D}{s^4 - 1}. \end{aligned} \tag{19}$$

Equating coefficients of like powers of s in the numerators gives the linear equations

$$\begin{aligned} s^3: &\quad 1 = A + B + C + D, \\ s^2: &\quad 0 = -A + B - iC + iD, \\ s: &\quad 0 = A + B - C - D, \\ 1: &\quad 0 = -A + B + iC - iD, \end{aligned}$$

solution of which (for instance by Gauss elimination) gives $A = B = C = D = 1/4$. Thus,

$$\begin{aligned} y(x) &= \frac{1}{4}e^{-x} + \frac{1}{4}e^{x} + \frac{1}{4}e^{-ix} + \frac{1}{4}e^{ix} \\ &= \frac{1}{2}(\cosh x + \cos x) \end{aligned} \tag{20}$$

is the desired particular solution. ∎

EXAMPLE 3. Solve the first-order initial-value problem

$$x' + px = q(t); \qquad x(0) = x_0 \tag{21}$$

for $x(t)$, where p is a constant and $q(t)$ is any prescribed forcing function. Application of the Laplace transform gives

$$X(s) = \frac{x_0}{s+p} + \frac{Q(s)}{s+p} \tag{22}$$

and hence the particular solution

$$\begin{aligned} x(t) &= x_0 e^{-pt} + \int_0^t e^{-p(t-\tau)}\, q(\tau)\, d\tau \\ &= e^{-pt}\left[x_0 + \int_0^t q(\tau)\, e^{p\tau}\, d\tau\right]. \end{aligned} \tag{23}$$

COMMENT. Alternatively, let us begin by integrating the differential equation on t, from 0 to t:

$$x(t)\Big|_0^t + p\int_0^t x(\tau)\, d\tau = \int_0^t q(\tau)\, d\tau \tag{24}$$

or, since $x(0) = x_0$,

$$x(t) + p\int_0^t x(\tau)\, d\tau = x_0 + \int_0^t q(\tau)\, d\tau. \tag{25}$$

Of course, (25) is not the solution of the differential equation because the unknown $x(t)$ is under the integral sign. Thus, (25) is an example of an **integral equation**. Although we will not study integral equations systematically in this text, it will be useful to at least introduce them. Observe, first, that (25) is equivalent to both the differential equation and the initial condition, for they led to (25); conversely, the derivative of (25) gives back $x' + px = q(t)$

[if $q(t)$ is continuous], and putting $t = 0$ in (25) gives back $x(0) = x_0$. That is, unlike the differential equation, the integral equation version has the initial condition "built in."

Further, we can solve (25) by the Laplace transform conveniently because each integral is of convolution type: the first is $1 * x(t)$, and the second is $1 * q(t)$. Thus, taking a Laplace transform of (25), and noting that $L\{1 * x(t)\} = L\{1\}L\{x(t)\} = (1/s)X(s)$ and $L\{1 * q(t)\} = (1/s)Q(s)$, gives

$$X(s) + p\frac{1}{s}X(s) = \frac{x_0}{s} + \frac{1}{s}Q(s), \tag{26}$$

which, once again, gives (22) and hence the solution (23). ∎

Closure. In this section we describe the application of the Laplace transform to the solution of linear differential equations with constant coefficients, homogeneous or nonhomogeneous. In a sense, the method is of comparable difficulty to the solution methods studied in Chapter 3 in that one still needs to be able to factor the characteristic polynomial, which can be difficult if the equation is of high order. However, the Laplace transform method has a number of advantages. First, the method reduces a linear differential equation to a linear algebraic equation. Second, the hardest part, namely the inversion of the transform of the unknown, can often be accomplished with the help of tables or computer software, as well as with several additional theorems that are given in the final section of this chapter. Third, any initial conditions that are given become built in, in the process of taking the transform of the differential equation, so they do not need to be applied separately, at the end, as they were in Chapter 3.

We also saw, in the final example, that the Laplace transform is convenient to use in solving integral equations (equations in which the unknown function appears under an integral sign), provided that the integrals therein are Laplace convolution integrals; additional discussion of this idea is left for the exercises. In fact, it might be noted that the Laplace transform itself, $F(s) = \int_0^\infty f(t)\, e^{-st}\, dt$, is really an integral equation for $f(t)$ if $F(s)$ is known. Although that integral equation was studied by Laplace, it was *Simeon-Denis Poisson* (1781 – 1840) who discovered the solution $f(t) = \frac{1}{2\pi i}\int_{\gamma-i\infty}^{\gamma+i\infty} F(s)\, e^{st}\, ds$, namely, the Laplace inversion formula. Poisson was one of the great nineteenth century analysts and a professor at the Ecole Polytechnique.

Also left for the exercises is discussion of the application of the method to a limited class of nonconstant differential equations.

EXERCISES 5.4

1. Use the Laplace transform to find the general solution, or the particular solution if initial conditions are given.

(a) $x' + 2x = 4t^2$
(b) $3x' + x = 6e^{2t}$; $x(0) = 0$
(c) $x' - 6x = e^{-t}$; $x(0) = 4$
(d) $x'' = 6t$; $x(0) = 2$, $x'(0) = -1$
(e) $x'' + 5x' = 10$
(f) $x'' - x' = 1 + t + t^2$

(g) $x'' - 3x' + 2x = 0; \quad x(0) = 3,\ x'(0) = 1$
(h) $x'' - 4x' - 5x = 2 + e^{-t}; \quad x(0) = x'(0) = 0$
(i) $x'' - x' - 12x = t; \quad x(0) = -1,\ x'(0) = 0$
(j) $x'' + 6x' + 9x = 1; \quad x(0) = 0,\ x'(0) = -2$
(k) $x'' - 2x' + 2x = -2t; \quad x(0) = 0,\ x'(0) = -5$
(l) $x'' - 2x' + 3x = 5; \quad x(0) = 1,\ x'(0) = -1$
(m) $x''' - x'' + 2x' = t^2; \quad x(0) = 1,\ x'(0) = x''(0) = 0$
(n) $x''' + x'' - 2x' = 1 + e^t; \quad x(0) = x'(0) = x''(0) = 0$
(o) $x''' + 5x'' = t^4; \quad x(0) = x'(0) = 0,\ x''(0) = 1$
(p) $x''' + 3x'' + 3x' + x = e^{2t}$
(q) $x''' - x'' - x' + x = 0; \quad x(0) = 2,\ x'(0) = x''(0) = 0$
(r) $x^{(iv)} = 2\sin t$
(s) $x^{(iv)} + 3x''' = 0; \quad x(0) = x'(0) = 0,$
$x''(0) = x'''(0) = 3$
(t) $x^{(iv)} + 8x'' + 16x = 4$
(u) $x^{(iv)} - x = 1; \quad x(0) = x'(0) = x''(0) = 0,$
$x'''(0) = 4$
(v) $x^{(iv)} - x = 4; \quad x(0) = x'(0) = x''(0) = 1,$
$x'''(0) = 0$
(w) $x^{(iv)} - 16x = -32; \quad x(0) = 0,\ x'(0) = 2,$
$x''(0) = x'''(0) = 0$

2. (a) Show that for a constant-coefficient linear homogeneous differential equation of order n, the Laplace transform $X(s)$ of the solution $x(t)$ is necessarily of the form

$$X(s) = P(s)/Q(s), \tag{2.1}$$

where $Q(s)$ and $P(s)$ are polynomials in s, with Q of degree n and P of degree less than n.

(b) Show that if $Q(s) = 0$ has n distinct roots $r_1, \ldots r_n$, then

$$x(t) = \sum_{k=1}^{n} \frac{P(r_k)}{Q'(r_k)}\, e^{r_k t}. \tag{2.2}$$

3. Our purpose, in this exercise, is to follow up on Example 3 in showing a connection between differential equations and integral equations, and in considering the solution of certain integral equations by the Laplace transform method.

(a) Convert the initial–value problem

$$\begin{aligned} &mx'' + kx = f(t), \qquad (0 \le t < \infty) \\ &x(0) = x_0,\ \ x'(0) = x_0' \end{aligned} \tag{3.1}$$

to an integral equation, as follows. Integrate the differential equation from 0 to t twice. Using the initial conditions, show that those steps give

$$\begin{aligned} mx(t) - mx_0 - mx_0' t + k\int_0^t \int_0^{t'} x(\tau)\, d\tau\, dt' \\ = \int_0^t \int_0^{t'} f(\tau)\, d\tau\, dt'. \end{aligned} \tag{3.2}$$

Show that, by interchanging the order of integration, the double integrals can be reduced to single integrals, so that the integral equation (3.2) can be simplified to the form

$$\begin{aligned} mx(t) - mx_0 - mx_0' t + k\int_0^t (t-\tau)\, x(\tau)\, d\tau \\ = \int_0^t (t-\tau)\, f(\tau)\, d\tau. \end{aligned} \tag{3.3}$$

(b) Taking a Laplace transform of (3.3), obtain

$$X(s) = \frac{sx(0) + x'(0)}{s^2 + \omega^2} + \frac{F(s)}{m(s^2+\omega^2)}, \qquad \left(\omega = \sqrt{k/m}\right)$$

which is the same as equation (10).

4. Convert the initial-value problem

$$\begin{aligned} &mx'' + cx' + kx = f(t) \qquad (0 \le t < \infty) \\ &x(0) = x_0,\ \ x'(0) = x_0' \end{aligned}$$

to an integral equation, analogous to (3.3) in Exercise 3. Then, solve that integral equation for $x(t)$ by using the Laplace transform.

5. (*Variable-coefficient equation*) Consider the problem

$$\begin{aligned} &tx'' + x' + tx = 0 \qquad (0 \le t < \infty) \\ &x(0) = 1,\ \ x'(0) = 0, \end{aligned} \tag{5.1}$$

where our special interest lies in seeing whether or not we can solve (5.1) by the Laplace transform method even though the differential equation has nonconstant coefficients.

(a) Take the Laplace transform of the differential equation. Note that the transforms of $t\,x''(t)$ and $t\,x(t)$,

$$L\{t\,x''(t)\} = \int_0^\infty t\,x''\,e^{-st}\,dt,$$
$$L\{t\,x(t)\} = \int_0^\infty t\,x\,e^{-st}\,dt,$$

present a difficulty in that we cannot express them in terms of $X(s)$ the way we can express $L\{x'(t)\} = sX(s) - x(0)$ and $L\{x''(t)\} = s^2X(s) - sx(0) - x'(0)$. Nevertheless, these terms can be handled as follows. Observe that

$$\begin{aligned} L\{t\,x''(t)\} &= \int_0^\infty t\,x''\,e^{-st}\,dt = -\int_0^\infty \frac{d}{ds}\left(x''e^{-st}\right)dt \\ &= -\frac{d}{ds}\int_0^\infty x''e^{-st}\,dt \\ &= -\frac{d}{ds}\left[s^2X(s) - sx(0) - x'(0)\right] \\ &= -\frac{d}{ds}\left[s^2X(s) - s\right], \end{aligned} \tag{5.2}$$

if we assume that the unknown $x(t)$ is sufficiently well behaved for the third equality (where we have interchanged the order of two limit processes, the s differentiation and the t integration) to be justified. Handling the $L\{t\,x(t)\}$ term in the same way, show that application of the Laplace transform to (5.1) leads to the equation

$$(s^2+1)\frac{dX}{ds} + sX = 0 \tag{5.3}$$

on $X(s)$. Note that whereas the Laplace transform method reduces constant-coefficient differential equations to linear algebraic equations on $X(s)$, here the nonconstant coefficients result in the equation on $X(s)$ being itself a linear *differential* equation! However, it is a simple one. Solving (5.3), show that

$$X(s) = \frac{C}{\sqrt{s^2+1}}. \tag{5.4}$$

(b) From Appendix C, we find the inverse as $x(t) = CJ_0(t)$, where J_0 is the Bessel function of the first kind, of order zero. Appying the initial condition once again gives $x(0) = 1 = CJ_0(0) = C$, so $C = 1$, and the desired solution of (5.1) is $x(t) = J_0(t)$. Here, however, we ask you to proceed as though you don't know about Bessel functions. Specifically, re-express (5.4) as

$$X(s) = \frac{C}{s\sqrt{1+(1/s^2)}} = \frac{C}{s}\left(1 - \frac{1}{2}\frac{1}{s^2} + \cdots\right), \tag{5.5}$$

where the last equality amounts to the Taylor expansion of $\sqrt{1+r}$ in the quantity r, about $r = 0$, where $r = 1/s^2$. Carry that expansion further; invert the resulting series term by term (assuming that that step is valid), and thus show that

$$x(t) = C\left[1 - \frac{t^2}{2^2} + \frac{1}{(2!)^2}\frac{t^4}{2^4} - \frac{1}{(3!)^2}\frac{t^6}{2^6} + \cdots\right].$$

Setting $x(0) = 1$ gives $C = 1$, and the result is that we have obtained the solution in power series form. Of course, that power series is the Taylor series of the Bessel function $J_0(t)$. NOTE: Observe that rather than pulling an s out of the square root in (5.5), and then expanding $1/\sqrt{1+(1/s^2)}$ in powers of $1/s^2$, we could have expanded (5.4) directly in powers of s as $X(s) = C(1 - \frac{1}{2}s^2 + \cdots)$. However, positive powers of s are not invertible, so this form is of no use. [We will see, in Theorem 5.7.6, that to be invertible a transform must tend to zero as $s \to \infty$. Positive powers of s do not satisfy this condition, but negative powers do.] Also, observe that the degree to which nonconstant-coefficient differential equations are harder than constant-coefficient ones can be glimpsed from the fact that coefficients proportional to t cause the equation on $X(s)$ to be a first-order differential equation; coefficients proportional to t^2 will cause the equation on $X(s)$ to be a second-order differential equation, and so on.

6. It is found that the integral equation

$$C(T) = \int_0^\infty e^{-0.0744\nu^2/T^2}\rho(\nu)\,d\nu \tag{6.1}$$

is an approximate relation between the frequency spectrum $\rho(\nu)$ and the specific heat $C(T)$ of a crystal, where T is the temperature. Solve for $\rho(\nu)$ if

(a) $C(T) = T$ (b) $C(T) = Te^{-1/T}$

HINT: By a suitable change of variables, the integral can be made to be a Laplace transform.

7. We have seen that two crucial properties of the Laplace transform are its linearity and the property that $L\{f'(t)\} = s\overline{f}(s) - f(0)$; that is, the transform of the derivative is of the simple form $L\{f'(t)\} = a\overline{f}(s) + b$. With these properties in mind, consider the general integral transform

$$F(s) = \int_c^d K(t,s)f(t)\,dt \tag{7.1}$$

[equation (1) in Section 5.1] from a "design" point of view: how to choose the limits c, d and the kernel $K(t,s)$ to achieve these properties. Since $0 < t < \infty$, it is reasonable to choose $c = 0$ and $d = \infty$. Further, (7.1) automatically satisfies the linearity property $L\{\alpha u(t)+\beta v(t)\} = \alpha L\{u(t)\}+\beta L\{v(t)\}$ because the right side of (7.1) is an integral and integrals satisfy the property of linearity. Thus, we simply ask you for a logical derivation of the choice $K(t,s) = e^{-st}$ so that $L\{f'(t)\}$ is of the form $a\overline{f}(s) + b$.

5.5 Discontinuous Forcing Functions; Heaviside Step Function

Although we show in Section 5.2 that a given function has a Laplace transform if it is piecewise continuous on $0 \le t < A$ for every A and of exponential order as $t \to \infty$, we have thus far avoided functions with discontinuities. In applications, however, systems are often subjected to discontinuous forcing functions. For instance, a circuit might be subjected to an applied voltage that is held constant at 12 volts for a minute and then shut off (i.e., reduced to zero for all subsequent time). In this section we study systems with forcing functions that are discontinuous, although we still assume that they are piecewise continuous on $0 \le t < A$ for every A and of exponential order as $t \to \infty$, so that they are Laplace transformable.

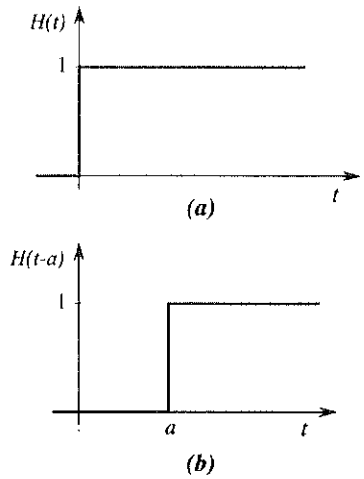


Figure 1. Unit step function.

We begin by defining the **Heaviside step function*** or **unit step function** (Fig. 1a),

$$\boxed{H(t) = \begin{cases} 0, & t < 0 \\ 1, & t \ge 0 \end{cases}} \tag{1}$$

which is a basic building block for our discussion. The value of $H(t)$ at $t = 0$ (i.e., at the jump discontinuity) is generally inconsequential in applications. We have chosen $H(0) = 1$ somewhat arbitrarily, and do not show the value of $H(t)$ at $t = 0$ in Fig. 1a to suggest that it is unimportant in this discussion.

Since $H(t)$ is a unit step at $t = 0$, $H(t - a)$ is a unit step shifted to $t = a$, as shown in Fig. 1b. In fact, the step function is useful in building up more complicated cases. We begin with the **rectangular pulse** shown in Fig. 2. Denoting that function as $P(t; a, b)$, we have

$$P(t; a, b) = H(t - a) - H(t - b). \tag{2}$$

More generally, observe that any piecewise continuous function

$$f(t) = \begin{cases} f_1(t), & 0 < t < t_1 \\ f_2(t), & t_1 < t < t_2 \\ \vdots & \vdots \\ f_n(t), & t_n < t < \infty \end{cases} \tag{3}$$

defined on $0 < t < \infty$ (which is the interval of interest in Laplace transform applications) can be given by the single expression

$$f(t) = f_1(t)\,P(t; 0, t_1) + \cdots + f_{n-1}(t)\,P(t; t_{n-1}, t_n) + f_n(t)\,H(t - t_n). \tag{4}$$

*Oliver Heaviside (1850–1925), initially a telegraph and telephone engineer, is best known for his contributions to vector field theory and to the development of a systematic Laplace transform methodology for the solution of differential equations. Note the spelling: Heaviside, not Heavyside.

In $0 < t < t_1$, for instance, each P function in (4) is zero except for the first, which equals unity in that interval; also, $H(t-t_n)$ is zero there, so (4) gives $f(t) = f_1(t)$. Similarly, in $t_1 < t < t_2, \ldots,$ and $t_{n-1} < t < t_n$. In $t_n < t < \infty$, each P function is zero and the $H(t-t_n)$ is unity, so (4) gives $f(t) = f_n(t)$ there.

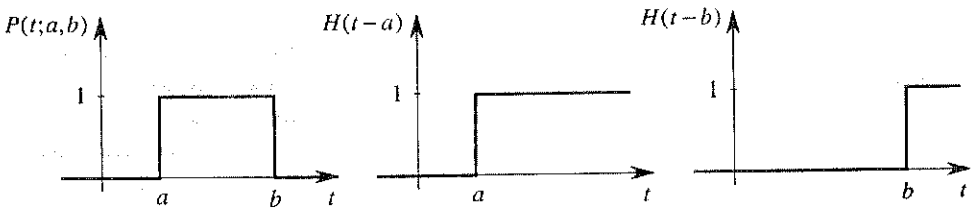


Figure 2. Rectangular pulse.

Note that (3) does not define $f(t)$ at the endpoints $0, t_1, \ldots, t_n$. The Laplace transform of f will be the same no matter what those values are (assuming that they are finite) since the transform is an integral, an integral represents area, and there is no area under a finite number of points. Thus, those values will be inconsequential; hence we don't even specify them in (3).

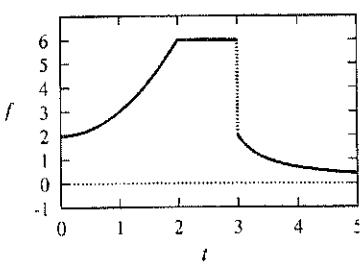


Figure 3. $f(t)$ of Example 1.

EXAMPLE 1. The function

$$f(t) = \begin{cases} 2+t^2, & 0 < t < 2 \\ 6, & 2 < t < 3 \\ 2/(2t-5) & 3 < t < \infty \end{cases} \tag{5}$$

shown in Fig. 3, can be expressed, according to (4), as

$$f(t) = (2+t^2)\,[H(t) - H(t-2)] + 6\,[H(t-2) - H(t-3)] + \frac{2}{2t-5}\,H(t-3). \tag{6}$$

Actually, since the interval is $0 < t < \infty$ we cannot distinguish between $H(t)$ and unity, so we could replace the $H(t)$ in the first term by 1. ∎

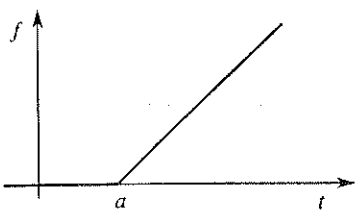


Figure 4. The ramp function of Example 2.

EXAMPLE 2. *Ramp function.* The function

$$f(t) = \begin{cases} 0, & 0 < t < a \\ t-a & a < t < \infty \end{cases} \tag{7}$$

shown in Fig. 4 is called a **ramp** function and, according to (4), it can be expressed as $f(t) = (t-a)H(t-a)$. ∎

Before considering applications, observe that

$$L\{H(t-a)\} = \int_0^\infty H(t-a)\,e^{-st}\,dt = \int_a^\infty e^{-st}\,dt = \frac{e^{-as}}{s},$$

so the Laplace transform of $H(t-a)$ is

$$\boxed{L\{H(t-a)\} = \frac{e^{-as}}{s}.} \tag{8}$$

Also important to us is the result

$$\boxed{L\{H(t-a)f(t-a)\} = e^{-as}F(s)} \tag{9a}$$

or, equivalently,

$$\boxed{L^{-1}\left\{e^{-as}F(s)\right\} = H(t-a)f(t-a)} \tag{9b}$$

for any (Laplace-transformable) function $f(t)$. Proof is as follows:

$$\begin{aligned} L\{H(t-a)f(t-a)\} &= \int_0^\infty H(t-a)\, f(t-a)\, e^{-st}\, dt \\ &= \int_a^\infty f(t-a)\, e^{-st}\, dt = \int_0^\infty f(\tau)\, e^{-s(\tau+a)}\, d\tau \\ &= e^{-as}\int_0^\infty f(\tau)\, e^{-s\tau}\, d\tau = e^{-as}F(s), \end{aligned} \tag{10}$$

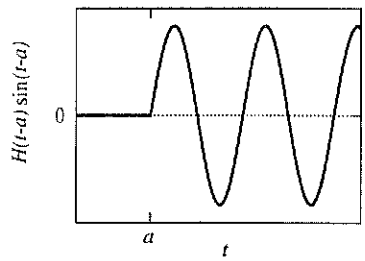


Figure 5. Delay significance of $H(t-a)f(t-a)$.

where the third equality follows the change of variables $t - a = \tau$. In words, $H(t-a)f(t-a)$ is the function $f(t)$ delayed by a time interval a, as illustrated in Fig. 5 for the function $f(t) = \sin t$.

EXAMPLE 3. *LC Circuit.* We saw in Section 2.3 that the differential equation governing the charge $Q(t)$ on the capacitor in the circuit shown (Fig. 6) is $LQ'' + RQ' + (1/C)Q = E(t)$. Let $R = 0$ and let $E(t)$ be the rectangular pulse shown in Fig. 2, of magnitude E_0, and with $a = 2$ and $b = 5$, for definiteness. Thus, $E(t) = E_0\,[H(t-2) - H(t-5)]$. If $Q(0) = Q_0$ and $Q'(0) = 0$, then we have the initial-value problem

$$LQ'' + \frac{1}{C}Q = E_0\,[H(t-2) - H(t-5)]\,, \tag{11a}$$

$$Q(0) = Q_0, \qquad Q'(0) = 0 \tag{11b}$$

on $Q(t)$. [Since the current $i(t)$ is dQ/dt, $Q'(0) = 0$ means that $i(0) = 0$ so we can think of a switch being open until time $t = 0$, and closed at that instant.] We wish to solve (11) for $Q(t)$. Be careful: we will need to distinguish the inductance L from the Laplace transform L by the context.

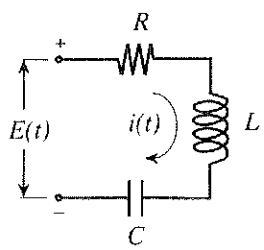


Figure 6. *RLC* circuit.

Taking a Laplace transform of (11a), and using (11b) and (8), gives

$$L\left(s^2\overline{Q}(s) - sQ_0\right) + \frac{1}{C}\overline{Q}(s) = E_0\left(\frac{e^{-2s}}{s} - \frac{e^{-5s}}{s}\right) \tag{12}$$

so

$$\overline{Q}(s) = \frac{Q_0 s}{s^2+\omega^2} + \frac{E_0}{L}\,\frac{1}{s(s^2+\omega^2)}\left(e^{-2s} - e^{-5s}\right), \tag{13}$$

where $\omega = 1/\sqrt{LC}$. [Generally, we use the notation $L\{f(t)\} = F(s)$, but in $Q(t)$ the Q is already capitalized, so we use $L\{Q(t)\} = \overline{Q}(s)$ instead.] To invert (13), we begin with

$$L^{-1}\left\{\frac{1}{s(s^2+\omega^2)}\right\} = L^{-1}\left\{\frac{1}{s}\right\} * L^{-1}\left\{\frac{1}{s^2+\omega^2}\right\} = 1 * \frac{\sin\omega t}{\omega}$$
$$= \int_0^t \frac{\sin\omega\tau}{\omega}\,d\tau = \frac{1-\cos\omega t}{\omega^2}. \tag{14}$$

Then, using (14) and (9b) and $L^{-1}\left\{s/\left(s^2+\omega^2\right)\right\} = \cos\omega t$ from Appendix C, we have

$$Q(t) = Q_0\cos\omega t + E_0 C\left\{H(t-2)\left[1-\cos\omega(t-2)\right]\right.$$
$$\left.-H(t-5)\left[1-\cos\omega(t-5)\right]\right\}, \tag{15}$$

which is shown in Fig. 7 for the representative case where $Q_0 = E_0 = L = C = 1$.

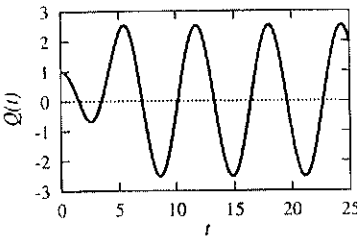


Figure 7. $Q(t)$ given by (15).

COMMENT 1. Most striking is the way the use of the Heaviside notation and the Laplace transform have enabled us to solve for $Q(t)$ on the entire t domain $(0 < t < \infty)$. In contrast, if we rely on the methods of Chapter 3 we need to break (11) into three separate problems:

$$\begin{array}{llll} 0 \le t \le 2: & LQ'' + (1/C)Q = 0, & Q(0) = Q_0, & Q'(0) = 0 \\ 2 \le t \le 5: & LQ'' + (1/C)Q = E_0, & Q(2) = ?, & Q'(2) = ? \\ 5 \le t < \infty: & LQ'' + (1/C)Q = 0, & Q(5) = ?, & Q'(5) = ? \end{array} \tag{16a,b,c}$$

First, we solve (16a) for $Q(t)$ on $0 \le t \le 2$. The final values from that solution, $Q(2)$ and $Q'(2)$, then serve as the initial conditions for the next problem, (16b). Then we solve for $Q(t)$ on $2 \le t \le 5$ and use the final values from that solution, $Q(5)$ and $Q'(5)$, as the initial conditions for the next problem, (16c). Clearly, this approach is more tedious than the Laplace transform approach that led to (15).

COMMENT 2. A fundamental question comes to mind: Does the discontinuous nature of the input $E(t)$ result in the output $Q(t)$ being discontinuous as well? We can see from the graph in Fig. 7 that the answer is no. The continuity of $Q(t)$ may be surprising from the solution form (15), because of the presence of the two Heaviside functions in (15). However, the jump discontinuity implied by the $H(t-2)$ is eliminated by its $1-\cos\omega(t-2)$ factor since the latter vanishes at $t = 2$. Similarly, the jump discontinuity that is implied by the $H(t-5)$ is eliminated by its $1-\cos\omega(t-5)$ factor since the latter vanishes at $t = 5$. ∎

To better understand how a discontinuous input can produce a continuous output, consider the following simplified situation, the equation

$$Q''(t) = H(t-a) \tag{17}$$

with discontinuous right-hand side. Integrating (17) gives

$$Q'(t) = (t-a)H(t-a) + A, \tag{18}$$

because the derivative of the right-hand side is indeed $H(t-a)$, as can be seen from Fig. 4. Integrating again, we obtain

$$Q(t) = \frac{(t-a)^2}{2} H(t-a) + At + B, \tag{19}$$

and these results are shown in Fig. 8 (for the case where $A = B = 0$, say). The idea is that a differential equation is solved by a process which, essentially, involves integration, and integration is a smoothing process! For observe that whereas $Q''(t) = H(t-a)$ is discontinuous at $t = a$, $Q'(t) = (t-a)H(t-a)$ is continuous but with a "kink," and $Q(t) = (t-a)^2 H(t-a)/2$ is continuous and smooth (differentiable) as well.

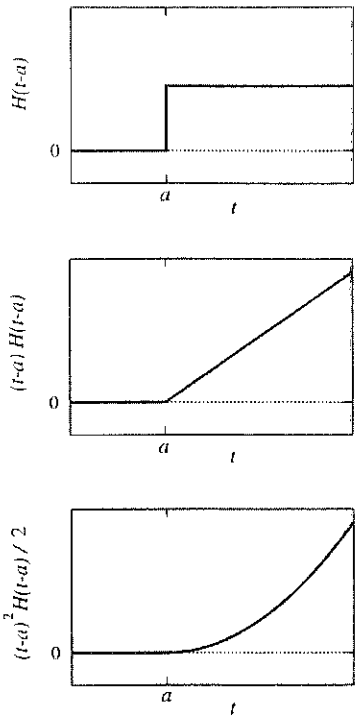


Figure 8. The smoothing effect of integration.

EXAMPLE 4. *RC Circuit.* In Example 3 we took $R = 0$ in the circuit shown in Fig. 6, and considered the resulting LC circuit. Here, let us take $L = 0$ instead, and consider the resulting RC circuit, governed by the first-order equation $RQ' + (1/C)Q = E(t)$. Further, let $Q(0) = 0$ and let $E(t) = 50t$ on $0 < t < 2$ and $E(t) = 40$ on $2 < t < \infty$ (sketch it). According to (4) then, $E(t) = 50t\,[1 - H(t-2)] + 40H(t-2)$. Let $R = C = 1$, for simplicity. Then the initial-value problem on $Q(t)$ is

$$Q' + Q = 50t + (40 - 50t)H(t-2), \tag{20a}$$

$$Q(0) = 0. \tag{20b}$$

Laplace transforming (20a),

$$s\overline{Q}(s) + \overline{Q}(s) = \frac{50}{s^2} + L\left\{(40 - 50t)H(t-2)\right\}, \tag{21}$$

where

$$\begin{aligned} L\left\{(40-50t)H(t-2)\right\} &= L\left\{[-60 - 50(t-2)]\,H(t-2)\right\} \\ &= -60L\left\{H(t-2)\right\} - 50L\left\{(t-2)H(t-2)\right\} \\ &= -60\frac{e^{-2s}}{s} - 50e^{-2s}L\{t\} = -60\frac{e^{-2s}}{s} - 50\frac{e^{-2s}}{s^2}. \end{aligned} \tag{22}$$

Putting (22) into (21) and solving for $\overline{Q}(s)$ gives

$$\overline{Q}(s) = \frac{50}{s^2(s+1)} - \frac{60}{s(s+1)}e^{-2s} - \frac{50}{s^2(s+1)}e^{-2s}, \tag{23}$$

which we now need to invert. Taking one term at a time,

$$L^{-1}\left\{\frac{1}{s(s+1)}\right\} = L^{-1}\left\{\frac{1}{s}\right\} * L^{-1}\left\{\frac{1}{s+1}\right\} = 1 * e^{-t} = \int_0^t e^{-\tau}\,d\tau = 1 - e^{-t}, \tag{24a}$$

$$\begin{aligned} L^{-1}\left\{\frac{1}{s^2(s+1)}\right\} &= L^{-1}\left\{\frac{1}{s^2}\right\} * L^{-1}\left\{\frac{1}{s+1}\right\} = t * e^{-t} \\ &= \int_0^t (t-\tau)\,e^{-\tau}\,d\tau = t - 1 + e^{-t}, \end{aligned} \tag{24b}$$

$$L^{-1}\left\{\frac{1}{s(s+1)}e^{-2s}\right\} = \left(1-e^{-(t-2)}\right)H(t-2), \quad \text{from (9b), (24a)} \tag{24c}$$

$$L^{-1}\left\{\frac{1}{s^2(s+1)}e^{-2s}\right\} = \left[(t-2)-1+e^{-(t-2)}\right]H(t-2), \quad \text{from (9b), (24b)} \tag{24d}$$

so

$$\begin{aligned} Q(t) &= 50\left(t-1+e^{-t}\right) - 60\left(1-e^{-(t-2)}\right)H(t-2) \\ &\quad -50\left[(t-2)-1+e^{-(t-2)}\right]H(t-2) \\ &= 50\left(t-1+e^{-t}\right) + \left(90-50t+10e^{2-t}\right)H(t-2) \end{aligned}$$

is the desired particular solution of (20). ∎

Computer software. The *Maple* name for $H(t)$ is Heaviside(t).

Closure. We introduce the Heaviside step function $H(t)$ and show, in equation (4), how to use it in representing piecewise-continuous functions. A key result involving the Heaviside function is given by (9). Finally, we show how convenient the Heaviside notation is, together with the Laplace transform, in solving differential equations with piecewise continuous forcing functions.

EXERCISES 5.5

1. Use (4) to give a single expression for $f(t)$, and give a labeled sketch of its graph, as well. From that expression, evaluate the transform $F(s)$ of $f(t)$.

(a) $f(t) = t$ on $0 < t < 2$, $4 - t$ on $2 < t < 4$, and 0 on $t > 4$
(b) $f(t) = e^{-t}$ on $0 < t < 1$, 0 on $t > 1$
(c) $f(t) = 2$ on $0 < t < 5$, -3 on $5 < t < 7$, 1 on $t > 7$
(d) $f(t) = t^3 - t$ on $0 < t < 1$, -6 on $t > 1$
(e) $f(t) = 2 - t$ on $0 < t < 2$, $2t - 6$ on $2 < t < 5$, t on $t > 5$
(f) $f(t) = t^3$ on $0 < t < 10$, $3t^2 - 2t$ on $10 < t < 20$, $5t$ on $t > 20$
(g) $f(t) = \sin t$ on $0 < t < 5\pi$, 0 on $t > 5\pi$
(h) $f(t) = \cos t$ on $0 < t < \pi$, -1 on $t > \pi$

2. Draw a labeled sketch of the graph of each function.

(a) $H(t-1)e^{t-1}$
(b) $H(t-2\pi)\cos(t-2\pi)$
(c) $(1+t)H(t-2)$
(d) $(2+t)\left[H(t-2)-H(t-3)\right]$
(e) $t\left[H(t-1)-H(t-2)+H(t-3)\right]$
(f) $t^2\left[2H(t-1)-H(t-3)-H(t-4)\right]$
(g) $\left[H(t-\pi/2)-H(t-\pi)\right]\sin t$
(h) $1+H(t-1)+H(t-2)+H(t-3)+H(t-4)$
(i) $H(t-3)[H(t-2)-H(t-1)]$

3. Evaluate in terms of Heaviside functions. You may use these results for the definite and indefinite integrals of the Heaviside function:

$$\boxed{\int_0^t H(\tau)\,d\tau = \begin{cases} 0, & t<0 \\ t, & t>0 \end{cases} = tH(t).} \tag{3.1}$$

and

$$\boxed{\int^t H(\tau)\,d\tau = tH(t) + \text{constant}.} \tag{3.2}$$

(a) $\int_{-5}^t [H(\tau)-2]\,d\tau$
(b) $\int_0^t \tau H(\tau-2)\,d\tau$
(c) $\int_0^t [1-H(\tau-5)]\,d\tau$
(d) $\int_0^t [H(\tau-a)-H(\tau-b)]\,d\tau \qquad (b>a)$
(e) $\int_t^\infty [H(\tau-2)-H(\tau-3)]\,d\tau$
(f) $\int_t^{2t} H(\tau-1)\,d\tau$

(g) $\int_0^5 H(\tau - t)\, d\tau$
(h) $t * H(t-1)$
(i) $\sin t * [H(t-1) - H(t-2)]$
(j) $e^{-t} * H(t-5)$
(k) $1 * H(t-1)$

4.(a)–(k) Evaluate the integral in the corresponding part of Exercise 3 using computer software such as the *Maple* int command.

5. Solve $x' - x = f(t)$, where $x(0) = 0$, by the methods of this section, where $f(t)$ is:

(a) $H(t-1)$
(b) $e^{-t}H(t-3)$
(c) t on $0 < t < 2$, 2 on $t > 2$
(d) 0 on $0 < t < 5$, 10 on $5 < t < 7$, 0 on $t > 7$
(e) 0 for $t \neq 5$, 100 for $t = 5$
(f) $1 - e^{-t}$ on $0 < t < 6$, 0 on $t > 6$
(g) 0 on $0 < t < 1$, 1 on $1 < t < 2$, 2 on $2 < t < 3$, 1 on $3 < t < 4$, 0 on $t > 4$
(h) t on $0 < t < 1$, $2 - t$ on $1 < t < 2$, 0 on $t > 2$
(i) e^{-t} on $0 < t < 1$, e^{1-t} on $t > 1$
(j) 20 on $0 < t < 1$, 10 on $1 < t < 2$, 0 on $t > 2$

6.(a)–(j) Same as Exercise 5, but using computer software such as the *Maple* dsolve command.

7.(a)–(j) Same as Exercise 5, but for $x'' - x = f(t)$, $x(0) = x'(0) = 0$.

5.6 Impulsive Forcing Functions; Dirac Impulse Function (Optional)

Besides forcing functions that are discontinuous, we are interested in ones that are impulsive – that is, sharply focused in time. For instance, consider the forced mechanical oscillator governed by the differential equation

$$mx'' + cx' + kx = f(t), \tag{1}$$

where $f(t)$ is the force applied to the mass. If the force is due to a hammer blow, for instance, initiated at time $t = 0$, then we expect $f(t)$ to be somewhat as sketched in Fig. 1a. However, we do not know the functional form of $f(t)$ corresponding to such an event as a hammer blow, so the problem that we pose is how to proceed with the solution of (1) without knowing f. Of course we can solve (1) in terms of f, but eventually we need to know f to find the response $x(t)$.

In working with impulsive forces one normally tries to avoid dealing with the detailed shape of f and tries to limit one's concern to a global quantity known as the **impulse** I of f, the area under its graph. The idea is that if ϵ really is small, then the response $x(t)$, while sensitive to I, should be rather insensitive to the detailed shape of f. That is, if we vary the shape of f but keep its area the same, then we expect little change in the response $x(t)$. This idea suggests that we replace the unknown f by a simple rectangular pulse having the correct impulse as shown in Fig. 1b: $f(t) = I/\epsilon$ for $0 \le t \le \epsilon$, and 0 for $t > \epsilon$. With f thus simplified we can proceed to solve for the response $x(t)$. But even so, the solution still depends upon ϵ, and the latter is probably not known, just as the actual shape of f is not known. Thus, we adopt one more idealization: we suppose that since ϵ is very small, we might as well take the limit of the solution as $\epsilon \to 0$, to eliminate ϵ.

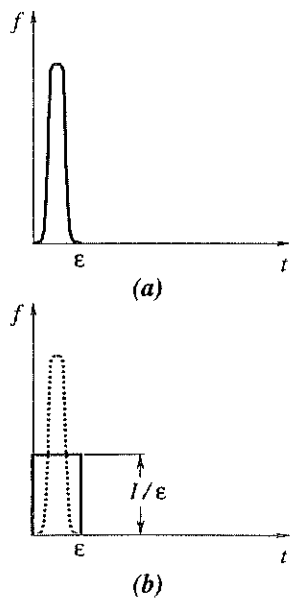


Figure 1. Impulsive force at $t = 0$.

Let us denote such a rectangular pulse having a unit impulse ($I = 1$) as $D(t;\epsilon)$:

$$D(t;\epsilon) = \begin{cases} 1/\epsilon, & 0 \le t \le \epsilon \\ 0, & t > \epsilon, \end{cases} \tag{2}$$

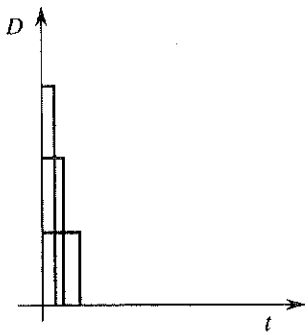


Figure 2. Letting $\epsilon \to 0$ in (2).

where we use D (after the physicist *P. A. M. Dirac*, who developed the idea of impulsive forces in 1929). As $\epsilon \to 0$, D becomes taller and narrower as shown in Fig. 2, in such a way as to maintain its unit area. Of course, the limit

$$\lim_{\epsilon\to 0} D(t;\epsilon) = \begin{cases} \infty, & t = 0 \\ 0, & t > 0 \end{cases} \tag{3}$$

does not exist, because ∞ is not an acceptable value, but Dirac showed that it is nevertheless useful to think of that limiting case as representing an idealized point unit impulse focused at $t = 0$.

To explain, we first prove that

$$\lim_{\epsilon\to 0} \int_0^\infty g(\tau)\, D(\tau;\epsilon)\, d\tau = g(0) \tag{4}$$

for any function g that is continuous at the origin. To begin our proof, write

$$\lim_{\epsilon\to 0} \int_0^\infty g(\tau)\, D(\tau;\epsilon)\, d\tau = \lim_{\epsilon\to 0} \int_0^\epsilon g(\tau)\, \frac{1}{\epsilon}\, d\tau. \tag{5}$$

Suppose that g is continuous on $0 \le \tau \le b$ for some positive b. We can assume that $\epsilon < b$ because we are letting $\epsilon \to 0$. Thus, g is continuous on the integration interval $0 \le \tau \le \epsilon$, so the mean value theorem of the integral calculus tells us that there is a point τ_1 in $[0, \epsilon]$ (i.e., the closed interval $0 \le \tau \le \epsilon$) such that $\int_0^\epsilon g(\tau)\, d\tau = g(\tau_1)\epsilon$. Thus, (5) gives

$$\lim_{\epsilon\to 0} \int_0^\infty g(\tau)\, D(\tau;\epsilon)\, d\tau = \lim_{\epsilon\to 0} \frac{1}{\epsilon} g(\tau_1)\epsilon = \lim_{\epsilon\to 0} g(\tau_1) = g(0), \tag{6}$$

where the last equality holds since τ_1 is in the interval $[0, \epsilon]$, and ϵ is going to zero. Finally, since b is arbitrarily small, we only need the continuity of g *at* $\tau = 0$. This completes our proof of (4).

For brevity, it is customary to dispense with calling attention to the ϵ limit and to express (4) as

$$\int_0^\infty g(\tau)\, \delta(\tau)\, d\tau = g(0), \tag{7}$$

where $\delta(\tau)$ is known as the **Dirac delta function**, or **unit impulse function**. We can think of $\delta(\tau)$ as being zero everywhere except at the origin and infinite at the origin, in such a way as to have unit area, but it must be noted that that definition is not satisfactory within the framework of ordinary function theory. To create a legitimate place for the delta function, one needs to extend the concept of function.

That was done by *L. Schwartz*, and the result is known as the *theory of distributions*, but that theory is well beyond our present scope.

Let us illustrate the application of the delta function with an example.

EXAMPLE 1. Consider (1), with $m = k = 1$ and $c = 0$; let $f(t)$ correspond to a hammer blow as sketched in Fig. 1a, and let $x(0) = x'(0) = 0$, so that before the blow the mass is at rest. The solution of the problem

$$x'' + x = f(t), \qquad x(0) = x'(0) = 0 \tag{8}$$

is found, for instance, by using the Laplace transform, to be

$$x(t) = \int_0^t \sin(t-\tau)\, f(\tau)\, d\tau. \tag{9}$$

As outlined above, the idea is to replace $f(\tau)$ by a rectangular pulse $ID(\tau;\epsilon)$ having the same area I as $f(\tau)$ and then to take the limit as $\epsilon \to 0$:

$$\begin{aligned} x(t) &= \lim_{\epsilon\to 0} \int_0^t \sin(t-\tau)\, ID(\tau;\epsilon)\, d\tau = \lim_{\epsilon\to 0} \int_0^\epsilon \sin(t-\tau)\, \frac{I}{\epsilon}\, d\tau \\ &= \lim_{\epsilon\to 0} I \frac{\cos(t-\epsilon) - \cos t}{\epsilon} = I \sin t, \end{aligned} \tag{10}$$

where the last equality follows from l'Hôpital's rule.

Alternatively and more simply, let $f(\tau) = I\delta(\tau)$ in (9), where the scale factor I is needed since the delta function is a unit impulse whereas we want the impulse to be I. Then property (7) of the delta function gives

$$x(t) = \int_0^t \sin(t-\tau)\, I\delta(\tau)\, d\tau = I \sin(t-\tau)\Big|_{\tau=0} = I \sin t, \tag{11}$$

as obtained previously in (10). You may be concerned that we have applied (7) even though the upper integration limits in (7) and (9) are not the same. However, in (5) we see that the ∞ was immediately changed to ϵ, and then we let ϵ tend to zero. Thus, (7) holds for any positive upper limit; we used ∞ just for definiteness. ∎

Let us review the idea. Since, generally, we know neither the exact shape nor the duration of an impulsive forcing function f, we do two things to solve for the response. We replace f by an equivalent rectangular pulse (i.e., having the same impulse, or area, as f), solve for $x(t)$, and then we let the width of the pulse, ϵ, tend to zero. Equivalently and more simply, we take f to be a Dirac delta function and evaluate the resulting integral using the fundamental property (7) of the delta function. The latter procedure is more efficient because one no longer needs to take the limit of the integral as $\epsilon \to 0$; the limit was already carried out, once and for all, in our derivation of (4).

Since $\delta(t)$ is focused at $t = 0$, it follows that $\delta(t-a)$ is focused at $t = a$, and (7) generalizes to

$$\int_0^\infty g(\tau)\, \delta(\tau - a)\, d\tau = g(a). \tag{12}$$

Here we continue to use the $0, \infty$ limits, but it should be understood that the result is $g(a)$ for any limits A, B (with $B > A$) such that the point of action of the delta function is contained within the interval of integration. If the point $t = a$ falls outside the interval, then the integral is zero. Thus, for reference, we give the following more complete result:*

$$\boxed{\int_A^B g(\tau)\,\delta(\tau - a)\,d\tau = \begin{cases} g(a), & A \le a < B \\ 0, & a < A \quad \text{or} \quad a \ge B. \end{cases}} \tag{13}$$

EXAMPLE 2. *RC Circuit.* Recall from Section 5.5 that the charge $Q(t)$ on the capacitor of the RC circuit is governed by the differential equation $RQ' + (1/C)Q = E(t)$. Let $E(t)$ be an impulsive voltage, with impulse I acting at $t = T$, and let $Q(0) = Q_0$. We wish to solve for $Q(t)$. Expressing $E(t) = I\delta(t - T)$, the initial-value problem is

$$Q' + \kappa Q = I\delta(t - T), \qquad Q(0) = Q_0, \tag{14}$$

where $\kappa = 1/(RC)$. Taking the Laplace transform of (14) gives $s\overline{Q} - Q_0 + \kappa\overline{Q} = IL\{\delta(t - T)\}$ so

$$\overline{Q} = \frac{Q_0}{s + \kappa} + I\frac{1}{s + \kappa}L\{\delta(t - T)\}, \tag{15}$$

and

$$\begin{aligned} Q(t) &= Q_0 e^{-\kappa t} + Ie^{-\kappa t} * \delta(t - T) \\ &= Q_0 e^{-\kappa t} + I\int_0^t e^{-\kappa(t-\tau)}\delta(\tau - T)\,d\tau \\ &= Q_0 e^{-\kappa t} + \begin{cases} 0, & t < T \\ Ie^{-\kappa(t-T)}, & t > T \end{cases} \\ &= Q_0 e^{-\kappa t} + IH(t - T)e^{-\kappa(t-T)}, \end{aligned} \tag{16}$$

where the third equality follows from (13). ∎

Observe that we do not need to know the transform of the delta function in Example 2; we merely call its transform $L\{\delta(t - T)\}$, and inversion by the convolution theorem gives us back the $\delta(t - T)$ that we started with. Nonetheless, for reference, let us work out its Laplace transform. According to (12),

$$L\{\delta(t - a)\} = \int_0^\infty \delta(t - a)\,e^{-st}\,dt = e^{-st}\Big|_{t=a} = e^{-as}. \tag{17}$$

*Following (12), we state that the result is $g(a)$ if the delta function acts within the integration interval. How then do we interpret the integral when a is at an endpoint (A or B)? We've met that case in equation (7). Since the $D(\tau;\epsilon)$ sequence (Fig. 2) is defined on $[0, \epsilon]$, the delta function acts essentially to the right of $\tau = 0$, hence within the interval of integration, and the result of the integration is $g(0)$. To be consistent, let us suppose that the D sequence is always to the right of the point $\tau = a$. Then the integral in (13) will be $g(a)$ if $A \le a < B$ and 0 if $a < A$ or $a \ge B$.

In particular, $L\{\delta(t)\} = 1$.

Since this section is about the Laplace transform, the independent variable has been the time t, so the delta function has represented actions that are focused in time. But the argument of the delta function need not be time. For instance, if $w(x)$ is the load distribution on a beam (Fig. 3a), in pounds per unit length, then $\delta(x-a)$ represents a point unit load (i.e., one pound) at $x = a$ (Fig. 3b).

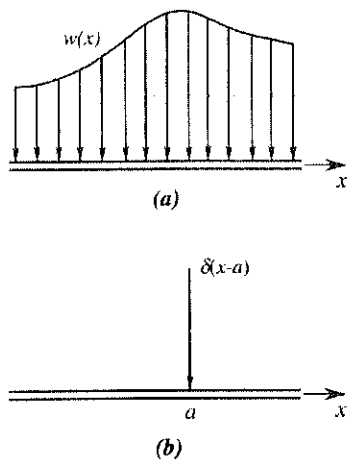


Figure 3. Load distribution on a beam.

Let us close this discussion with a comment on the delta function idealization from a modeling point of view. Consider a metal plate, extending over $-\infty < x < \infty$ and $0 < y < \infty$, loaded by pressing a metal coin against it, at the origin, with a force P (Fig. 4a). If one is to determine (from the theory of elasticity) the stress distribution within the plate, one needs to know the load distribution $w(x)$ along the edge of the plate (namely, the x axis). Because the coin will flatten slightly, at the point of contact, the load $w(x)$ will be distributed over a short interval, say from $x = -\epsilon$ to $x = \epsilon$. However, the function $w(x)$ is not known a priori and its determination is part of the problem. Whether one needs to determine the exact $w(x)$ distribution or if it suffices to represent it simply as an idealized point force of magnitude P, $w(x) = P\delta(x)$, depends upon whether one is interested in the "near field" or the "far field." By the near field we mean that part of the plate within several ϵ lengths of the point of the load application – for instance, within the dashed semicircle shown in Fig. 4b. The far field is the region beyond. If we are concerned only with the far field, then it should suffice to use

$$w(x) = P\delta(x), \tag{18}$$

but if concerned with the near field then the approximation (18) will lead to large errors. A ball bearing manufacturer, for instance, is primarily interested with the near field induced by a loaded ball bearing due to concern regarding wear and surface damage. Within the theory of elasticity, the insensitivity of the far field to the detailed shape of $w(x)$ [given that the area under the $w(x)$ graph is held fixed] is an example of *Saint Venant's principle*.

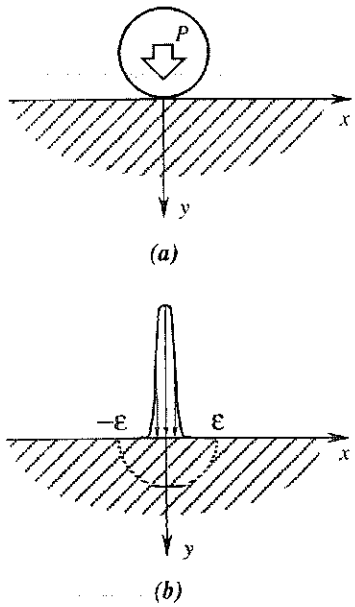


Figure 4. Delta function idealization.

Computer software. The *Maple* name for $\delta(t)$ is Dirac(t).

Closure. We introduce the delta function out of a need to deal effectively with impulsive forcing functions, functions that are highly focused in time or space. Often we know neither the precise form of such a function nor the precise interval of application. If that interval is short enough one can model the force as an idealized point force, represented mathematically as a delta function $\delta(t)$. One is not so much interested in the numerical values of $\delta(t)$ [indeed, one says that $\delta(0) = \infty$] as in the effect of integration upon a delta function, and that effect is expressed by (13), which we regard as the most important formula in this section.

EXERCISES 5.6

1. Solve for $x(t)$, on $0 \le t < \infty$.

(a) $x'' - x = \delta(t-2)$; $x(0) = x'(0) = 0$
(b) $x'' - 4x = 6\delta(t-1)$; $x(0) = 0$, $x'(0) = -3$
(c) $x'' - 3x' + 2x = 2 + \delta(t-5)$; $x(0) = x'(0) = 0$
(d) $x'' + x' = 1 + \delta(t-2)$; $x(0) = 0$, $x'(0) = 3$
(e) $x'' + 2x' + x = 10\delta(t-5)$; $x(0) = x'(0) = 0$
(f) $2x'' - x' = \delta(t-1) - \delta(t-2)$; $x(0) = x'(0) = 0$
(g) $x'' - 3x' + 2x = 100\delta(t-3)$; $x(0) = 4$, $x'(0) = 0$
(h) $x''' = 2\delta(t-5)$; $x(0) = x'(0) = x''(0) = 0$
(i) $x''' + 3x'' + 2x' = \delta(t-5)$; $x(0) = x'(0) = x''(0) = 0$
(j) $x'''' - 4x'' = 3\delta(t-1)$; $x(0) = x'(0) = x''(0) = 0$, $x'''(0) = 1$
(k) $x'''' - 5x'' + 4x = 6\delta(t-2)$; $x(0) = x'(0) = x''(0) = x'''(0) = 0$
(l) $x'''' - x = \delta(t-1)$; $x(0) = x'(0) = x''(0) = x'''(0) = 0$

2. Show that the delta function has these properties, where κ is a nonzero real constant, and the function $f(t)$ is continuous at the origin. NOTE: Recall that the delta function is defined by its integral behavior. Thus, by an equation such as $\delta(-t) = \delta(t)$ we mean that

$$\int_{-\infty}^{\infty} g(t)\,\delta(-t)\,dt = \int_{-\infty}^{\infty} g(t)\,\delta(t)\,dt \tag{2.1}$$

for every function $g(t)$ that is continuous at the origin. The right side of (2.1) is $g(0)$, so to show that $\delta(-t) = \delta(t)$, in part (a), you need to verify that the left side of (2.1) is $g(0)$ too.

(a) $$\delta(-t) = \delta(t) \tag{2.2}$$

(b) $$\delta(\kappa t) = \frac{1}{|\kappa|}\,\delta(t) \quad (\kappa \neq 0) \tag{2.3}$$

(c) $$f(t)\delta(t) = \begin{cases} f(0)\delta(t), & f(0) \neq 0 \\ 0, & f(0) = 0 \end{cases} \tag{2.4}$$

For instance, $(3t+2)\delta(t) = 2\delta(t)$, $(\sin t)\delta(t) = 0$, $(3t+2)\delta(t-1) = 5\delta(t-1)$, and $(t^2+t-2)\delta(t-1) = 0$. Formally, the first part of (2.4) makes sense as follows: $\delta(t)$ is nonzero only at $t = 0$, so there is no difference between $f(t)\delta(t)$ and $f(0)\delta(t)$.

(d) $$\boxed{\int_{-\infty}^{t} \delta(\tau)\,d\tau = H(t)} \tag{2.5}$$

3. The result (2.5), above, reveals the close relation between the delta and Heaviside functions. Alternatively, we can write that relation as

$$\boxed{H'(t) = \delta(t).} \tag{3.1}$$

The latter follows from (2.5) only in a formal sense, but is quite useful, along with (2.2)–(2.5). For instance, suppose we wish to verify that $x(t) = H(t-1)\sin(t-1)$ satisfies the initial-value problem $x'' + x = \delta(t-1)$; $x(0) = x'(0) = 0$. Differentiating $x(t)$ gives

$$\begin{aligned} x'(t) &= H'(t-1)\sin(t-1) + H(t-1)\cos(t-1) \\ &= \delta(t-1)\sin(t-1) + H(t-1)\cos(t-1) \\ &= 0 + H(t-1)\cos(t-1), \end{aligned} \tag{3.2}$$

and

$$\begin{aligned} x''(t) &= H'(t-1)\cos(t-1) - H(t-1)\sin(t-1) \\ &= \delta(t-1)\cos(t-1) - H(t-1)\sin(t-1) \\ &= \delta(t-1) - H(t-1)\sin(t-1) \end{aligned} \tag{3.3}$$

so $x'' + x$ does give $\delta(t-1)$. In the second equality in (3.2) we used (3.1), and in the third we used (2.4): $\delta(t-1)\sin(t-1) = \delta(t-1)\sin 0 = 0$. In the second equality in (3.3) we used (3.1), and in the third we used (2.4): $\delta(t-1)\cos(t-1) = \delta(t-1)\cos 0 = \delta(t-1)$. Further, we see that $x(0) = 0$ and, from (3.2), that $x'(0) = 0$. Here is the problem: In the same manner as above, verify the following solutions that are given in the Answers to the Selected Exercises.

(a) exercise 1(a)
(b) exercise 1(d)
(c) exercise 1(g)

5.7 Additional Properties

In Section 5.3 we establish the linearity of the transform and its inverse, the transform of the derivative $f'(t)$, and the Laplace convolution theorem, results that we deem essential in applying the Laplace transform to the solution of differential equations. In this final section of Chapter 5 we present several additional useful properties of the Laplace transform.

THEOREM 5.7.1 *s-Shift*
If $L\{f(t)\} = F(s)$ exists for $s > s_0$, then for any real constant a,

$$L\left\{e^{-at}f(t)\right\} = F(s+a) \tag{1}$$

for $s + a > s_0$ or, equivalently,

$$L^{-1}\{F(s+a)\} = e^{-at}f(t). \tag{2}$$

Proof:

$$\begin{aligned} L\left\{e^{-at}f(t)\right\} &= \int_0^\infty e^{-at}f(t)\,e^{-st}\,dt \\ &= \int_0^\infty f(t)\,e^{-(s+a)t}\,dt = F(s+a). \quad \blacksquare \end{aligned} \tag{3}$$

EXAMPLE 1. Determine $L\left\{t^3e^{5t}\right\}$. From Appendix C, $L\left\{t^3\right\} = 6/s^4$ so it follows from Theorem 5.7.1 that

$$L\left\{t^3e^{5t}\right\} = \frac{6}{(s-5)^4}. \quad \blacksquare \tag{4}$$

EXAMPLE 2. We can invert $(2s+1)/(s^2+2s+4)$ by partial fractions, but it is simpler to note that

$$\begin{aligned} L^{-1}\left\{\frac{2s+1}{s^2+2s+4}\right\} &= L^{-1}\left\{\frac{2s+1}{(s+1)^2+3}\right\} = L^{-1}\left\{\frac{2(s+1)-1}{(s+1)^2+3}\right\} \\ &= 2L^{-1}\left\{\frac{(s+1)}{(s+1)^2+3}\right\} - L^{-1}\left\{\frac{1}{(s+1)^2+3}\right\} \\ &= 2e^{-t}\cos\sqrt{3}t - e^{-t}\frac{\sin\sqrt{3}t}{\sqrt{3}}, \end{aligned} \tag{5}$$

where in the last step we use entries 3 and 4 in Appendix C and Theorem 5.7.1. $\blacksquare$

THEOREM 5.7.2 *t-Shift*
If $L\{f(t)\} = F(s)$ exists for $s > s_0$, then for any constant $a > 0$

$$L\{H(t-a)f(t-a)\} = e^{-as}F(s) \tag{6}$$

for $s > s_0$ or, equivalently,

$$L^{-1}\left\{e^{-as}F(s)\right\} = H(t-a)f(t-a). \tag{7}$$

Equations (6) and (7) are already given in Section 5.5, where we studied the Heaviside step function, but we repeat them here because the t-shift results seem a natural companion for the s-shift results given in Theorem 5.7.1.

THEOREM 5.7.3 *Multiplication by* $1/s$
If $L\{f(t)\} = F(s)$ exists for $s > s_0$, then

$$L\left\{\int_0^t f(\tau)\,d\tau\right\} = \frac{F(s)}{s} \tag{8}$$

for $s > \max\{0, s_0\}$ or, equivalently,

$$L^{-1}\left\{\frac{F(s)}{s}\right\} = \int_0^t f(\tau)\,d\tau. \tag{9}$$

Proof: This theorem is but a special case of the convolution theorem. Specifically, $\int_0^t f(\tau)\,d\tau = 1 * f$ so, according to that theorem,

$$L\left\{\int_0^t f(\tau)\,d\tau\right\} = L\{1 * f\} = L\{1\}\,L\{f\} = \frac{1}{s}F(s), \tag{10}$$

as asserted. ■

EXAMPLE 3. To evaluate $L^{-1}\left\{1/[s(s^2+1)]\right\}$, for example, we identify $F(s)$ as $1/(s^2+1)$. Since $f(t) = L^{-1}\left\{1/(s^2+1)\right\} = \sin t$,

$$L^{-1}\left\{\frac{1}{s(s^2+1)}\right\} = \int_0^t \sin\tau\,d\tau = 1 - \cos t. \tag{11}$$

Alternatively, we could have used partial fractions. ■

Next, we obtain two useful theorems by differentiating and integrating the definition

$$F(s) = \int_0^\infty f(t)\,e^{-st}\,dt \tag{12}$$

with respect to s. First, we state without proof that if the integral in (12) converges for $s > s_0$, then

$$\frac{dF(s)}{ds} = \frac{d}{ds}\int_0^\infty f(t)\,e^{-st}\,dt = \int_0^\infty \frac{\partial}{\partial s}\left[f(t)\,e^{-st}\right]\,dt$$
$$= -\int_0^\infty t\,f(t)\,e^{-st}\,dt = -L\left\{t\,f(t)\right\}, \tag{13}$$

for $s > s_0$, and

$$\int_a^b F(s)\,ds = \int_a^b \int_0^\infty f(t)\,e^{-st}\,dt\,ds = \int_0^\infty f(t)\left(\int_a^b e^{-st}\,ds\right)\,dt \tag{14}$$

for $b \geq a \geq s_0$. The key step in (13) is the second equality, where we have inverted the order of the integration with respect to t and the differentiation with respect to s. In (14), the key is again the second equality, where we have inverted the order of integration with respect to t and the integration with respect to s.

In particular, if $a = s$ and $b = \infty$, then (14) becomes

$$\int_s^\infty F(s')\,ds' = \int_0^\infty f(t)\left(\int_s^\infty e^{-s't}\,ds'\right)\,dt$$
$$= \int_0^\infty \frac{f(t)}{t}\,e^{-st}\,dt = L\left\{\frac{f(t)}{t}\right\} \tag{15}$$

for all $s > s_0$.

For reference, we state the results (13) and (15) as theorems.

THEOREM 5.7.4 *Differentiation with Respect to s*
If $L\left\{f(t)\right\} = F(s)$ exists for $s > s_0$, then

$$L\left\{t\,f(t)\right\} = -\frac{dF(s)}{ds} \tag{16}$$

for $s > s_0$ or, equivalently,

$$L^{-1}\left\{\frac{dF(s)}{ds}\right\} = -t\,f(t). \tag{17}$$

EXAMPLE 4. From the known transform

$$L\left\{\sin at\right\} = \frac{a}{s^2 + a^2}, \qquad (s > 0) \tag{18}$$

we may use (16) to infer the additional results

$$L\left\{t\,\sin at\right\} = -\frac{d}{ds}\frac{a}{s^2 + a^2} = \frac{2as}{(s^2 + a^2)^2}, \qquad (s > 0) \tag{19}$$

$$L\left\{t^2 \sin at\right\} = -\frac{d}{ds}\frac{2as}{(s^2+a^2)^2} = 2a\frac{3s^2-a^2}{(s^2+a^2)^3}, \qquad (s>0) \tag{20}$$

and so on. ∎

THEOREM 5.7.5 *Integration with Respect to s*
If there is a real number s_0 such that $L\{f(t)\} = F(s)$ exists for $s > s_0$, and $\lim_{t\to 0} f(t)/t$ exists, then

$$L\left\{\frac{f(t)}{t}\right\} = \int_s^\infty F(s')\,ds' \tag{21}$$

for $s > s_0$ or, equivalently,

$$L^{-1}\left\{\int_s^\infty F(s')\,ds'\right\} = \frac{f(t)}{t}. \tag{22}$$

EXAMPLE 5. To evaluate

$$L^{-1}\left\{\ln\frac{s-a}{s-b}\right\}, \tag{23}$$

where a and b are real numbers, note that

$$\frac{d}{ds}\ln\frac{s-a}{s-b} = \frac{1}{s-a} - \frac{1}{s-b}$$

so

$$\ln\frac{s-a}{s-b} - \ln\frac{s_1-a}{s_1-b} = \int_{s_1}^{s}\left(\frac{1}{s'-a} - \frac{1}{s'-b}\right)ds' \tag{24}$$

for any s_1. Letting $s_1 \to \infty$ and recalling that $\ln 1 = 0$, (24) gives

$$\ln\frac{s-a}{s-b} = -\int_s^\infty\left(\frac{1}{s'-a} - \frac{1}{s'-b}\right)ds'. \tag{25}$$

Thus, identify $F(s)$ in (21) as $-1/(s-a) + 1/(s-b)$ (which does exist for $s > \max\{a,b\} \equiv s_0$). Then

$$f(t) = L^{-1}\left\{-\frac{1}{s-a} + \frac{1}{s-b}\right\} = e^{bt} - e^{at}. \tag{26}$$

Furthermore,

$$\lim_{t\to 0}\frac{f(t)}{t} = \lim_{t\to 0}\frac{e^{bt}-e^{at}}{t} = b-a \tag{27}$$

does exist, the last equality following from l'Hôpital's rule, so (22) gives the desired inverse as

$$L^{-1}\left\{\ln\frac{s-a}{s-b}\right\} = \frac{e^{bt}-e^{at}}{t}. \quad ∎ \tag{28}$$

THEOREM 5.7.6 *Large s Behavior of $F(s)$*
Let $f(t)$ be piecewise continuous on $0 \le t \le t_0$ for each finite t_0 and of exponential order as $t \to \infty$. Then
(i) $F(s) \to 0$ as $s \to \infty$,
(ii) $sF(s)$ is bounded as $s \to \infty$.

Proof: Since $f(t)$ is of exponential order as $t \to \infty$, there exist real constants K and c, with $K \ge 0$ such that $|f(t)| \le K e^{ct}$ for all $t \ge t_0$ for some sufficiently large t_0. And since $f(t)$ is piecewise continuous on $0 \le t \le t_0$, there must be a finite constant M such that $|f(t)| \le M$ on $0 \le t \le t_0$. Then

$$\begin{aligned}
|F(s)| &= \left| \int_0^\infty f(t)\, e^{-st}\, dt \right| \le \int_0^\infty |f(t)|\, e^{-st}\, dt \\
&= \int_0^{t_0} |f(t)|\, e^{-st}\, dt + \int_{t_0}^\infty |f(t)|\, e^{-st}\, dt \\
&\le \int_0^{t_0} M\, e^{-st}\, dt + \int_{t_0}^\infty K\, e^{-(s-c)t}\, dt \\
&= M \frac{1 - e^{-st_0}}{s} + K \left. \frac{e^{-(s-c)t}}{-(s-c)} \right|_{t_0}^\infty < \frac{M}{s} + \frac{K}{s-c}
\end{aligned} \tag{29}$$

for all $s > c$. It follows from this result that $F(s) \to 0$ as $s \to \infty$, and that $sF(s)$ is bounded as $s \to \infty$. ■

For instance, for each of the entries 1–7 in Appendix C we do have $F(s) \to 0$ and $sF(s)$ bounded as $s \to \infty$. For entry 8 we do too, unless $-1 < p < 0$, in which case $F(s) \to 0$ but $sF(s)$ is not bounded. However, in this case $f(t) = t^p$ is not piecewise continuous since $f(t) \to \infty$ as $t \to 0$ if p is negative.

THEOREM 5.7.7 *Initial-Value Theorem*
Let f be continuous and f' be piecewise continuous on $0 \le t \le t_0$ for each finite t_0, and let f and f' be of exponential order as $t \to \infty$. Then

$$\lim_{s \to \infty} [s\, F(s)] = f(0). \tag{30}$$

Proof: With the stated assumptions on f and f', it follows from Theorem 5.3.3 that

$$L\{f'(t)\} = s\, L\{f(t)\} - f(0). \tag{31}$$

Since f' satisfies the conditions of Theorem 5.5.7, it follows that $L\{f'(t)\} \to 0$ as $s \to \infty$. Thus, letting $s \to \infty$ in (31) gives the result stated in (30). ■

Normally, we invert $F(s)$ and obtain $f(t)$. However, if it is only $f(0)$ that we desire, not $f(t)$, we do not need to invert $F(s)$; all we need to do, according to (30), is to determine the limit of $sF(s)$ as $s \to \infty$.

As our final item, we show how to transform a periodic function, which is important in applications. First, we define what is meant by a periodic function. If there exists a positive constant T such that

$$f(t+T) = f(t) \tag{32}$$

for all $t \geq 0$, then we say that f is **periodic**, with **period** T.

EXAMPLE 6. The function $\sin t$ is periodic with period 2π since $\sin(t+2\pi) = \sin t \cos 2\pi + \sin 2\pi \cos t = \sin t$, for all t. ∎

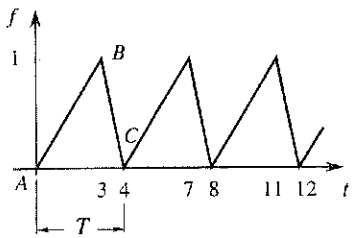


Figure 1. Periodic function f.

EXAMPLE 7. The function f shown in Fig. 1 is, by inspection, seen to be periodic with period $T = 4$, for if we "stamp out" the segment ABC repeatedly, then we generate the graph of f. ∎

Notice that if f is periodic with period T, then it is also periodic with period $2T$, $3T$, $4T$, and so on. For instance, it follows from (32) that

$$f(t+2T) = f((t+T)+T) = f(t+T) = f(t)$$

so that if f is periodic with period T then it is also periodic with period $2T$. If there is a smallest period, it is called the **fundamental period**. Thus, $\sin t$ in Example 6 is periodic with period $2\pi, 4\pi, 6\pi, \ldots$, so its fundamental period is 2π; $f(t)$ in Example 7 is periodic with period $4, 8, 12, \ldots$, so its fundamental period is 4. In contrast, $f(t) = 3$ (i.e., a constant) is periodic with period T for *every* $T > 0$. Thus, there is no smallest period, and hence this f does not have a fundamental period.

To evaluate the Laplace transform of a periodic function $f(t)$, with period T (which is normally taken to be the fundamental period of f, if f has a fundamental period), it seems like a good start to break up the integral on t as

$$L\{f(t)\} = \int_0^\infty f(t)\, e^{-st}\, dt = \int_0^T f(t)\, e^{-st}\, dt + \int_T^{2T} f(t)\, e^{-st}\, dt + \cdots. \tag{33}$$

Next, let $\tau = t$ in the first integral on the right side of (33), $\tau = t - T$ in the second, $\tau = t - 2T$ in the third, and so on. Thus,

$$\begin{aligned} L\{f(t)\} = &\int_0^T f(\tau)\, e^{-s\tau}\, d\tau + \int_0^T f(\tau+T)\, e^{-s(\tau+T)}\, d\tau \\ &+ \int_0^T f(\tau+2T)\, e^{-s(\tau+2T)}\, d\tau + \cdots, \end{aligned} \tag{34}$$

but $f(\tau+T)=f(\tau)$, $f(\tau+2T)=f(\tau)$, and so on, so (34) becomes

$$L\{f(t)\}=\left(1+e^{-sT}+e^{-2sT}+\cdots\right)\int_0^T f(\tau)\,e^{-s\tau}\,d\tau. \tag{35}$$

Unfortunately, this expression is an infinite series. However, observe that

$$1+e^{-sT}+e^{-2sT}+\cdots=1+\left(e^{-sT}\right)+\left(e^{-2sT}\right)^2+\cdots \tag{36}$$

is a geometric series $1+z+z^2+\cdots$, with $z=e^{-sT}$, and the latter is known to have the sum $1/(1-z)$ if $|z|<1$. Since $|z|=\left|e^{-sT}\right|=e^{-sT}<1$ if $s>0$, we can sum the parenthetic series in (35) as $1/(1-e^{-sT})$.

Finally, if we ask that f be piecewise continuous on $0\le t\le T$, to ensure the existence of the integral in (35), then we can state the result as follows.

THEOREM 5.7.8 *Transform of Periodic Function*
If f is periodic with period T on $0\le t<\infty$ and piecewise continuous on one period, then

$$L\{f(t)\}=\frac{1}{1-e^{-sT}}\int_0^T f(t)\,e^{-st}\,dt \tag{37}$$

for $s>0$.

The point, of course, is that (37) requires integration only over one period, rather than over $0\le t<\infty$, and gives the transform in closed form rather than as an infinite series.

EXAMPLE 8. If f is the sawtooth wave shown in Fig. 2, then $T=2$, and

$$\int_0^T f(t)\,e^{-st}\,dt=\int_0^2 2t\,e^{-st}\,dt=2\frac{1-(1+2s)e^{-2s}}{s^2}, \tag{38}$$

so

$$L\{f(t)\}=\frac{1}{1-e^{-2s}}\,\frac{2\left[1-(1+2s)e^{-2s}\right]}{s^2}=\frac{2}{s^2}-\frac{4}{s}\,\frac{e^{-2s}}{1-e^{-2s}} \tag{39}$$

for $s>0$.

A more interesting question is the reverse: What is the inverse of

$$F(s)=\frac{2}{s^2}-\frac{4}{s}\,\frac{e^{-2s}}{1-e^{-2s}}, \tag{40}$$

where we pretend no advance knowledge of the sawtooth wave in Fig. 2, or even the knowledge that $f(t)$ is periodic? The key is to proceed in reverse – that is, to expand the $1/(1-e^{-2s})$ in a geometric series in powers of e^{-2s}. Thus

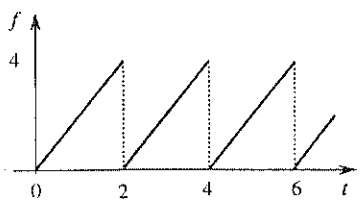


Figure 2. Sawtooth wave.

$$\begin{aligned}F(s)&=\frac{2}{s^2}-\frac{4}{s}e^{-2s}\left(1+e^{-2s}+e^{-4s}+\cdots\right)\\&=\frac{2}{s^2}-\frac{4}{s}\left(e^{-2s}+e^{-4s}+e^{-6s}+\cdots\right).\end{aligned} \tag{41}$$

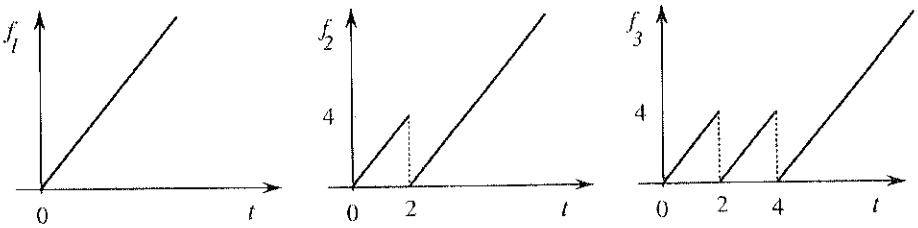


Figure 3. Partial sums of (42).

Assuming that the series can be inverted termwise,

$$f(t) = 2t - 4\left[H(t-2) + H(t-4) + H(t-6) + \cdots\right]. \tag{42}$$

The first few partial sums,

$$\begin{aligned} f_1(t) &= 2t, \\ f_2(t) &= 2t - 4H(t-2), \\ f_2(t) &= 2t - 4H(t-2) - 4H(t-4) \end{aligned}$$

are sketched in Fig. 3, and it is easy to infer that (42) gives the periodic sawtooth wave shown in Fig. 2.

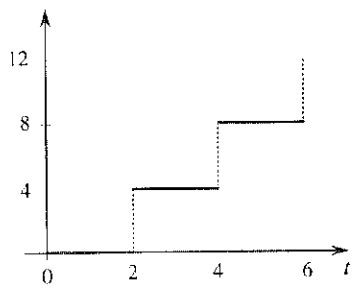


Figure 4. The staircase $4[H(t-2) + H(t-4) + \cdots]$.

COMMENT. Observe that the presence of $1 - e^{-sT}$ in the denominator of a transform does not suffice to imply that the inverse is periodic. For example, the inverse of $4e^{-2s}/[s(1 - e^{-2s})]$, in (40), is the nonperiodic "staircase" shown in Fig. 4. It is only when this staircase is subtracted from $2t$ that a periodic function results. ∎

This completes our discussion of the Laplace transform. Just as we used it, in this chapter, to solve linear ordinary differential equation initial-value problems, in later chapters we will use it to solve linear partial differential equation initial-value problems.

Closure. It would be difficult to pick out the one or two most important results in this section, since there are eight theorems, all of comparable importance. Most of these theorems are included as entries within Appendix C.

EXERCISES 5.7

1. Invert each of the following by any method. Cite any items from Appendix C and any theorems that you use. If it is applicable, verify the initial-value theorem, namely, that $sF(s) \to f(0)$ as $s \to \infty$; if it is not, then state why it is not.

(a) $\dfrac{1}{(s^2+a^2)^2}$ (b) $\dfrac{1}{(s^2-a^2)^2}$ (c) $\dfrac{s}{(s-2)^2}$

(d) $\dfrac{s^2}{(s+1)^3}$ (e) $\dfrac{1}{\sqrt{s+1}}$ (f) $\dfrac{s}{(s-a)^{3/2}}$

(g) $\dfrac{e^{-s}}{s^5}$ (h) $\dfrac{s}{(s+4)^6}$ (i) $\dfrac{e^{-2s}}{(s+1)^2}$

(j) $\ln\left(1+\dfrac{a^2}{s^2}\right)$ (k) $\dfrac{e^{-\pi s}}{(s+2)^4}$ (l) $\ln\left(1-\dfrac{a^2}{s^2}\right)$

(m) $\dfrac{e^{-s}}{s^2+s+1}$ (n) $\dfrac{e^{-3s}}{s^2+2s-4}$ (o) $\dfrac{e^{-2s}}{s^2(s^2-2s-2)}$

(p) $\ln\left(\dfrac{s^2+1}{s^2+s}\right)$ (q) $\dfrac{1}{s(1-e^{-s})}$ (r) $\dfrac{1}{s(1+e^{-s})}$

(s) $\dfrac{1}{s}\tanh s$ (t) $\dfrac{1}{s^2(1-e^{-s})}$ (u) $\dfrac{1}{(s+1)(1-e^{-2s})}$

(v) $\dfrac{e^{-s}}{s-8}$ (w) $\dfrac{s}{(s^2+1)^{3/2}}$ (x) $\dfrac{1}{s(1-e^{s})}$

2 (a)–(x) Invert the transform given in the corresponding part of Exercise 1, using computer software.

3. (a) In the simple case where $f(t) = \sin t$, show that (37) does indeed give

$$L\{\sin t\} = \frac{1}{s^2+1}.$$

(b) In the case where $f(t) = \cos t$, show that (37) does indeed give

$$L\{\cos t\} = \frac{s}{s^2+1}.$$

4. (*Scale changes*) Show that if $L\{f(t)\} = F(s)$, then

(a) $L\{f(at)\} = \dfrac{1}{a}F\left(\dfrac{s}{a}\right)$ (b) $L^{-1}\{F(as)\} = \dfrac{1}{a}f\left(\dfrac{t}{a}\right)$

5. Determine the Laplace transform of the function $f(t)$ that is periodic and defined on one period as follows.

(a) $\sin t, \quad 0 \le t \le \pi$ (b) $\begin{cases} 1, & 0 \le t < 2 \\ 0, & 2 \le t < 3 \end{cases}$

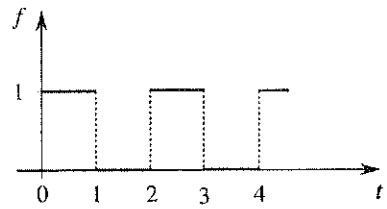
(c) $\sin 2t, \quad 0 \le t \le \pi$ (d) $e^{-t}, \quad 0 \le t < 2$

(e) $\begin{cases} t, & 0 \le t < 1 \\ 0, & 1 \le t < 2 \end{cases}$ (f) $\begin{cases} t, & 0 \le t < 1 \\ t-2, & 1 \le t < 2 \end{cases}$

(g) $\begin{cases} 2, & 0 \le t < 1 \\ 4, & 1 \le t < 2 \\ 1, & 2 \le t < 3 \end{cases}$

6. (a) Solve $x' + x = f(t)$ by the Laplace transform, where $x(0) = x_0$ and $f(t)$ is the square wave shown, and show that the solution is

$$\begin{aligned} x(t) = x_0e^{-t} &+ \left[1-e^{-(t-0)}\right]H(t-0) \\ &- \left[1-e^{-(t-1)}\right]H(t-1) \\ &+ \left[1-e^{-(t-2)}\right]H(t-2) - \cdots. \end{aligned} \tag{6.1}$$

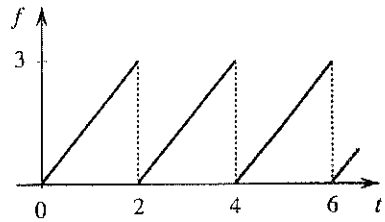


HINT: It would be wasteful to determine $F(s)$ because the solution can be expressed as a convolution integral involving $f(t)$ directly. In that integral, express $f(t)$ as an infinite series of Heaviside functions.

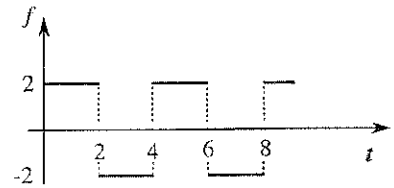
(b) Sketch and label the graph of $x(t)$ over $0 < t < 3$, say, for the case where $x_0 = 0$. Is $x(t)$ periodic? If not, is there a value of x_0 such that $x(t)$ is periodic? Explain.

7. Solve $x' + x = f(t)$ by the Laplace transform, where $x(0) = x_0$ and $f(t)$ is the periodic function shown. HINT: Read Exercise 6.

(a)



(b)



8. Solve $x'' + x = f(t)$ by the Laplace transform, where $x(0) = x_0$, $x'(0) = x_0'$, and $f(t)$ is the square wave shown in Exercise 6. Evaluate $x(5)$ if $x_0 = x_0' = 1$.

Chapter 5 Review

The Laplace transform has a variety of uses, but its chief application is in the solution of linear ordinary and partial differential differential equations. In this chapter our focus is on its use in solving linear ordinary differential equations with constant coefficients. The power of the method is that it reduces such a differential equation, homogeneous or not, to a linear algebraic one. The hardest part, the inversion, is often accomplished with the help of tables, a considerable number of theorems, and computer software. Also, any initial conditions that are given become built in, in the process of transforming the differential equation, so they do not need to be applied at the end.

Chief properties given in Section 5.3 are:

Linearity of the transform and its inverse

$$L\{\alpha u(t) + \beta v(t)\} = \alpha L\{u(t)\} + \beta L\{v(t)\},$$

$$L^{-1}\{\alpha U(s) + \beta V(s)\} = \alpha L^{-1}\{U(s)\} + \beta L^{-1}\{V(s)\},$$

Transform of derivatives

$$L\{f'\} = sF(s) - f(0), \quad L\{f''\} = s^2F(s) - sf(0) - f'(0), \quad \cdots,$$

Convolution Theorem

$$L\{(f * g)(t)\} = F(s)G(s)$$

or

$$L^{-1}\{F(s)G(s)\} = (f * g)(t),$$

where

$$(f * g)(t) = \int_0^t f(\tau)\, g(t-\tau)\, d\tau$$

is the Laplace convolution of f and g.

In Sections 5.5 and 5.6 we introduce the step and impulse functions $H(t-a)$ and $\delta(t-a)$, defined by

$$H(t-a) = \begin{cases} 0, & t < a \\ 1, & t \geq a \end{cases}$$

and

$$\int_A^B g(t)\,\delta(t-a)\,dt = \begin{cases} g(a), & A \leq a < B \\ 0, & a < A \quad \text{or} \quad a \geq B, \end{cases}$$

to model piecewise-defined and impulsive forcing functions.

Finally, in Section 5.7 we derive additional properties:

s-shift

$$L\{e^{-at}f(t)\} = F(s+a)$$

or

$$L^{-1}\{F(s+a)\} = e^{-at}f(t).$$

t-shift

$$L\{H(t-a)f(t-a)\} = e^{-as}F(s)$$

or

$$L^{-1}\left\{e^{-as}F(s)\right\} = H(t-a)f(t-a).$$

Multiplication by $1/s$

$$L\left\{\int_0^t f(\tau)\,d\tau\right\} = \frac{F(s)}{s}$$

or

$$L^{-1}\left\{\frac{F(s)}{s}\right\} = \int_0^t f(\tau)\,d\tau.$$

Differentiation with respect to s

$$L\{t\,f(t)\} = -\frac{dF(s)}{ds}$$

or

$$L^{-1}\left\{\frac{dF(s)}{ds}\right\} = -t\,f(t).$$

Integration with respect to s

$$L\left\{\frac{f(t)}{t}\right\} = \int_s^\infty F(s')ds'$$

or

$$L^{-1}\left\{\int_s^\infty F(s')ds'\right\} = \frac{f(t)}{t}.$$

Large s behavior of $F(s)$

$$F(s) \to 0 \quad \text{as} \quad s \to \infty,$$
$$sF(s) \quad \text{bounded as} \quad s \to \infty.$$

Transform of periodic function of period T

$$L\{f(t)\} = \frac{1}{1-e^{-sT}}\int_0^T f(t)e^{-st}dt.$$

NOTE: The preceding list is intended as an overview so, for brevity, the various conditions under which these results hold have been omitted.

Chapter 6

Quantitative Methods: Numerical Solution of Differential Equations

6.1 Introduction

Following the introduction in Chapter 1, Chapters 2–5 cover both the underlying theory of differential equations and analytical solution techniques as well. That is, the objective thus far has been to find an analytical solution – in closed form if possible or as an infinite series if necessary. Unfortunately, a great many differential equations encountered in applications, and most nonlinear equations in particular, are simply too difficult for us to find analytical solutions.

Thus, in Chapters 6 and 7 our approach is fundamentally different, and complements the analytical approach adopted in Chapters 2–5: in Chapter 6 we develop *quantitative* methods, and in Chapter 7 our view is essentially *qualitative*. More specifically, in this chapter we "discretize" the problem and seek, instead of an analytical solution, the numerical values of the dependent variable at a discrete set of values of the independent variable so that the result is a table or graph, with those values determined approximately (but accurately), rather than exactly.

Perhaps the greatest drawback to numerical simulation is that whereas an analytical solution explicitly displays the dependence of the dependent variable(s) on the various physical parameters (such as spring stiffnesses, driving frequencies, electrical resistances, inductances, and so on), one can carry out a numerical solution only for a specific set of values of the system parameters. Thus, parametric studies (i.e., studies of the qualitative and quantitative effects of the various parameters upon the solution) can be tedious and unwieldy, and it is useful to reduce the number of parameters as much as possible (by nondimensionalization, as discussed in Section 2.4.4) before embarking upon a numerical study.

The numerical solution of differential equations covers considerable territory so the present chapter is hardly complete. Rather, we aim at introducing the funda-

mental ideas, concepts, and potential difficulties, as well as specific methods that are accurate and readily implemented. We do mention computer software that carries out these computations automatically, but our present aim is to provide enough information so that you will be able to select a specific method and program it. In contrast, in Chapter 7, where we look more at qualitative issues, we rely heavily upon available software.

6.2 Euler's Method

In this section and the two that follow, we study the numerical solution of the first-order initial-value problem

$$y' = f(x, y); \quad y(a) = b \tag{1}$$

on $y(x)$.

To motivate the first and simplest of these methods, Euler's method, consider the problem

$$y' = y + 2x - x^2; \quad y(0) = 1 \quad (0 \le x < \infty) \tag{2}$$

with the exact solution (Exercise 1)

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

Of course, in practice one wouldn't solve (2) numerically because we can solve it analytically and obtain the solution (3), but we will use (2) as an illustration.

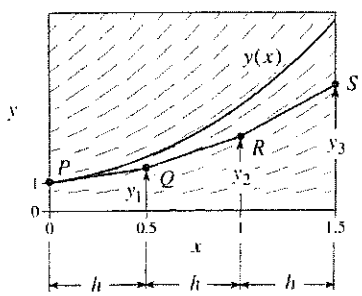


Figure 1. Direction field motivation of Euler's method, for the initial-value problem (2).

In Fig. 1 we display the direction field defined by $f(x, y) = y + 2x - x^2$, as well as the exact solution (3). In graphical terms, Euler's method amounts to using the direction field as a road map in developing an approximate solution to (2). Beginning at the initial point P, namely $(0, 1)$, we move in the direction dictated by the lineal element at that point. As seen from the figure, the farther we move along that line, the more we expect our path to deviate from the exact solution. Thus, the idea is not to move very far. Stopping at $x = 0.5$, say, for the sake of illustration, we revise our direction according to the slope of the lineal element at that point Q. Moving in that new direction until $x = 1$, we revise our direction at R, and so on, moving in x increments of 0.5. We call the x increment the **step size** and denote it as h. In Fig. 1, h is 0.5.

Let us denote the y values at $Q, R, \ldots$ as $y_1, y_2, \ldots$. They are computed as $y_1 = y_0 + f(x_0, y_0)h$, $y_2 = y_1 + f(x_1, y_1)h, \ldots$, where (x_0, y_0) is the initial point P. Expressed as a numerical algorithm, the **Euler method** is therefore as follows:

$$\boxed{y_{n+1} = y_n + f(x_n, y_n)h, \qquad (n = 0, 1, 2, \ldots)} \tag{4}$$

where f is the function on the right side of the given differential equation (1), $x_0 = a$, $y_0 = b$, h is the chosen step size, and $x_n = x_0 + nh$.

Euler's method is also known as the **tangent-line method** because the first straight-line segment of the approximate solution is tangent to the exact solution

$y(x)$ at P, and each subsequent segment emanating from (x_n, y_n) is tangent to the solution curve through that point.

Apparently, the greater the step size the less accurate the results, in general. For instance, the first point Q deviates more and more from the exact solution as the step size is increased – that is, as the segment PQ is extended. Conversely, we expect the approximate solution to approach the exact solution curve as h is reduced. This expectation is supported by the results shown in Table 1 for the initial-

Table 1. Comparison of numerical solution of (2) using Euler's method, with the exact solution.

x	$h = 0.5$	$h = 0.1$	$h = 0.02$	$y(x)$
0.5	1.5000	1.7995	1.8778	1.8987
1.0	2.6250	3.4344	3.6578	3.7183
1.5	4.4375	6.1095	6.5975	6.7317

value problem (2), obtained by Euler's method with step sizes of $h = 0.5, 0.1$, and 0.02; we have included the exact solution $y(x)$, given by (3), for comparison. With $h = 0.5$, for instance,

$$y_1 = y_0 + \left(y_0 + 2x_0 - x_0^2\right) h = 1 + (1 + 0 - 0)(0.5) = 1.5,$$
$$y_2 = y_1 + \left(y_1 + 2x_1 - x_1^2\right) h = 1.5 + (1.5 + 1 - 0.25)(0.5) = 2.625,$$
$$y_3 = y_2 + \left(y_2 + 2x_2 - x_2^2\right) h = 2.625 + (2.625 + 2 - 1)(0.5) = 4.4375.$$

With $h = 0.1$, the values tabulated at $x = 0.5, 1.0, 1.5$ are y_5, y_{10}, y_{15}, with the intermediate computed y values omitted for brevity.

Scanning each row of the tabulation, we can see that the approximate solution appears to be converging to the exact solution as $h \to 0$ (though we cannot be certain from such results no matter how small we make h), and that the convergence is not very rapid, for even with $h = 0.02$ the computed value at $x = 1.5$ is in error by 2%.

As strictly computational as this sounds, two important theoretical questions present themselves: Does the method converge to the exact solution as $h \to 0$ and, if so, how fast? By a method being **convergent** we mean that for any fixed x value in the x interval of interest the sequence of y values, obtained using smaller and smaller step size h, tends to the exact solution $y(x)$ as $h \to 0$.

Let us see whether the Euler method is convergent. Observe that there are two sources of error in the numerical solution. One is the tangent-line approximation upon which the method is based, and the other is the accumulation of numerical roundoff errors within the computing machine since a machine can carry only a finite number of significant figures, after which it rounds off (or chops off, depending upon the machine). In discussing convergence, one ignores the presence of such roundoff error and considers it separately. Thus, in this discussion we imagine our computer to be perfect, carrying an infinite number of significant figures.

Local truncation error. Although we are interested in the accumulation of error after a great many steps have been carried out, to reach any given x, it seems best to begin by investigating the error incurred in a single step, from x_{n-1} to x_n (or from x_n to x_{n+1}, it doesn't matter). We need to distinguish between the exact and approximate solutions so let us denote the exact solution at x_n as $y(x_n)$ and the approximate numerical solution at x_n as y_n. These are given by the Taylor series

$$\begin{aligned} y(x_n) &= y(x_{n-1}) + y'(x_{n-1})\,(x_n - x_{n-1}) + \frac{y''(x_{n-1})}{2!}\,(x_n - x_{n-1})^2 + \cdots \\ &= y(x_{n-1}) + y'(x_{n-1})h + \frac{y''(x_{n-1})}{2!}h^2 + \cdots \end{aligned} \tag{5}$$

and the Euler algorithm

$$y_n = y_{n-1} + f\,(x_{n-1}, y_{n-1})\,h, \tag{6}$$

respectively. It is important to understand that the Euler method (6) amounts to retaining only the first two terms of the Taylor series in (5). Thus, it replaces the actual function by its tangent-line approximation.

We suppose that $y(x_{n-1})$ and y_{n-1} are identical, and we ask how large the error $e_n \equiv y(x_n) - y_n$ is after making that single step, from x_{n-1} to x_n. We can get an expression for e_n by subtracting (6) from (5), but the right side will be an infinite series. Thus, it is more convenient to use, in place of the (infinite) Taylor series (5), the (finite) Taylor's formula with remainder,

$$y(x_n) = y(x_{n-1}) + y'(x_{n-1})h + \frac{y''(\xi)}{2!}h^2, \tag{7}$$

where ξ is some point in the interval $[x_{n-1}, x_n]$. Now, subtracting (6) from (7), and noting that $y'(x_{n-1}) = f\,[x_{n-1}, y(x_{n-1})] = f(x_{n-1}, y_{n-1})$ because of our supposition that $y(x_{n-1}) = y_{n-1}$, gives

$$e_n = \frac{y''(\xi)}{2}h^2. \tag{8}$$

The latter expression for e_n is awkward to apply since we don't know ξ, except that $x_{n-1} \le \xi \le x_n$.* However, (8) is of more interest in that it shows how the single-step error e_n varies with h. Specifically, since $x_{n-1} \le \xi \le x_{n-1} + h$, we see that as $h \to 0$ we have $\xi \to x_{n-1}$, so (8) gives $e_n \sim \dfrac{y''(x_{n-1})}{2}h^2 = Ch^2$ as $h \to 0$, where C is a constant. Accordingly, we say that e_n is of order h^2 and write

$$\boxed{e_n = O(h^2)} \tag{9}$$

*It also appears that we do not know the y'' function, but it follows from (2) that $y'' = y' + 2 - 2x = (y + 2x - x^2) + 2 - 2x = y + 2 - x^2$.

as $h \to 0$. [The big oh notation is defined in Section 4.5, and (9) simply means that $e_n \sim Ch^2$ as $h \to 0$ for some nonzero constant C.]

Since the error e_n is due to truncation of the Taylor series it is called the truncation error – more specifically, the **local truncation error** because it is the truncation error incurred in a single step.

Accumulated truncation error and convergence. Of ultimate interest, however, is the truncation error that has accumulated over *all* of the preceding steps since that error is the difference between the exact solution and the computed solution at any given x_n. We denote it as $E_n \equiv y(x_n) - y_n$ and call it the **accumulated truncation error**. If it seems strange that we have defined both the local and accumulated truncation errors as $y(x_n) - y_n$, it must be remembered that the former is that which results from a single step (from x_{n-1} to x_n) whereas the latter is that which results from the entire sequence of steps (from x_0 to x_n).

We can estimate E_n at a fixed x location (at least insofar as its order of magnitude) as the local truncation error e_n times the number of steps n. Since $e_n = O(h^2)$, this idea gives

$$E_n = O(h^2) \cdot n = O(h^2)\frac{n\,h}{h} = O(h^2)\frac{x_n}{h} = O(h) \cdot x_n = O(h), \qquad (10)$$

where the last step follows from the fact that the selected location x_n is held fixed as $h \to 0$. Since the big oh notation is insensitive to scale factors, the x_n factor can be absorbed by the $O(h)$ so

$$\boxed{E_n = O(h),} \qquad (11)$$

which result tells us how fast the numerical solution converges to the exact solution (at any fixed x location) as $h \to 0$. Namely, $E_n \sim Ch$ for some nonzero constant C. To illustrate, consider the results shown in Table 1, and consider $x = 1.5$, say, in particular. According to $E_n \sim Ch$, if we reduce h by a factor of five, from 0.1 to 0.02, then likewise we should reduce the error by a factor of five. We find that $(6.7317 - 6.1095)/(6.7317 - 6.5875) \approx 4.6$, which is indeed close to five. We can't expect it to be exactly five for two reasons: First, (11) holds only as $h \to 0$, whereas we have used $h = 0.1$ and 0.02. Second, we obtained the values in Table 1 using a computer, and a computer introduces an additional error, due to roundoff, which has not been accounted for in our derivation of (11). Probably it is negligible in this example.

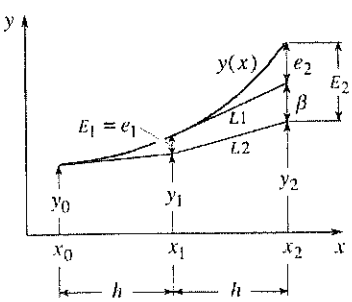


Figure 2. The global truncation error.

While (11) can indeed be proved rigorously, be aware that our reasoning in (10) was only heuristic. To understand the shortcoming of our logic, consider the diagram in Fig. 2, where we show only two steps, for simplicity.

Our reasoning, in writing $E_n = O(h^2) \cdot n$ in (10), is that the accumulated truncation error E_n is (at least insofar as order of magnitude) the sum of the n single-step errors. However, that is not quite true. We see from Fig. 2 that E_2 is $e_2 + \beta$, not the sum of the single-step errors $e_2 + e_1$, and β is not identical to e_1. The difference between β and e_1 is the result of the *slightly* different slopes of $L1$ and $L2$ acting over the *short* distance h, and that difference can be shown to be a higher-order effect that does not invalidate the final result that $E_n = O(h)$, provided that

f is well enough behaved (for example, if f, f_x, and f_y are all continuous on the x, y region of interest).

In summary, (11) shows that the Euler method (4) is convergent because the accumulated truncation error tends to zero as $h \to 0$. More generally if, for a given method, $E_n = O(h^p)$ as $h \to 0$, then the method is convergent if $p > 0$, and we say that it is **of order p**. Thus, *the Euler method is a first-order method.*

Although convergent and easy to implement, Euler's method is usually too inaccurate for serious computation because it is only a first-order method. That is, since the accumulated truncation error is proportional to h to the first power, we need to make h extremely small if the error is to be extremely small. Why can't we do that? Why can't we merely let $h = 10^{-8}$, say? There are two important reasons. One is that with $h = 10^{-8}$, it would take 10^8 steps to generate the Euler solution over a unit x interval. That number of steps might simply be impractically large in terms of computation time and expense.

Second, besides the truncation error that we have discussed there is also machine roundoff error, and that error can be expected to grow with the number of calculations. Thus, as we diminish the step size h and increase the number of steps, to reduce the truncation error, we inflict a roundoff error penalty that diminishes the intended increase in accuracy. In fact, we can anticipate the existence of an optimal h value so that to decrease h below that value is counterproductive. Said differently, a given level of accuracy may prove unobtainable because of the growth in the roundoff error as h is reduced. Further discussion of this point is contained in the exercises.

Finally, there is an important practical question not yet addressed: How do we know how small to choose h? We will have more to say about this later, but for now let us give a simple procedure, namely, reducing h until the results settle down to the desired accuracy. For instance, suppose we solve (2) by Euler's method using $h = 0.5$, as a first crack. Pick any fixed point x in the interval of interest, such as $x = 1.5$. The computed solution there is 4.4375. Now reduce h, say to 0.01, and run the program again. The result this time, at $x = 1.5$, is 6.1095. Since those results differ considerably, reduce h again, say to 0.02, and run the program again. Simpy repeat that procedure until the solution at $x = 1.5$ settles down to the desired number of significant figures. Accept the results of the final run, and discard the others. (Of course, one will not have an exact solution to compare with as we did in Table 1.)

The foregoing idea is merely a rule of thumb, and is the same idea that we use in computing an infinite series: keep adding more and more terms until successive partial sums agree with the desired number of significant figures.

Closure. The Euler method is embodied in (4). It is easy to implement, either using a hand-held calculator or programming it to be run on a computer. The method is convergent but only of first order and hence is not very accurate. Thus, it is important to develop more accurate methods, and we do that in the next section.

We also use our discussion of the Euler method to introduce the concept of the local and accumulated truncation errors e_n and E_n, respectively, which are

due to the approximate discretization of the problem and which have nothing to do with additional errors that enter due to machine roundoff. The former is the error incurred in a single step, and the latter is the accumulated error over the entire caclulation. Finally, we define the method to be convergent if the accumulated truncation error E_n tends to zero at any given fixed point x, as the step size h tends to zero, and of order p if $E_n = O(h^p)$ as $h \to 0$. The Euler method is convergent and of order one.

EXERCISES 6.2

1. Derive the particular solution (3) of the initial-value problem (2).

2. Use the Euler method to compute, by hand, y_1, y_2, and y_3 for the specified initial-value problem using $h = 0.2$.

(a) $y' = -y; \quad y(0) = 1$
(b) $y' = 2xy; \quad y(0) = 0$
(c) $y' = 3x^2y^2; \quad y(0) = 0$
(d) $y' = 1 + 2xy^2; \quad y(1) = -2$
(e) $y' = 2xe^{-y}; \quad y(1) = -1$
(f) $y' = x^2 - y^2; \quad y(3) = 5$
(g) $y' = x \sin y; \quad y(0) = 0$
(h) $y' = \tan(x + y); \quad y(1) = 2$
(i) $y' = 5x - 2\sqrt{y}; \quad y(0) = 4$
(j) $y' = \sqrt{x + y}; \quad y(0) = 3$

3. Program and run Euler's method for the initial-value problem $y' = f(x, y)$, with $y(0) = 1$ and $h = 0.1$, through y_{10}. Print $y_1, \ldots, y_{10}$ and the exact solution $y(x_1), \ldots, y(x_{10})$ as well. (Six significant figures will suffice.) Evaluate E_{10}. Use the $f(x, y)$ specified below.

(a) $2x$	(b) $-6y^2$	(c) $x + y$
(d) $y \sin x$	(e) $(y^2 + 1)/2$	(f) $4xe^{-y}$
(g) $1 + x^2 + y$	(h) $-y \tan x$	(i) e^{x+y}

4. (a)–(h) Program and run Euler's method for the initial-value problem $y' = f(x, y)$ (with f given in the corresponding part of Exercise 3), and print out the result at $x = 0.5$. Use $h = 0.1$, then 0.05, then 0.01, then 0.005, then 0.001, and compute the accumulated truncation error at $x = 0.5$ for each case. Is the rate of decrease of the accumulated truncation error, as h decreases, consistent with the fact that Euler's method is a first-order method? Explain.

5. Thus far we have taken the step h to be positive, and therefore developed a solution to the right of the initial point. Is Euler's method valid if we use a negative step, $h < 0$, and hence develop a solution to the left? Explain.

6. We have seen that by discretizing the problem, we can approximate the solution $y(x)$ of a differential equation $y' = f(x, y)$ by a discrete variable y_n by solving

$$y_{n+1} = y_n + f(x_n, y_n)h \tag{6.1}$$

sequentially, for $n = 0, 1, 2, \ldots$. Besides being a numerical algorithm for the calculation of the y_n's, (6.1) is an example of a **difference equation** governing the sequence of y_n's, just as $y' = f(x, y)$ is a differential equation governing $y(x)$. If f is simple enough it may be possible to solve (6.1) for y_n analytically, and that idea is the focus of this exercise. Specifically, consider $y' = Ay$, where A is a given constant. Then (6.1) becomes

$$y_{n+1} = (1 + Ah)y_n. \tag{6.2}$$

(a) Derive the solution

$$y_n = C(1 + Ah)^n \tag{6.3}$$

of (6.2), where C is the initial value y_0, if one is specified.
(b) Show that as $h \to 0$ (6.3) does converge to the solution Ce^{Ax} of the original equation $y' = Ay$. HINT: Begin by expressing $(1 + Ah)^n$ as $e^{\ln(1+Ah)^n}$. NOTE: Thus, for the simple differential equation $y' = Ay$ we have been able to prove the convergence of the Euler method by actually solving (6.2) for y_n, in closed form, then taking the limit of that result as $h \to 0$.
(c) Use computer software to obtain the solution (6.3) of the difference equation (6.2). On *Maple*, for instance, use the rsolve command.

7. In this section we have taken the step size h to be a constant from one step to the next. Is there any reason why we could not vary h from one step to the next? Explain.

6.3 Improvements: Midpoint Rule and Runge–Kutta

Our objective in this section is to develop more accurate methods than the first-order Euler method – namely, higher-order methods. In particular, we are aiming at the widely used fourth-order Runge–Kutta method, which is an excellent general-purpose differential equation solver. To bridge the gap between these two methods, we begin with some general discussion about how to develop higher-order methods.

6.3.1. Midpoint rule. To derive more accurate differential equation solvers, Taylor series (better yet, Taylor's formula with remainder) offers a useful line of approach. To illustrate, consider the Taylor's formula with remainder,

$$y(x) = y(a) + y'(a)(x-a) + \frac{y''(\xi)}{2!}(x-a)^2, \tag{1}$$

where ξ is some point in $[a, x]$. If we let $x = x_{n+1}$, $a = x_n$, and $x - a = x_{n+1} - x_n = h$, then (1) becomes

$$y(x_{n+1}) = y(x_n) + y'(x_n)h + \frac{y''(\xi)}{2!}h^2. \tag{2}$$

Since $y' = f(x, y)$, we can replace the $y'(x_n)$ in (2) by $f(x_n, y(x_n))$. Also, the last term in (2) can be expressed more simply as $O(h^2)$ so we have

$$y(x_{n+1}) = y(x_n) + f(x_n, y(x_n))h + O(h^2). \tag{3}$$

If we neglect the $O(h^2)$ term and call attention to the approximation thereby incurred by replacing the exact values $y(x_{n+1})$ and $y(x_n)$ by the approximate values y_{n+1} and y_n, respectively, then we have the Euler method

$$y_{n+1} = y_n + f(x_n, y_n)h. \tag{4}$$

Since the term that we dropped in passing from (3) to (4) was $O(h^2)$, the local truncation error is $O(h^2)$, and the accumulated truncation error is $O(h)$.

One way to obtain a higher-order method is to retain more terms in the Taylor's formula. For instance, begin with

$$y(x_{n+1}) = y(x_n) + y'(x_n)h + \frac{y''(x_n)}{2}h^2 + \frac{y'''(\eta)}{6}h^3 \tag{5}$$

in place of (2) or, since $y'' = \dfrac{d}{dx}f(x, y(x)) = f_x + f_y y' = f_x + f_y f$,

$$\begin{aligned} y(x_{n+1}) &= y(x_n) + f(x_n, y(x_n))h \\ &\quad + \frac{1}{2}\left[f_x(x_n, y(x_n)) + f_y(x_n, y(x_n))f(x_n, y(x_n))\right]h^2 + O(h^3). \end{aligned} \tag{6}$$

If we truncate (6) by dropping the $O(h^3)$ term, and change $y(x_{n+1})$ and $y(x_n)$ to y_{n+1} and y_n, respectively, then we have the method

$$y_{n+1} = y_n + f(x_n, y_n)h + \frac{1}{2}\left[f_x(x_n, y_n) + f_y(x_n, y_n)f(x_n, y_n)\right]h^2 \tag{7}$$

with a local truncation error that is $O(h^3)$ and an accumulated truncation error that is $O(h^2)$; that is, we now have a *second*-order method.

Why do we say that (7) is a second-order method? Following the same heuristic reasoning as in Section 6.2, the accumulated truncation error E_n is of the order of the local truncation error times the number of steps so

$$E_n = O(h^3) \cdot n = O(h^3)\frac{n\,h}{h} = O(h^3)\frac{x_n}{h} = O(h^2) \cdot x_n = O(h^2),$$

as claimed. In fact, as a simple rule of thumb it can be said that if the local truncation error is $O(h^p)$, with $p > 1$, then the accumulated truncation error is $O(h^{p-1})$, and one has a $(p-1)$th-order method.

Although the second-order convergence of (7) is an improvement over the first-order convergence of Euler's method, the attractiveness of (7) is diminished by an approximately threefold increase in the computing time per step since it requires three function evaluations (f, f_x, f_y) per step whereas Euler's method requires only one (f). It's true that to carry out one step of Euler's method we need to evaluate f, multiply that by h, and add the result to y_n, but we can neglect the multiplication by h and addition of y_n on the grounds that a typical $f(x, y)$ involves many more arithmetic steps than that. Thus, as a rule of thumb, one compares the computation time per step of two methods by comparing only the number of function evaluations per step.

Not satisfied with (7) because it requires three function evaluations, let us return to Taylor's formula (5). If we replace h by $-h$, that amounts to making a backward step so the term on the left will be $y(x_{n-1})$ instead of $y(x_{n+1})$. Making those changes, and also copying (5), for comparison, we have

$$y(x_{n-1}) = y(x_n) - y'(x_n)h + \frac{y''(x_n)}{2}h^2 - \frac{y'''(\zeta)}{6}h^3, \tag{8a}$$

$$y(x_{n+1}) = y(x_n) + y'(x_n)h + \frac{y''(x_n)}{2}h^2 + \frac{y'''(\eta)}{6}h^3, \tag{8b}$$

respectively, where ζ is some point in $[x_{n-1}, x_n]$ and η is some point in $[x_n, x_{n+1}]$. Now we can eliminate the bothersome y'' terms by subtracting (8a) from (8b). Doing so gives

$$y(x_{n+1}) - y(x_{n-1}) = 2y'(x_n)h + \frac{y'''(\eta) + y'''(\zeta)}{6}h^3$$

or

$$y(x_{n+1}) = y(x_{n-1}) + 2f\left(x_n, y(x_n)\right)h + O\left(h^3\right).$$

Finally, if we drop the $O\left(h^3\right)$ term and change $y(x_{n+1})$, $y(x_{n-1})$, $y(x_n)$ to y_{n+1}, y_{n-1}, y_n, respectively, we have

$$\boxed{y_{n+1} = y_{n-1} + f(x_n, y_n)(2h),} \tag{9}$$

which method is known as the **midpoint rule**. Like (7), the midpoint rule is a second-order method but, unlike (7), it requires only one function evaluation per step. It is an example of a **multi-step method** because it uses information from more than one of the preceding points – namely, from two: the nth and $(n-1)$th. Thus, it is a two-step method whereas Euler's method and the Taylor series method given by (7) are single-step methods.

A disadvantage of the midpoint rule (and other multi-step methods) is that it is not self-starting. That is, the first step gives y_1 in terms of x_0, y_0, y_{-1}, but y_{-1} is not defined. Thus, (9) applies only for $n \geq 1$, and to get the method started we need to compute y_1 by a different method. For instance, we could use Euler's method to obtain y_1 and then switch over to the midpoint rule (9). Of course, if we do that we should do it not in a single Euler step but in many so as not to degrade the subsequent second-order accuracy.

EXAMPLE 1. Consider the same "test problem" as in Section 6.2,

$$y' = y + 2x - x^2; \qquad y(0) = 1, \qquad (0 \leq x < \infty) \tag{10}$$

with the exact solution $y(x) = x^2 + e^x$. Let us use the midpoint rule with $h = 0.1$. To get it started, carry out ten steps of Euler's method with $h = 0.01$. The result of those steps is the approximate solution 1.11358 at $x = 0.1$, which we now take as y_1. Then proceeding with the midpoint rule we obtain from (9)

$$\begin{aligned} y_2 &= y_0 + 2\left(y_1 + 2x_1 - x_1^2\right)h \\ &= 1 + 2\left(1.11358 + 0.2 - 0.01\right)(0.1) = 1.26072 \\ y_3 &= y_1 + 2\left(y_2 + 2x_2 - x_2^2\right)h \\ &= 1.11358 + 2\left(1.26072 + 0.4 - 0.04\right)(0.1) = 1.43772, \end{aligned}$$

and so on. The results are shown in Table 1 and contrasted with the less accurate Euler results using the same step size, $h = 0.1$. ∎

Before leaving the midpoint rule, it is interesting to interpret the improvement in accuracy, from Euler to midpoint, graphically. If we solve

$$y(x_{n+1}) \approx y(x_n) + y'(x_n)h \qquad \text{(Euler)} \tag{11}$$

and

$$y(x_{n+1}) \approx y(x_{n-1}) + 2y'(x_n)h \qquad \text{(midpoint)} \tag{12}$$

for $y'(x_n)$, we have

$$y'(x_n) \approx \frac{y(x_{n+1}) - y(x_n)}{h} \qquad \text{(Euler)} \tag{13}$$

Table 1. Comparison of Euler, midpoint rule, and exact solutions of the initial-value problem (10), with $h = 0.1$.

x	Euler	Midpoint	Exact
0.10	1.10000	1.11358	1.11517
0.20	1.22900	1.26072	1.26140
0.30	1.38790	1.43772	1.43986
0.40	1.57769	1.65026	1.65182
0.50	1.79946	1.89577	1.89872

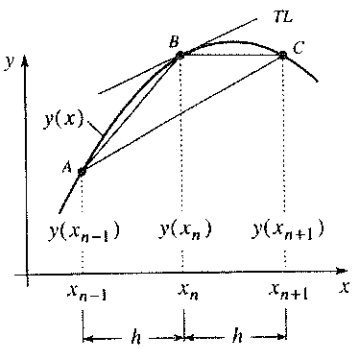


Figure 1. Graphical interpretation of midpoint rule versus Euler.

and

$$y'(x_n) \approx \frac{y(x_{n+1}) - y(x_{n-1})}{2h}, \qquad \text{(midpoint)} \tag{14}$$

which are difference quotient approximations of the derivative $y'(x_n)$. In Fig. 1, we can interpret (14) and (13) as approximating $y'(x_n)$ by the slopes of the chords AC and BC, respectively, while the exact $y'(x_n)$ is the slope of the tangent line TL at x_n. We can see from the figure that AC gives a more accurate approximation than BC.

6.3.2. Second-order Runge–Kutta. The Runge–Kutta methods are developed somewhat differently. Observe that the low-order Euler method $y_{n+1} = y_n + f(x_n, y_n)h$ amounts to an extrapolation away from the initial point (x_n, y_n) using the slope $f(x_n, y_n)$ at that point. Expecting an average slope to give greater accuracy, one might try the algorithm

$$y_{n+1} = y_n + \frac{1}{2}\left[f(x_n, y_n) + f(x_{n+1}, y_{n+1})\right] h, \tag{15}$$

which uses an average of the slopes at the initial and final points. Unfortunately, the formula (15) does not give y_{n+1} explicitly since y_{n+1} appears not only on the left-hand side but also in the argument of $f(x_{n+1}, y_{n+1})$. Intuition tells us that we should still do well if we replace that y_{n+1} by an estimated value, say, the Euler estimate $y_{n+1} = y_n + f(x_n, y_n)h$. Then the revised version of (15) is

$$y_{n+1} = y_n + \frac{1}{2}\left\{f(x_n, y_n) + f\left[x_{n+1}, y_n + f(x_n, y_n)h\right]\right\} h. \tag{16}$$

Thus, guided initially by intuition we can put the idea on a rigorous basis by considering a method of the form

$$y_{n+1} = y_n + \left\{af(x_n, y_n) + bf\left[x_n + \alpha h, y_n + \beta f(x_n, y_n)h\right]\right\} h \tag{17}$$

and choosing the adjustable parameters a, b, α, β so as to make the order of the method (17) as high as possible; α, β determine the second slope location, and a, b

determine the "weights" of the two slopes. That is, we seek a, b, α, β so that the left- and right-hand sides of

$$\begin{aligned} y(x_{n+1}) \approx y(x_n) &+ \{af[x_n, y(x_n)] \\ &+ bf[x_n + \alpha h, y(x_n) + \beta f[x_n, y(x_n)]h]\}h \end{aligned} \tag{18}$$

agree to as high a degree in h as possible. Thus, expand the left-hand side (LHS) and right-hand side (RHS) of (18) in Taylor series in h:

$$\begin{aligned} \text{LHS} &= y(x_n) + y'(x_n)h + \frac{y''(x_n)}{2}h^2 + \cdots \\ &= y + fh + \frac{1}{2}(f_x + f_y f)h^2 + \cdots \end{aligned} \tag{19}$$

where y means $y(x_n)$ and the arguments of f, f_x, f_y are $x_n, y(x_n)$. Similarly (Exercise 9),

$$\text{RHS} = y + (a+b)fh + (\alpha f_x + \beta f f_y)bh^2 + \cdots. \tag{20}$$

Matching the h terms requires that

$$a + b = 1. \tag{21a}$$

Matching the h^2 terms, for any function f requires that

$$\alpha b = \frac{1}{2} \quad \text{and} \quad \beta b = \frac{1}{2}. \tag{21b}$$

The outcome then is that any method (17), with a, b, α, β chosen so as to satisfy (21), has a local truncation error that is $O(h^3)$ and is therefore a second-order method [subject to mild conditions on f such as the continuity of f and its first- and second-order partial derivatives so that we can justify the steps in deriving (19) and (20)]. These are the **Runge–Kutta methods of second order**.†

For instance, with $a = b = 1/2$ and $\alpha = \beta = 1$ we have

$$y_{n+1} = y_n + \frac{1}{2}\{f(x_n, y_n) + f[x_{n+1}, y_n + f(x_n, y_n)h]\}h, \tag{22}$$

which is usually expressed in the computationally convenient form

$$\boxed{\begin{aligned} y_{n+1} &= y_n + \tfrac{1}{2}(k_1 + k_2), \\ k_1 = hf(x_n, y_n), &\qquad k_2 = hf(x_{n+1}, y_n + k_1). \end{aligned}} \tag{23}$$

†The Runge–Kutta method was originated by *Carl D. Runge* (1856–1927), a German physicist and mathematician, and extended by the German aerodynamicist and mathematician *M. Wilhelm Kutta* (1867–1944). Kutta is well known for the Kutta–Joukowski formula for the lift on an airfoil, and for the "Kutta condition" of classical airfoil theory.

To understand this result, note that Euler's method would give $y_{n+1} = y_n + f(x_n, y_n)\,h$. If we denote that Euler estimate as y_{n+1}^{Euler}, then (22) can be expressed as

$$y_{n+1} = y_n + \frac{f(x_n, y_n) + f\left(x_{n+1}, y_{n+1}^{\text{Euler}}\right)}{2}\,h.$$

That is, we take a tentative step using Euler's method, then we average the slopes at the initial point x_n, y_n and at the Euler estimate of the final point $x_{n+1}, y_{n+1}^{\text{Euler}}$, and then make another Euler step, this time using the improved (average) slope. For this reason (23) is also known as the **improved Euler method.**

A different choice, $a = 0, b = 1, \alpha = \beta = 1/2$, gives what is known as the **modified Euler method**.

EXAMPLE 2. Let us proceed through the first two steps of the improved Euler method (23) for the same test problem as was used in Example 1,

$$y' = y + 2x - x^2; \qquad y(0) = 1, \qquad (0 \le x < \infty) \tag{24}$$

with $h = 0.1$; a more detailed illustration is given in Section 6.3.3 below. Here, $f(x, y) = y + 2x - x^2$.

$n = 0$:

$$\begin{aligned} k_1 &= hf(x_0, y_0) = 0.1\left[1 + 0 - (0)^2\right] = 0.1, \\ k_2 &= hf(x_1, y_0 + k_1) = 0.1\left[(1 + 0.1) + 2(0.1) - (0.1)^2\right] = 0.129, \\ y_1 &= y_0 + \tfrac{1}{2}(k_1 + k_2) = 1 + 0.5\,(0.1 + 0.129) = 1.1145; \end{aligned}$$

$n = 1$:

$$\begin{aligned} k_1 &= hf(x_1, y_1) = 0.1\left[1.1145 + 2(0.1) - (0.1)^2\right] = 0.13045, \\ k_2 &= hf(x_2, y_1 + k_1) \\ &= 0.1\left[(1.1145 + 0.13045) + 2(0.2) - (0.2)^2\right] = 0.160495, \\ y_2 &= y_1 + \tfrac{1}{2}(k_1 + k_2) = 1.1145 + 0.5\,(0.13045 + 0.160495) = 1.2600, \end{aligned}$$

compared with the values $y(x_1) = y(0.1) = 1.1152$ and $y(x_2) = y(0.2) = 1.2614$ obtained from the known exact solution $y(x) = x^2 + e^x$. ∎

6.3.3. Fourth-order Runge–Kutta. Using this idea of a weighted average of slopes at various points in the x, y plane, with the weights and locations determined so as to maximize the order of the method, one can derive higher-order Runge–Kutta methods as well, although the derivations are quite tedious. One of the most

commonly used is the **fourth-order Runge–Kutta method**:

$$\boxed{\begin{aligned} &y_{n+1} = y_n + \tfrac{1}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right) \\ &k_1 = hf\left(x_n, y_n\right), \qquad k_2 = hf\left(x_n + \tfrac{h}{2}, y_n + \tfrac{1}{2}k_1\right), \\ &k_3 = hf\left(x_n + \tfrac{h}{2}, y_n + \tfrac{1}{2}k_2\right), \quad k_4 = hf\left(x_{n+1}, y_n + k_3\right), \end{aligned}} \tag{25}$$

which we give without derivation. Here the effective slope used is a weighted average of the slopes at the four points (x_n, y_n), $(x_n + h/2, y_n + k_1/2)$, $(x_n + h/2, y_n + k_2/2)$ and $(x_{n+1}, y_n + k_3)$ in the x, y plane, an average because the sum of the coefficients $1/6$, $2/6$, $2/6$, $1/6$ that multiply the k's is 1. Similarly, the sum of the coefficients $1/2, 1/2$ in the second-order version (23) is 1 as well.

EXAMPLE 3. As a summarizing illustration, we solve another "test problem,"

$$y' = -y; \qquad y(0) = 1 \tag{26}$$

by each of the methods considered, using a step size of $h = 0.05$ and single precision arithmetic (on the computer used that amounts to carrying eight significant figures; double precision would carry 16). The results are given in Table 2, together with the exact solution $y(x) = e^{-x}$ for comparison; 0.529E + 2, for instance, means 0.529×10^2. The value of y_1 for the midpoint rule was obtained by Euler's method with a reduced step size of 0.0025.

To illustrate the fourth-order Runge–Kutta calculation, let us proceed through the first step:

$$\begin{aligned} n = 0: \\ k_1 &= hf(x_0, y_0) = -hy_0 = -0.05(1) = -0.05, \\ k_2 &= hf\left(x_0 + \tfrac{h}{2}, y_0 + \tfrac{1}{2}k_1\right) = -h\left(y_0 + \tfrac{1}{2}k_1\right) \\ &= -0.05(1 - 0.025) = -0.04875, \\ k_3 &= hf\left(x_0 + \tfrac{h}{2}, y_0 + \tfrac{1}{2}k_2\right) = -h\left(y_0 + \tfrac{1}{2}k_2\right) \\ &= -0.05(1 - 0.024375) = -0.04878125, \\ k_4 &= hf\left(x_1, y_0 + k_3\right) = -h\left(y_0 + k_3\right) \\ &= -0.05(1 - 0.04878125) = -0.047560938, \\ y_1 &= y_0 + \tfrac{1}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right) = 0.95122943, \end{aligned}$$

which final result does agree with the corresponding entry in Table 2. Actually, there is a discrepancy of 2 in the last digit, but an error of that size is not unreasonable in view of the fact that the machine used carried only eight significant figures.

Most striking is the excellent accuracy of the fourth-order Runge–Kutta method, with six significant figure accuracy over the entire calculation.

COMMENT. We see that the midpoint rule and the second-order Runge–Kutta method yield comparable results initially, but the midpoint rule eventually develops an error that

Table 2. Comparison of the Euler, midpoint rule, second-order Runge–Kutta, fourth-order Runge–Kutta, and exact solutions of the initial-value problem (26), with $h = 0.05$.

x	Euler	Midpoint	2nd-order Runge–Kutta	4th-order Runge–Kutta	Exact $= e^{-x}$
0.00	1.00000000 E+1	1.00000000 E+1	1.00000000 E+1	1.00000000 E+1	1.00000000 E+1
0.05	0.94999999 E+1	0.95116991 E+1	0.95125002 E+1	0.95122945 E+1	0.95122945 E+1
0.10	0.90249997 E+1	0.90488303 E+1	0.90487659 E+1	0.90483743 E+1	0.90483743 E+1
0.15	0.85737497 E+1	0.86068159 E+1	0.86076385 E+1	0.86070800 E+1	0.86070800 E+1
0.20	0.81450623 E+1	0.81881487 E+1	0.81880164 E+1	0.81873077 E+1	0.81873077 E+1
0.25	0.77378094 E+1	0.77880013 E+1	0.77888507 E+1	0.77880079 E+1	0.77880079 E+1
0.30	0.73509192 E+1	0.74093485 E+1	0.74091440 E+1	0.74081820 E+1	0.74081820 E+1
⋮					
2.00	0.12851217 E+0	0.13573508 E+0	0.13545239 E+0	0.13533530 E+0	0.13533528 E+0
2.05	0.12208656 E+0	0.12853225 E+0	0.12884909 E+0	0.12873492 E+0	0.12873492 E+0
2.10	0.11598223 E+0	0.12288185 E+0	0.12256770 E+0	0.12245644 E+0	0.12245644 E+0
2.15	0.11018312 E+0	0.11624406 E+0	0.11659253 E+0	0.11648417 E+0	0.11648415 E+0
2.20	0.10467397 E+0	0.11125745 E+0	0.11090864 E+0	0.11080316 E+0	0.11080315 E+0
2.25	0.99440269 E−1	0.10511832 E+0	0.10550185 E+0	0.10539923 E+0	0.10539922 E+0
2.30	0.94468258 E−1	0.10074562 E+0	0.10035863 E+0	0.10025885 E+0	0.10025885 E+0
⋮					
5.00	0.59205294 E−2	0.12618494 E−1	0.67525362 E−2	0.67379479 E−2	0.67379437 E−2
5.05	0.56245029 E−2	0.25511871 E−3	0.64233500 E−2	0.64093345 E−2	0.64093322 E−2
5.10	0.53432779 E−2	0.12592983 E−1	0.61102118 E−2	0.60967477 E−2	0.60967444 E−2
5.15	0.50761141 E−2	-0.10041796 E−2	0.58123390 E−2	0.57994057 E−2	0.57994043 E−2
5.20	0.48223082 E−2	0.12693400 E−1	0.55289874 E−2	0.55165654 E−2	0.55165626 E−2
5.25	0.45811930 E−2	-0.22735195 E−2	0.52594491 E−2	0.52475194 E−2	0.52475161 E−2
5.30	0.43521333 E−2	0.12920752 E−1	0.50030509 E−2	0.49915947 E−2	0.49915928 E−2
⋮					
9.70	0.47684727 E−4	0.64383668 E+0	0.61541170 E−4	0.61283507 E−4	0.61283448 E−4
9.75	0.45300490 E−4	-0.67670959 E+0	0.58541038 E−4	0.58294674 E−4	0.58294663 E−4
9.80	0.43035467 E−4	0.71150762 E+0	0.55687164 E−4	0.55451608 E−4	0.55451590 E−4
9.85	0.40883693 E−4	-0.74786037 E+0	0.52972413 E−4	0.52747200 E−4	0.52747171 E−4
9.90	0.38839509 E−4	0.78629363 E+0	0.50390008 E−4	0.50174691 E−4	0.50174654 E−4
9.95	0.36897534 E−4	-0.82648975 E+0	0.47933496 E−4	0.47727641 E−4	0.47727597 E−4
10.00	0.35052657 E−4	0.86894262 E+0	0.45596738 E−4	0.45399935 E−4	0.45399931 E−4

oscillates in sign, from step to step, and grows in magnitude. The reason for this strange (and incorrect) behavior will be studied in Section 6.5. ∎

Of course, in real applications we do not have the exact solution to compare with the numerical results. In that case, how do we know whether or not our results are sufficiently accurate? A useful rule of thumb, mentioned in Section 6.2, is to redo the entire calculation, each time with a smaller step size, until the results "settle down" to the desired number of significant digits.

Thus far we have taken h to be a constant, for simplicity, but there is no reason why it cannot be varied from one step to the next. In fact, there may be a compelling reason to do so. For instance, consider the equation $y' + y = \tanh 20x$ on $-10 \leq x \leq 10$. The function $\tanh 20x$ is almost a constant, except near the origin, where it varies dramatically approximately from -1 to $+1$. Thus, we need a very fine step size h near the origin for good accuracy, but to use that h over the entire x interval would be wasteful in terms of computer time and expense.

One can come up with a rational scheme for varying the step size to maintain a consistent level of accuracy, but such refinements are already available within existing software. For example, the default numerical differential equation solver in *Maple* is a "fourth-fifth order Runge–Kutta–Fehlberg method" denoted as RKF45 in the literature. According to RKF45, a tentative step is made, first using a fourth-order Runge–Kutta method, and then again using a fifth-order Runge–Kutta method. If the two results agree to a prespecified number of significant digits, then the fifth-order result is accepted. If they agree to more than that number of significant digits, then h is increased and the next step is made. If they agree to less than that number of significant digits, then h is decreased and the step is repeated.

6.3.4. Empirical estimate of the order. (Optional) The relative accuracies achieved by the different methods, as seen from the results in Table 2, strikingly reveal the importance of the order of the method. Thus, it is important to know how to verify the order of whatever method we use, if only as a partial check on the programming.

Recall that by a method being of order p we mean that at any chosen x the error behaves as Ch^p for some constant C:

$$y_{\text{exact}} - y_{\text{comp}} \sim Ch^p \tag{27}$$

as $h \to 0$. Suppose we wish to check the order of a given method. Select a test problem such as the one in Example 3, and use the method to compute y at any x point such as $x = 1$, for two different h's say h_1 and h_2. Letting $y_{\text{comp}}^{(1)}$ and $y_{\text{comp}}^{(2)}$ denote the y's computed at $x = 1$ using step sizes of h_1 and h_2, respectively, we have

$$y_{\text{exact}} - y_{\text{comp}}^{(1)} \approx Ch_1^p,$$
$$y_{\text{exact}} - y_{\text{comp}}^{(2)} \approx Ch_2^p.$$

Dividing one equation by the other, to cancel the unknown C, and solving for p,

gives

$$p \approx \frac{\ln\left[\frac{y_{\text{exact}} - y_{\text{comp}}^{(1)}}{y_{\text{exact}} - y_{\text{comp}}^{(2)}}\right]}{\ln\left[\frac{h_1}{h_2}\right]}. \tag{28}$$

To illustrate, let us run the test problem (26) with Euler's method, with $h_1 = 0.1$ and $h_2 = 0.05$. The results at $x = 1$ are

$$\begin{aligned} h_1 &= 0.1, & y_{\text{comp}}^{(1)} &= 0.348678440100, \\ h_2 &= 0.05, & y_{\text{comp}}^{(2)} &= 0.358485922409, \end{aligned}$$

and since $y_{\text{exact}}(1) = 0.367879441171$, (28) gives $p \approx 1.03$, which is respectably close to 1. We should be able to obtain a more accurate estimate of p by using smaller h's since (27) holds only as $h \to 0$. In fact, using $h_1 = 0.05$ and $h_2 = 0.02$ gives $p \approx 1.01$. Using those same step sizes, we also obtain $p \approx 2.05, 2.02$, and 4.03 for the midpoint rule, second-order Runge–Kutta, and fourth-order Runge–Kutta methods, respectively.

Why not use even smaller h's to determine p more accurately? Two difficulties arise. One is that as the h's are decreased the computed solutions become more and more accurate, and the $y_{\text{exact}} - y_{\text{comp}}^{(1)}$ and $y_{\text{exact}} - y_{\text{comp}}^{(2)}$ differences in (28) are known to fewer and fewer significant figures, due to cancelation. This is especially true for a high-order method. The other difficulty is that (27) applies to the truncation error alone so, implicit in our use of (27) is the assumption that roundoff errors are negligible. If we make h too small, that assumption may become invalid. For both of these reasons it is important to use extended precision for such calculations, as we have for the preceding calculations.

6.3.5. Multi-step and predictor-corrector methods. (Optional) We've already called attention to the multi-step nature of the midpoint rule. Our purpose in this optional section is to give a brief overview of a class of multi-step methods known as **Adams–Bashforth methods**, obtained by integrating $y' = f(x, y)$ from x_n to x_{n+1}:

$$\int_{x_n}^{x_{n+1}} y'\,dx = \int_{x_n}^{x_{n+1}} f\left(x, y(x)\right) dx \tag{29}$$

or

$$y(x_{n+1}) = y(x_n) + \int_{x_n}^{x_{n+1}} f\left(x, y(x)\right) dx. \tag{30}$$

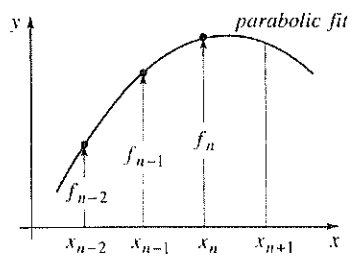


Figure 2. Adams–Bashforth interpolation of f for the case $m = 2$.

To evaluate the integral, we fit $f\left(x, y(x)\right)$ with a polynomial of degree m, which is readily integrated. The polynomial interpolates (i.e., takes on the same values as) $f\left(x, y(x)\right)$ at the $m + 1$ points $x_{n-m}, \ldots, x_{n-1}, x_n$ as illustrated in Fig. 2 for the case $m = 2$. As the simplest case, let $m = 0$. Then the zeroth degree polynomial approximation of $f\left(x, y(x)\right)$ on $[x_n, x_{n+1}]$ is $f\left(x, y(x)\right) \approx f_n$, where f_n denotes $f(x_n, y_n)$, and (30) gives the familiar Euler method $y_{n+1} = y_n + f_n h$. Omitting

the steps in this overview we state that with $m = 3$ one obtains the fourth-order **Adams–Bashforth method**

$$y_{n+1} = y_n + (55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3})\frac{h}{24} \tag{31a}$$

with a local truncation error of

$$(e_n)_{AB} = \frac{251}{720}y^{(v)}(\xi)h^5 \tag{31b}$$

for some ξ in $[x_{n-3}, x_n]$.We can see that (31a) is not self-starting; it applies only for $n = 3, 4, \ldots$, so the first three steps (for $n = 0, 1, 2$) need to be carried out by some other method.

Suppose that instead of interpolating f at $x_{n-m}, \ldots, x_{n-1}, x_n$ we interpolate at $x_{n-m+1}, \ldots, x_n, x_{n+1}$. With $m = 3$, again, one obtains the fourth-order **Adams–Moulton method**

$$y_{n+1} = y_n + (9f_{n+1} + 19f_n - 5f_{n-1} + f_{n-2})\frac{h}{24} \tag{32a}$$

with a local truncation error of

$$(e_n)_{AM} = -\frac{19}{720}y^{(v)}(\xi)h^5, \tag{32b}$$

where the ξ's in (31b) and (32b) are different, in general.

Although both methods are fourth order, the Adams–Moulton method is more accurate because the constant factor in $(e_n)_{AM}$ is roughly thirteen times smaller than the constant in $(e_n)_{AB}$. This increase in accuracy occurs because the points $x_{n-m+1}, \ldots, x_n, x_{n+1}$ are more centered on the interval of integration (from x_n to x_{n+1}) than the points $x_{n-m}, \ldots, x_{n-1}, x_n$. On the other hand, the term $f_{n+1} = f(x_{n+1}, y_{n+1})$ in (32a) is awkward because the argument y_{n+1} is not yet known! [If f is linear in y, then (32a) can be solved for y_{n+1} by simple algebra, and the awkwardness disappears.] Thus, the method (32a) is said to be of **closed** type, whereas (31a) and all of our preceding methods have been of **open** type.

To handle this difficulty, it is standard practice to solve closed formulas by iteration. Using superscripts to indicate the iterate, (32a) becomes

$$y_{n+1}^{(k+1)} = y_n + \left[9f\left(x_{n+1}, y_{n+1}^{(k)}\right) + 19f_n - 5f_{n-1} + f_{n-2}\right]\frac{h}{24}. \tag{33}$$

To start the iteration, we compute $y_{n+1}^{(0)}$ from a **predictor formula**, with subsequent corrections made by the **corrector formula** (33). It is recommended that the predictor and corrector formulas be of the same order (certainly, the corrector should never be of lower order than the predictor) with the corrector applied only once. Thus, the Adams–Bashforth and Adams–Moulton methods constitute a natural **predictor-corrector** pair with "AB" as the predictor and "AM" as the corrector. Why might we choose the fourth-order AB–AM predictor-corrector over

the Runge–Kutta method of the same order or vice versa? On the negative side, AB–AM is not self-starting, it requires the storage of f_{n-3}, f_{n-2}, and f_{n-1}, and is more tedious to program. On the other hand, it involves only two function evaluations per step (namely, f_n and f_{n+1}) if the corrector is applied only once, whereas Runge–Kutta involves four. Thus, if $f(x, y)$ is reasonably complicated then we can expect AB–AM to be almost twice as fast. In large-scale computing, the savings can be significant.

Closure. Motivated to seek higher-order methods than the first-order Euler method, we use a Taylor series approach to obtain the second-order midpoint rule. Though more accurate, a disadvantage of the midpoint rule is that it is not self-starting. Pursuing a different approach, we look at the possibility of using a weighted average of slopes at various points in the x, y plane, with the weights and locations determined so as to maximize the order of the method. We thereby derive the second-order Runge–Kutta method and present, without derivation, the fourth-order Runge–Kutta method. The latter is widely used because it is accurate and self-starting.

Because of the importance of the order of a given method, we suggest that the order be checked empirically using a test problem with a known exact solution. The resulting approximate expression for the order is given by (28).

In the final section we return to the idea of multistep methods and present a brief overview of the Adams–Bashforth methods, derived most naturally from an approximate integral approach. Though not self-starting, the fourth-order Adams-Bashforth method (31a) is faster than the Runge–Kutta method of the same order because it requires only one function evaluation per step (namely, f_n; the f_{n-1}, f_{n-2}, and f_{n-3} terms are stored from previous steps). A further refinement consists of predictor-corrector variations of the Adams–Bashforth methods. However, we stress that such refinements become worthwhile only if the scope of the computational effort becomes large enough to justify the additional inconvenience caused by such features as the absence of self-starting and predictor-corrector iteration. Otherwise, one might as well stick to a simple and accurate method such as fourth-order Runge–Kutta.

Computer software. Computer-software systems such as *Maple* include numerical differential equation solvers. In *Maple* one can use the dsolve command together with a numeric option. The default numerical solution method is the RKF45 method mentioned above. Note that with the numeric option of dsolve one does not specify a step size h since that choice is controlled within the program and, in general, is varied from step to step to maintain a certain level of accuracy. To specify the absolute error tolerance one can use an additional option called abserr, which is formatted as abserr = Float(1,2-digits) and which means 1 times 10 to the one- or two-digit exponent. For instance, to solve

$$y' = -y; \quad y(0) = 1$$

for $y(x)$ with an absolute error tolerance of 1×10^{-5}, and to print the results at $x = 2, 10$, enter

```
with(DEtools):
```

and return. Then enter

$$\text{dsolve}(\{\text{diff}(y(x), x) = -y(x),\ y(0) = 1\},\ y(x),\ \text{type} = \text{numeric},$$
$$\text{value} = \text{array}([2, 10]),\ \text{abserr} = \text{Float}(1, -5));$$

and return. The printed result is

$$\begin{bmatrix} [x, y(x)] \\ \begin{bmatrix} 2. & .1353337989380555 \\ 10. & .00004501989255717160 \end{bmatrix} \end{bmatrix}$$

For comparison, the exact solution is $y(2) = \exp(-2) = 0.1353352832$ and $y(10) = \exp(-10) = 0.0000453999$, respectively.

EXERCISES 6.3

1. Evaluate y_1 and y_2 by hand, by the second-order and fourth-order Runge–Kutta methods, with $h = 0.02$. Obtain the exact values $y(x_1)$ and $y(x_2)$ as well.

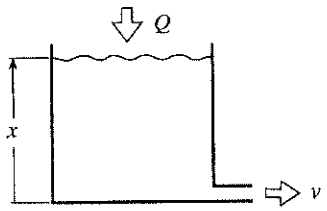
(a) $y' = 3000xy^{-2}$; $y(0) = 2$
(b) $y' = 40xe^{-y}$; $y(0) = 3$
(c) $y' = x + y$; $y(-1) = 5$
(d) $y' = -y \tan x$; $y(1) = -8$
(e) $y' = (y^2 + 1)/4$; $y(0) = 0$
(f) $y' = -2y \sin x$; $y(2) = 5$

2. (a)–(f) Program the second- and fourth-order Runge–Kutta methods and use them to solve the initial-value problem given in the corresponding part of Exercise 1 but with the initial condition $y(0) = 1$. Use $h = 0.05$. Print out all computed values of y, up to $x = 0.5$, as well as the exact solution.

3. Using the test problem (10), do an empirical evaluation of the order of the given method. Use (28), with $h = 0.1$ and 0.05, say. Do the evaluation at two different locations, such as $x = 1$ and $x = 2$. (The order should not depend upon x so your results at the two points should be almost identical.)

(a) Euler's method
(b) Second-order Runge–Kutta method
(c) Fourth-order Runge–Kutta method

4. (*Liquid level*) Liquid is pumped into a tank of horizontal cross-sectional area A (m^2) at a rate Q (liters/sec), and is drained by a valve at its base as sketched in the figure.



According to Bernoulli's principle, the efflux velocity $v(t)$ is approximately $\sqrt{2gx(t)}$, where g is the acceleration of gravity. Thus, a mass balance gives

$$\begin{aligned} Ax'(t) &= Q(t) - Bv(t) \\ &= Q(t) - B\sqrt{2gx(t)}, \end{aligned} \tag{4.1}$$

where B is the cross-sectional area of the efflux pipe. For definiteness, suppose that $A = 1$ and $B\sqrt{2g} = 0.01$ so

$$x' = Q(t) - 0.01\sqrt{x}. \tag{4.2}$$

We wish to know the depth $x(t)$ at the end of 10 minutes ($t = 600$ sec), 20 minutes, ..., up to one hour. Program the computer solution of (4.2) by the second-order Runge–Kutta method for the following cases, and use it to solve for those x values: $x(600), x(1200), \ldots, x(3600)$. (Using the rule of thumb given below Example 3, reduce h until those results settle down to four significant digits.)

(a) $Q(t) = 0.02$; $x(0) = 0$
(b) $Q(t) = 0.02$; $x(0) = 2$
(c) $Q(t) = 0.02$; $x(0) = 4$
(d) $Q(t) = 0.02$; $x(0) = 6$

(e) $Q(t) = 0.02\left(1 - e^{-0.004t}\right)$; $x(0) = 0$
(f) $Q(t) = 0.02\left(1 - e^{-0.004t}\right)$; $x(0) = 8$
(g) $Q(t) = 0.02t$; $x(0) = 0$
(h) $Q(t) = 0.02(1 + \sin 0.1t)$; $x(0) = 0$
NOTE: Surely, we will need h to be small compared to the period 200π of $Q(t)$ in part (h).

5. (a)–(h) (*Liquid level*) Same as Exercise 4, but use fourth-order Runge–Kutta instead of second order.

6. (a)–(h) (*Liquid level*) Same as Exercise 4, but use computer software to do the numerical solution of the differential equation. In *Maple*, for instance, the dsolve command uses the fourth-fifth order RKF45 method.

7. (*Liquid level*) (a) For the case where $Q(t)$ is a constant, derive the general solution of (4.2) in Exercise 4 as

$$Q - 0.01\sqrt{x} - Q\ln\left(Q - 0.01\sqrt{x}\right) = 0.00005t + C, \quad (7.1)$$

where C is the constant of integration.
(b) Evaluate C in (7.1) if $Q = 0.02$ and $x(0) = 0$. Then, solve (7.1) for $x(t)$ at $t = 600, 1200, \ldots, 3600$. NOTE: Unfortunately, (7.1) is in implicit rather than explicit form, but you can use computer software to solve. In *Maple*, for instance, the relevant command is fsolve.

8. Suppose that we have a convergent method, with $E_n \sim Ch^p$ as $h \to 0$. Someone offers to improve the method by either halving C or by doubling p. Which would you choose? Explain.

9. Expand the right-hand side of (18) in a Taylor series in h and show that the result is as given in (20). HINT: To expand the $f(x_n + \alpha h, y + \beta f h)$ term you need to use chain differentiation.

10. (a) Program the fourth-order Runge–Kutta method (25) and use it to run the test problem (10) and to compute y at $x = 1$ using $h = 0.05$ and then $h = 0.02$. From those values and the known exact solution, empirically verify that the method is fourth order.
(b) To see what harm a programming error can cause, change the $x_n + h/2$ in the formula for k_2 to x_n, repeat the two evaluations of y at $x = 1$ using $h = 0.05$ and $h = 0.02$, and empirically determine the order of the method. Is it still a fourth-order method?
(c) Instead of introducing the programming error suggested in part (b), suppose we change the coefficient of k_2 in $y_{n+1} = y_n + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)$ from 2 to 3. Do you think the method will still be convergent? Explain.

11. (*Rectangular, trapezoidal, and Simpson's rule*) Consider the special case where f in $y' = f$ is a function of x only. Integrating $y' = f(x)$ from x_n to x_{n+1}, we have

$$y(x_{n+1}) = y(x_n) + \int_{x_n}^{x_{n+1}} f(x)\,dx \quad (11.1)$$

If we fit $f(x)$, over $[x_n, x_{n+1}]$, with a *zeroth*-degree polynomial (i.e., a constant) that interpolates f at x_n, then we have $f(x) \approx f(x_n)$, and (11.1) gives $y(x_{n+1}) \approx y(x_n) + f(x_n)h$ and hence the Euler method $y_{n+1} = y_n + f(x_n)h$.

(a) Show that if we fit $f(x)$, over $[x_n, x_{n+1}]$, with a *first*-degree polynomial (a straight line) that interpolates f at x_n and x_{n+1}, then $f(x) \approx f(x_n) + [f(x_{n+1}) - f(x_n)](x - x_n)/h$. Putting that approximation into (11.1), derive the approximation

$$y(x_{n+1}) = y(x_n) + \frac{1}{2}\left[f(x_n) + f(x_{n+1})\right]h, \quad (11.2)$$

and show that (for the special case where f is a function of x only) (11.2) is identical to the second-order Runge–Kutta method (23).
(b) Show that if we fit $f(x)$, over $[x_n, x_{n+1}]$, with a second-degree polynomial (a parabola) that interpolates f at x_n, $x_n + h/2$, and $x_{n+1} = x_n + h$, and put that approximation into (11.1), then one obtains

$$y(x_{n+1}) = y(x_n) + \frac{1}{6}\left[f(x_n) + 4f(x_n + h/2) + f(x_{n+1})\right]h, \quad (11.3)$$

and show that (for the case where f is a function of x only) (11.3) is identical to the fourth-order Runge–Kutta method (25). NOTE: These three results amount to the well-known **rectangular, trapezoidal**, and **Simpson's** rules of numerical integration for a single interval of size h. If we sum over all of the intervals, they take the forms

$$\int_a^b f(x)\,dx = \begin{cases} [f(a) + f(a+h) + \cdots + f(b)]\,h, \\ [f(a) + 2f(a+h) + 2f(a+2h) \\ \qquad + \cdots + 2f(b-h) + f(b)]\,\frac{h}{2}, \\ [f(a) + 4f(a+h) + 2f(a+2h) + 4f(a+3h) \\ \qquad + \cdots + 4f(b-h) + f(b)]\,\frac{h}{6}; \end{cases} \quad (11.4)$$

respectively. In passing from (11.3) to the last line of (11.4) we have replaced $h/2$ by h everywhere in

(11.3), so that the revised (11.3) reads $y_{n+1} = y_n + [f(x_n) + 4f(x_{n+1}) + f(x_{n+2})]\,h/3$, where $x_n = a+nh$. For the rectangular and trapezoidal rules the number of subdivisions, $(b-a)/h$, can be even or odd, but for Simpson's rule it must be even. The order of the error for these integration methods is $O(h)$, $O(h^2)$, and $O(h^4)$, respectively.

12. (a) Using $m = 1$, derive from (30) the Adams–Bashforth method

$$y_{n+1} = y_n + (3f_n - f_{n-1})\,\frac{h}{2}. \tag{12.1}$$

(b) Determine the order of the method (12.1) empirically by using it to solve the test problem (10), at $x = 1$, with two different step sizes, and then using (28).

13. This exercise is to take you through the fourth-order AB–AM predictor-corrector scheme.
(a) For the problem $y' = 2xy^2, y(0) = 1$, compute y_1, y_2, y_3 from the exact solution, with $h = 0.1$, and use those as starting values to determine y_4, by hand, by means of the fourth-order AB–AM predictor-corrector scheme given by (31a) and (33). Apply the corrector three times.
(b) Continuing in the same way, determine y_5.

6.4 Application to Systems and Boundary-Value Problems

The methods developed in the preceding sections are for an initial-value problem with a single first-order differential equation, but what if we have a system of differential equations, a higher-order equation, or a boundary-value problem? In this section we extend the Euler and fourth-order Runge–Kutta methods to these cases.

6.4.1. Systems and higher-order equations. Consider the system of initial-value problems

$$u'(x) = f(x, u, v); \qquad u(a) = u_0 \tag{1a}$$

$$v'(x) = g(x, u, v); \qquad v(a) = v_0 \tag{1b}$$

on $u(x)$ and $v(x)$. To extend Euler's method to such a system, we merely apply it to each of the problems (1a) and (1b) as follows:

$$\begin{aligned} u_{n+1} &= u_n + f(x_n, u_n, v_n)h, \\ v_{n+1} &= v_n + g(x_n, u_n, v_n)h \end{aligned} \tag{2}$$

for $n = 0, 1, 2, \ldots$. Equations (2) are coupled (since each involves both u and v), as were equations (1), but that coupling causes no complication because it occurs in the right-hand sides of the equations, and the values u_n, v_n are already known from the preceding step.

EXAMPLE 1. Consider the system

$$\begin{aligned} u' &= x + v; \qquad u(0) = 0 \\ v' &= uv^2; \qquad v(0) = 1. \end{aligned} \tag{3}$$

The latter looks fairly simple, but it is not. It is nonlinear because of the uv^2 term. Turning to numerical solution using the Euler method (2), let $h = 0.1$, say, and let us go through the first couple of steps. First, $u_0 = 0$ and $v_0 = 1$ from the initial conditions. Then,

$$
\begin{aligned}
n = 0:& \\
&u_1 = u_0 + (x_0 + v_0)h = 0 + (0+1)(0.1) = \underline{0.1}, \\
&v_1 = v_0 + u_0 v_0^2 h = 1 + (0)(1)^2(0.1) = \underline{1}. \\
n = 1:& \\
&u_2 = u_1 + (x_1 + v_1)h = 0.1 + (0.1+1)(0.1) = \underline{0.21}, \\
&v_2 = v_1 + u_1 v_1^2 h = 1 + (0.1)(1)^2(0.1) = \underline{1.01},
\end{aligned}
$$

and so on. ∎

Similarly, if the system contains more than two equations.

Next, we show how to adapt the fourth-order Runge–Kutta method to the system (1). Recall that for the single equation

$$
y' = f(x, y); \qquad y(a) = y_0 \tag{4}
$$

the algorithm is

$$
\begin{aligned}
&y_{n+1} = y_n + \tfrac{1}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right), \\
&k_1 = hf(x_n, y_n), \qquad k_2 = hf\left(x_n + \tfrac{h}{2}, y_n + \tfrac{1}{2}k_1\right), \\
&k_3 = hf\left(x_n + \tfrac{h}{2}, y_n + \tfrac{1}{2}k_2\right), \quad k_4 = hf\left(x_{n+1}, y_n + k_3\right).
\end{aligned} \tag{5}
$$

For the system (1) it becomes

$$
\begin{aligned}
u_{n+1} &= u_n + \frac{1}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right), \\
v_{n+1} &= v_n + \frac{1}{6}\left(l_1 + 2l_2 + 2l_3 + l_4\right),
\end{aligned}
$$

$$
\begin{aligned}
k_1 &= hf(x_n, u_n, v_n), \\
l_1 &= hg(x_n, u_n, v_n), \\
k_2 &= hf\left(x_n + \frac{h}{2}, u_n + \frac{1}{2}k_1, v_n + \frac{1}{2}l_1\right), \\
l_2 &= hg\left(x_n + \frac{h}{2}, u_n + \frac{1}{2}k_1, v_n + \frac{1}{2}l_1\right), \\
k_3 &= hf\left(x_n + \frac{h}{2}, u_n + \frac{1}{2}k_2, v_n + \frac{1}{2}l_2\right),
\end{aligned} \tag{6}
$$

$$
\begin{aligned}
l_3 &= hg\left(x_n + \frac{h}{2}, u_n + \frac{1}{2}k_2, v_n + \frac{1}{2}l_2\right), \\
k_4 &= hf\left(x_{n+1}, u_n + k_3, v_n + l_3\right), \\
l_4 &= hg\left(x_{n+1}, u_n + k_3, v_n + l_3\right),
\end{aligned}
$$

and similarly for systems containing more than two equations.

EXAMPLE 2. Let us illustrate (6) using the same system as in Example 1,

$$
\begin{aligned}
u' &= x + v; \qquad u(0) = 0 \\
v' &= uv^2; \qquad v(0) = 1.
\end{aligned}
\tag{7}
$$

With $h = 0.1$, say, (6) gives

$n = 0$:

$$
\begin{aligned}
k_1 &= h(x_0 + v_0) = (0.1)(0 + 1) = 0.1, \\
l_1 &= hu_0v_0^2 = (0.1)(0)(1)^2 = 0, \\
k_2 &= h\left[\left(x_0 + \frac{h}{2}\right) + \left(v_0 + \frac{l_1}{2}\right)\right] = (0.1)\left[(0 + 0.05) + (1 + 0)\right] = 0.105, \\
l_2 &= h\left(u_0 + \frac{k_1}{2}\right)\left(v_0 + \frac{l_1}{2}\right)^2 = (0.1)\left(0 + 0.05\right)\left(1 + 0\right)^2 = 0.005, \\
k_3 &= h\left[\left(x_0 + \frac{h}{2}\right) + \left(v_0 + \frac{l_2}{2}\right)\right] \\
&= (0.1)\left[(0 + 0.05) + (1 + 0.0025)\right] = 0.10525, \\
l_3 &= h\left(u_0 + \frac{k_2}{2}\right)\left(v_0 + \frac{l_2}{2}\right)^2 \\
&= (0.1)\left(0 + 0.0525\right)\left(1 + 0.0025\right)^2 = 0.005276, \\
k_4 &= h\left[x_1 + (v_0 + l_3)\right] \\
&= (0.1)\left[0.1 + (1 + 0.005276)\right] = 0.110528, \\
l_4 &= h\left(u_0 + k_3\right)\left(v_0 + l_3\right)^2 \\
&= (0.1)\left(0 + 0.10525\right)\left(1 + 0.005276\right)^2 = 0.010636, \\
u_1 &= u_0 + \frac{1}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right) \\
&= 0 + \frac{1}{6}\left(0.1 + 0.21 + 0.2105 + 0.110528\right) = \underline{0.105171}, \\
v_1 &= v_0 + \frac{1}{6}\left(l_1 + 2l_2 + 2l_3 + l_4\right) \\
&= 1 + \frac{1}{6}\left(0 + 0.01 + 0.010552 + 0.010636\right) = \underline{1.005198}.
\end{aligned}
$$

$n = 1$:

$$\begin{aligned} k_1 &= 0.110520, & l_1 &= 0.010627, \\ k_2 &= 0.116051, & l_2 &= 0.016382, \\ k_3 &= 0.116339, & l_3 &= 0.016760, \\ k_4 &= 0.122196, & l_4 &= 0.023134, \\ u_2 &= \underline{0.221420}, & v_2 &= \underline{1.021872}, \end{aligned}$$

and so on for $n = 2, 3, \ldots$. We suggest that you fill in the details for the calculation of the $k_1, \ldots, v_2$ values shown above for $n = 1$. ∎

Of course, the idea is to carry out such calculations on a computer, not by hand. The calculations shown in Examples 1 and 2 are merely intended to clarify the methods.

What about higher-order equations? The key is to re-express an nth-order equation as an equivalent system of n first-order equations.

EXAMPLE 3. The problem

$$y''' - xy'' + y' - 2y^3 = \sin x; \qquad y(1) = 2,\ y'(1) = 0,\ y''(1) = -3 \tag{8}$$

can be converted to an equivalent system of three first-order equations as follows. Define $y' = u$ and $y'' = v$ (hence $u' = v$). Then (8) can be re-expressed in the form

$$\begin{aligned} y' &= u; & y(1) &= 2 \\ u' &= v; & u(1) &= 0 \\ v' &= \sin x + 2y^3 - u + xv; & v(1) &= -3. \end{aligned} \tag{9a,b,c}$$

Of the three differential equations in (9), the first two merely serve to introduce the auxiliary dependent variables u and v, and since v' is y''' the third one is a restated version of the given equation $y''' - xy'' + y' - 2y^3 = \sin x$. Equation (9a) is the y equation, so the initial condition is on $y(1)$, namely, $y(1) = 2$, as given in (8). Equation (9b) is the u equation, so the initial condition is on $u(1)$, and we have $u(1) = y'(1) = 0$, from (8). Similarly, for equation (9c).

The system (9) can now be solved by the Euler or fourth-order Runge–Kutta methods or any other such algorithm. To illustrate, let us carry out the first two steps using Euler's method, taking $h = 0.2$, say.

$n = 0$:

$$\begin{aligned} y_1 &= y_0 + u_0 h = 2 + (0)(0.2) = \underline{2}, \\ u_1 &= u_0 + v_0 h = 0 + (-3)(0.2) = -0.6, \\ v_1 &= v_0 + \left(\sin x_0 + 2y_0^3 - u_0 + x_0 v_0\right) h \\ &= -3 + \left[\sin 1 + 2(2)^3 - 0 + (1)(-3)\right](0.2) = -0.231706. \end{aligned}$$

$n = 1$:

$$y_2 = y_1 + u_1 h = 2 + (-0.6)(0.2) = \underline{1.88},$$

$$
\begin{aligned}
u_2 &= u_1 + v_1 h = -0.6 + (-0.231706)(0.2) = -0.646341, \\
v_2 &= v_1 + \left(\sin x_1 + 2y_1^3 - u_1 + x_1 v_1\right) h \\
&= -0.231706 + \left[\sin 0.2 + 2(2)^3 - (-0.6) + (0.2)(-0.231706)\right](0.2) \\
&= 3.118760,
\end{aligned}
$$

and so on for $n = 2, 3, \ldots$.

COMMENT. Observe that at each step we compute y, u, and v, yet we are not really interested in the auxiliary variables u and v. Perhaps we could just compute $y_1, y_2, \ldots$ and not the u, v values? No; equations (9) are coupled so we need to bring all three variables along together. Of course, we don't need to print or plot u and v, but we do need to compute them. ∎

EXAMPLE 4. Examples 1 and 2 involve a system of first-order equations, and Example 3 involve a single higher-order equation. As a final example, consider a combination of the two such as the initial-value problem

$$
\begin{aligned}
u'' - 3xuv &= \sin x; \quad u(0) = 4, \quad u'(0) = -1 \\
v'' + 2u - v &= 5x; \quad v(0) = 7, \quad v'(0) = 0.
\end{aligned}
\tag{10}
$$

The idea is exactly the same as before. We need to recast (10) as a system of first-order initial value problems. We can do so by introducing auxiliary dependent variables w and z according to $u' = w$ and $v' = z$. Then (10) becomes

$$
\begin{aligned}
u' &= w; & u(0) &= 4 \\
w' &= \sin x + 3xuv; & w(0) &= -1 \\
v' &= z; & v(0) &= 7 \\
z' &= 5x - 2u + v; & z(0) &= 0
\end{aligned}
\tag{11}
$$

which system can now be solved by Euler's method or any other such numerical differential equation solver. ∎

6.4.2. Linear boundary-value problems. Our discussion is based mostly upon the following example.

EXAMPLE 5. Consider the third-order boundary-value problem

$$
y''' - x^2 y = -x^4; \qquad y(0) = 0, \; y'(0) = 0, \; y(2) = 4. \tag{12}
$$

To solve numerically, we begin by recasting (12) as the first-order system:

$$
\begin{aligned}
y' &= u; & y(0) &= 0, \quad y(2) = 4 \\
u' &= v; & u(0) &= 0 \\
v' &= x^2 y - x^4.
\end{aligned}
\tag{13a,b,c}
$$

However, we cannot apply the numerical integration techniques that we have discussed because the problem (13c) does not have an initial condition so we cannot get the solution started. Whereas (13c) is missing an initial condition on v, (13a) has an extra condition – the right end condition $y(2) = 4$, but that condition is of no help in developing a numerical integration scheme that develops a solution beginning at $x = 0$.

Nevertheless, the linearity of (12) saves the day and permits us to work with an initial-value version instead. Specifically, suppose that we solve (numerically) the four initial-value problems

$$\begin{aligned} L[Y_1] &= 0, & Y_1(0) &= 1, & Y_1'(0) &= 0, & Y_1''(0) &= 0, \\ L[Y_2] &= 0, & Y_2(0) &= 0, & Y_2'(0) &= 1, & Y_2''(0) &= 0, \\ L[Y_3] &= 0, & Y_3(0) &= 0, & Y_3'(0) &= 0, & Y_3''(0) &= 1, \\ L[Y_p] &= -x^4, & Y_p(0) &= 0, & Y_p'(0) &= 0, & Y_p''(0) &= 0, \end{aligned} \tag{14}$$

where $L = d^3/dx^3 - x^2$ is the differential operator in (12). The nine initial conditions in the first three of these problems were chosen so as to have a nonzero determinant so that Y_1, Y_2, Y_3 comprise a fundamental set of solutions (i.e., a linearly independent set of solutions) of the homogeneous equation $L[Y] = 0$. The three initial conditions on the particular solution Y_p were chosen as zero for simplicity; any values will do since any particular solution will do. Suppose we imagine that the four initial-value problems in (14) have now been solved by the methods discussed above. Then Y_1, Y_2, Y_3, Y_p are known functions of x over the interval of interest $[0, 2]$, and we have the general solution

$$y(x) = C_1Y_1(x) + C_2Y_2(x) + C_3Y_3(x) + Y_p(x) \tag{15}$$

of $L[y] = -x^4$. Finally, we evaluate the integration constants C_1, C_2, C_3 by imposing the boundary conditions given in (12):

$$\begin{aligned} y(0) &= 0 = C_1 + 0 + 0 + 0, \\ y'(0) &= 0 = 0 + C_2 + 0 + 0, \\ y(2) &= 4 = C_1Y_1(2) + C_2Y_2(2) + C_3Y_3(2) + Y_p(2). \end{aligned} \tag{16}$$

Solving (16) gives $C_1 = C_2 = 0$ and $C_3 = [4 - Y_p(2)]/Y_3(2)$, so we have the desired solution of (12) as

$$y(x) = \frac{4 - Y_p(2)}{Y_3(2)}\, Y_3(x) + Y_p(x). \tag{17}$$

In fact, since $C_1 = C_2 = 0$ the functions $Y_1(x)$ and $Y_2(x)$, have dropped out so we don't need to calculate them. All we need are $Y_3(x)$ and $Y_p(x)$, and these are found by the numerical integration of the initial-value problems

$$\begin{aligned} Y_3' &= U_3, & Y_3(0) &= 0, \\ U_3' &= V_3, & U_3(0) &= 0, \\ V_3' &= x^2Y_3, & V_3(0) &= 1, \end{aligned} \tag{18}$$

and

$$\begin{aligned} Y_p' &= U_p, & Y_p(0) &= 0, \\ U_p' &= V_p, & U_p(0) &= 0, \\ V_p' &= x^2 Y_p - x^4, & V_p(0) &= 0, \end{aligned} \tag{19}$$

respectively.

COMMENT. Remember that whereas initial-value problems have unique solutions (if the functions involved are sufficiently well behaved), boundary-value problems can have no solution, a unique solution, or even an infinite number of solutions. How do these possibilities work out in this example? The clue is that (17) fails if $Y_3(2)$ turns out to be zero. The situation is seen more clearly from (16), where all of the possibilities come into view. Specifically, if $Y_3(2) \neq 0$, then we can solve uniquely for C_3, and we have a unique solution, given by (17). If $Y_3(2)$ does vanish, then there are two possiblities as seen from (16): if $Y_p(2) \neq 4$, then there is no solution, and if $Y_p(2) = 4$ then there are an infinite number of solutions of (12), namely,

$$y(x) = C_3 Y_3(x) + Y_p(x), \tag{20}$$

where C_3 remains arbitrary. ∎

We see that boundary-value problems are more difficult than initial-value problems. From Example 5 we see that a nonhomogeneous nth-order linear boundary-value problem generally involves the solution of $n + 1$ initial-value problems, although in Example 5 (in which $n = 3$) we were lucky and did not need to solve for two of the four unknowns, Y_1 and Y_2.

Nonlinear boundary-value problems are more difficult still, because we cannot use the idea of finding a fundamental set of solutions plus a particular solution and thus forming a general solution, as we did in Example 5, and which idea is based upon linearity. One viable line of approach comes under the heading of **shooting methods**. For instance, to solve the nonlinear boundary-value problem

$$y'' + \sin y = 3x; \qquad y(0) = 0,\ y(5) = 2 \tag{21}$$

we can solve the initial-value problem

$$\begin{aligned} y' &= u, & y(0) &= 0 \\ u' &= 3x - \sin y, & u(0) &= u_0 \end{aligned} \tag{22}$$

iteratively. That is, we can guess at the initial condition u_0 [which is the initial slope $y'(0)$] and solve (22) for y(x) and u(x). Next, we compare the computed value of $y(5)$ with the boundary condition $y(5) = 2$ (which we have not yet used). If the computed value is too high, then we return to (22), reduce the value of u_0, and solve again. Comparing the new computed value of $y(5)$ with the prescribed value $y(5) = 2$, we again revise our value of u_0. If these revisions are done in

a rational way, one can imagine obtaining a convergent scheme. Such a scheme is called a *shooting method* because of the obvious analogy with the shooting of a projectile, with the intention of having the projectile strike the ground at some distant prescribed point.

Thus, we can see the increase in difficulty as we move away from linear initial-value problems. For a linear boundary-value problem of order n we need to solve not one problem but $n+1$ of them. For a nonlinear boundary-value problem we need to solve an infinite sequence of them, in principle; in practice, we need to carry out only enough iterations to produce the desired accuracy.

Closure. In Section 6.4.1 we extend the Euler and fourth-order Runge–Kutta solution methods to cover systems of equations and higher-order equations. In that discussion it is more convenient to use n-dimensional vector notation because of its compactness, but that notation is not be introduced until Chapters 9 and 10. Nonetheless, let us indicate the result, if only for the Euler method, for future reference. The idea is that we can express the system

$$\begin{aligned} y_1'(x) &= f_1(x, y_1(x), \ldots, y_n(x)); \quad & y_1(a) &= y_{10}, \\ &\vdots & &\vdots \\ y_n'(x) &= f_n(x, y_1(x), \ldots, y_n(x)); \quad & y_n(a) &= y_{n0}, \end{aligned} \tag{23}$$

in the vector form

$$\mathbf{y}'(x) = \mathbf{f}(x, \mathbf{y}(x)); \qquad \mathbf{y}(a) = \mathbf{y_0}, \tag{24}$$

where the boldface letters denote "n-dimensional column vectors:"

$$\mathbf{y}(x) = \begin{bmatrix} y_1(x) \\ \vdots \\ y_n(x) \end{bmatrix}, \quad \mathbf{y}'(x) = \begin{bmatrix} y_1'(x) \\ \vdots \\ y_n'(x) \end{bmatrix}, \quad \mathbf{f}(x, \mathbf{y}(x)) = \begin{bmatrix} f_1(x, \mathbf{y}(x)) \\ \vdots \\ f_n(x, \mathbf{y}(x)) \end{bmatrix}, \tag{25}$$

and where $f_j(x, \mathbf{y}(x))$ is simply a shorthand notation for $f_j(x, y_1(x), \ldots, y_n(x))$. Then the Euler algorithm corresponding to (24) is

$$\mathbf{y_{n+1}} = \mathbf{y_n} + \mathbf{f}(x_n, \mathbf{y_n})\, h. \tag{26}$$

In Section 6.4.2 we turn to boundary-value problems, but only linear ones. Given an nth-order linear boundary-value problem $L[y] = f(x)$ on an interval $[a, b]$ plus n boundary conditions the idea is to solve the problems

$$\begin{aligned} &L[Y_1] = 0; \quad Y_1(a) = 1,\ Y_1'(a) = \cdots = Y_1^{(n-1)}(a) = 0, \\ &\vdots \qquad\qquad\qquad\qquad \vdots \\ &L[Y_n] = 0; \quad Y_n(a) = Y_n'(a) = \cdots = Y_n^{(n-2)}(a) = 0,\ Y_n^{(n-1)}(a) = 1, \\ &L[Y_p] = f; \quad Y_p(a) = \cdots = Y_p^{(n-1)}(a) = 0 \end{aligned} \tag{27}$$

for $Y_1(x), \ldots, Y_n(x), Y_p(x)$ and to form the general solution as

$$y(x) = C_1 Y_1(x) + \cdots + C_n Y_n(x) + Y_p(x). \tag{28}$$

Finally, application of the original boundary conditions to (28) yields n linear algebraic equations for $C_1, \ldots, C_n$, which equations will have a unique solution, no solution, or an infinity of solutions.

Computer software. No new software is needed for the methods described in this section. For instance, we can use the *Maple* command dsolve, with the numeric option, to solve the problem

$$u' = x + v; \quad u(0) = 0$$
$$v' = -5uv; \quad v(0) = 1$$

and to print the results at $x = 1, 2$, and 3. First, enter

with(DEtools):

and return. Then enter

dsolve({diff($u(x), x$) = $x + v(x)$, diff($v(x), x$) = $-5 * u(x) * v(x)$,
$u(0) = 0$, $v(0) = 1$}, {$u(x), v(x)$}, type = numeric,
value = array ([1, 2, 3]));

and return. The printed result is

$$\left[\begin{array}{c} [x, u(x), v(x)] \\ \left[\begin{array}{lll} 1. & 1.032499017614234 & .07285274036469075 \\ 2. & 2.544584704578166 & .00001413488345836790 \\ 3. & 5.044585755162072 & -.3131443346304622 \times 10^{-9} \end{array} \right] \end{array} \right]$$

The only differences between the command above and the one given at the end of Section 6.3 is that here we have entered two differential equations, two initial conditions, and two dependent variables, and we have omitted the abserr option.

Observe that to solve a differential equation, or system of differential equations, numerically, we must first express the equations as a system of first-order equations, as illustrated in Example 4. However, to use the *Maple* dsolve command we can leave the original higher-order equations intact.

EXERCISES 6.4

1. In Example 2 we gave $k_1, l_1, k_2, l_2, k_3, l_3, k_4, l_4$ for $n = 1$, and the resulting values of u_2 and v_2, but did not show the calculations. Provide those calculations, as we did for the step $n = 0$.

2. As we did in Example 1, work out y_1, z_1, by hand. Use three methods: Euler, second-order Runge–Kutta, and fourth-order Runge–Kutta, and take $h = 0.2$. These problems are rigged so as to have simple closed-form solutions, some of which are given in brackets. Compare your results with the exact solution.

(a) $y' = z; \quad y(0) = 1$
$z' = -y; \quad z(0) = 0$

(b) $y' = 4z; \quad y(2) = 5$
$z' = -y; \quad z(2) = 0$

(c) $y' = -z^3/y; \quad y(0) = 1 \quad [y(x) = e^{-x}]$
$z' = -y; \quad z(0) = 0 \quad [z(x) = e^{-x}]$

(d) $y' = 2xz^2/y; \quad y(1) = 1 \quad [y(x) = x^2]$
$z' = y/z^2; \quad z(1) = 1 \quad [z(x) = x]$

(e) $y' = (x + y)z - 1; \quad y(1) = 1 \quad [y(x) = x]$
$z' = -yz^3; \quad z(1) = 1 \quad [z(x) = 1/x]$

3. (a)–(e) First, read Exercise 2. Use computer software to solve the initial-value problem given in the corresponding part of Exercise 2, for $y(x)$ and $z(x)$ at $x = 3, 5$, and 10, and compare those results with the exact solution at those points.

4. (a) Just as (2) and (6) give the Euler and fourth-order Runge-Kutta algorithms for the second-order system (1), write down the analogous Euler, second-order Runge–Kutta, and fourth-order Runge–Kutta algorithms for the third-order system

$$\begin{aligned} x'(t) &= f(t, x, y, z), \quad x(a) = x_0 \\ y'(t) &= g(t, x, y, z), \quad y(a) = y_0 \\ z'(t) &= f(t, x, y, z), \quad z(a) = z_0. \end{aligned} \tag{4.1}$$

Use the Euler and second-order Runge–Kutta algorithms to work out x_1, y_1, z_1 and x_2, y_2, z_2, by hand, for the case where f, g, h are $y - 1, z, t + x + 3(z - y + 1)$, respectively, with the initial conditions $x(0) = -3$, $y(0) = 0$, $z(0) = 2$ using $h = 0.3$.

(b) Same as (a), but with $x(0) = y(0) = z(0) = 0$.

(c) Same as (a), but with $x(0) = 1, y(0) = 0, z(0) = 0$

(d) Same as (a), but with $x(0) = y(0) = 0, z(0) = 10$.

5. We re-expressed (8) and (10) as the equivalent systems of first-order initial-value problems (9) and (11), respectively. Do the same for the problem given. You need go no further.

(a) $mx'' + cx' + kx = f(t); \quad x(0) = x_0, \quad x'(0) = x_0'$

(b) $Li'' + Ri' + (1/C)i = E'(t); \quad i(0) = i_0, \quad i'(0) = i_0'$

(c) $y'' - xyy' = \sin x; \quad y(1) = 5, \quad y'(1) = -1$

(d) $y'' + y' - 4y = 3x; \quad y(0) = 2, \quad y'(0) = 1$

(e) $y''' - 2\sin y' = 3x; \quad y(-2) = 7, \quad y'(-2) = 4$
$y''(-2) = 0$

(f) $y''' + xy = \cos 2x; \quad y(1) = 3, \quad y'(1) = 2, \quad y''(1) = 0$

(g) $x'' + 2x - 3y = 10\cos 3t; \quad x(0) = 2, \quad x'(0) = -1$
$y'' - x + 5y = 0; \quad y(0) = 4, y'(0) = 3$

(h) $y''' + xy'z = f(x); \quad y(3) = 2, \quad y'(3) = -1$
$y''(3) = 6$
$z'' + y - z = g(x); \quad z(3) = 0, z'(3) = 8$

(i) $x'' - 3xz = \sin t; \quad x(1) = x'(1) = 0, \quad x''(1) = 3$
$y' + 2x + y - 5z = 0; \quad y(1) = 6$
$z' - 4xy = e^{2t}; \quad z(1) = 2$

(j) $y''' = yz; \quad y(0) = 1, \quad y'(0) = y''(0) = 0$
$z'' = -xy + z; \quad z(0) = 5, z'(0) = 1$

6. Use computer software to solve the given system numerically, and print out the solution for $y(x)$ and $z(x)$ at $x = 1, 2$.

(a) $y'' - 2xz' = 5x; \quad y(0) = 2, \quad y'(0) = -1$
$z' + yz^3 = -3; \quad z(0) = 1$

(b) $y' + 3xz = x^2; \quad y(0) = 1$
$z''' - y^2z' + z = 0; \quad z(0) = z'(0) = 2, \quad z''(0) = -1$

(c) $y' = z' + x; \quad y(2) = 1$
$z'' + y^2z^2 = 3; \quad z(2) = -1, \quad z'(2) = 0$

(d) $y' - z'' = x; \quad y(1) = -1$
$z''' - y^2z = 3; \quad z(1) = 1, \quad z'(1) = z''(1) = 0$

7. Complete the solution of Example 5 by using computer software to solve (18) for $Y_3(x)$ and (19) for $Y_p(x)$, at $x = 2, 4, 6$, and then using (17) to determine $y(x)$ at those points.

8. Use the method explained in Example 5 to reduce the given linear boundary-value problem to a system of linear initial-value problems. Then complete the solution and solve for the specified quantity, either using computer software or by programming any of the numerical solution methods that we have studied. Obtain accuracy to five significant figures, and indicate why you believe that you have achieved that accuracy.

If you believe that there is no solution or that it exists but is nonunique, then give your reasoning. HINT: You can specify homogeneous initial conditions for the Y_p problem, as we did in Example 5, but be aware that you do not *have* to use homogeneous conditions, and that you may be able to reduce your labor by a more optimal choice of those conditions.

(a) $y'' - 2xy' + y = 3\sin x; \quad y(0) = 1, \quad y(2) = 3.$
Determine $y(1)$.
(b) $y'' + (\cos x)y = 0; \quad y(0) = 1, \quad y(10) = 2.$
Determine $y(2)$.
(c) $y'' - [\ln(x+1)]\,y' - y = 2\sin 3x + 1;$
$y(0) = 3, \quad y(2) = -1.$
Determine $y(x)$ at $x = 0.5, 1.0, 1.5$.
(d) $y'' + y' - xy = x^3; \quad y(0) = 1, \quad y(5) = 2.$
Determine $y(x)$ at $x = 1, 2, 3, 4$.
(e) $y''' + xy = 2x^3; \quad y(1) = y'(1) = 0, \quad y''(2) = -3.$
Determine $y(2)$.
(f) $y''' + xy' + y = x; \; y(1) = 2, \; y'(1) = 0.4, \; y'(5) = 3.$
Determine $y(x)$ at $x = 4, 5$.

6.5 Stability and Difference Equations

6.5.1. Introduction. In progressing from the simple Euler method to the more sophisticated higher-order methods our aim was improvement in accuracy. However, there are cases where the results obtained not only fail to be sufficiently accurate but are grossly incorrect, as illustrated in the two examples to follow. The second one introduces the idea of stability, and in Section 6.5.2 we concentrate on that topic.

EXAMPLE 1. The initial-value problem

$$y' - 2y = -6e^{-4x}; \quad y(0) = 1 \tag{1}$$

has the exact solution $y(x) = \exp(-4x)$. If we solve it by the fourth-order Runge–Kutta method for the step sizes $h = 0.1, 0.05$, and 0.01, we obtain in Table 1 the results shown, at

Table 1. Runge–Kutta solution of (1).

x	$h = 0.1$	$h = 0.05$	$h = 0.01$	Exact
1	0.179006 E−1	0.182893 E−1	0.183153 E−1	0.183156 E−1
4	−0.167842 E+0	−0.106538 E−1	−0.146405 E−3	0.112535 E−6
8	−0.500286 E+3	−0.317586 E+2	−0.436763 E+0	0.126642 E−13
12	−0.149120 E+7	−0.946704 E+5	−0.130197 E+4	0.142516 E−20

the representative points $x = 1, 4, 8$, and 12. Since the Runge–Kutta method is convergent, the results should converge to the exact solution at any given x as h tends to zero, but that convergence is hard to see in the tabulated results except for $x = 1$. In fact, it is doubtful that we could ever come close to the exact values at $x = 8$ or 12 since we might need to make h so small that roundoff errors might come to dominate before the accumulated truncation error is sufficiently reduced.

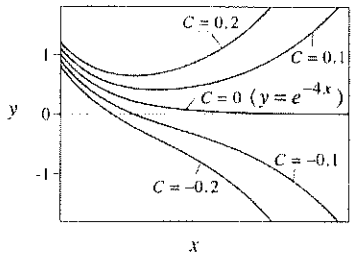


Figure 1. Solution curves for the equation $y' - 2y = -6e^{-4x}$.

More central to the purpose of this example is to see that with h fixed the results diverge dramatically from the exact solution as x increases so as to become grossly incorrect. We cannot blame this strange and unexpected result on complications due to nonlinearity because (1) is linear.

To understand the source of the difficulty, note that the general solution of the differential equation is $y(x) = \exp(-4x) + C\exp(2x)$, where C is an arbitrary constant. The initial condition implies that $C = 0$, leaving the particular solution $y(x) = \exp(-4x)$. In Fig. 1 we show several solution curves for values of C close to and equal to zero, and we can see the rapid divergence of neighboring curves from the solution $y(x) = \exp(-4x)$. Thus, the explanation of the difficulties found in the tabulated numerical results is that even a very small numerical error shifts us from the exact solution curve to a neighboring curve, which then diverges from the true solution. ∎

EXAMPLE 2. In Example 3 of Section 6.3 we solved the equation $y' = -y$, with initial condition $y(0) = 1$, by several methods – from the simple Euler method to the more accurate and sophisticated fourth-order Runge–Kutta method, and we gave the results in Table 2. Since the midpoint rule and the second-order Runge–Kutta methods are both of second order we expected their accuracy to be comparable. Indeed they were initially, but the midpoint rule eventually developed an error that oscillated in sign from step to step and grew in magnitude (see Table 2 in Section 6.3). Let us solve the similar problem

$$y' = -2y; \quad y(0) = 1 \tag{2}$$

by the midpoint rule, with $h = 0.05$. Since the midpoint rule is not self-starting, we use ten Euler steps from $x = 0$ to $x = 0.05$ before switching over to the midpoint rule. We have plotted the results in Fig. 2, along with the exact solution, $y(x) = \exp(-2x)$. Once again, we see that the midpoint rule results follow the exact solution initially, but they develop an error that oscillates in sign and grows such that the results are soon hopelessly incorrect.

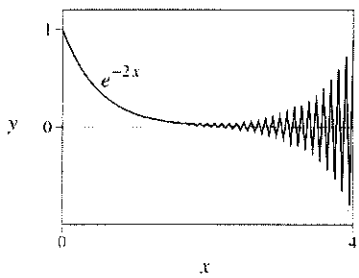


Figure 2. Illustration of numerical instability associated with the midpoint rule, for the initial-value problem (2).

This numerical difficulty is different from the one found above in Example 1, for rather than being due to an extreme sensitivity to initial conditions, it is associated with machine roundoff error and is an example of numerical instability. ∎

6.5.2. Stability. Let us analyze the phenomenon of numerical instability that we encountered in Example 2. Recall that we denote the exact solution of a given initial-value problem as $y(x_n)$ and the numerical solution as y_n. Actually, the latter is not quite the same as the computer printout because of the inevitable presence of machine roundoff errors. Thus, let us distinguish further between the numerical solution y_n that would be generated on a perfect computer, and the solution y_n^* that is generated on a real machine and which includes the effects of numerical roundoff – that is, the truncation of numbers after a certain number of significant figures.

It is useful to decompose the total error, at any nth step, as

$$\boxed{\begin{aligned}\text{Total error} &= y(x_n) - y_n^* = [y(x_n) - y_n] + [y_n - y_n^*] \\ &= [\text{accum. truncation error}] + [\text{accum. roundoff error}].\end{aligned}} \tag{3}$$

We ask two things of a method: first, that the accumulated truncation error tend to zero at any fixed x as the step size h tends to zero and, second, that the accumulated roundoff error remain small compared to the exact solution. The first is the issue of **convergence**, discussed earlier in this chapter, and the second is the issue of **stability**, our present concern.

We have already noted that the midpoint rule can produce the strange behavior shown in Fig. 2, so let us study the application of that method to the standard "test problem,"

$$y' = Ay; \quad y(0) = 1, \tag{4}$$

where it is useful to include the constant A as a parameter. The midpoint rule generates y_n according to the algorithm

$$\begin{aligned} y_{n+1} &= y_{n-1} + 2hf(x_n, y_n) \\ &= y_{n-1} + 2Ay_n; \qquad y_0 = 1 \end{aligned} \tag{5}$$

for $n = 0, 1, 2, \ldots$.

To determine whether a solution algorithm, in this case (5), is stable, it is customary to "inject" a roundoff error at any step in the solution, say at $n = 0$, and to see how much the perturbed solution differs from the exact solution as n increases, assuming that no further roundoff errors occur. Thus, in place of (5), consider the perturbed problem

$$y^*_{n+1} = y^*_{n-1} + 2Ahy^*_n; \quad y^*_0 = 1 - \epsilon, \tag{6}$$

say, where ϵ is the (positive or negative) roundoff error in the initial condition. Defining the error $e_n \equiv y_n - y^*_n$ and subtracting (6) from (5), gives

$$e_{n+1} = e_{n-1} + 2Ahe_n, \tag{7}$$

with the initial condition $e_0 = \epsilon$, as governing the evolution of e_n. We call (7) a **difference equation**. Just as certain differential equations can be solved by seeking solutions in the form $y(x) = e^{\lambda x}$, the appropriate form for the difference equation (7) is

$$e_n = \rho^n, \tag{8}$$

where ρ is to be determined. Putting this expression into (7) gives

$$\rho^{n+1} - 2Ah\rho^n - \rho^{n-1} = 0 \tag{9}$$

or

$$\left(\rho^2 - 2Ah\rho - 1\right)\frac{1}{\rho^n} = 0. \tag{10}$$

Since $1/\rho^n$ is not zero, it follows from (10) that we must have $\rho^2 - 2Ah\rho - 1 = 0$, so we have the two roots

$$\rho = Ah + \sqrt{1 + A^2h^2} \quad \text{and} \quad \rho = Ah - \sqrt{1 + A^2h^2}. \tag{11}$$

By considerations analogous to those for differential equations, we have

$$e_n = C_1\left(Ah + \sqrt{1 + A^2h^2}\right)^n + C_2\left(Ah - \sqrt{1 + A^2h^2}\right)^n \tag{12}$$

as the general solution of (7).

If we let $h \to 0$, then

$$\left(Ah + \sqrt{1 + A^2h^2}\right)^n \sim (Ah + 1)^n = e^{n\ln(1+Ah)} \sim e^{nAh} = e^{Ax_n}, \tag{13}$$

where we have used the identity $a = e^{\ln a}$, the Taylor expansion $\ln(1 + x) = x - x^2/2 + \cdots \sim x$, and the fact that $x_n = nh$. Similarly,

$$\begin{aligned}\left(Ah - \sqrt{1 + A^2h^2}\right)^n &\sim (Ah - 1)^n \\ &= (-1)^n(1 - Ah)^n = (-1)^n e^{n\ln(1-Ah)} \\ &\sim (-1)^n e^{-nAh} = (-1)^n e^{-Ax_n},\end{aligned} \tag{14}$$

so (12) becomes

$$e_n = C_1 e^{Ax_n} + C_2(-1)^n e^{-Ax_n} \tag{15}$$

as $h \to 0$. Since there are two arbitrary constants, C_1 and C_2, two initial conditions are appropriate, whereas we have attached only the single condition $e_0 = \epsilon$ in (7). With no great loss of generality let us specify as a second initial condition $e_1 = 0$. Imposing these conditions on (15), we have

$$\begin{aligned}e_0 &= \epsilon = C_1 + C_2, \\ e_1 &= 0 = C_1 e^{4x_1} - C_2 e^{-4x_1}.\end{aligned}$$

Finally, solving for C_1 and C_2 and inserting these values into (15), gives

$$e_n = \frac{\epsilon}{2\cosh Ax_1}\left[e^{A(x_n - x_1)} + (-1)^n e^{-A(x_n - x_1)}\right]. \tag{16}$$

To infer from (16) whether the method is stable or not, we consider the cases $A > 0$ and $A < 0$ separately. If $A > 0$, then the second term in (16) decays to zero, and even though the first term grows exponentially, it remains small compared to the exact solution $y(x_n) = \exp(Ax_n)$ as n increases because ϵ is very small (for example, on the order of 10^{-10}). We conclude, formally, that if $A > 0$ then the midpoint rule is stable.

On the other hand, if $A < 0$, then the second term starts out quite small, due to the ϵ factor, but grows exponentially with x_n and oscillates due to the $(-1)^n$, whereas the exact solution is $\exp(-x)$. This is precisely the sort of behavior that was observed in Example 2 (where A was -2), and we conclude that if $A < 0$, then the midpoint rule is unstable.

Since the stability of the midpoint rule depends upon the sign of A in the test equation $y' = Ay$ (stability for $A > 0$ and instability for $A < 0$), we say that the

midpoint rule is only **weakly stable**. If, instead, a method is stable independent of the sign of A, then we classify it as **strongly stable**.

Having found that the midpoint rule when applied to the equation $y' = Ay$ is stable for $A > 0$ and unstable for $A < 0$, what about the stability of the midpoint rule when it is applied to an equation $y' = f(x, y)$ that is more complicated? Observing that A is the partial derivative of Ay (i.e., the right-hand side of $y' = Ay$) with respect to y, we expect, as a rule of thumb, the midpoint rule to be stable if $\partial f/\partial y > 0$ and unstable if $\partial f/\partial y < 0$ over the x, y domain of interest. For instance, if $y' = e^{xy}$ on $x > 0$, then we can expect the midpoint rule to be stable because $\partial(e^{xy})/\partial y = xe^{xy} > 0$ on $x > 0$, but if $y' = e^{-xy}$ on $x > 0$, then we can expect the midpoint rule to be unstable on $x > 0$ because $\partial(e^{-xy})/\partial y = -xe^{-xy} < 0$ on $x > 0$.

Besides arriving at the above-stated conclusions as to the stability of the midpoint rule for the test equation $y' = Ay$, we can now understand the origin of the instability, for notice that the difference equations $y_{n+1} - 2Ahy_n - y_{n-1} = 0$ and $e_{n+1} - 2Ahe_n - e_{n-1} = 0$, governing y_n and e_n, are identical. Thus, analogous to (15) we must have

$$y_n = B_1 e^{Ax_n} + B_2(-1)^n e^{-Ax_n} \tag{17}$$

for arbitrary constants B_1, B_2, as h tends to zero. The first of these terms coincides with the exact solution of the original equation $y' = Ay$, and the second term (which gives rise to the instability if $A < 0$) is an extraneous solution that enters because we have replaced the original first-order differential equation by a second-order difference equation (second-order because the difference between the subscripts $n+1$ and $n-1$ is 2). Single-step methods (e.g., Euler and Runge–Kutta) are strongly stable (i.e., independent of the sign of A) because the resulting difference equation is only first order so there are no extraneous solutions. Thus, we can finally see why, in Example 3 of Section 6.3, the midpoint rule proved unstable but the other methods were stable.

Understand that these stability claims are based upon analyses in which we let h tend to zero, whereas in practice h is, of course, finite. To illustrate what can happen as h is varied, let us solve

$$y' = -1000(y - x^3) + 3x^2; \quad y(0) = 0 \tag{18}$$

by Euler's method. The exact solution is simply $y(x) = x^3$ so that at $x = 1$, for instance, we have $y(1) = 1$. By comparison, the values computed by Euler's method are as given in Table 2.

Even from this limited data we can see that we do have the stability claimed above for the single-step Euler method, but only when h is made sufficiently small. To understand this behavior, consider the relevant test equation $y' = -1000y$, namely, $y' = Ay$, where $A = \partial[-1000(y - x^3) + 3x^2]/\partial y = -1000$. Then Euler's method for that test equation is $y_{n+1} = y_n - 1000hy_n$. Similarly, $y^*_{n+1} = y^*_n - 1000hy^*_n$. Subtracting these two equations, we find that the roundoff error $e_n = y_n - y^*_n$ satisfies the simple difference equation

$$e_{n+1} = (1 - 1000h)e_n. \tag{19}$$

Table 2. Finite-h stability.

h	Computed $y(1)$
0.2500	2.3737566×10^5
0.1000	8.7725049×10^{14}
0.0100	Exponential overflow
0.0010	0.99999726
0.0001	0.99999970

Letting $n = 0, 1, 2, \ldots$ in (19) reveals that the solution of (19) is

$$e_n = (1 - 1000h)^n e_0, \tag{20}$$

where e_0 is the initial roundoff error. If we take the limit as $h \to 0$, then

$$e_n = (1 - 1000h)^n e_0 = e_0 e^{n \ln(1-1000h)} \sim e_0 e^{-1000nh} = e_0 e^{-1000x_n}, \tag{21}$$

which is small compared to the exact solution $y_n = e^{-1000x_n}$ because of the e_0 factor, so the method is stable. This result is in agreement with the numerical results given in Table 2: as $h \to 0$ the scheme is stable. However, in a real calculation h is, of course, finite and it appears, from the tabulation that there is some critical value, say h_{cr}, such that the guaranteed stability is realized only if $h < h_{cr}$. To see this, let us retain (20) rather than let $h \to 0$. It is seen from (20) that if $|1 - 1000h| < 1$, then $e_n \to 0$ as $n \to \infty$, and if $|1 - 1000h| > 1$, then $e_n \to \infty$ as $n \to \infty$. Thus, for stability we need $|1 - 1000h| < 1$ or $-1 < 1 - 1000h < 1$. The right-hand inequality imposes no restriction on h because it is true for all h's (provided that h is positive, as is normally the case), and the left-hand inequality is true only for $h < 0.002$. Hence $h_{cr} = 0.002$ in this example, and this result is consistent with the tabulated results, which show instability for the h's greater than that value, and stability for the h's smaller than that value. Thus, when we say that the Euler method is strongly stable, what we should really say is that it is strongly stable for *sufficiently small* h. Likewise for the Runge–Kutta and other single-step methods.

6.5.3. Difference equations. (Optional) Difference equations are important in their own right, and the purpose of this Section 6.5.3 is not only to clarify some of the steps in Section 6.5.2, but also to take this opportunity to present some basics regarding the theory and solution of such equations.

To begin, we define a **difference equation of order** N as a relation involving $y_n, y_{n+1}, \ldots, y_{n+N}$. As we have seen, one way in which difference equations arise is in the numerical solution of differential equations. For instance, if we discretize the differential equation $y' = -y$ and solve by Euler's method or the midpoint rule, then in place of the differential equation we have the first- and second-order difference equations $y_{n+1} = y_n - hy_n = (1 - h)y_n$ and $y_{n+1} = y_{n-1} - 2hy_n$, or

$$y_{n+1} - (1 - h)y_n = 0 \tag{22}$$

and

$$y_{n+1} + 2hy_n - y_{n-1} = 0, \tag{23}$$

respectively. In case it is not clear that (23) is of second order, we could let $n - 1 = m$ and obtain $y_{m+2} + 2hy_{m+1} - y_m = 0$ instead, which equation is more clearly of second order. That is, the order is always the difference between the largest and smallest subscripted indices.

Analogous to differential equation terminology, we say that (22) and (23) are **linear** because they are of the form

$$a_0(n)y_{n+N} + a_1(n)y_{n+N-1} + \cdots + a_N(n)y_n = f(n), \tag{24}$$

homogeneous because $f(n)$ is zero in each case, and of **constant-coefficient** type because their a_j's are constants rather than functions of n. By a solution of (24) is meant any sequence y_n that reduces (24) to a numerical identity for each n under consideration, such as $n = 0, 1, 2, \ldots$.

The theory of difference equations is analogous to that of differential equations. For instance, just as one seeks solutions to a linear homogeneous differential equation with constant coefficients in the form $y(x) = e^{\lambda x}$, one seeks solutions to a linear homogeneous difference equation with constant coefficients in the form

$$y_n = \rho^n \tag{25}$$

as we did in Section 6.5.2. [In case these forms don't seem analogous, observe that $e^{\lambda x} = (e^{\lambda})^x$ is a constant to the power x, just as ρ^n is a constant to the power n.] Putting (25) into such an Nth-order difference equation gives an Nth-degree polynomial equation on ρ, the characteristic equation corresponding to the given difference equation, and if the N roots $(\rho_1, \ldots, \rho_N)$ are distinct, then

$$y_n = C_1\rho_1^n + \cdots + C_N\rho_2^N, \tag{26}$$

where the C_j's are arbitrary constants, can be shown to be a **general solution** of the difference equation in the sense that every solution is of the form (26) for some specific choice of the C_j's. For an Nth-order linear differential equation, N initial conditions (y and its first $N - 1$ derivatives at the initial point) are appropriate for narrowing a general solution down to a particular solution. Likewise for a linear difference equation N initial conditions are appropriate – namely, the first N values $y_0, y_1, \ldots, y_{N-1}$.

EXAMPLE 3. Solve the difference equation

$$y_{n+1} - 4y_n = 0. \tag{27}$$

Since (27) is linear, homogeneous and of constant-coefficient type, seek solutions in the form (25). Putting that form into (27) gives

$$\rho^{n+1} - 4\rho^n = (\rho - 4)\rho^n = 0 \tag{28}$$

so that if $\rho \neq 0$ then $\rho - 4 = 0$, $\rho = 4$, and the general solution of (27) is

$$y_n = C(4)^n. \tag{29}$$

For example, if an initial condition $y_0 = 3$ is specified, then $y_0 = 3 = C(4)^0 = C$ gives $C = 3$ and hence the particular solution $y_n = 3(4)^n$.

Actually, (29) is simple enough to solve more directly since (for $n = 0, 1, \ldots$) it gives $y_1 = 4y_0$, $y_2 = 4y_1 = 4^2 y_0$, $y_3 = 4y_2 = 4^3 y_0$, and so on, so one can see that $y_n = y_0(4)^n$ or, if y_0 is not specified, $y_n = C(4)^n$, which is the same as (29). ∎

EXAMPLE 4. Solve the difference equation

$$y_{n+2} - y_{n+1} - 6y_n = 0. \tag{30}$$

Seeking solutions in the form (25) gives the characteristic equation $\rho^2 - \rho - 6 = 0$ with roots -2 and 3 so the general solution of (30) is

$$y_n = C_1(-2)^n + C_2(3)^n. \tag{31}$$

If initial conditions are prescribed, say $y_0 = 4$ and $y_1 = -13$, then

$$\begin{aligned} y_0 &= 4 = C_1 + C_2, \\ y_1 &= -13 = -2C_1 + 3C_2 \end{aligned}$$

give $C_1 = 5$ and $C_2 = -1$. ∎

If the characteristic polynomial has a pair of complex conjugate roots, say $\rho_1 = \alpha + i\beta$ and $\rho_2 = \alpha - i\beta$, then the solution can still be expressed in real form, for if we express ρ_1 and ρ_2 in polar form (as is explained in Section 22.2 on complex numbers) as

$$\rho_1 = r\,e^{i\theta} \quad \text{and} \quad \rho_2 = r\,e^{-i\theta}, \tag{32}$$

where $r = \sqrt{\alpha^2 + \beta^2}$ and $\theta = \tan^{-1}(\beta/\alpha)$, then

$$\begin{aligned} C_1\rho_1^n + C_2\rho_2^n &= C_1 r^n e^{in\theta} + C_2 r^n e^{-in\theta} \\ &= r^n \left(C_1 e^{in\theta} + C_2 e^{-in\theta}\right) \\ &= r^n \left[C_1 (\cos n\theta + i \sin n\theta) + C_2 (\cos n\theta - i \sin n\theta)\right] \\ &= r^n \left(C_3 \cos n\theta + C_4 \sin n\theta\right), \end{aligned} \tag{33}$$

where $C_3 = C_1 + C_2$ and $C_4 = i(C_1 - C_2)$ are arbitrary constants.

EXAMPLE 5. Solve the difference equation

$$y_{n+2} + 4y_n = 0. \tag{34}$$

The characteristic roots are $\pm 2i$ so (32) becomes $\rho_1 = 2e^{i\pi/2}$ and $\rho_2 = 2e^{-i\pi/2}$, and (33) gives the general solution

$$y_n = 2^n \left(A \cos \frac{n\pi}{2} + B \sin \frac{n\pi}{2} \right), \tag{35}$$

where A, B are arbitrary constants. ∎

As we have seen, one way in which difference equations arise is in the numerical solution of differential equations, wherein the continuous process (described by the differential equation) is approximated by a discrete one (described by the difference equation). However, they also arise directly in modeling discrete processes. To illustrate, let p_n be the principal in a savings account at the end of the nth year. We say that the process is discrete in that p is a function of the integer variable n rather than a continuous time variable t. If the account earns an annual interest of I percent, then the growth of principal from year to year is governed by the difference equation

$$p_{n+1} = \left(1 + \frac{I}{100} \right) p_n, \tag{36}$$

which is of the same form as (22).

In fact, discrete processes governed by nonlinear difference equations are part of the modern theory of *dynamical systems*, in which theory the phenomenon of chaos plays a prominent role. Let us close with a brief mention of one such discrete process that is familiar to those who study dynamical systems and chaos. Let x_n, y_n be a point in a Cartesian x, y plane, and let its polar coordinates be r and θ_n. Consider a simple process, or **mapping**, which sends that point into a point x_{n+1}, y_{n+1} at the same radius but at an incremented angle $\theta_n + \alpha$. Then, recalling the identity $\cos(A + B) = \cos A \cos B - \sin A \sin B$, we can express $x_{n+1} = r \cos(\theta_n + \alpha) = r \cos \theta_n \cos \alpha - r \sin \theta_n \sin \alpha = x_n \cos \alpha - y_n \sin \alpha$ and, recalling the identity $\sin(A + B) = \sin A \cos B + \sin B \cos A$, we can express $y_{n+1} = r \sin(\theta_n + \alpha) = r \sin \theta_n \cos \alpha + r \cos \theta_n \sin \alpha = y_n \cos \alpha + x_n \sin \alpha$. Thus, the process is described by the system of linear difference equations

$$\begin{aligned} x_{n+1} &= (\cos \alpha) x_n - (\sin \alpha) y_n, \\ y_{n+1} &= (\sin \alpha) x_n + (\cos \alpha) y_n. \end{aligned} \tag{37}$$

Surely, if one plots such a sequence of points it will fall on the circle of radius r centered at the origin. Suppose that we now modify the process by including two quadratic x_n^2 terms, so that we have the *non*linear system

$$\begin{aligned} x_{n+1} &= (\cos \alpha) x_n - (\sin \alpha) \left(y_n - x_n^2 \right), \\ y_{n+1} &= (\sin \alpha) x_n + (\cos \alpha) \left(y_n - x_n^2 \right). \end{aligned} \tag{38}$$

This system, studied initially by *M. Henòn*, turns out to be remarkably complex and interesting by virtue of the nonlinearity. For a discussion of the main results, we

highly recommend the little book *Mathematics and the Unexpected*, by Ivar Ekeland (Chicago: University of Chicago Press, 1988).

Closure. This section is primarily about the concept of stability in the numerical solution of differential equations. A scheme is stable if the roundoff error remains small compared to the exact solution. Normally, one establishes the stability or instability of a method with respect to the simple test equation $y' = Ay$. Assuming that roundoff enters in the initial condition and that the computer is perfect thereafter, one can derive a difference equation governing the roundoff error e_n, and solve it analytically to see if e_n remains small. Doing so, we show that the midpoint rule is only weakly stable: stable if $A > 0$ and unstable if $A < 0$. As a rule of thumb, we suggest that for a given differential equation $y' = f(x, y)$ we can expect the midpoint rule to be stable if $\partial f/\partial y > 0$ and unstable if $\partial f/\partial y < 0$ over the x, y region of interest.

To explain the source of the instability in the midpoint rule, we observe that the exact solution (17) of the midpoint rule difference equation corresponding to the test equation $y' = Ay$ contains two terms, one that corresponds to the exact solution of $y' = Ay$ and the other extraneous. The latter enters because the midpoint rule difference equation is of second order, whereas the differential equation is only of first order, and it is that extraneous term that leads to the instability. Single-step methods such as the Euler and Runge–Kutta methods, however, are strongly stable, provided that h is sufficiently small.

Observe that the only multi-step method that we examine is the midpoint rule; we neither show nor claim that all multi-step methods exhibit such instability. For instance, it is left for the exercises to show that the multi-step fourth-order Adams–Moulton method is strongly stable (for sufficiently small h). Thus, the idea is that the extraneous terms in the solution, that arise because the difference equation is of a higher order than the differential equation, can, but need not, cause trouble.

We close the section with a brief study of difference equations, independent of any connection with differential equations and stability since they are important in their own right in the modeling of discrete systems. We stress how analogous are the theories governing differential and difference equations that are linear, homogeneous, and with constant coefficients.

Computer software. Just as many differential equations can be solved analytically by computer-algebra systems, so can many difference equations. Using *Maple*, for instance, the relevant command is **rsolve**. For instance, to solve the difference equation $y_{n+2} - y_{n+1} - 6y_n = 0$ (from Example 4), enter

$$\text{rsolve}(y(n+2) - y(n+1) - 6 * y(n) = 0,\ y(n));$$

and return. The result is

$$\left(\frac{3}{5}y(0) - \frac{1}{5}y(1)\right)(-2)^n - \left(-\frac{2}{5}y(0) - \frac{1}{5}y(1)\right)(3)^n$$

the correct solution for any initial values $y(0)$ and $y(1)$. Of course, we could re-

express the latter as

$$C_1(-2)^n + C_2(3)^n$$

if we wish. If we had initial conditions $y_0 = 4$ and $y_1 = -7$, say, then we would have entered

rsolve($\{y(n+2) - y(n+1) - 6 * y(n) = 0,\ y(0) = 4,\ y(1) = -7\}$, $y(n)$);

and would have obtained

$$\frac{19}{5}(-2)^n + \frac{1}{5}(3)^n$$

as the desired particular solution.

EXERCISES 6.5

1. If the given initial-value problem were to be solved by the fourth-order Runge–Kutta method (and we are not asking you to do that), do you think accurate results could be obtained? Explain. The x domain is $0 \le x < \infty$.

(a) $y' = 2y - 8x + 4;\quad y(0) = 0$
(b) $y' = y - 2e^{-x};\quad y(0) = 1$
(c) $y' = y + 5e^{-4x};\quad y(0) = -1$
(d) $y' = 1 + 3(y - x);\quad y(0) = 0$

2. It is natural to wonder how well we would fare trying to solve (1) using computer software. Using the *Maple* dsolve command with the abserr option, see if you can obtain accurate results at the points $x = 1, 4, 8, 12$ listed in Table 1.

3. One can see if a computed solution exhibits instability, as did the solution obtained by the midpoint rule and plotted in Fig. 2, when we have the exact solution to compare it with. In practice, of course, we don't have the exact solution to compare with; if we did, then we would not be solving numerically in the first place. Thus, when a computed solution exhibits an oscillatory behavior how do we know that it is incorrect; perhaps the exact solution has exactly that oscillatory behavior? One way to find out is to rerun the solution with h halved. If the oscillatory behavior is part of the exact solution, then the new results will oscillate every two steps rather than every step. Using this idea, run the case shown in Fig. 2 twice, for $h = 0.05$ and $h = 0.025$, and comment on whether the results indicate a true instability or not.

4. We derived the solution (12) of the difference equation (7) in the text. Verify, by direct substitution, that (12) does satisfy (7) for any choice of the arbitrary constants C_1 and C_2.

5. In (13) we showed that $\left(Ah + \sqrt{1 + A^2h^2}\right)^n \sim e^{Ax_n}$ as $h \to 0$, yet it would appear that $\left(Ah + \sqrt{1 + A^2h^2}\right)^n \sim \left(\sqrt{1}\right)^n = 1$. Explain the apparent contradiction.

6. The purpose of this exercise is to explore the validity of the rule of thumb that we gave regarding the solution of the equation $y' = f(x, y)$ by the midpoint rule – namely, that the method should be stable if $\partial f/\partial y > 0$ and unstable if $\partial f/\partial y < 0$ over the region of interest. Specifically, in each case apply the rule of thumb and draw what conclusions you can about the stability of the midpoint rule solution of the given problem. Then, program and run the midpoint rule with $h = 0.05$, say, over the given x interval. Discuss the numerical results and whether the rule of thumb was correct. (Since the midpoint rule is not self-starting, use ten Euler steps from $x = 0$ to $x = 0.05$ to get the method started.)

(a) $y' = e^{x+y};\quad y(0) = 1, 0 \le x \le 4$
(b) $y' = e^{x-y};\quad y(0) = 1, 0 \le x \le 4$
(c) $y' = (4 - x)y;\quad y(0) = 1, 0 \le x \le 10$
(d) $y' = (x - 1)y;\quad y(0) = 1, 0 \le x \le 5$
(e) $y' = (x + y)/(1 + x);\quad y(0) = 2, 0 \le x \le 4$

7. We stated in the text that the results in Table 2 are consistent with a critical h value of 0.002 because the calculations change from unstable to stable as h decreases from 0.01 to 0.001. Program and carry out the Euler calculation of the solution to the initial-value problem (18) using $h = 0.0021$ and 0.0019, out to around $x = 1$, and see if these h values continue to bracket the change from unstable to stable. (You may try to bracket h_{cr} even more tightly if you wish.)

8. (*Stability of second-order Runge–Kutta methods*) In Sec-

tion 6.3 we derived the general second-order Runge–Kutta method, which includes these as special cases: the *improved Euler method*,

$$y_{n+1} = y_n + \frac{h}{2}\left\{f(x_n, y_n) + f\left[x_{n+1}, y_n + hf(x_n, y_n)\right]\right\}, \tag{8.1}$$

and the *modified Euler method*,

$$y_{n+1} = y_n + hf[x_n + \frac{h}{2}, y_n + \frac{h}{2}f(x_n, y_n)]. \tag{8.2}$$

(a) For the test equation $y' = Ay$, show that the improved Euler method is strongly stable for sufficiently small h (i.e., as $h \to 0$). For the case where $A < 0$, show that that stability is achieved only if $h < h_{cr} = 2/|A|$.
(b) For the test equation $y' = Ay$, show that the modified Euler method is strongly stable for sufficiently small h (i.e., as $h \to 0$). For the case where $A < 0$, show that that stability is achieved only if $h < h_{cr} = 2/|A|$.

9. (*Strong stability of the multi-step Adams–Bashforth method*) Recall, from Section 6.3, the fourth-order Adams–Bashforth method

$$y_{n+1} = y_n + \frac{h}{24}\left(55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3}\right), \tag{9.1}$$

where $f_n = f(x_n, y_n)$. This exercise is to show that the "AB" method is strongly stable for sufficiently small h (i.e., as $h \to 0$) even though it is a multi-step method.

(a) Consider the test equation $y' = Ay$, where the constant A can be positive or negative; that is, let $f(x, y) = Ay$ be a solution of the fourth-order difference equation (9.1) in the form $y_n = \rho^n$, show that ρ must satisfy the fourth-degree characteristic equation

$$\rho^4 - (1 + 55\alpha)\rho^3 + 59\alpha\rho^2 - 37\alpha\rho + 9\alpha = 0, \tag{9.2}$$

where $\alpha = Ah/24$.
(b) Notice that as h tends to zero so does α, and (9.2) reduces to $\rho^4 - \rho^3 = 0$, with the roots $= 0, 0, 0, 1$. Thus, if we denote the roots of (9.2) as $\rho_1, \ldots, \rho_4$, then we see that the first three of these tend to zero and the last to unity as $h \to 0$, and the general solution for y_n behaves as

$$y_n = C_1\rho_1^n + C_2\rho_2^n + C_3\rho_3^n + C_4\rho_4^n \sim C_4 1^n \tag{9.3}$$

as $h \to 0$. Since p^n tends to zero, unity, or infinity, depending upon whether p is smaller than, equal to, or greater than unity, we need to examine the 1^n term in (9.3) more closely. Specifically, seeking ρ_4 in the power series form $\rho_4 = 1 + a\alpha + \cdots$, put that form into (9.2). Equating coefficients of α on both sides of that equation through first-order terms, show that $a = 24$. Thus, in place of (9.3) we have the more informative statement

$$y_n \sim C_4(1 + 24\alpha)^n = C_4(1 + Ah)^n \sim C_4 e^{Ax_n} \tag{9.4}$$

as $n \to \infty$. Show why the final step in (9.4) is true. Since the right-hand side of (9.4) is identical to the exact solution $y(x)$ of the given differential equation, whether A is positive or negative, we conclude that the AB method is strongly stable for sufficiently small h.

10. (*Strong stability of the multi-step Adams–Moulton method*) Recall, from Section 6.3, the fourth-order Adams–Moulton method,

$$y_{n+1} = y_n + \frac{h}{24}\left(9f_{n+1} + 19f_n - 5f_{n-1} + f_{n-2}\right). \tag{10.1}$$

Proceeding along the same lines as outlined in Exercise 9, show that the AM method is stongly stable for sufficiently small h.

11. Derive the general solution of the given difference equation. If initial conditions are specified, then find the corresponding particular solution. In each case $n = 0, 1, 2, \ldots$.

(a) $y_{n+1} - 4y_n = 0$; $y_0 = 5$
(b) $y_{n+2} - y_n = 0$; $y_0 = 1$, $y_1 = 3$
(c) $y_{n+2} + y_{n+1} - 6y_n = 0$; $y_0 = 9$, $y_1 = -2$
(d) $y_{n+2} - 4y_{n+1} + 3y_n = 0$; $y_0 = 3$, $y_1 = 1$
(e) $y_{n+2} + 3y_{n+1} + 2y_n = 0$; $y_0 = 5$, $y_1 = 7$
(f) $y_{n+3} - y_{n+2} - 4y_{n+1} + 4y_n = 0$; $y_0 = 3$, $y_1 = 5$, $y_2 = 9$
(g) $y_{n+4} - 5y_{n+2} + 4y_n = 0$
(h) $y_{n+4} - 6y_{n+2} + 8y_n = 0$

12. (a)–(h) Use computer software to obtain the general solution of the corresponding problem in Exercise 11, and the particular solution as well, if initial conditions are given.

13. (*Repeated roots*) Recall from the theory of linear homogeneous differential equations with constant coefficients that if, when we seek $y(x) = e^{\lambda x}$, λ is a root of multiplicity k of the characteristic equation, then it gives rise to the solutions $y(x) = \left(C_1 + C_2 x + \cdots + C_k x^{k-1}\right)e^{\lambda x}$, where $C_1, \ldots, C_k$ are arbitrary constants. An analogous result holds for difference equations. Specifically, verify that the characteristic equation of $y_{n+2} - 2by_{n+1} + b^2 y_n = 0$ has the root b with multiplicity 2, and that $y_n = (C_1 + C_2 n)b^n$.

14. Show that if $y_n^{(1)}$ and $y_n^{(2)}$ are solutions of the second-order linear homogeneous difference equation $a_0(n)y_{n+2} + a_1(n)y_{n+1} + a_2(n)y_n = 0$, and Y_n is any particular solution of the nonhomogeneous equation $a_0(n)y_{n+2} + a_1(n)y_{n+1} + a_2(n)y_n = f_n$, then $y_n = C_1 y_n^{(1)} + C_2 y_n^{(2)} + Y_n$ is a solution of the nonhomogeneous equation.

15. (*Nonhomogeneous difference equations*) For the given equation, first find the general solution of the homogeneous equation. Then adapt the method of undetermined coefficients from the theory of differential equations and find a particular solution. Finally, give the general solution of the given nonhomogeneous difference equation. (First, read Exercise 14.)

(a) $y_{n+1} - 3y_n = n$
(b) $y_{n+1} - 2y_n = 3\sin n$
(c) $y_{n+1} - y_n = 2 + \cos n$
(d) $y_{n+2} - 5y_{n+1} + 6y_n = 2n^2 - 6n - 1$
(e) $y_{n+2} + y_{n+1} - 2y_n = n^2$
(f) $y_{n+2} - 4y_n = 6n^2 - 1$
(g) $y_{n+2} - y_n = e^n$
(h) $y_{n+4} - 7y_{n+2} + 12y_n = n + 6$

16. (a)–(h) Use computer software to obtain the general solution of the corresponding problem in Exercise 15.

Chapter 6 Review

To solve a differential equation $y' = f(x, y)$ numerically, one begins by discretizing the problem and working with the discrete variables x_n, y_n in place of the continuous variables $x, y(x)$. The solution is accomplished using a numerical algorithm that gives y_{n+1} in terms of y_n in the case of a single-step method, or in terms of y_n and one or more of the preceding values $y_{n-1}, y_{n-2}, \ldots$ in the case of a multi-step method. If the algorithm gives y_{n+1} explicitly, then it is said to be of **open** type; if not it is of **closed** type. All methods considered in this chapter are of open type, except for the Adams–Moulton method (Section 6.3.5).

Decomposing the error as

$$\text{Total error} = y(x_n) - y_n^* = [y(x_n) - y_n] + [y_n - y_n^*],$$

where $y(x_n)$ is the exact solution of the differential equation, y_n is the exact solution of the numerical algorithm (i.e., carried out on a perfect machine, having no roundoff error), and y_n^* is the actual solution of the algorithm (carried out on a real machine), we call $y(x_n) - y_n$ the **accumulated truncation error** and $y_n - y_n^*$ the **accumulated roundoff error**. If the former is of order $O(h^p)$ at a fixed x point as $h \to 0$, then the method is said to be of **order** p. If $p > 0$, then the accumulated truncation error tends to zero and the method is described as **convergent**. The greater p is, the more accurate the method is for a given step size h. Besides requiring of a method that it be convergent, we also require that it be **stable**; that is, we require that the accumulated roundoff error (which is inevitable in a real machine) remain small compared to the exact solution of the differential equation.

We begin with the **Euler method** $y_{n+1} = y_n + f(x_n, y_n)h$, which is simple but not very accurate because it is only a first-order method. Desiring greater accuracy, we introduce the second-order **midpoint rule** and second- and fourth-order **Runge–Kutta methods**, the latter providing us with an accurate general-purpose differential equation solver.

One's interest in higher-order methods is not just a matter of accuracy because, in principle, one could rely exclusively on the simple and easily programmed Euler method, and make h small enough to achieve any desired accuracy. There are two problems with that idea. First, as h is decreased the number of steps increases and one can expect the numerical roundoff error to grow, so that it may not be possible to achieve the desired accuracy. Second, there is the question of economy. For instance, while the fourth-order Runge–Kutta method (for example) is about four times as slow as the Euler method (because it requires four function evaluations per step compared to one for the Euler method), the gain in accuracy that it affords is so great that we can use a step size much more than four times that needed by the Euler method for the same accuracy, thereby resulting in greater economy.

Naturally, higher-order methods are more complex and hence more tedious to program. Thus, we strongly urge (in Section 6.3.4) the empirical estimation of the order, if only as a check on the programming and implementation of the method.

In Section 6.4 we showed that the methods developed for the single equation $y' = f(x, y)$ can be used to solve systems of equations and higher-order equations as well. There we also study boundary-value problems, and find them to be significantly more difficult than initial-value problems. However, we show how to use the principle of superposition to convert a boundary-value problem to one or more problems of initial-value type, provided that the problem is linear.

Finally, in Section 6.5 we look at "what can go wrong," mostly insofar as numerical instability due to the growth of roundoff error, and an analytical approach is put forward for predicting whether a given method is stable. Actually, stability depends not only on the solution algorithm but also on the differential equation, and our analyses are for the simple test equation $y' = Ay$ rather than for the general case $y' = f(x, y)$. We find that whereas the differential equation $y' = Ay$ is of first order, the difference equation expressed by the algorithm is of higher order if the method is of multi-step type. Thus, it has among its solutions the exact solution (as $h \to 0$) and one or more extraneous solutions as well. It is those extraneous solutions that can cause instability. For instance, the midpoint rule is found to be stable if $A > 0$ and unstable if $A < 0$; we classify it as weakly stable because its stability depends upon the sign of A. However, the fourth-order Adams–Bashforth and Adams–Moulton methods are stable, even though they are multistep methods because the extraneous solutions do not grow. Single-step methods such as Euler and those of Runge–Kutta type do not give rise to extraneous solutions and are stable.

Finally, we stress that even if a method is stable as $h \to 0$, h needs to be reduced below some critical value for that stability to be manifested.

Chapter 7

Qualitative Methods: Phase Plane and Nonlinear Differential Equations

7.1 Introduction

This is the final chapter on ordinary differential equations, although we do return to the subject in Chapter 11, where we reconsider systems of linear differential equations using matrix methods.

Interest in nonlinear differential equations is virtually as old as the subject of differential equations itself, which dates back to Newton, but little progress was made until the late 1880's when the great mathematician and astronomer *Henri Poincaré* (1854–1912) took up a systematic study of the subject in connection with celestial mechanics. Realizing that nonlinear equations are rarely solvable analytically, and not yet having the benefit of computers to generate solutions numerically, he sidestepped the search for solutions altogether and instead sought to answer fundamental questions about the qualitative and topological nature of solutions of nonlinear differential equations without actually finding them.

The entire chapter reflects either his methods, such as the use of the so-called "phase plane" and focusing attention upon the "singular points" of the equation, or the spirit of his approach. In addition, however, we can now rely heavily upon computer simulation. Thus, our approach in this chapter is a blend of a qualitative, topological, and geometric approach, with quantitative results obtained readily with computer software.

Though Poincaré's work was motivated primarily by problems of celestial mechanics, the subject began to attract broader attention during and following World War II, especially in connection with nonlinear control theory. In the postwar years, interest was stimulated further by the publication in English of N. Minorsky's *Nonlinear Mechanics* (Ann Arbor, MI: J. W. Edwards) in 1947. With that and other books, such as A. Andronov and C. Chaikin's *Theory of Oscillations* (Princeton:

Princeton University Press, 1949) and J. J. Stoker's *Nonlinear Vibrations* (New York: Interscience, 1950) available as texts, the subject appeared in university curricula by the end of the 1950's. With that base, and the availability of digital computers by then, the subject of nonlinear dynamics, generally known now as *dynamical systems*, has blossomed into one of the most active research areas, with applications well beyond celestial mechanics and engineering – to biological systems, the social sciences, economics, and chemistry. The shift from the orderly determinism of Newton to the often nondeterministic chaotic world of Mitchell Feigenbaum, E. N. Lorenz, Benoit Mandelbrot, and Stephen Smale has been profound. For a wonderful historical discussion of these changes we suggest the little book *Mathematics and the Unexpected* by Ivar Ekeland (Chicago: University of Chicago Press, 1988).

7.2 The Phase Plane

To introduce the phase plane, consider the system

$$mx'' + kx = 0 \tag{1}$$

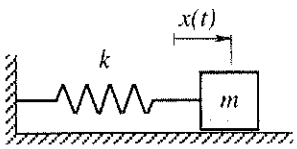


Figure 1. Simple harmonic mechanical oscillator.

governing the free oscillation of the simple harmonic mechanical oscillator shown in Fig. 1. Of course we can readily solve (1) and obtain the general solution $x(t) = C_1 \cos \omega t + C_2 \sin \omega t$, where $\omega = \sqrt{k/m}$ is the natural frequency – or, equivalently,

$$x(t) = A \sin(\omega t + \phi), \tag{2}$$

where A and ϕ are the amplitude and phase angle, respectively. To present this result graphically, one can plot x versus t and obtain any number of sine waves of different amplitude and phase, but let us proceed differently.

We begin by re-expressing (1), equivalently, as the system of first-order equations

$$\frac{dx}{dt} = y, \tag{3a}$$

$$\frac{dy}{dt} = -\frac{k}{m}x, \tag{3b}$$

as is discussed in Section 3.9. The auxiliary variable y, defined by (3a), happens to have an important physical significance, it is the velocity, but having such significance; is not necessary. Next, we deviate from the ideas presented in Section 3.9 and divide (3b) by (3a), obtaining

$$\frac{dy}{dx} = -\frac{k}{m}\frac{x}{y} \qquad \text{or} \qquad my\,dy + kx\,dx = 0, \tag{4}$$

integration of which gives

$$\frac{1}{2}my^2 + \frac{1}{2}kx^2 = C. \tag{5}$$

Since $y = dx/dt$, (5) is a first-order differential equation. We could solve for y (i.e., dx/dt), separate variables, integrate again, and eventually arrive at (2) once again. Instead, let us take (5) as our end result and plot the one-parameter family of ellipses that it defines (Fig. 2), the parameter being the integration constant C. In this example C happens to the total energy (kinetic energy of the mass plus potential energy of the spring); $C = 0$ gives the "point ellipse" $x = y = 0$ and the greater the value of C, the larger the ellipse.

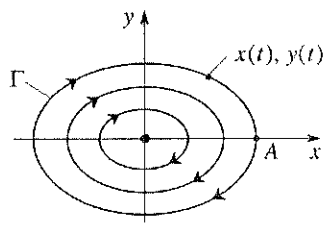


Figure 2. Phase portrait of (1).

It is customary to speak of the x, y plane as the **phase plane**. Each integral curve represents a possible motion of the mass, and each point on a given curve represents an instantaneous state of the mass (the horizontal coordinate being the displacement and the vertical coordinate being the velocity). Observe that the time t enters only as a parameter, through the parametric representation $x = x(t)$, $y = y(t)$. So we can visualize the representative point $x(t), y(t)$ as moving along a given curve as suggested by the arrows in Fig. 2. The direction of the arrows is implied by the fact that $y = dx/dt$, so that $y > 0$ implies that $x(t)$ is increasing and $y < 0$ implies that $x(t)$ is decreasing. One generally calls the integral curves **phase trajectories**, or simply **trajectories**, to suggest the idea of movement of the representative point. A display of a number of such trajectories in the phase plane is called a **phase portrait** of the original differential equation, in this case (1). Of course, there is a trajectory through each point of the phase plane, so if we showed *all* possible trajectories we would simply have a black picture; the idea is to plot enough trajectories to establish the key features of the phase portrait.

What are the advantages of presenting results in the form of a phase portrait, rather than as traditional plots of $x(t)$ versus t? One advantage of the phase portrait is that it requires only a "first integral" of the original second-order equation such as equation (5) in this example, and sometimes we can obtain the first integral even when the original differential equation is nonlinear. For instance, let us complicate (1) by supposing that the spring force is not given by the linear function $F_s = kx$, but by the nonlinear function $F_s = ax + bx^3$, and suppose that $a > 0$ and $b > 0$ so that the spring is a "hard" spring: it grows stiffer as x increases (Fig. 3), as does a typical rubber band. If we take $a = b = m$, say, for definiteness and for simplicity, then in place of (1) we have the nonlinear equation

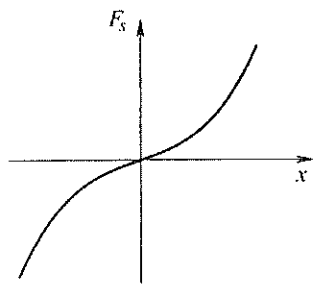


Figure 3. Hard spring.

$$x'' + x + x^3 = 0. \tag{6}$$

Proceeding as before, we re-express (6) as the system

$$x' = y, \tag{7a}$$
$$y' = -x - x^3. \tag{7b}$$

Division gives

$$\frac{dy}{dx} = -\frac{x + x^3}{y} \qquad \text{or} \qquad y\,dy + (x + x^3)dx = 0, \tag{8}$$

which yields the first integral

$$\frac{1}{2}y^2 + \frac{1}{2}x^2 + \frac{1}{4}x^4 = C. \tag{9}$$

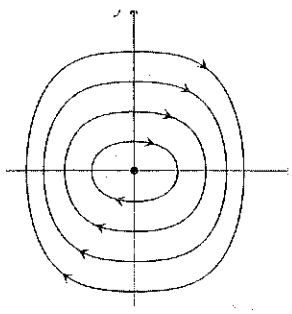


Figure 4. Phase portrait of (6) (hard spring).

In principle, if we plot these curves for various values of C we can obtain the phase portrait shown in Fig. 4. More conveniently, we generated the figure by using the *Maple* phaseportrait command discussed at the end of this section. A comparable phase portrait plotting capability is provided in numerous other computer software systems.

To repeat, one advantage of the phase portrait presentation is that it requires only a first integral. In the present case (6) was nonlinear due to the x^3 term, yet its first integral (9) was readily obtained.

A second attractive feature of the phase portrait is its compactness. For instance, observe that the single phase trajectory Γ in Fig. 2 corresponds to an entire family of oscillations of amplitude A, several of which are shown in Fig. 5 since any point on Γ can be designated as the initial point ($t = 0$): if the initial point on Γ is $(A, 0)$, then we get the curve #1 in Fig. 5; if the initial point on Γ is a bit counterclockwise of $(A, 0)$ then we get the curve #2; and so on. Passing from the x, t plane to the x, y plane, the infinite family of curves shown in Fig. 5 collapse onto the single trajectory Γ in Fig. 2. Put differently, whereas the solution (2) of equation (1) is a two-parameter family of curves in x, t space (the parameters being A and ϕ), (5) is only a one-parameter family of curves in the x, y plane (the parameter being C). That compactness can be traced to the division of (3b) by (3a) or (7b) by (7a) for that step essentially eliminates the time t.

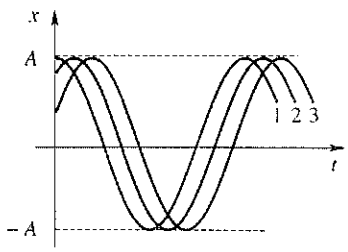


Figure 5. Solutions $x(t)$ corresponding to the trajectory Γ.

To learn about nonlinear systems, it is useful to contrast the phase portraits of the linear oscillator governed by (1) and the nonlinear oscillator governed by (6), and given in Figs. 2 and 4, respectively. The phase portrait in Fig. 2 is extremely simple in the sense that all the trajectories are geometrically similar, differing only in scale. That is, if a trajectory is given by $x = X(t)$, $y = Y(t)$, then $x = \kappa X(t)$, $y = \kappa Y(t)$ is also a trajectory for every possible scale factor κ, be it positive, negative, or zero. That result holds not only for the system (3) but for any constant-coefficient linear homogeneous system

$$\begin{aligned} x' &= ax + by, \\ y' &= cx + dy. \end{aligned} \tag{10}$$

In contrast, consider the phase portrait of the nonlinear equation $x'' + \alpha x + \beta x^3 = 0$ shown in Fig. 4. In that case the trajectories are not mere scalings of each other; there is distortion of shape from one to another, and that distortion is due entirely to the nonlinearity of the differential equation. The innermost trajectories approach ellipses [because as smaller and smaller motions are considered the x^4 becomes more and more negligible compared to the other terms in (9)], and the outer ones become more and more distorted as the effect of the x^4 term grows in (9).

Thus, whereas the phase portrait of the linear equation (1) amounts to a single kind of trajectory, repeated endlessly through scalings, that of the nonlinear equation (6) is made up of an infinity of different kinds of trajectories. That richness is a hallmark of nonlinear equations, as we shall see in the next example and in the sections that follow.

Before turning to the next example, let us complement the phase portrait in Fig. 4 with representative plots of $x(t)$ versus t. We choose the two sets of initial

conditions: $x(0) = 0.5$, $x'(0) = 0$ and $x(0) = 1$, $x'(0) = 0$. The results are shown in Fig. 6, together with the corresponding solutions of the linear equation $x'' + x = 0$ (shown as dotted) for reference. Besides the expected distortion we also observe that the frequency of the oscillation is amplitude dependent for the nonlinear case: the frequency increases as the amplitude increases. In contrast, for the linear equation (1) the frequency $\omega = \sqrt{k/m} = 1$ is a constant, independent of the amplitude.

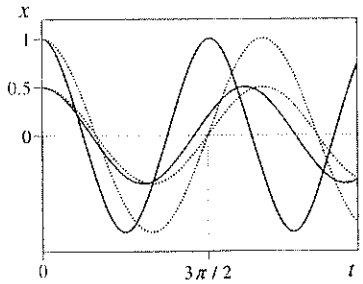


Figure 6. Effects of nonlinearity on x(t).

Above, we mentioned the richness of the sets of solutions to nonlinear differential equations. A much more striking example of that richness is obtained if we reconsider the nonlinear oscillator, this time with a "soft" spring – that is, with $F_s = ax - bx^3$ ($a > 0$ and $b > 0$) as sketched in Fig. 7. Again setting $a = b = m$ we have, in place of (6),

$$x'' + x - x^3 = 0. \tag{11}$$

In place of (9) we have

$$\frac{1}{2}y^2 + \frac{1}{2}x^2 - \frac{1}{4}x^4 = C, \tag{12}$$

and in place of the phase portrait shown in Fig. 4 we obtain the strikingly different one shown in Fig. 8. We continue to study this example in Section 7.3, but even now we can make numerous interesting observations. First, whereas all of the motions revealed in the phase plane for the hard spring (Fig. 4) are qualitatively similar oscillatory motions – give or take some distortion from one to another we see in Fig. 8 a number of qualitatively different types of motion, and these are

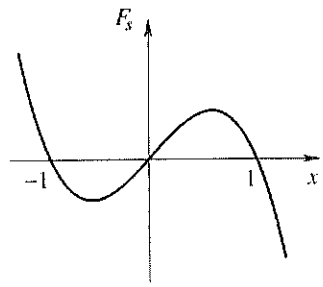


Figure 7. Soft spring.

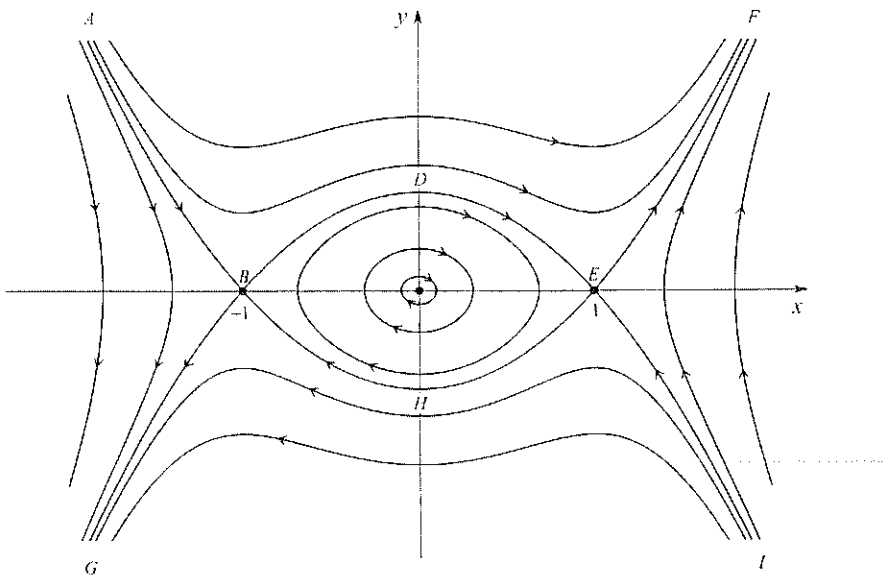


Figure 8. Phase portrait of (11) (soft spring).

separated by the trajectories $ABHEF$ and $GBDEI$ which, together, are called a **separatrix**.

Before examining these different types of motion let us distinguish between the physical velocity $dx/dt = x' = y$ of the mass and the **phase velocity**

$$\frac{ds}{dt} = s' = \sqrt{x'^2 + y'^2} \tag{13}$$

of the representative point $P = (x(t), y(t))$ in the phase plane as it moves along a trajectory, where s is arc length along that trajectory. Observe from (13) that $s' = 0$, and the representative point P is stationary if and only if both $x' = 0$ and $y' = 0$ at that point. If $x' = 0$ and $y' = 0$ at a point $P_0 = (x_0, y_0)$, then we call P_0 an **equilibrium point**, or **fixed point**, because then $x(t) = x_0$ and $y(t) = y_0$ remain constant; if we start at that point we remain at that point for all $t > 0$.

Let us identify the equilibrium points (if any) for the present soft-spring example. Solving

$$x' = y = 0, \tag{14a}$$
$$y' = -x + x^3 = 0 \tag{14b}$$

gives $y = 0$ and $x = 0, \pm 1$, so there are three equilibrium points:

$$(-1, 0),\ (0, 0),\ (1, 0), \tag{15}$$

which correspond to B, the origin, and E in Fig. 8. Physically, $x = 0$ and ± 1 are the three x values for which the spring force F_s is zero (Fig. 7).

Consider the motions inside the $BDEHB$ "football." First, each closed loop trajectory Γ inside $BDEHB$ corresponds to a periodic solution $x(t)$ (likewise for each closed loop trajectory in Figs. 2 and 4), and is therefore called a **closed orbit**. [Recall from Section 5.7 that $x = x(t)$ is periodic with period T if $x(t+T) = x(t)$ for all $t \geq 0$.] Actually, how can we be sure that each Γ represents a periodic motion? Well, the representative point does trace out the given trajectory Γ over and over; it cannot stop because there are no equilibrium points on Γ. Further, the phase speed s' at any point (x, y) on Γ is the same, from one traversal to the next, because

$$s' = \sqrt{x'^2 + y'^2} = \sqrt{y^2 + (x - x^3)^2} \tag{16}$$

is a function only of x and y, not t. Thus, as P traverses Γ, over and over, each time in the same manner, the projection $x(t)$ of P onto the x axis generates a periodic function $x(t)$.

Next, consider part of the separatrix itself. To obtain the equation of the separatrix we use (12) and require the curve to go through the point E (or, equivalently, B) – namely, $x = 1$ and $y = 0$. That step gives $C = 1/4$, so $ABHEF$ and $GBDEI$ are given by

$$y = \pm\sqrt{\frac{1}{2} - x^2 + \frac{1}{2}x^4}, \tag{17}$$

respectively. Beginning at D, say, the representative point P moves rightward on DE and approaches the equilibrium point E. Does it reach E in finite time and then remain there, or does it approach E asymptotically as $t \to \infty$? To answer that question we use (16). Let the tangent line to the curve DEI, at E, be $y = m(x-1)$. [We could determine the slope m by differentiating (17), but the value of m will not be important.] Then, since $ds = \sqrt{1+(dy/dx)^2}\,dx \sim \sqrt{1+m^2}\,dx$ as $P \to E$, we can replace s' in (16) by $\sqrt{1+m^2}\,dx/dt$, and y by $m(x-1)$, so (16) becomes

$$\begin{aligned}\sqrt{1+m^2}\,\frac{dx}{dt} &\sim \sqrt{m^2(x-1)^2 + x^2(x+1)^2(x-1)^2} \\ &\sim -\sqrt{m^2+4}\,(x-1),\end{aligned} \tag{18}$$

where the negative square root has been chosen since $dx/dt > 0$ as $P \to E$ on DE, whereas $x - 1 < 0$ on DE, and where the last step in (18) is left for the exercises. The upshot is that

$$\frac{dx}{dt} \sim \gamma(1-x) \tag{19}$$

as $P \to E$, for some finite positive constant γ. Thus,

$$\frac{dx}{1-x} \sim \gamma\,dt \tag{20}$$

so

$$-\ln(1-x) \sim \gamma t + \text{constant}, \tag{21}$$

and we can now see that $t \to \infty$ as $P \to E$ (i.e., as $x \to 1$). Thus, P does not reach E in finite time but only asymptotically as $t \to \infty$. Similarly, if we begin at point H and go backward in time, then we reach E only as $t \to -\infty$.

Let us return now to the region inside of the football and consider any closed orbit Γ. As the size of Γ shrinks to zero, Γ tends to the elliptical (actually circular because of our choice $a = m$) shape $y^2 + x^2 =$ constant, and the period of the motion tends to 2π [since the solution of the linearized problem is $x(t) = A\sin(t+\phi)$]. At the other extreme, as Γ gets larger it approaches the pointed shape $BDEHB$. Bearing in mind that it takes infinite time to reach E along an approach from D, it seems evident that the period of the Γ motion must tend to infinity as Γ approaches $BDEHB$; we will ask you to explore this point in the exercises. From a physical point of view, the idea is that not only is the "flow" zero *at* E (where $x' = y' = 0$), it is very slow in the neighborhood of E. If Γ is any closed trajectory that is just barely inside of $BDEHB$, then part of Γ falls within that stagnant neighborhood of E (similarly at B). The representative point P moves very slowly there, hence the period is very large. We reiterate that although each closed loop inside the football corresponds to a periodic motion, the closed loop $BDEHB$ does not. In fact, although BDE and EHB meet at B and E they are distinct trajectories. On BDE, t varies from $-\infty$ at B to $+\infty$ at E; likewise, on EHB t varies from $-\infty$ at E to $+\infty$ at B. Thus, if we begin on BDE we can never get to EHB, and vice versa.

Finally, it should be evident that every trajectory that is *not* within the football corresponds to a nonperiodic motion. Thus, BDE and EHB form a transition from nonperiodic to periodic motions.

Thus far we have studied the three differential equations (1), (6), and (11). In each case we have changed the single second-order differential equation to a system of two first-order equations by setting $x' = y$ and then studied them in the x, y phase plane. More generally, we consider in this chapter systems of the form

$$x' = P(x, y), \tag{22a}$$
$$y' = Q(x, y). \tag{22b}$$

That is, $P(x, y)$ need not equal y, and the system need not be a restatement of a single second-order equation. Rather, it might arise directly in the form (22).

For instance, suppose that two species of fish coexist in a lake, say bluegills and bass. The bluegills, with population x, feed on vegetation which is available in unlimited quantity, and the bass, with population y, feed exclusively on the bluegills. If the two species were separated, their populations could be assumed to be governed approximately by the rate equations

$$x' = \alpha x, \qquad y' = -\beta y, \tag{23}$$

where the populations $x(t)$ and $y(t)$ are considered to be large enough so that they can be approximated as continuous rather than discrete (integer valued) variables, and α, β are (presumably positive) constants that reflect net birth/death rates. The species are not separated, however, so we expect the effective α to decrease as y increases and the effective β to decrease as x increases. An approximate revised model might then be expressed as

$$x' = (\alpha - \gamma y)x, \tag{24a}$$
$$y' = -(\beta - \delta x)y, \tag{24b}$$

which system is indeed of the form (22). This ecological problem is well known as *Volterra's problem*, and we shall return to it later.

The system (22) is said to be **autonomous** because there is no explicit dependence on the independent variable (the time t here but which could have some other physical or nonphysical significance). Surely not all systems are autonomous, but that class covers a great many cases of important interest, and that is the class that is considered in phase plane analysis and in this chapter. Because (22a,b) are autonomous, any explicit reference to t (namely, the t derivatives) can be suppressed by dividing one equation by the other and obtaining

$$\frac{dy}{dx} = \frac{P(x, y)}{Q(x, y)}, \tag{25}$$

where we now change our point of view and regard y as a function of x in the x, y phase plane, rather than x and y as functions of t. If the system (22) were not

autonomous; that is, if it were of the form $x' = P(x, y, t)$ and $y' = Q(x, y, t)$, one could still make it autonomous by re-expressing it, equivalently, as

$$\begin{aligned} x' &= P(x, y, z), \\ y' &= Q(x, y, z), \\ z' &= 1, \end{aligned}$$

but then the trajectories exist in the three-dimensional x, y, z space, and that case is more complicated. In this chapter we continue to consider the autonomous case (22) and the two-dimensional x, y phase plane.

Closure. As explained immediately above, our program in this section is to show the advantages of recasting an autonomous system (22) (which could, but need not, arise from a single second- order equation by letting x' be an auxiliary dependent variable y) in the form (25) and then study the solutions of that equation in the two-dimensional x, y phase plane. One advantage is that (25) can sometimes be solved analytically even if (22a,b) are nonlinear. Indeed, our primary interest in Chapter 7 is in the nonlinear case. We find that the phase portrait provides a remarkable overview of the system dynamics, and the hard- and soft-spring oscillator examples begin to reveal some of the phenomenological richness of nonlinear systems. We do not suggest the use of the phase plane as a substitute for obtaining and plotting solutions of (22) in the more usual way, x versus t and y versus t. Rather, we suggest that to understand a complex nonlinear system one needs to combine several approaches, and for autonomous systems the phase plane is one of the most valuable. Finally, we ask you to observe how the phase plane discussion is more qualitative and topological than lines of approach developed in the preceding chapters. For instance, regarding Fig. 8 we distinguish the qualitatively different types of motion such as the periodic orbits within $BDEHB$, the transitional motions on the separatrix itself, and the nonperiodic motions as well.

We distinguish between the physical velocity $x'(t)$ of the mass, in the preceding examples, and the phase velocity $s'(t)$, which is the velocity of the representative point $x(t), y(t)$ in the phase plane. It is useful, conceptually, to think of the $x'(t), y'(t)$ velocity field as the velocity field of a "flow" such as a fluid flow in the phase plane.

Finally, we mention that in the hard- and soft-spring oscillators, (6) and (8), we meet special cases of the extremely important **Duffing equation**, to which we return in a later section.

Computer software. Here is how we generate the phase portrait shown in Fig. 8 using the *Maple* **phaseportrait** command. First, enter

```
with(DEtools):
```

and return, to gain access to the phaseportrait command. Note the colon, whereas *Maple* commands are followed by semicolons. Next, enter

```
phaseportrait([y, -x + x^3], [t, x, y], t = -20..20, {[0, 0, 0.1], [0, 0, 0.3],
```

$[0,0,0.6],[0,0,0.70710781],[0,0,-0.70710781],[0,0,0.9],[0,0,-0.9],$
$[0,0,1.25],[0,0,-1.25],[0,1.5,0.8838834761],[0,1.5,-0.8838834761],$
$[0,1.4,0],[0,1.8,0],[0,-1.5,0.8838834761],[0,-1.5,-0.8838834761],$
$[0,-1.4,0],[0,-1.8,0]\}$, stepsize $= 0.05$, $y = -1.8..1.8$, $x = -2..2$,
scene $= [x,y]$);

and return. In $[y, -x + x\hat{}3]$ the items are the right-hand sides of the first and second differential equations, respectively; $[t,x,y]$ are the independent variable and dependent variables; $t = -20..20$ is the range of integration of the differential equations; within { } are the initial t,x,y points chosen in order to generate the trajectories shown in Fig. 8. After those points the remaining items are optional: stepsize= 0.05 sets the stepsize h in the Runge–Kutta–Fehlberg integration because the default value would be (final t − initial t)/20 = (20 + 20)/20 = 2, which would give too coarse a plot (as found by experience); $y = -1.8..1.8, x = -2..2$ gives a limit to the x,y region, with the t integrations terminated once a trajectory leaves that region; scene $= [x,y]$ specifies the plot to be a two-dimensional plot in the x,y plane, the default being to give a three-dimensional plot in t,x,y space. There are additional options that we have not used, one especially useful option for phase plane work being the arrow option, which gives a lineal element grid. The elements can be barbed or not. To include thin barbed arrows, type a comma and then arrows=THIN after the last option. Thus, we would have ... scene $= [x,y]$, arrows =THIN);. In place of THIN type SLIM, or THICK for thicker arrows. For unbarbed lineal elements, use LINE in place of these. The order of the options is immaterial.

Observe that the separatrix must be generated separately as BDE, EHB, AB, GB, EF, and IE. To generate BDE, for instance, we determine the coordinates of D. The equation of the entire separatrix is given by (12), where C is determined by using the x,y pair $1,0$ (namely, the point E, which is a known point on the separatrix). Thus, putting $x = 1$ and $y = 0$ into the left side of (12) gives $C = 1/4$. Next, put $x = 0$ and solve for y, obtaining $y = 1/\sqrt{2} = 0.770710781$ at D. Then, with D as the initial point we need t to go from $-\infty$ to $+\infty$ to generate BDE. By trial, we find that -20 to $+20$ suffices for this segment and all others; similarly, we generate EF by using a point on EF at $x = 1.5$, and determine y at that point from the separatrix equation. That calculation gives $y = 0.8838834761$.

Notice that to generate Fig. 8 with phaseportrait we need to already know something about the phaseportrait – the equation of the separatrix, (12), so that we can choose suitable initial points on AB, GB, BDE, EHB, EF, and IE.

Suppose that we desire only the lineal element field, over $0 < x < 4$ and $0 < y < 4$, say. We can get it from phaseportrait as follows:

phaseportrait($[y, -x + x\hat{}3]$, $[t,x,y]$, $t = 0..1$, $\{[0,0,0]\}$, $x = 0..4$, $y = 0..4$,
scene $= [x,y]$, arrows = THIN, grid $= [20,20]$);

because the trajectory through $[0,0,0]$ gives simply the single point $x = y = 0$ in the x,y phase plane. We have included the one initial point $[0,0,0]$ because the

program calls for at least one. If the default grid is too coarse or too fine we can define it through the grid option, where grid = [20, 20] calls for a 20 by 20 grid.

Besides generating the x, y phase plane, phaseportrait can also generate plots of x or y versus t using the scene option. For instance,

phaseportrait([$y, -x + x$^3], [t, x, y], $t = 0..5$, {[0, 0.2, 0], [0, 0.8, 1.3]},
stepsize = 0.05, scene = [t, x]);

gives two plots of $x(t)$ versus t over $0 \le t \le 5$, one with initial conditions $x(0) = 0.2$ and $y(0) = x'(0) = 0$, and the other with initial conditions $x(0) = 0.8$ and $y(0) = x'(0) = 1.3$.

EXERCISES 7.2

1. We stated, below (5) that if we solve (5) for y (i.e., dx/dt), separate variables, and integrate, we obtain the general solution $x(t) = A\sin(\omega t + \phi)$ of (1). Here we ask you to do that, to carry out those steps.

2. Supply the steps missing between the first and second lines of (18).

3. We found in Fig. 6 that for the hard-spring oscillator the frequency increases with the amplitude. Explain, in simple terms, why that result makes sense.

4. Determine the equation of the phase trajectories for the given system, and sketch several representative trajectories. Use arrows to indicate the direction of movement along those trajectories.

(a) $x' = y, \quad y' = -x$ (b) $x' = xy, \quad y' = -x^2$
(c) $x' = y^2, \quad y' = -xy$

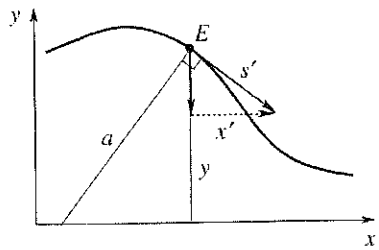
5. Determine the equation of the phase trajectories and sketch enough representative trajectories to show the essential features of the phase portrait. Use arrows to indicate the direction of movement along those trajectories.

(a) $x' = y, \quad y' = -y$ (b) $x' = y, \quad y' = y$
(c) $x' = y, \quad y' = x$ (d) $x' = y, \quad y' = 9x$
(e) $x' = x, \quad y' = x$ (f) $x' = x, \quad y' = -4x$

6. (*Period, for soft-spring oscillator*) In the paragraph below (21), we suggest that the period T of the periodic motions inside of $BDEHB$ (Fig. 8) tends to 2π in the limit as the amplitude A of the motion tends to zero, and to ∞ as $A \to 1$. Here we ask you to explore that claim with calculations. Specifically, use phaseportrait (or other software) to solve $x'' + x - x^3 = 0$ subject to the initial conditions $x(0) = A$, $x'(0) = 0$, for $A = 0.05, 0.3, 0.6, 0.9, 0.95, 0.99$. From your results, obtain the period T for each case and plot T versus A for those values of A (and additional ones if you wish). Does the claim made in the first sentence appear to be correct?

7. We stated, in our discussion of Fig. 8, that all trajectories outside of the "football" region correspond to nonperiodic motions. Explain why that is true.

8. (*Graphical determination of phase velocity*) (a) For the system (22), consider the special case where $P(x, y) = y$, as occurred in (3) and (7), for instance. From the accompanying sketch, show that in that case the phase velocity s' can be



interpreted graphically as

$$s' = a, \tag{8.1}$$

where a is the perpendicular distance from E to the x axis.
(b) Consider a rectangular phase trajectory $ABCDA$, where the corner points have the x, y coordinates $A = (-1, 1)$, $B = (3, 1)$, $C = (3, -1)$, $D = (-1, -1)$. Using (8.1), plot the graph of $x(t)$ versus t, from $t = 0$ through $t = 20$, if the representative point E is at A at $t = 0$.

(c) Consider a phase trajectory ABC consisting of straight-line segments from $A = (-1, 0)$ to $B = (0, 1)$ to $C = (1, 0)$ with E at B at $t = 0$. Using (8.1), sketch the graph of $x(t)$ versus t over $-\infty < t < \infty$. Also, give $x(t)$ analytically over $-\infty < t < 0$ and $0 < t < \infty$.
(d) Consider a straight-line phase trajectory from $A = (0, 5)$ to $B = (10, -5)$. Using (8.1), sketch the graph of $x(t)$ versus t over $0 < t < \infty$, if E is at A at $t = 0$.
(e) Same as (d), but with E at B at $t = 0$.

9. (a) Reduce the equation $x'' + 2x^3 = 0$ to a system of equations by setting $x' = y$. Find the equation of the phase trajectories and sketch several of them by hand. Show that for larger and larger motions the trajectories are flatter and flatter over $-1 < x < 1$.
(b) Use the *Maple* phaseportrait command (or other software) to generate the phase portrait and, on the same plot, the lineal element field, using barbed arrows to show the flow direction. You will need to make decisions, with some experimentation, as to the t-interval, the step size, the initial points, and so on, so as to obtain good results.

10. Reduce the equation $x'' + x^2 = 0$ to a system of equations by setting $x' = y$. Find the equation of the trajectories and carefully sketch seven or eight of them, so as to show clearly the key features of the phase portrait. Pay special attention to the one through the origin, and give its equation.

11. (*Volterra problem*) Consider the Volterra problem (24), with $\alpha = \beta = \gamma = \delta = 1$. Determine any fixed points. Use phaseportrait (or other software) to obtain the lineal element field, with barbed arrows, over the region $0 < x < 4$ and $0 < y < 4$, say. (Of course, x and y need to be positive because they are populations.). On that plot, sketch a number of representative trajectories. You should find a circulatory motion about the point $(1, 1)$. Can you tell, from the lineal element field, whether the trajectories circulate in closed orbits or whether they spiral in (or away from) that point?

7.3 Singular Points and Stability

7.3.1. Existence and uniqueness. Before we go any further, we need to return to the question of the existence and uniqueness of solutions. Theorem 2.4.1 gives conditions on f for the equation $y' = f(x, y)$ to admit a unique solution through a given initial point $y(a) = b$. Theorem 3.9.1 does likewise for a system of such equations, but covered only linear equations. In the present chapter, however, our principal interest is in systems of nonlinear equations so we give the following result, which is essentially a generalization of Theorem 2.4.1 to a system of equations.

THEOREM 7.3.1 *Existence and Uniqueness*
Let $f(x, y, t)$ and $g(x, y, t)$ and each of the partial derivatives f_x, f_y, g_x, g_y exist and be continuous in some neighborhood of the point (x_0, y_0, t_0) in Cartesian x, y, t space. Then the initial-value problem

$$\frac{dx}{dt} = f(x, y, t); \qquad x(t_0) = x_0 \tag{1a}$$

$$\frac{dy}{dt} = g(x, y, t); \qquad y(t_0) = y_0 \tag{1b}$$

has a solution $x(t)$, $y(t)$ on some open t interval containing $t = t_0$ and that

solution is unique.

More general and more powerful theorems could be given but this one will suffice for our purposes in this chapter. For such theorems and proofs we refer you to texts on differential equations such as G. Birkhoff and G.-C. Rota, *Ordinary Differential Equations*, 2nd ed. (New York: John Wiley, 1969).

For instance, consider the soft-spring oscillator equation $x'' + x - x^3 = 0$ or, equivalently, the system

$$x' = y, \tag{2a}$$
$$y' = -x + x^3 \tag{2b}$$

that we studied in Section 7.2. In this case $f(x, y, t) = y$, $g(x, y, t) = -x + x^3$, $f_x = 0$, $f_y = 1$, $g_x = -1 + 3x^2$, $g_y = 0$ are continuous for all values of x, y, and t, so Theorem 7.3.1 assures us that no matter what initial condition is chosen there is a unique solution through it. The extent of the t interval over which that solution exists is not predicted by the theorem, which is a "local" theorem like Theorem 2.4.1. But it is understood that that interval is not merely the point t_0 itself, for how could dx/dt and dy/dt make sense if $x(t)$ and $y(t)$ were defined only at a single point? Linear differential equations are simpler, and for them we have "global" theorems such as Theorem 3.9.1.

If f and g satisfy the conditions of Theorem 7.3.1 at (x_0, y_0, t_0) in x, y, t space, then there does *exist* a solution curve, or trajectory, through that point, and there is only *one* such trajectory. Geometrically, it follows that trajectories in x, y, t space cannot touch or cross each other at a point of existence and uniqueness. However, what about the possibility of crossings of trajectories in the x, y phase plane? Be careful, because whereas the theorem precludes crossings in three-dimensional x, y, t space, the phase plane shows only the projection of the three-dimensional trajectories onto the two-dimensional x, y plane. For instance, choose any point P_0 on a closed orbit inside the "football" in Fig. 8 of Section 7.2. As the representative point P goes round and round on that orbit it passes through P_0 an infinite number of times, yet that situation does not violate the theorem because if that trajectory is viewed in three-dimensional x, y, t space, we see that it is actually helical, and there are no self-crossings. The only points of serious concern in Fig. 8 are $(1, 0)$ and $(-1, 0)$. But here too there is no violation of the theorem because there is only the unique trajectory $x(t) = 1$ and $y(t) = 0$ through any initial point $(1, 0, t_0)$ – namely, a straight-line trajectory which is perpendicular to the x, y plane and which extends from $-\infty$ to $+\infty$ in the t direction. The trajectories DE and IE, in the x, y, t space, approach that line asymptotically as $t \to \infty$, and the trajectories FE and HE approach it asymptotically (both in x, y, t space and in the x, y phase plane) as $t \to -\infty$, but they never reach it. Similarly for $(-1, 0, t_0)$.

Recall, from Section 7.2, that we proceeded to divide (2b) by (2a), obtaining

$$\frac{dy}{dx} = \frac{-x + x^3}{y}. \tag{3}$$

Thus, apart from the question of existence and uniqueness for the system (2) in x, y, t space, there is the separate question as to the existence and uniqueness of the solution of the single equation (3) through a given initial point in the x, y plane, to which question we now turn. Here, $f(x, y) = (-x + x^3)/y$ and $f_y = (x - x^3)/y^2$ are continuous everywhere except on the line $y = 0$, that is, on the x-axis. However, since we are not aiming at finding $y(x)$ but at finding the phase trajectories, then it doesn't matter if, at any given location, we regard y as a function of x or vice versa. Thus, if we re-express (3) as $dx/dy = y/(-x+x^3)$, then we find that the right side and its partial derivative with respect to the dependent variable x are continuous everywhere except at $x = 0, 1$, and -1. Thus, the only points in the phase plane at which there can possibly be breakdowns of existence and uniqueness of a trajectory passing through that point are the intersections of the horizontal line $y = 0$ and the vertical lines $x = 0, 1$, and -1, namely, the points $(-1, 0), (0, 0)$, and $(1, 0)$, and we call these "singular points" of (3). At each of these points dy/dx and dx/dy are of the form $0/0$, so there is no unique slope defined through that point. Indeed we do find breakdowns of uniqueness at the points $(-1, 0)$ and $(1, 0)$ in Fig. 8 because at each point we find two crossing trajectories. (Remember that we are in x, y space now, not x, y, t space, and that time does not enter explicitly.) And at $(0, 0)$ we find a failure of existence because there is no solution that passes through that point.

The upshot is that even though the system (2) has a unique solution in x, y, t space, through any point $(-1, 0, t_0)$, $(0, 0, t_0)$, or $(1, 0, t_0)$, the equation (3) does *not* have a unique solution through any of the points $(-1, 0), (0, 0)$, and $(1, 0)$. We therefore say that these three points are singular points of (3), which concept we now formalize.

7.3.2. Singular points. Thus, in considering the general case

$$\frac{dy}{dx} = \frac{f(x, y)}{g(x, y)}, \tag{4}$$

obtained from $x' = dx/dt = f(x, y)$ and $y' = dy/dt = g(x, y)$ by division, we say that any point x_s, y_s at which both $f(x, y)$ and $g(x, y)$ vanish is a **singular point** (or **critical point**) of (4). Recall that from a dynamical point of view a singular point is an equilibrium point, or fixed point, because if we start at such a point then we remain there for all t because $x' = y' = 0$ there.

Singular points are of great importance in phase plane analysis. They need not be isolated. For example, if $f(x, y) = x$ and $g(x, y) = x$, then every point on the y axis (i.e., the line $x = 0$) is a singular point. To study a singular point of a nonlinear system, we focus attention on its immediate neighborhood by expanding $f(x, y)$ and $g(x, y)$ in Taylor series about that point and linearizing, that is, cutting them off after the first-order terms. For instance, to study the singular point of (2) at $(1, 0)$, write

$$x' = y, \tag{5a}$$

$$y' = -x + x^3 = 0 + 2(x - 1) + \frac{6}{2!}(x - 1)^2 + \frac{6}{3!}(x - 1)^3. \tag{5b}$$

Close to $(1,0)$ we neglect the higher-order terms and consider the linearized version

$$x' = y, \tag{6a}$$
$$y' = 2(x-1) \tag{6b}$$

or, moving our coordinate system to $(1,0)$ for convenience by setting $X = x - 1$ and $Y = y$,

$$X' = Y, \tag{7a}$$
$$Y' = 2X. \tag{7b}$$

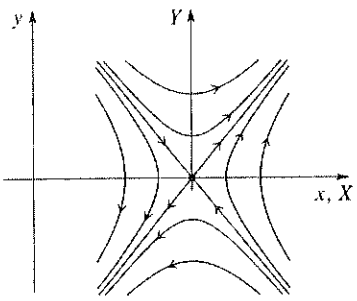


Figure 1. The flow near the singular point $(1,0)$.

Dividing these gives $dY/dX = 2X/Y$, with the solution $Y^2 = 2X^2 + C$; $C = 0$ gives $Y = \pm\sqrt{2}X$; and for $C \neq 0$ we have $Y = \pm\sqrt{(2X^2 + C)}$. These curves are shown in Fig. 1. The pairs crossing the X axis correspond to increasingly negative values of C, and those crossing the Y axis correspond to increasingly positive values of C.

Similarly, to study the singular point $(0,0)$ observe that the right-hand sides of (2a) and (2b) are already Taylor series expansions about $x = 0$ and $y = 0$. Thus, keeping terms only up to first order gives the linearized version

$$x' = y, \tag{8a}$$
$$y' = -x, \tag{8b}$$

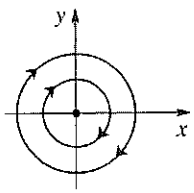


Figure 2. The flow near the singular point (0,0).

with $dy/dx = -x/y$ giving the family of circles $y^2 + x^2 = C$ and hence the trajectories shown in Fig. 2.

Treating the singular point $(-1,0)$ in the same manner, we find the same behavior there as at $(1,0)$. If we show the three results together, in Fig. 3, it is striking how those three localized phenomena appear to set up the global flow in the whole plane. That is, if we fill in what's missing, by hand, we obtain – at least in a qualitative or topological sense – the same picture as in Fig. 8 of Section 7.2.*

From this example we can see some things and raise some questions. We see that by virtue of our Taylor expansions of $f(x,y)$ and $g(x,y)$ about the singular point (x_s, y_s) and their linearization, we are always going to end up with linearized equations of the form

$$X' = aX + bY, \tag{9a}$$
$$Y' = cX + dY \tag{9b}$$

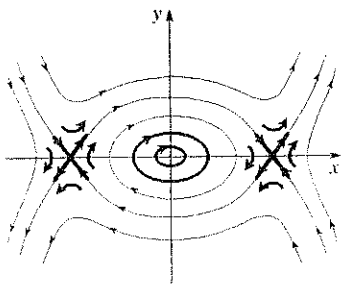


Figure 3. Global flow determined, qualitatively, by the singular points.

to study, where $X \equiv x - x_s$ and $Y \equiv y - y_s$ are Cartesian coordinate axes located at the singular point. Thus, we might as well study the general system (9) once and for all. Evidently, for different combinations of a, b, c, d there can be different types of singular points, for from Fig. 3 it seems clear that the ones at $(1,0)$ and $(-1,0)$ are different from the one at $(0,0)$. How many different types are there?

*It may appear inconsistent that the trajectories near the origin in Fig. 3 look elliptical, whereas they are circles in Fig. 2. That distortion, from circles to ellipses, is merely the result of stretching the x axis relative to y axis for display purposes.

What are they?

7.3.3. The elementary singularities and their stability. We wish to solve (9), examine the results, and classify them into types. If we equate the right-hand sides of (9) to zero, we have the unique solution $X = Y = 0$ only if the determinant $ad - bc$ is nonzero. If that determinant vanishes, then the solution $X = Y = 0$ is nonunique, and there is either an entire line of solutions through the origin or the entire plane of solutions. For instance, if $a = b = 0$ and c and d are not both zero, then every point on the line $cX + dY = 0$ is a singular point of (9), and if $a = b = c = d = 0$, then every point in the plane is a singular point of (9). Wishing to study the generic case, where the origin is an *isolated* singular point, we will require of a, b, c, d that $\mathbf{ad - bc \neq 0}$.

If we solve (9), for instance by elimination, we find that the solution is of the form

$$X(t) = C_1 e^{\lambda_1 t} + C_2 e^{\lambda_2 t}, \quad Y(t) = C_3 e^{\lambda_1 t} + C_4 e^{\lambda_2 t}, \tag{10a,b}$$

where C_1, C_2, C_3, C_4 are not independent, and where λ_1, λ_2 are the roots

$$\lambda = \frac{a + d \pm \sqrt{(a-d)^2 + 4bc}}{2} \tag{11}$$

of the characteristic equation

$$\lambda^2 - (a + d)\lambda + (ad - bc) = 0. \tag{12}$$

Since $ad - bc \neq 0$, zero is not among the roots. There are exactly four possibilities:

(1) purely imaginary roots (CENTER),
(2) complex conjugate roots (FOCUS),
(3) real roots of the same sign (NODE),
(4) real roots of opposite sign (SADDLE).

These cases lead to four different types of singularity: center, focus, node, and saddle, as we note within parentheses, and we will discuss these in turn. In doing so, it is important to examine each in terms of its stability, which concept we define before continuing.

A singular point $S = (x_s, y_s)$ of the autonomous system (4) is said to be **stable** if motions (i.e., trajectories) that start out sufficiently close to S remain close to S. To make that intuitively stated definition mathematically precise, let $d(P_1, P_2)$ denote the distance* between any two points $P_1 = (x_1, y_1)$ and $P_2 = (x_2, y_2)$. Further, we continue to let $P(t) = (x(t), y(t))$ denote the representative point in the phase plane corresponding to (4). Then, a singular point S is stable if, given any $\epsilon > 0$ (i.e., as small as we wish) there is a $\delta > 0$ such that $d(P(t), S) < \epsilon$ for all $t > 0$ if $d(P(0), S) < \delta$. (See Fig. 4a.) If S is not stable, then it is **unstable**.

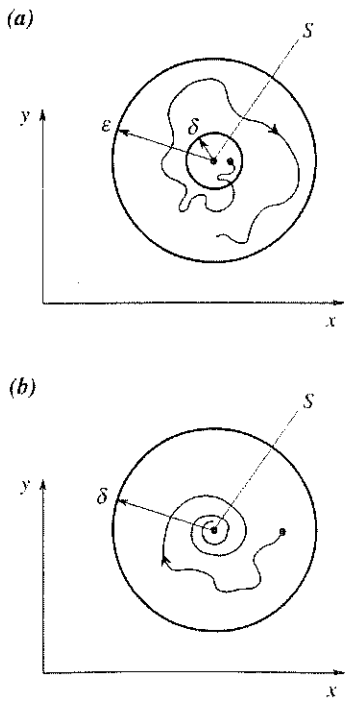


Figure 4. Stability and asymptotic stability.

*In the Euclidean sense, the distance $d(P_1, P_2)$ is defined as $\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$, but one can define distance in other ways. Here, we understand it in the Euclidean sense.

Further, we say that S is not only stable but **asymptotically stable** if motions that start out sufficiently close to S not only stay close to S but actually approach S as $t \to \infty$. That is, if there is a $\delta > 0$ such that $d(P(t), S) \to 0$ as $t \to \infty$ whenever $d(P(0), S) < \delta$, then S is asymptotically stable. (See Fig. 4b.)

Now let us return to the four cases listed. The most inciteful way to study these cases is by seeking solutions in exponential form and dealing with the "eigenvalue problem" that results. However, the eigenvalue problem is not discussed until Chapter 11, so in the present section we rely on an approach that should suffice but which is in some ways less satisfactory. In Section 11.5 we return to this problem and deal with it as an eigenvalue problem. If you are already sufficiently familiar with linear algebra, we suggest that you study that section immediately following this one.

It is convenient to use the physical example of the mechanical oscillator, with the governing equation $mx'' + px' + kx = 0$, or

$$x' = y, \tag{13a}$$

$$y' = -\frac{k}{m}x - \frac{p}{m}y \tag{13b}$$

as a unifying example because by suitable choice of m, p, k we can obtain each of the four cases, and because this application has already been discussed in Section 3.5. [Here we use p instead of c for the damping coefficient to avoid confusion with the c in (9b).]

Purely imaginary roots. (CENTER) Let $p = 0$ so there is no damping. Then $a = 0$, $b = 1$, $c = -k/m$, $d = 0$; (11) gives the purely imaginary roots $\lambda = \pm i\sqrt{k/m}$ and $dy/dx = -(k/m)x/y$ gives the family of ellipses

$$\frac{1}{2}my^2 + \frac{1}{2}kx^2 = C \tag{14}$$

sketched in Fig. 5a. The singular point at $(0, 0)$ is called a **center** because it is surrounded by closed orbits corresponding to periodic motions. For instance, with $\lambda_1 = +i\omega$ (where $\omega = \sqrt{k/m}$ is the natural frequency) and $\lambda_2 = -i\omega$, (10a) gives $x(t) = C_1 \exp(i\omega t) + C_2 \exp(i\omega t)$ or, equivalently,

$$x(t) = A \sin(\omega t + \phi). \tag{15}$$

(Here, $X = x$ and $Y = y$ because the singular point is at $x = y = 0$.) In Fig. 5a the principal axes of the elliptical orbits coincide with the x, y coordinate axes. More generally, they need not. For instance, for the system

$$x' = \frac{\sqrt{8}}{3}x + \frac{4}{3}y, \tag{16a}$$

$$y' = -\frac{11}{3}x - \frac{\sqrt{8}}{3}y \tag{16b}$$

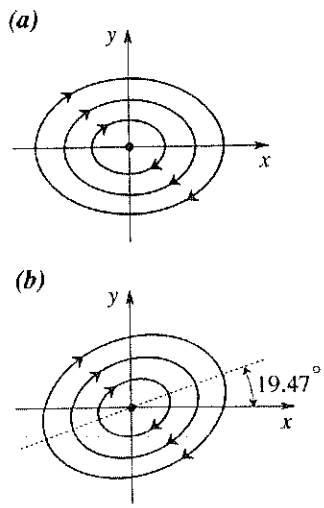


Figure 5. A center at $(0, 0)$.

(11) again gives purely imaginary roots, $\lambda = \pm i/\sqrt{3}$ so the solutions are harmonic oscillations with frequency $1/\sqrt{3}$, but the principal axes of the elliptical orbits are at an angle of $\sin^{-1}(1/3) = 19.47°$ with respect to the x, y axes as shown in Fig. 5b (see Exercise 5). [The system (16) is, of course, not a special case of (13), it is a separate example.]

We see that a center is stable but not asymptotically stable.

Complex conjugate roots. (FOCUS) This time let p be positive in (13), but small enough so that $p < \sqrt{4km}$. According to the terminology introduced in Section 3.8, we say that the damping is subcritical because $p_{cr} = \sqrt{4km}$. Then $a = 0$, $b = 1$, $c = -k/m$, $d = -p/m$; (11) gives the complex conjugate roots

$$\lambda = -\frac{p}{2m} \pm \sqrt{\left(\frac{p}{2m}\right)^2 - \frac{k}{m}} = -\frac{p}{2m} \pm i\sqrt{\frac{k}{m} - \left(\frac{p}{2m}\right)^2},$$

and (10a) gives the solution

$$\begin{aligned} x(t) &= e^{-pt/2m}\left[A\cos\left(\sqrt{\frac{k}{m} - \left(\frac{p}{2m}\right)^2}\right)t + B\sin\left(\sqrt{\frac{k}{m} - \left(\frac{p}{2m}\right)^2}\right)t\right] \\ &= Ce^{-pt/2m}\sin\left(\sqrt{\frac{k}{m} - \left(\frac{p}{2m}\right)^2}\,t + \phi\right), \end{aligned} \tag{17}$$

where C and ϕ are arbitrary constants. As we discussed in Section 3.8, this solution differs from the undamped version (15) in two ways. First, the frequency of the sinusoid is diminished from the natural frequency $\omega = \sqrt{k/m}$ to $\sqrt{k/m - (p/2m)^2}$, and the $\exp(-pt/2m)$ factor modulates the amplitude, reducing it to zero as $t \to \infty$. In terms of the phase portrait, one obtains a family of spirals such as the one shown in Fig. 6a. If we imagine the representative point P moving along that curve, we see that the projection onto the x axis is indeed a damped oscillation. We call the singularity at the origin a **focus** because trajectories "focus" to the origin as $t \to \infty$; the term **spiral** is also used.

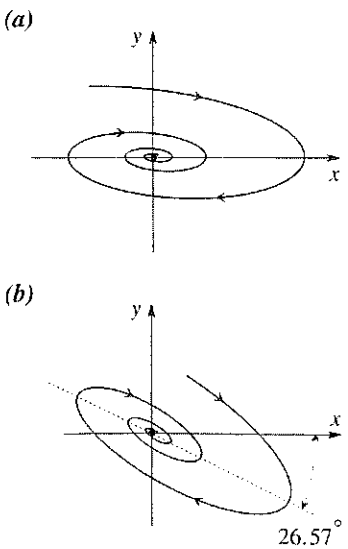


Figure 6. A stable focus at $(0, 0)$.

In Fig. 6a the principal axes of the orbits, which would be elliptical if not for the damping, coincide with the x, y coordinate axes. More generally, they need not. For instance, for the system

$$x' = \frac{1}{3}x + \frac{7}{9}y, \tag{18a}$$

$$y' = -\frac{1}{3}x - \frac{4}{9}y \tag{18b}$$

we obtain similar results but with the principal axes rotated clockwise by an angle of $\sin^{-1}(1/\sqrt{5}) = 26.57°$ as shown in Fig. 6b.

In each case (Fig. 6a and 6b) we see that the focus is stable and, indeed, asymptotically stable as well. However, one can have foci that wind outward instead of

inward, and these will be unstable. For instance, if we return to the solution (17) of the damped oscillator system (13), but this time imagine p to be negative (without concerning ourselves with how that might be arranged physically), and smaller in magnitude than $\sqrt{4km}$ as before, then in place of the clockwise inward flow shown in Fig. 5a, we obtain the counterclockwise outward flow shown in Fig. 7, and we classify the singularity at the origin as an unstable focus. Note that a stable singular point can, additionally, be asymptotically stable or not, but an unstable singular point is simply unstable.

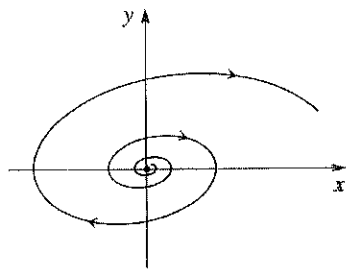


Figure 7. An unstable focus at$(0,0)$.

Real roots of the same sign. (NODE) We've seen that without damping the mechanical oscillator (13) gives pure oscillations, elliptical orbits, and a center. With light damping (i.e., $0 < p < p_{cr}$) it gives damped oscillations and a stable focus. If we now increase p so as to exceed p_{cr}, then the oscillations disappear altogether, and we have the solution form

$$x(t) = C_1 e^{\lambda_1 t} + C_2 e^{\lambda_2 t} \tag{19a}$$

with

$$\lambda_1 = -\frac{p}{2m} + \sqrt{\left(\frac{p}{2m}\right)^2 - \frac{k}{m}}, \qquad \lambda_2 = -\frac{p}{2m} - \sqrt{\left(\frac{p}{2m}\right)^2 - \frac{k}{m}}.$$

Because of the way we have numbered the λ's, we have $\lambda_2 < \lambda_1 < 0$. Since $x' = y$ in this application we have from (19a),

$$y(t) = \lambda_1 C_1 e^{\lambda_1 t} + \lambda_2 C_2 e^{\lambda_2 t}. \tag{19b}$$

We can see from (19) that

$$x(t) \sim C_1 e^{\lambda_1 t}, \qquad y(t) \sim \lambda_1 C_1 e^{\lambda_1 t} \tag{20}$$

and

$$y \sim \lambda_1 x \tag{21}$$

as $t \to \infty$, provided that the initial conditions do not give $C_1 = 0$. If they do give $C_1 = 0$, then

$$x(t) = C_2 e^{\lambda_2 t}, \qquad y(t) = \lambda_2 C_2 e^{\lambda_2 t} \tag{22}$$

and

$$y = \lambda_2 x \tag{23}$$

as $t \to \infty$.

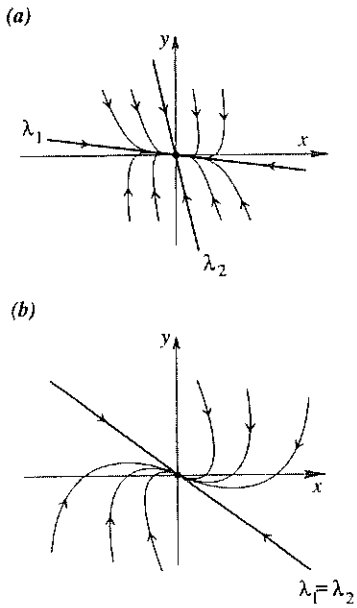


Figure 8. A stable improper node at $(0,0)$: distinct roots and repeated roots, respectively.

The resulting phase portrait is shown in Fig. 8a. We call the singularity at $(0,0)$ a **node** – more specifically, an **improper node** ("improper" is explained below). In accord with the preceding discussion, observe from Fig. 8a that in the exceptional case in which the initial point lies on the line of slope λ_2 the representative point approaches the origin along that line. In all other cases, the approach is asymptotic to the line of slope λ_1. If we let p tend to p_{cr}, then the two lines coalesce

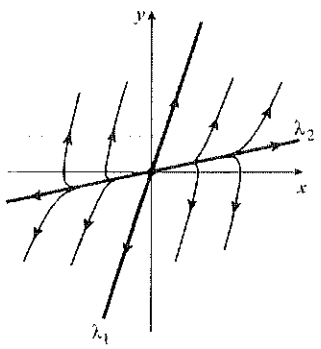


Figure 9. An unstable improper node at $(0,0)$: distinct roots.

and we obtain the portrait shown in Fig. 8b, which is, likewise, an improper node (Exercise 6).

If p is negative and greater in magnitude than p_{cr}, then we see from the expressions given above for the λ's, that $\lambda_1 > \lambda_2 > 0$, and the phase portrait is as shown in Fig. 9. The nodes shown in Fig. 8a,b are both stable, asymptotically stable, and the one shown in Fig. 9 (which is analogous to the one in Fig. 8a) is unstable.

Our mechanical oscillator example is but one example leading to a node. We could make up additional ones by writing down equations $x' = ax + by$ and $y' = cx + dy$ if we choose the coefficients a, b, c, d so that the two λ's are of the same sign, but the results will be one of the types shown in Fig. 8 and 9. There is, however, another type of node, which we illustrate by the problem

$$x' = ax, \tag{24a}$$
$$y' = ay. \tag{24b}$$

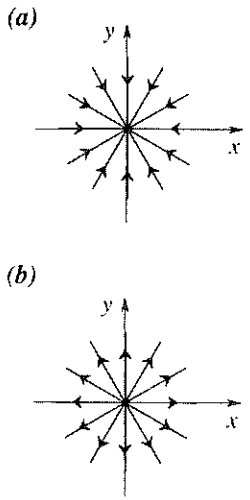


Figure 10. Stable and unstable proper nodes at $(0,0)$.

That is, $b = c = 0$ and $a = d$. Then $\lambda_1 = \lambda_2 = a$, and we have the solution

$$x(t) = Ae^{at}, \tag{25a}$$
$$y(t) = Be^{at}, \tag{25b}$$

where A, B are arbitrary. In this case y/x is not only asymptotic to a constant as $t \to \infty$, it is *equal* to a constant for all t. Thus, the phase portrait is as shown in Fig. 10a if $a < 0$, and as in Fig. 10b if $a > 0$. The former is an asymptotically stable node, and the latter is an unstable node. But this time we call them **proper nodes** (or **stars**) because every trajectory is a straight line through the singular point $(0,0)$, not just one or two of them.

Real roots of opposite sign. (SADDLE) Consider, once again, the undamped mass/spring system governed by the equation $mx'' + kx = 0$, or

$$x' = y, \tag{26a}$$
$$y' = -\frac{k}{m}x. \tag{26b}$$

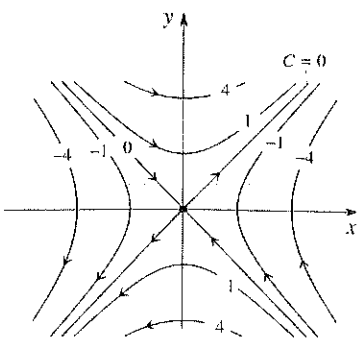


Figure 11. A saddle at $(0,0)$.

This time, imagine k to be negative (without concern about how that case could occur physically) and set $-k/m \equiv h^2$. Then (26) gives $dy/dx = h^2x/y$ so

$$y^2 = h^2x^2 + C \qquad \text{or} \qquad y(x) = \pm\sqrt{h^2x^2 + C}, \tag{27}$$

which trajectories are shown in Fig. 11 for various values of C. In particular, $C = 0$ gives the two straight lines through the singular point, namely, $y = \pm hx$, with the flow approaching the origin on one ($y = -hx$) and moving away from it on the other ($y = +hx$). Such a singular point is called a **saddle** and is always unstable. The two straight-line trajectories through the saddle, along which the flow is attracted and repelled, are called the **stable** and **unstable manifolds**, respectively.

Of course, (26) is not the only example of a linear system $x' = ax + by$ and $y' = cx + dy$ with a saddle. Any such system with real roots of opposite sign will have such a singularity. For instance,

$$x' = x + 2y, \tag{28a}$$
$$y' = 8x - 5y \tag{28b}$$

has the roots $\lambda = 3$ and $\lambda = -7$. Thus, it has a saddle, and we know that two straight-line solutions can be found through the origin. To find them, try $y = \kappa x$. Putting $y = \kappa x$ into (28) gives $x' = (1 + 2\kappa)x$ and $x' = (8 - 5\kappa)x/\kappa$ so it follows from these that we need $1 + 2\kappa = (8 - 5\kappa)/\kappa$, which equation gives the slopes $\kappa = 1$ and $\kappa = -4$. (If we would obtain $\kappa = \infty$, we would understand that, from $y = \kappa x$, to correspond to the x axis.) With $\kappa = 1$ the equation $x' = (1+2\kappa)x = 3x$ gives $x(t)$ proportional to $\exp(3t)$ [likewise for $y(t)$ because $y = \kappa x$], and with $\kappa = -4$ it gives $x(t)$ proportional to $\exp(-7t)$. Thus, the line trajectory $y = x$ is the unstable manifold (since x and y grow exponentially on it), and the line trajectory $y = -4x$ is the stable manifold (since x and y die out exponentially on it).

The same procedure, which we have just outlined and which should be clearly understood, can be used for a node as well, to find any straight-line trajectories through the node.

7.3.4. Nonelementary singularities. In this final subsection we turn from the elementary singularities to nonelementary ones, with two purposes in mind. First, one doesn't completely understand elementary singularities until one distinguishes elementary singularities from nonelementary ones and, second, nonelementary singularities do arise in applications.

Recall that

$$X' = aX + bY, \quad Y' = cX + dY \tag{29a,b}$$

has an elementary singularity at $(0, 0)$ if $ad - bc \neq 0$. Consider two examples.

EXAMPLE 1. The system

$$x' = y, \quad y' = y \tag{30a,b}$$

has the phase trajectories $y = x + C$ and the phase portrait shown in Fig. 12. Since $ad - bc = (0)(1) - (1)(0) = 0$, the singularity of (30) at $(0, 0)$ is nonelementary. It is nonisolated and, in fact, $y = 0$ is an entire line of singular points. ∎

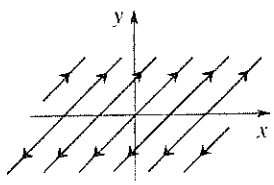


Figure 12. Nonelementary singularity of (30) at $(0, 0)$.

EXAMPLE 2. Consider the singularity of the system

$$x' = y, \quad y' = 1 - \cos x \tag{31a,b}$$

at $(0, 0)$. Expanding the right side of (31b) gives $1 - \cos x = \frac{1}{2}x^2 - \frac{1}{24}x^4 + \cdots$ so the linearized version of (31) is $x' = 0x + 1y$ and $y' = 0x + 0y$. Thus, $ad - bc = (0)(0) - (1)(0) = 0$ again and the singularity of (31) at the origin is nonelementary. The

difficulty this time is not that the singular point is not isolated; it *is*. The problem is that it is of higher order for when we linearize the expansion of $1-\cos x$ we simply have $0x+0y$. In not retaining at least the first nonvanishing term (namely, $x^2/2$) we have "thrown out the baby with the bathwater." To capture the local behavior of (31) near the origin, we need to retain that leading term and consider the system

$$x' = y, \quad y' = \frac{1}{2}x^2. \tag{32a,b}$$

Dividing (32b) by (32a) and integrating gives $y = \sqrt{\frac{x^3}{3} + C}$, several of which trajectories are shown in Fig. 13. We see that the singularity is, indeed, isolated, but that the phase portrait is not of one of the elementary types. ∎

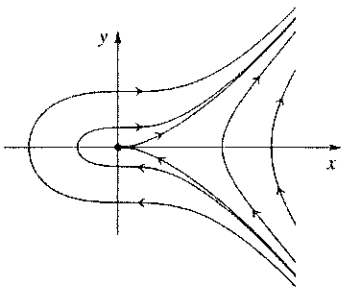


Figure 13. Nonelementary singularity of (31) at $(0,0)$.

Closure. In this section we establish a foundation for our use of the phase plane in studying nonlinear systems. We begin with the issue of existence and uniqueness of solutions, first in x, y, t space (Theorem 7.3.1), and then in the x, y phase plane. The latter leads us to introduce the concept of a singular point in the phase plane as a point at which both $x' = P(x,y) = 0$ and $y' = Q(x,y) = 0$. To study a singular point $S = (x_s, y_s)$ one focuses on the immediate neighborhood of S, in which neighborhood we work with the locally linearized equations $X' = aX + bY$ and $Y' = cX + dY$, where $X = x - x_s$ and $Y = y - y_s$, so X, Y is a Cartesian system with its origin at S. Studying that linearized system, we categorize the possible "flows" into four qualitatively distinct types – the center, focus, node, and saddle – and illustrate each through the mass/spring system $mx'' + px' + kx = 0$, with suitable choices of m, p, k, and other examples as well. These are the so-called elementary singularities that result when $ad - bc \neq 0$. In the next section we apply these results to several nonlinear systems, where we will see the role of such singular points in establishing the key features of the overall flow in the phase plane.

EXERCISES 7.3

1. Find all singular points of the given system. Are they isolated?

(a) $x' = 0$, $y' = 2x - y$

(b) $x' = 2x - 2y$, $y' = x - y$

(c) $x' = x^2 + y^2 - 1$, $y' = x - y$

(d) $x' = \sin y$, $y' = x + y$

(e) $x' = x^2 + y^2$, $y' = x + y$

(f) $x' = 1 - e^y$, $y' = 1 - x^2 - x\sin y$

(g) $x' = xy - 4$, $y' = x - 2y$

(h) $x' = \cos(x - y)$, $y' = xy - 1$

2. Is it possible for (4) to have *no* singular points? Explain.

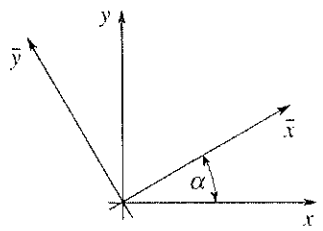
3. Derive the solution (10) of (9) by the method of elimination, and find C_3 and C_4 in terms of C_1 and C_2.

4. Suppose that we reverse the ϵ's and δ's in our definition of stability, so that the definition becomes: A singular point S is stable if, given any $\epsilon > 0$ (i.e., as small as we wish), there is a $\delta > 0$ such that $d(P(t), S) < \delta$ for all $t > 0$ if $d(P(0), S) < \epsilon$. Would that definition work? That is, would it satisfy the idea of motions that start out sufficiently close to S remaining close to S? Explain.

5. In this exercise we wish to elaborate on our claim below (16) that the principal axes of the elliptical orbits are at an an-

gle of 19.47° with respect to the x, y axes as shown in Fig. 5b.

(a) If x, y and $\overline{x}, \overline{y}$ coordinate systems are at an angle α, as shown here, show that the x, y and $\overline{x}, \overline{y}$ coordinates of any



given point are related according to

$$x = \overline{x}\cos\alpha - \overline{y}\sin\alpha, \tag{5.1a}$$

$$y = \overline{x}\sin\alpha + \overline{y}\cos\alpha \tag{5.1b}$$

(b) Putting (5.1) into (16), insist upon the result being of the form

$$\overline{x}' = \beta^2\overline{y}', \qquad \overline{y}' = -\gamma^2\overline{x}' \tag{5.2}$$

for some constants β and γ so as to yield elliptical orbits with $\overline{x}, \overline{y}$ as principal axes, and show that you obtain $\alpha = 19.47°$. If you obtain another α as well, explain its significance.

6. We claimed that if $p = p_{cr}$, then the two straight-line trajectories in Fig. 8a coalesce, as shown in Fig. 8b. Here, we ask you to verify that claim. Begin by recalling that if $\lambda_1 = \lambda_2 = \lambda$, then the general solution of (13) is

$$x(t) = (C_1 + C_2 t)e^{\lambda t}, \qquad y(t) = x'(t) = \text{etc.}$$

7. What does Fig. 8a look like in the limit as $p \to \infty$? Sketch it.

8. Given the presence of the saddles (i.e., saddle-type singularities) at $(-1, 0)$ and $(1, 0)$ and the center at $(0, 0)$, can you come up with any global flow patterns that are qualitatively different from the one sketched in Fig. 3? Explain. (Assume that these three are the only singularities.)

9. (*Saddles and nodes*) Classify the singularity at the origin, find the equations of any straight-line trajectories through the origin, and sketch the phase portrait, including flow direction arrows.

(a) $x' = x + y$, $y' = 4x + y$
(b) $x' = y$, $y' = -x - 4y$
(c) $x' = x + 2y$, $y' = x - 2y$
(d) $x' = -x + 3y$, $y' = x - y$
(e) $x' = 3x + y$, $y' = -x + y$
(f) $x' = -3x + y$, $y' = -x - y$
(g) $x' = 2x + y$, $y' = x + 2y$
(h) $x' = x + 3y$, $y' = 3x + y$
(i) $x' = x + y$, $y' = x + 2y$
(j) $x' = -3x + y$, $y' = x - 3y$

10. Prove that a linear system $x' = ax + by$, $y' = cx + dy$ can have one, two, or an infinite number of straight-line trajectories through the origin, but never a finite number greater than two.

11. Classify the singularity at the origin as a center, focus, node, or saddle. If it is a focus, node, or saddle, then classify it, further, as stable or unstable.

(a) $x' = x - 4y$, $y' = x + y$
(b) $x' = 2x + 3y$, $y' = 2x - y$
(c) $x' = x + y$, $y' = x - 4y$
(d) $x' = x + 3y$, $y' = -x - y$
(e) $x' = -2x - 3y$, $y' = 3x + 2y$
(f) $x' = -x + y$, $y' = -x - 2y$
(g) $x' = 2x - y$, $y' = -x + 3y$
(h) $x' = -2x - y$, $y' = -x - 3y$

12. (a)–(h) Use computer software to obtain the phase portrait for the corresponding system in Exercise 11. Be sure to include any key trajectories – namely, any straight-line trajectories through the origin. From the phase portrait, classify the singularity as a center, focus, node, or saddle, state whether it is stable, asymptotically stable, or unstable, and use arrows to show the flow direction.

7.4 Applications

In Sections 7.2 and 7.3 we established the phase plane concept, the idea of singularities, the center, focus, node, and saddle singularities of linear systems, and their

stability or instability. With those fundamentals in hand, we can undertake some interesting applications.

7.4.1. Singularities of nonlinear systems. We are interested in the autonomous system

$$x' = P(x, y), \tag{1a}$$
$$y' = Q(x, y). \tag{1b}$$

The x, y points at which both $P(x, y) = 0$ and $Q(x, y) = 0$ are the singular points of (1). Suppose we have determined those singular points, if any. We have emphasized the importance of determining the local flow near each such point. To do that, it should suffice to expand P and Q in Taylor series about the given singular point, say $S = (x_s, y_s)$, and to retain terms only up to the first order. Thus, instead of studying the complete equations (1) near S, we intend to simplify them by linearizing them about S.

Though familiar with the Taylor series of a function of one variable, from the calculus, you may not be familiar with Taylor series for functions of two or more variables. The Taylor series expansion of a given function of *two* variables, $f(x, y)$, about any point (a, b), is

$$\begin{aligned} f(x,y) = f(a,b) &+ \frac{1}{1!}\left[f_x(a,b)(x-a) + f_y(a,b)(y-b)\right] \\ &+ \frac{1}{2!}\left[f_{xx}(a,b)(x-a)^2 + 2f_{xy}(a,b)(x-a)(y-b) + f_{yy}(a,b)(y-b)^2\right] \\ &+ \text{terms of third order and higher,} \end{aligned} \tag{2}$$

where $(x-a)^m(y-b)^n$ is said to be of order $m+n$. One way to derive (2) is to expand $f(x, y)$ in one variable at a time. Thus, holding y fixed for the moment, and expanding in x about $x = a$, we know from the Taylor series formula for a function of a single variable that

$$f(x,y) = f(a,y) + f_x(a,y)(x-a) + \frac{1}{2!}f_{xx}(a,y)(x-a)^2 + \cdots. \tag{3}$$

The coefficients are functions of y alone, and each can now be expanded in y about $y = b$:

$$\begin{aligned} f(a,y) &= f(a,b) + f_y(a,b)(y-b) + \frac{1}{2!}f_{yy}(a,b)(y-b)^2 + \cdots, \\ f_x(a,y) &= f_x(a,b) + f_{xy}(a,b)(y-b) + \frac{1}{2!}f_{xyy}(a,b)(y-b)^2 + \cdots, \end{aligned} \tag{4}$$

and so on. Putting these into (3) and arranging terms in ascending order produces (2).

If we approximate f near (a, b) by cutting off the Taylor series (2) after the first-order terms, we have

$$f(x,y) \approx f(a,b) + f_x(a,b)(x-a) + f_y(a,b)(y-b). \tag{5}$$

That step is called **linearization** because the result is linear in x and y. Geometrically, (5) amounts to approximating the f surface, plotted above the x, y plane, by its **tangent plane** at (a, b) just as $f(x) = f(a) + f'(a)(x-a) + \frac{1}{2}f''(a)(x-a)^2 + \cdots \approx f(a) + f'(a)(x-a)$ amounts to approximating the graph of f versus x by its **tangent line** at $x = a$ in the case of a function of a single variable.

Returning to (1), the idea is to expand P and Q about the singular point of interest, (x_s, y_s), and linearize. Using (5) to do that, we obtain

$$P(x, y) \approx P(x_s, y_s) + P_x(x_s, y_s)(x - x_s) + P_y(x_s, y_s)(y - y_s), \tag{6a}$$
$$Q(x, y) \approx Q(x_s, y_s) + Q_x(x_s, y_s)(x - x_s) + Q_y(x_s, y_s)(y - y_s). \tag{6b}$$

But $P(x_s, y_s)$ and $Q(x_s, y_s)$ are zero because (x_s, y_s) is a singular point, so we have the approximate (linearized) equations

$$\boxed{\begin{aligned} x' &= P_x(x_s, y_s)(x - x_s) + P_y(x_s, y_s)(y - y_s), \\ y' &= Q_x(x_s, y_s)(x - x_s) + Q_y(x_s, y_s)(y - y_s). \end{aligned}} \tag{7}$$

Finally, it is convenient, though not essential, to move the origin to S by letting $X \equiv x - x_s$ and $Y \equiv y - y_s$, and calling $P_x(x_s, y_s) = a$, $Q_x(x_s, y_s) = c$, $P_y(x_s, y_s) = b$, and $Q_y(x_s, y_s) = d$, for brevity, in which case (7) becomes

$$X' = aX + bY, \tag{8a}$$
$$Y' = cX + dY, \tag{8b}$$

which system is studied in Section 7.3. There, we classify the singularity at $X = Y = 0$ as a center, focus, node, or saddle, depending upon the roots of the characteristic equation

$$\lambda^2 - (a + d)\lambda + (ad - bc) = 0, \tag{9}$$

namely,

$$\lambda = \frac{(a + d) \pm \sqrt{(a - d)^2 + 4bc}}{2}. \tag{10}$$

Knowing the numerical values of a, b, c, d, for the singular point in question, we can find the λ roots and determine whether the singular point is a center, focus, node, or saddle.

Understand that we are trying to ascertain the local behavior of the flow corresponding to the original nonlinear system (1) near the singular point S by studying the simpler linearized version of (1) at S, namely, (8). That program begs this question: Is the nature of the singularity of the nonlinear system (1), at S, truly captured by its linearized version (8)? To answer that question it is helpful to present the singularity classification, developed in Section 7.3, in a graphical form – as we have in Fig. 1. It is more convenient to deal with the two quantities $\mathrm{p} \equiv \mathrm{a} + \mathrm{d}$ and

$\mathbf{q} \equiv \mathbf{ad} - \mathbf{bc}$ (which are the axes in Fig. 1) than with the four quantities a, b, c, d since $a + d$ and $ad - bc$ determine the roots of (10) and hence the singularity type. In terms of p and q, (10) simplifies to $\lambda = (p \pm \sqrt{p^2 - 4q})/2$.

In the figure there are five regions separated by the p axis, the parabola $p^2 = 4q$, and the positive q axis. [Of these boundaries, the p axis can be discounted since the case $q = 0$ is ruled out of consideration in Section 7.3 because then there is a line of singular points through the origin rather than the origin being an isolated singular point. That is, our (p, q) point will not fall on the boundary between saddles and improper nodes – namely, the p axis.]

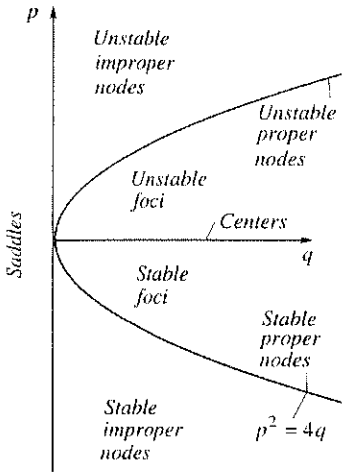


Figure 1. Singularity diagram; $p = a + d$, $q = ad - bc$.

The **Hartman–Grobman theorem** tells us that if a, b, c, d are such that the point (p, q) is *within* one of those regions, then the singularity types of the nonlinear system and its linearized version are identical. For instance, if the linearized system has a saddle, then so does the nonlinear system. Essentially, we can think of retaining the higher-order terms (which we drop in linearizing the nonlinear differential equations) as equivalent to infinitesimally perturbing the values of the coefficients a, b, c, d in the linearized equations and hence the values of p and q. If the point (p, q) is within one of the five regions, then the perturbed point will be within the same region, so the singularity type will be the same for the nonlinear system as for the linearized one.

However, we can imagine that for a borderline case, where (p, q) is on the parabola $p^2 = 4q$ or on the positive q axis, such a perturbation can push the point into one of the neighboring regions, thus changing the type. In fact, that is the way it turns out. For instance, if (p, q) is on the positive q axis (as occurs in the example to follow), then the nonlinear system could have an unstable focus or a center or a stable focus.

EXAMPLE 1. *Oscillator with Cubic Damping.* The equation $x'' + \epsilon x'^3 + x = 0$ models a harmonic oscillator with cubic damping – that is, with a damping term proportional to the velocity cubed. The equivalent system

$$\begin{aligned} x' &= y, \\ y' &= -x - \epsilon y^3 \end{aligned} \tag{11}$$

has one singular point, a center at $x = y = 0$. The linearized version is

$$\begin{aligned} X' &= Y = 0X + 1Y, \\ Y' &= -X = -1X + 0Y \end{aligned} \tag{12}$$

so $a = d = 0$, $b = 1$, and $c = -1$ hence $p = 0$ and $q = 1$. Thus, $(p, q) = (0, 1)$ so (from Fig. 1) the linearized system (12) has a center (no surprise, since the solutions of the linearized equation $x'' + x = 0$ are simple harmonic motions). However, it turns out (Exercise 1) that the nonlinear system (11) has a stable focus. ∎

To summarize, in general the linearized system faithfully captures the singularity type of the original nonlinear system. For the borderline cases, where (p, q)

is on the $p^2 = 4q$ parabola or on the positive q axis, however, we have these possibilities:

LINEARIZED		NONLINEAR
stable proper node	$\Longleftrightarrow$	stable focus, or stable proper node, or stable improper node
center	$\Longleftrightarrow$	unstable focus, or center, or stable focus
unstable proper node	$\Longleftrightarrow$	unstable improper node, or unstable proper node, or unstable focus

7.4.2. Applications. Consider some physical applications.

EXAMPLE 2. *Pendulum.* Recall from an introductory physics course that for a rigid body undergoing pure rotation about a pivot axis the inertia about the pivot O times the angular acceleration is equal to the applied torque. For the pendulum shown in Fig. 2, the inertia about O is ml^2, the angle from the vertical is $x(t)$, the angular acceleration is $x''(t)$, and the downward gravitational force mg gives a torque of $-mgl \sin x$. If the air resistance is proportional to the velocity lx', say clx', then it gives an additional torque $-cl^2x'$, so the equation of motion is $ml^2x'' = -mgl \sin x - cl^2x'$ or

$$x'' + rx' + \frac{g}{l} \sin x = 0, \tag{13}$$

where $r \equiv c/m$. The rx' term is a damping term. For definiteness, let $g/l = 1$ and consider the undamped case, where $r = 0$.

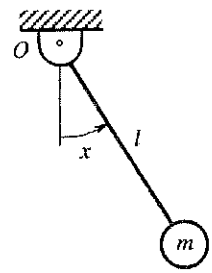


Figure 2. Pendulum.

For small motions we can approximate $\sin x$ by the first term of its Taylor series, $\sin x \approx x$, so that we have the simple harmonic oscillator equation $x'' + x = 0$ or

$$x' = y \tag{14a}$$

$$y' = -x; \tag{14b}$$

(14) has a center at $x = y = 0$, and its by now familiar phase portrait is shown in Fig. 3.

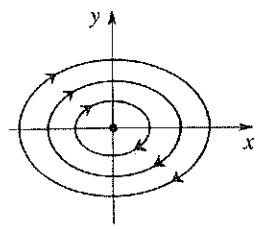


Figure 3. Phase portrait for the linearized system $x'' + x = 0$.

To study larger motions, suppose that we approximate $\sin x$ by the first *two* terms of its Taylor series instead: $\sin x \approx x - x^3/6$. Then we have the nonlinear, but still approximate, equation of motion

$$x'' + x - \frac{1}{6}x^3 = 0. \tag{15}$$

The latter is of the same form as the equation governing the rectilinear motion of a mass restrained by a "soft spring," which is studied in Section 7.2. The system

$$x' = y, \tag{16a}$$

$$y' = -x + \frac{1}{6}x^3 \tag{16b}$$

has a center at $(0,0)$ and saddles at $(\pm\sqrt{6},0)$ in the (x,y) phase plane, as discussed in Section 7.2, and its phase portrait is shown here in Fig. 4.

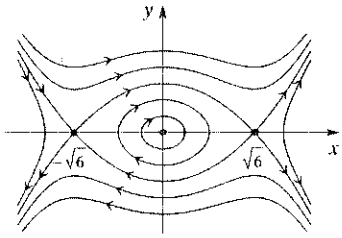


Figure 4. Phase portrait for the improved model (15).

Finally, if we retain the entire Taylor series of $\sin x$ (i.e., if we keep $\sin x$ intact), then we have the full nonlinear system (13), with $r = 0$, or

$$x' = y, \tag{17a}$$
$$y' = -\sin x, \tag{17b}$$

with singular points at $(x,y) = (n\pi, 0)$ for $n = 0, \pm 1, \pm 2, \ldots$. To classify these singularities, let us linearize equations (17) about the singular point $(n\pi, 0)$ using (7). Doing so, knowing that $\sin n\pi = 0$ and $\cos n\pi = (-1)^n$, and setting $X = x - n\pi$ and $Y = y - 0 = y$, the linearized version of (17) is

$$X' = Y = 0X + 1Y \tag{18a}$$
$$Y' = (-1)^{n+1}X = (-1)^{n+1}X + 0Y. \tag{18b}$$

In the notation of equation (8), $a = d = 0$, $b = 1$, and $c = (-1)^{n+1}$ so $p = a + d = 0$ and $q = ad - bc = (-1)^n$. Thus, these singular points are on the q axis in the p, q plane. For even integers n they are on the positive q axis and correspond to centers; for odd integers n they are on the negative q axis and correspond to saddles. In turn, the latter correspond to saddles of the nonlinear system (17), but the former could be centers or foci of (17), as is discussed in Section 7.4.1. The computer-generated phase portrait in Fig. 5 reveals that they are centers; we have centers at $x = 0, \pm 2\pi, \pm 4\pi, \ldots$, and saddles at $x = \pm\pi, \pm 3\pi, \ldots$, on the x axis.

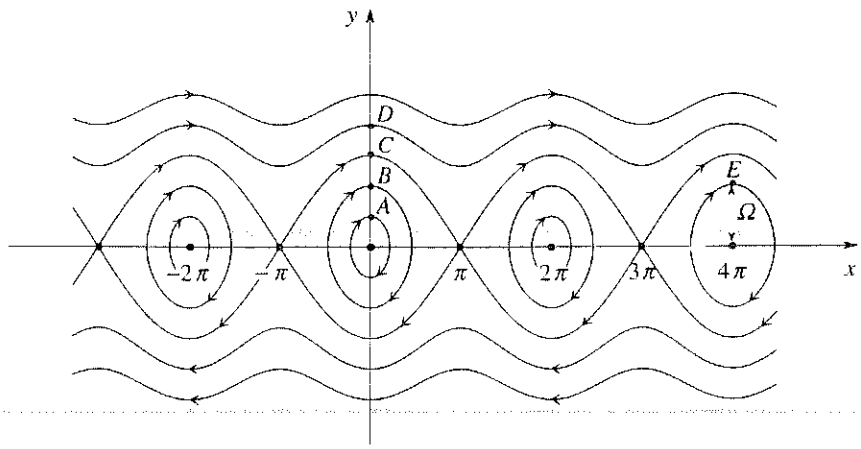


Figure 5. Phase portrait of the full nonlinear system $x'' + \sin x = 0$.

To understand the phase portrait, suppose (for definiteness) that the pendulum is hanging straight down ($x = 0$) initially, and that we impart an initial angular velocity $y(0)$, so that the initial point is A, B, C, or D in Fig. 5. If we start at A, then we follow a closed

orbit that is very close to elliptical, and the motion is very close to simple harmonic motion at frequency $\omega = 1$. If we start at B, the orbit is not so elliptical, there is an increase in the period, and the motion deviates somewhat from simple harmonic. If we start at C, then we approach the saddle at $x = \pi$ as $t \to \infty$; that is, the pendulum approaches the inverted position as $t \to \infty$. If we impart more energy by starting at D, then – even though it slows down as it approaches the inverted position – it has enough energy to pass through that position and to keep going round and round indefinitely. Though the trajectory in the phase plane is not closed, the motion is nonetheless physically periodic since the positions x and $x + 2n\pi$ (for any integer n) are physically identical.

How can we gain access to one of the other closed orbits such as E? That's easy: we "crank" the pendulum through two rotations by hand so that while hanging straight down it is now at $x = 4\pi$. Then we impart an initial angular velocity $y(0) = \Omega$.

What is the equation of the trajectories? Dividing (17b) by (17a) and integrating, gives

$$\frac{1}{2}y^2 - \cos x = \text{constant} \equiv C. \tag{19}$$

Do we really need (19)? After all, we turned the phase portrait generation over to the computer. Yes, to help us choose initial conditions that will enable us to generate the separatrix (the trajectories through the saddles) on the computer. With $x = \pi$ and $y = 0$, we find that $C = 1$, so the equation of the separatrix is $y^2 = 2(1 + \cos x)$. Be careful, because the initial point $x = 0$ and $y = 2$ will not generate the entire separatrix, but only the segment through that point, from $x = -\pi$ to $x = \pi$. To generate the next segment we could use an initial point $x = 2\pi$ and $y = 2$, and so on.

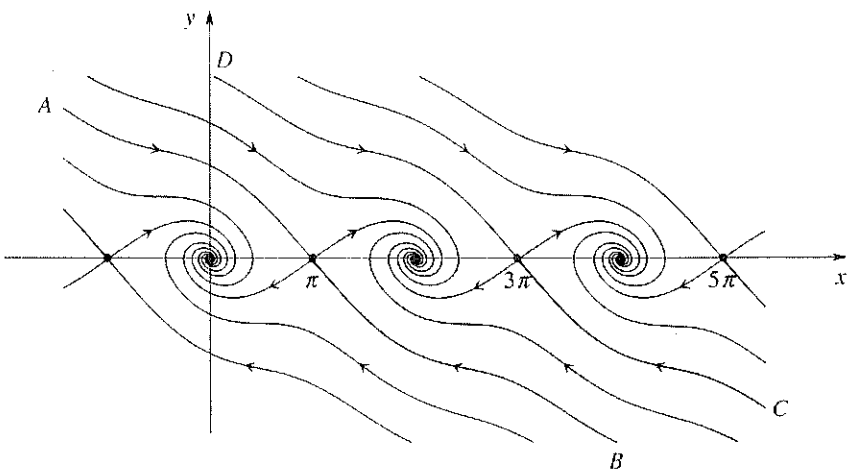


Figure 6. Phase portrait for the subcritically damped pendulum.

COMMENT 1. Recall that Fig. 3–5 correspond to taking $\sin x \approx x$, $\sin x \approx x - x^3/6$ and retaining $\sin x$ without approximation, respectively, in (13). Thus, and not surprisingly, as we retain more terms in the Taylor series approximation of $\sin x$ about $x = 0$ we capture the flow more accurately and completely.

COMMENT 2. What happens if we include some damping? It turns out that the singu-

larities are at $(n\pi, 0)$, as before. If $r < r_{cr}$, where $r_{cr} = 2$, then the singularities are still saddles if n is odd, but if n is even we now have stable foci rather than centers as seen in Fig. 6 (for $r = 0.5$). One calls the lightly shaded region (not including the boundaries AB and CD) the **basin of attraction** for the stable focus at $(2\pi, 0)$, the basin of attraction of an attracting singular point S being the set of all initial points P_0 such that the representative point $P(t)$ tends to S as $t \to \infty$ if $P(0) = P_0$. Similarly, each of the other stable foci has its own basin of attraction.

COMMENT 3. We have spoken, in this section, of a nonlinear system having the same type of singularity (or not), at a particular singular point S, as the system linearized about S. Let us use the present example to clarify that idea. By their singularities being of the same type, we mean that their phase portraits are *topologically equivalent* in the neighborhood of S. Intuitively, that means that one can be obtained from the other by a continuous deformation, with the direction of the arrows preserved. The situation is illustrated in Fig. 7, where we show both the nonlinear (solid) and linearized (dashed) portraits in the neighborhood of the saddle at $(\pi, 0)$. [In more mathematical terms, suppose that our system $x' = P(x, y)$ and $y' = Q(x, y)$ has a singular point at the origin and that

$$P(x, y) = ax + by + \text{higher-order terms} \equiv aX + bY, \tag{20a}$$

$$Q(x, y) = cx + dy + \text{higher-order terms} \equiv cX + dY \tag{20b}$$

define X and Y as continuous functions of x and y, and vice versa. Such a relationship between x, y and X, Y is called a **homeomorphism** and is what we mean by a continuous deformation.]

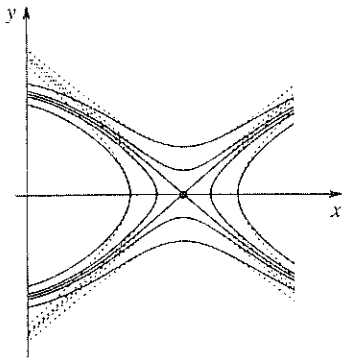


Figure 7. Continuous deformation of the portrait near $(\pi, 0)$. Solid lines correspond to the nonlinear system, dashed to the linearized version.

COMMENT 4. It turns out that the nonlinear pendulum equation is also prominent in connection with a superconducting device known as a *Josephson junction*. For discussion of the Josephson junction within the context of nonlinear dynamics, we recommend the book *Nonlinear Dynamics and Chaos* (Reading, MA: Addison–Wesley, 1994) by Steven H. Strogatz. ∎

In the preceding example we were unable to classify the singularities at $(n\pi, 0)$ with certainty, for n even and $r = 0$, as is discussed in the paragraph below (18). We relied on the computer-generated phase portrait, which show them to be centers, not foci. More satisfactorily, we could have used the fact that $x'' + \sin x = 0$ is a "conservative system," which idea we now explain.

In general, suppose that the application of Newton's second law gives

$$mx'' = F(x); \tag{21}$$

that is, where the force F happens to be an explicit function only of x, not of x' or t. Defining $V(x)$ by $F(x) = -V'(x)$, (21) becomes $mx'' + V'(x) = 0$. Let us multiply the latter by dx and integrate on x. Since, from the calculus,

$$x''dx = \frac{dx'}{dt}dx = \frac{dx'}{dt}\frac{dx}{dt}dt = x'\frac{dx'}{dt}dt = x'dx', \tag{22}$$

we obtain $mx'dx' + V'(x)dx = 0$, integration of which gives

$$\frac{1}{2}mx'^2 + V(x) = \text{constant}. \tag{23}$$

$V(x)$ is called the *potential energy* associated with the force $F(x)$. For a linear spring, for instance, the force is $F(x) = -kx$, and its potential is $V(x) = kx^2/2$. The upshot is that (23) tells us that the total energy (kinetic plus potential) is conserved, it remains constant over time, so we say that any system of the form (21) is **conservative**. The pendulum equation

$$ml^2x'' = -mgl\sin x \tag{24}$$

is of that form, and multiplying by dx and integrating on x gives

$$\frac{1}{2}m(lx')^2 - mgl\cos x = \text{constant}, \tag{25}$$

where $m(lx')^2/2$ is the kinetic energy and $-mgl\cos x$ is the potential energy associated with the gravitational force (with the pivot point chosen as the reference level).

For any conservative system (21), the total energy is $E(x, x') = mx'^2/2 + V(x)$. If we plot $E(x, x')$ above the x, x' phase plane, then the x, x' locations of maxima and minima of E are found from

$$\frac{\partial E}{\partial x} = V'(x) = 0 \qquad \text{and} \qquad \frac{\partial E}{\partial x'} = mx' = 0, \tag{26}$$

and these are precisely the singular points of the system

$$\begin{aligned} x' &= y, \\ y' &= F(x)/m = -V'(x)/m \end{aligned}$$

corresponding to (21). To illustrate, the point S beneath the minimum of E (Fig. 8) is a singular point. Furthermore, since E is constant on each trajectory, the phase trajectories are the projections of the closed curves of intersection of the E surface and the various horizontal planes, as sketched. Evidently (and we do not claim that this intuitive discussion constitutes a rigorous proof), a trajectory Γ very close to S must be a closed orbit. The only way Γ could fail to correspond to a periodic motion is if there is a singular point on Γ, for the flow would stop there. But if S is an isolated singular point, and Γ is small enough, then there can be no singular points on Γ, and we can conclude that S must be a center. By that reasoning, we could have known that the singularities at $(n\pi, 0)$ (for n even) must be centers, not foci. More generally, we state that *conservative systems do not have foci among their singularities.*

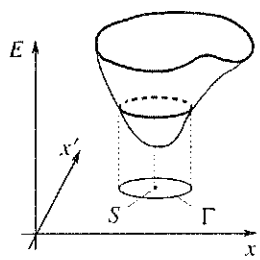


Figure 8. Occurrence of a center for a conservative system.

EXAMPLE 3. *Volterra's Predator-Prey Model.* The **Volterra model** (also known as the **Lotka–Volterra model**) of the ecological problem of two coexisting species, one the predator and the other its prey, is introduced in Section 7.2. Recall that if $x(t)$, $y(t)$ are the populations of prey and predator, respectively, then the governing equations are of the form

$$x' = \mu(1 - y)x, \tag{27a}$$

$$y' = -\nu(1 - x)y, \tag{27b}$$

where μ, ν are positive empirical constants. Setting the right-hand sides of (27) equal to zero reveals that there are two singular points: $(0,0)$ and $(1,1)$.

Linearizing (27) about $(0,0)$ gives

$$x' = \mu x, \qquad y' = -\nu y \tag{28}$$

so $a = \mu$, $b = c = 0$, $d = -\nu$. Hence, $\lambda = \mu$ and $-\nu$, which are of opposite sign, so the singularity at the origin is a saddle. Clearly, the straight-line trajectories through $(0,0)$ are simply the x and y axes since $x = 0$, $y = Ae^{-\nu t}$, and $x = Be^{\mu t}$, $y = 0$ satisfy (28) and give trajectories that pass through the origin.

To linearize about $(1,1)$ we use (7) and obtain the approximations

$$x' = -\mu(y-1), \qquad y' = \nu(x-1) \tag{29}$$

or, with $X \equiv x - 1$ and $Y \equiv y - 1$,

$$X' = -\mu Y = 0X - \mu Y, \tag{30a}$$
$$Y' = \nu X = \nu X + 0Y. \tag{30b}$$

Thus, $a = d = 0$, $b = -\mu$, $c = \nu$ so $\lambda = \pm 2i\sqrt{\mu\nu}$. Hence, the linearized version (30) has a center, and (27) has either a center or a focus.

The phase portrait in Fig. 9 shows the singularity at $(1,1)$ to be a center, with every trajectory being a periodic orbit, except for the two coordinate axes. (Of course, we show only the first quadrant because x and y are populations and hence nonnegative.) The direction of the arrows follows from (27), which reveals that $x' > 0$ for $y < 1$, and $x' < 0$ for $y > 1$ (or, $y' < 0$ for $x < 1$ and $y' > 0$ for $x > 1$).

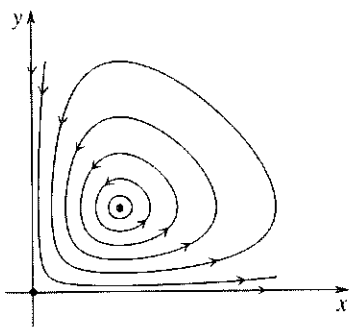


Figure 9. Phase portrait for Volterra problem (27).

COMMENT. Although the Volterra model is a useful starting point in the modeling process and is useful pedagogically, it is not regarded as sufficiently realistic for practical ecological applications. ∎

7.4.3. Bifurcations. As we have stressed, our approach in this chapter is largely qualitative. Of special importance, then, is the concept of bifurcations. That is, systems generally include one or more physical parameters (such as μ and ν in Example 3). As those parameters are varied continuously, one expects the system behavior to change continuously as well. For instance, if we vary μ in Example 3, then the eccentricity of the orbits close to the center at $(1,1)$ changes, and the overall flow field deforms, but – qualitatively – nothing dramatic happens. In other cases, there may exist certain critical values of one or more of the parameters such that the overall system behavior changes abruptly and dramatically as a parameter passes through such a critical value. We speak of such a result as a **bifurcation**. Let us illustrate the idea with an example.

EXAMPLE 4. *Saddle-Node Bifurcation.* The nonlinear system

$$x' = -rx + y, \tag{31a}$$
$$y' = \frac{x^2}{1+x^2} - y \tag{31b}$$

arises in molecular biology, where $x(t)$ and $y(t)$ are proportional to protein and messenger RNA concentrations, and r is a positive empirical constant, or parameter, associated with the "death rate" of protein in the absence of the messenger RNA [for if $y = 0$, then (31a) gives exponential decay of x, with rate constant r].

The singular points of (31) correspond to intersection points of $y = rx$ and $y = x^2/(1+x^2)$, as shown (solid curves) in Fig. 10. Equating these gives $x = y = 0$ and also the two distinct roots

$$x_\pm = \frac{1 \pm \sqrt{1-4r^2}}{2r}, \qquad y_\pm = rx_\pm = \frac{1 \pm \sqrt{1-4r^2}}{2}, \tag{32}$$

provided that $r < 1/2$. Thus, the critical slope of $y = rx$ is $r = 1/2$. If $r < 1/2$ we obtain the two intersections $S_+ \equiv (x_+, y_+)$ and $S_- \equiv (x_-, y_-)$ if $r = 1/2$ (dashed line in Fig. 10) these coalesce at $(1, 0.5)$, and if $r > 1/2$ they disappear and we have only the singular point at the origin.

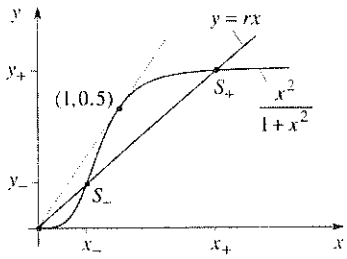


Figure 10. Determining the singular points of (31).

Let us study the three singular points, for $r < 1/2$. First $(0,0)$: we can see from (31) by inspection or Taylor series expansion, that the linearized equations are

$$x' = -rx + y, \qquad y' = -y \tag{33}$$

so $a = -r$, $b = 1$, $c = 0$, and $d = -1$. Thus, (10) gives $\lambda = -r$ and -1. Since both are negative, the singular point $(0,0)$ is a stable node.

In similar fashion (which calculations we leave to Exercise 6), we find that the singularity at S_- is a saddle, and that the singularity at S_+ is an unstable improper node. As r is increased, S_- and at S_+ approach each other along the curve $y = x^2/(1+x^2)$. When $r = 1/2$ they merge and form a singularity of some other type, and when r is increased beyond $1/2$ the singularity disappears altogether, leaving only the node at the origin. The bifurcation that occurs at $r = 1/2$ is an example of a "saddle-node bifurcation." From the way the singular points S_+ and S_- approach each other along the unstable manifold of the saddle, like "beads on a string" as Strogatz puts it, we see that the bifurcation process is essentially a one-dimensional event embedded within a higher-dimensional space (two-dimensional in this case). ∎

The saddle-node bifurcation illustrated above is but one type of bifurcation. A few others are discussed in the exercises and in the next section. For a more complete discussion of bifurcation theory, we recommend the book by Strogatz, referenced in Example 2.

Closure. In this section we got into the details of the phase plane analysis of autonomous nonlinear systems. Whether or not we generate the phase portrait by computer, it is essential to begin an analysis by finding any singular points and, by linearization, to determine the key features of the local flow near each singular point. That information is needed even if we turn to the computer to generate the phase portrait, as we discuss below, under "Computer software."

We also explored the correspondence between the type of a singularity of the nonlinear system and that of the linearized system and found that the type remains the same, except for the borderline cases corresponding to p, q points on the positive q axis or on the parabola $p^2 = 4q$ in Fig. 1. Those cases could "go either way." That

is, the higher-order terms in the nonlinear system can be thought of as perturbing the a, b, c, d coefficients in the linearized equations and hence p, q as well. The perturbed p, q point could remain on the boundary curve (positive q axis or the parabola $p^2 = 4q$), or it could be pushed slightly into either of the adjoining regions, thus changing the singularity type.

The other item of special importance is the notion of bifurcations, whereby dramatic and qualitative changes in the system behavior can result from a given parameter crossing a critical value known as a bifurcation point.

Computer software. Consider, for instance, the computer generation of the trajectories through the singular point $(\pi, 0)$ in Fig. 6. The idea is to express the governing equation $x'' + 0.5x' + \sin x = 0$ (we took $r = 0.5$ in Fig. 6) as the system

$$\begin{aligned} x' &= y, \\ y' &= -\sin x - 0.5y \end{aligned} \tag{34}$$

and to linearize (34) about $(\pi, 0)$ as

$$\begin{aligned} x' &= y, \\ y' &= -(\cos \pi)(x - \pi) - 0.5y \end{aligned} \tag{35}$$

or with $x - \pi \equiv X$ and $y - 0 \equiv Y$,

$$\begin{aligned} X' &= 0X + 1Y, \\ Y' &= 1X - 0.5Y. \end{aligned} \tag{36}$$

Next, seek straight-line solutions as $Y = \kappa X$, so (36) becomes

$$\begin{aligned} X' &= 0X + \kappa X, \\ \kappa X' &= 1X - 0.5\kappa X, \end{aligned} \tag{37}$$

comparison of which gives $0 + \kappa = (1 - 0.5\kappa)/\kappa$ and hence $\kappa = -1.2807764$ and 0.7807764. Thus, if we make very small steps along the line $Y = -1.2807764X$, away from $(\pi, 0)$, we can obtain initial points that are very close to being on the trajectories from A to $(\pi, 0)$ and from B to $(\pi, 0)$. For example, with $X = -0.02$ we obtain $Y = 0.02561553$. Similarly, $X = +0.02$ gives $x = 3.1615927$ and $y = -0.02561553$.

With these two initial points we can use the *Maple* phaseportrait command to obtain the desired trajectories from A to $(\pi, 0)$ and from B to $(\pi, 0)$, as follows. Remembering to first enter with(DEtools):, the phaseportrait command would be as follows:

```
phaseportrait([y, -.5 * y - sin (x)], [t, x, y], t = -10..10,
{[0, 3.1215927, .02561553], [0, 3.1615927, -.02561553]}, stepsize = .05,
x = -4.5..17, y = -6..6, scene = [x, y]);
```

The t range, from -10 to $+10$, is found to be sufficient to give the desired trajectories all the way to the borders $y = -6$ and $y = +6$, and the x and y ranges are simply chosen as the same ones used in Fig. 6.

EXERCISES 7.4

1. (a) In Example 1 we stated that the equation $x'' + \epsilon x'^3 + x = 0$ ($\epsilon > 0$) has a stable focus at the origin of the phase plane. Verify that claim by generating (with computer software) its phase portrait.
(b) Should the cubic damping result in the oscillation dying out more or less rapidly than for the case of linear damping, $x'' + \epsilon x' + x = 0$, for the same values of ϵ? Explain.
(c) Classify the singularity at $(0, 0)$ for the case where $\epsilon < 0$, and support your claim.

2. Determine all singular points, if any, and classify each insofar as possible.

(a) $x' = y, \quad y' = 1 - x^4$
(b) $x' = 1 - y^2, \quad y' = 1 - x$
(c) $x' = y, \quad y' = (1 - x^2)/(1 + x^2)$
(d) $x' = x - y, \quad y' = \sin(x + y)$
(e) $x' = (1 - x^2)y, \quad y' = -x - 2y$
(f) $x' = (1 - x^2)y, \quad y' = -x + 2y$
(g) $x' = -2x - y, \quad y' = x + x^3$
(h) $x' = -2x - y, \quad y' = \sin x$
(i) $x' = ye^x - 1, \quad y' = y - x - 1$
(j) $x' = x^2 - 2y, \quad y' = 2x - y$
(k) $x' = x^2 - y^2, \quad y' = x^2 + y - 2$
(l) $x' = y, \quad y' = -3\sin x$
(m) $x' = x + 2y, \quad y' = -x - \sin y$
(n) $x' = (x^6 + 1)y, \quad y' = x^2 - 4$

3. (a)–(n) Use computer software such as the *Maple* phaseportrait command, to generate the phase portrait of the corresponding system in Exercise 2.

4. Is the given system conservative? Explain.

(a) $x'' - 2x' + \sin x = 0$ (b) $x'' + x'^2 + x^2 = 0$
(c) $x'' + x^2 = 0$ (d) $x'' + x' + x = 0$

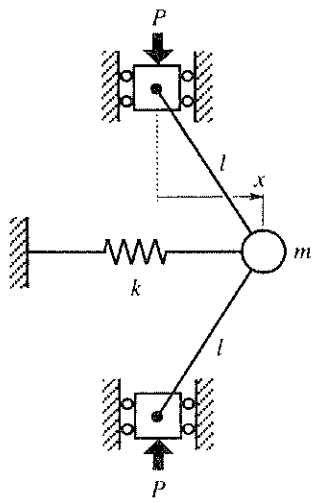
5. Use computer software such as the *Maple* phaseportrait command, to obtain the phase portrait of equation (13), over $-2 < x < 14$ and $-3 < y < 3$, showing enough trajectories to clearly portray the flow. Take $g/l = 1$, and

(a) $r = 0$.
(b) r to be any positive value that you wish but small enough for the damping to be subcritical. For instance, the singularity at $(0, 0)$ should be a stable focus.
(c) r to be any value that you wish but large enough for the damping to be supercritical. For instance, the singularity at $(0, 0)$ should be a stable node.

6. (*Example 4 continuation.*) In parts (a) and (b) below, let $r = 0.3$ in (31).

(a) Show that S_- is a saddle, and find the equations of the two straight-line trajectories through it. Show that S_+ is a stable improper node, and find two straight-line trajectories through it. Find two straight-line trajectories through the stable node at $(0, 0)$. Use these results to sketch the phase portrait of (31).
(b) For the critical case, $r = 1/2$, show that the singularity at $S_\pm$ is nonelementary.
(c) For the representative supercritical case $r = 1$, identify and classify any singularities, and use computer software to generate the phaseportrait of (31) in the rectangle $0 \le x \le 1.5$, $0 \le y \le 1.5$.

7. (*Dynamic formulation of a buckling problem*) Consider the buckling of the mechanical system shown in the figure, and



consisting of two massless rigid rods of length l pinned to a mass m and a lateral spring of stiffness k. That is, when the spring is neither stretched nor compressed $x = 0$ and the rods

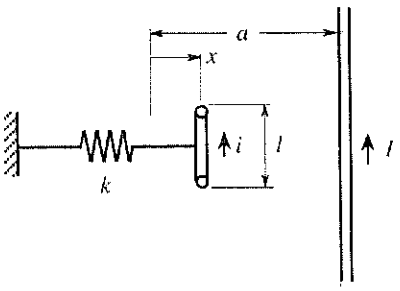
are aligned vertically. As we increase the downward load P nothing happens until we reach a critical value P_{cr}, at which value x increases (to one side or the other, we can't predict which) and the system collapses.

(a) Application of Newton's second law of motion gives

$$mx'' - \frac{2Px}{l}\left[1 - \left(\frac{x}{l}\right)^2\right]^{-1/2} + kx = 0 \qquad (7.1)$$

as governing the displacement $x(t)$. With $x' = y$, show that the singularity at the origin in the x, y phase plane changes its type as P is sufficiently increased. Discuss that change of type, show how it corresponds to the onset of buckling, and use it to show that the critical buckling load is $P_{cr} = kl/2$.
(b) Explain what the results of part (a) have to do with bifurcation theory.
(c) Use Newton's law to derive (7.1).

8. (*Motion of current-carrying wire*) A mutual force of attraction is exerted between parallel current-carrying wires. The infinite wire shown in the figure has current I, and the wire



of length l and mass m (with leads that are perpendicular to the paper) has current i in the same direction as I. According to the Biot-Savart law, the mutual force of attraction is $2Iil/(\text{separation}) = 2Iil/(a-x)$, where $x = 0$ is the position at which the spring force is zero, so the equation of motion of the restrained wire is

$$mx'' + k\left(x - \frac{r}{a-x}\right) = 0, \quad \text{where} \quad r = \frac{2Iil}{k}. \qquad (8.1)$$

Thinking of m, k, a, and l as fixed, and the currents I and i as variable, let us study the behavior of the system in terms of the parameter r. For definiteness, let $m = k = a = 1$.

(a) With $x' = y$, identify any singularities in the x, y phase plane and their types, and show that they depend upon whether r is less than, equal to, or greater than $1/4$. Suppose that $r < 1/4$. Find the equation of the phase trajectories and of the separatrix. Do a labeled sketch of the phase portrait.
(b) Let $r = 0.1$, say, and obtain a computer plot of the phase portrait.
(c) Next, consider the transitional case, where $r = 1/4$. Show that that case corresponds to the merging of the two singularities, and the forming of a single singularity of higher order (i.e., a nonelementary singularity). Do a labeled sketch of the phase portrait for that case.
(d) Let $r = 1/4$, and obtain a computer plot of the phase portrait.
(e) Next, consider the case where $r > 1/4$, and sketch the phase portrait.
(f) Let $r = 0.5$, say, and obtain a computer plot of the phase portrait.
(g) Discuss this problem from the point of view of bifurcations, insofar as the parameter r is concerned.

7.5 Limit Cycles, van der Pol Equation, and the Nerve Impulse

7.5.1. Limit cycles and the van der Pol equation. The **van der Pol equation**,

$$\boxed{x'' - \epsilon(1 - x^2)x' + x = 0,} \qquad (\epsilon > 0) \qquad (1)$$

was studied by *Balth van der Pol* (1889–1959), first in connection with current oscillations in a certain vacuum tube circuit and then in connection with the modeling

of the beating of the heart.* Usually, in applications the parameter ϵ is positive.

To study (1) in the phase plane we first re-express it as the system

$$x' = y, \tag{2a}$$
$$y' = -x + \epsilon(1 - x^2)y, \tag{2b}$$

which has one singular point: $(0, 0)$. Linearizing (2) about $(0, 0)$ gives

$$x' = y, \tag{3a}$$
$$y' = -x + \epsilon y, \tag{3b}$$

which has an unstable focus if $\epsilon < 2$ and an unstable node if $\epsilon > 2$. That result is not surprising since (3) is equivalent to $x'' - \epsilon x' + x = 0$ [equation (1) with the nonlinear $\epsilon x^2 x'$ term dropped], and the latter corresponds to a damped harmonic oscillator with *negative* damping. Near the origin in the x, y phase plane the flow is accurately described by (3) and is shown in Fig. 1a. As the motion increases, the neglected nonlinear term $\epsilon x^2 x'$ ceases to be negligible, and we wonder how the trajectory shown in Fig. 1a continues to develop as t increases. Since the "damping coefficient" $c = -\epsilon(1 - x^2)$, in (1), is negative throughout the vertical strip $|x| < 1$, we expect the spiral to continue to grow, with distortion as the $\epsilon x^2 x'$ term becomes more prominent. Eventually, the spiral will break out of the $|x| < 1$ strip (Fig. 1b). As the representative point $(x(t), y(t))$ spends more and more time outside that strip, where $c = -\epsilon(1 - x^2) > 0$, the effect of the positive damping in $|x| > 1$ increases, relative to the effect of the negative damping in $|x| < 1$, so it is natural to wonder if the trajectory might approach some limiting closed orbit as $t \to \infty$ over which the effects of the positive and negative damping are exactly in balance.

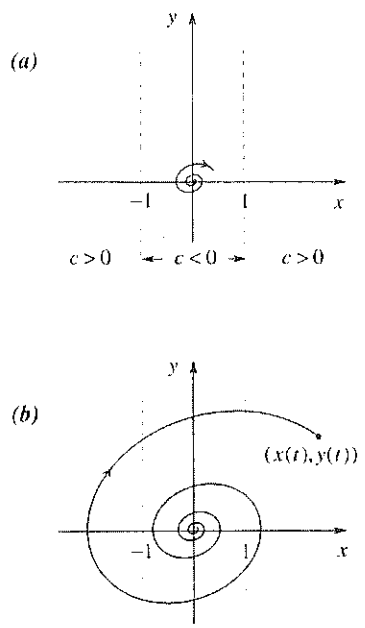


Figure 1. The unstable focus at$(0, 0)$.

We can use the following theorem, due to N. Levinson and O. K. Smith.

THEOREM 7.5.1 *Existence of Limit Cycle*
Let $f(x)$ be even $[f(-x) = f(x)]$ and continuous for all x. Let $g(x)$ be odd $[g(-x) = -g(x)]$ with $g(x) > 0$ for all $x > 0$, and $g'(x)$ be continuous for all x. With

$$\int_0^x f(\xi)\, d\xi \equiv F(x) \quad \text{and} \quad \int_0^x g(\xi)\, d\xi \equiv G(x), \tag{4}$$

suppose that (i) $G(x) \to \infty$ as $x \to \infty$ and (ii) there is an $x_0 > 0$ such that $F(x) < 0$ for $0 < x < x_0$, $F(x) > 0$ for $x > x_0$, and $F(x)$ is monotonically increasing for $x > x_0$ with $F(x) \to \infty$ as $x \to \infty$. Then the *generalized Liénard equation*

$$x'' + f(x)x' + g(x) = 0 \tag{5}$$

has a single periodic solution, the trajectory of which is a closed curve encircling the origin in the x, x' phase plane. All other trajectories (except the trajectory

---

*B. van der Pol, On "Relaxation Oscillations," *Philosophical Magazine*, Vol. 2, 1926, pp. 978–992, and B. van der Pol and J. van der Mark, The Heartbeat As a Relaxation Oscillation, and An Electrical Model of the Heart, *Philosophical Magazine*, Vol.6, 1928, pp. 763–775.

consisting of the single point at the origin) spiral toward the closed trajectory as $t \to \infty$.

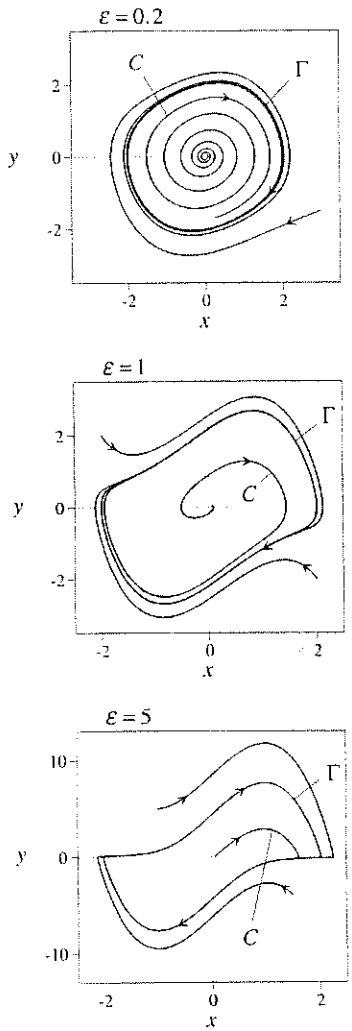


Figure 2. The van der Pol limit cycle, for $\epsilon = 0.2, 1$, and 5.

Applying this theorem to the van der Pol equation (1), $f(x) = -\epsilon(1 - x^2)$ is an even function of x and $F(x) = -\epsilon(x - x^3/3)$, which is less than zero for $0 < x < \sqrt{3}$, greater than zero for $x > \sqrt{3}$, and which increases monotonely to infinity as $x \to \infty$. Further, $g(x) = x$ is odd, and positive for $x > 0$, $g'(x) = 1$, and $G(x) = x^2/2 \to \infty$ as $x \to \infty$. Since the conditions of the theorem are met (for all $\epsilon > 0$), we conclude from the theorem that the van der Pol equation does admit a closed trajectory, a periodic solution, for every positive value of ϵ.

Computer results (using the *Maple* phaseportrait command) bear out this claim. The phase portraits are shown in Fig. 2 for the representative cases $\epsilon = 0.2, 1$, and 5, and $x(t)$ is plotted versus t in Fig. 3 for the trajectories labeled C. The closed trajectories labeled Γ, predicted by the theorem, are examples of **limit cycles** – namely, isolated closed orbits. By Γ being isolated we mean that neighboring trajectories through points arbitrarily close to Γ are not closed orbits. If we start on Γ we remain on Γ, but if we start on a neighboring trajectory, then we approach Γ as $t \to \infty$ (unless we start at the origin, which is an equilibrium point). Thus, we classify the van der Pol limit cycle as **stable** (or **attracting**). Clearly, that particular trajectory is of the greatest importance because every other trajectory (except the point trajectory $x = y = 0$) winds onto it as $t \to \infty$.

As one might suspect from Fig. 2 and as can be proved, the van der Pol limit cycle approaches a circle of radius 2 as $\epsilon \to 0$ through positive values. When ϵ becomes zero the singularity at the origin changes from a focus to a center, and while the circle of radius 2 persists as a trajectory it is joined by the whole family of circular orbits centered at the origin. If ϵ is diminished further and becomes negative, the origin becomes a stable focus and all closed orbits disappear and give way to inward-winding spirals. Thus, $\epsilon = 0$ is a bifurcation value of ϵ.

Observe the interesting extremes: as $\epsilon \to 0$, the steady-state oscillation (i.e., corresponding to the limit cycle) becomes a purely harmonic motion with amplitude 2. But as ϵ becomes large, the limit cycle distorts considerably and the steady-state oscillation $x(t)$ becomes "herky jerky." (In the exercises, we show that as $\epsilon \to \infty$ it even becomes discontinuous!) Such motions were dubbed as **relaxation oscillations** by van der Pol, and these are found all around us. Just a few, mentioned in the paper by van der Pol and van der Mark, are the singing of wires in a cross wind, the scratching noise of a knife on a plate, the squeaking of a door, the intermittent discharge of a capacitor through a neon tube, the periodic reoccurrence of epidemics and economic crises, the sleeping of flowers, menstruation, and the beating of the heart. Such oscillations are characterized by a slow buildup followed by a rapid discharge, then a slow buildup, and so on. Thus, there are two time scales present, a "slow time" associated with the slow buildup, and a "fast time" associated with the rapid discharge. In biological oscillators such as the heart, the period of oscillation provides a biological "clock."

Understand clearly that the *limit cycle phenomenon is possible only in nonlinear systems* for consider the case of small ϵ, say, where we have a limit cycle that is

approximately a circle of radius 2. If the system were linear, then the existence of that orbit would imply that the entire family of concentric circles would necessarily be trajectories as well, but they are not.

Besides the van der Pol example, other examples of differential equations exhibiting limit cycles are given in the exercises. In other cases a limit cycle can be **unstable (repelling)** in that other trajectories wind away from it, or **semistable** in exceptional cases, in that other trajectories wind toward it from the interior and away from it on the exterior, or vice versa.

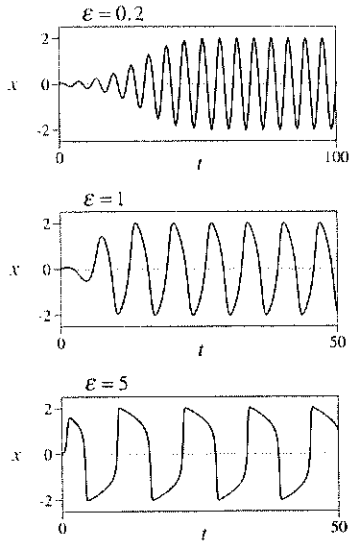


Figure 3. $x(t)$ versus t.

7.5.2. Application to the nerve impulse and visual perception. The brain contains about 10^{12} (a million million) neurons, with around 10^{14} to 10^{15} interconnections. Within this complex network, information is encoded and transmitted in the form of electrical impulses. The basic building block is the individual neuron, or nerve cell, and the functioning of a single neuron as an input/output device is of deep importance and interest. Our purposes in discussing the neuron here are in connection with relaxation oscillations, and especially with the key role of nonlinearity in the design and functioning of our central nervous system.

A typical neuron is comprised of a cell body that contains the nucleus and that emanates many dendrites and a single axon. The axon splits near its end into a number of teminals as shown schematically in Fig. 4. Dendrites are on the order of a millimeter long, and axons can be as short as that or as long as a meter. At the end of each terminal is a synapse, which is separated from a dendrite of an adjacent cell by a tiny synaptic gap. Electrical impulses, each of which is called an **action potential**, are generated near the cell body and travel down the axon. When an action potential arrives at a synapse, chemical signals in the form of neurotransmitter molecules are released into the synaptic gap and diffuse across that gap to a neighboring dendrite. These electrical signals to that neighboring neuron can be positive (excitatory) or negative (inhibitory). Each cell receives a great many such cues from other neurons. If the net excitation to a cell falls below some critical threshold value, then the cell will not fire – that is, it will not generate action potentials. If the net excitation is a bit above that threshold, then the cell will fire – not just once, but repeatedly and at a certain frequency.

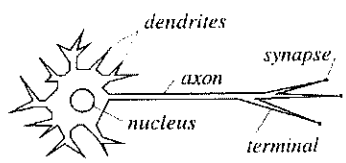


Figure 4. Typical neuron.

Let us consider briefly the generation of the nerve impulse. The nerve cell is surrounded by and also contains salt water. The salt molecules include sodium chloride (NaCl) and potassium chloride (KCl), and many of these molecules are ionized so that Na^+, K^+, and Cl^- ions are abundant both inside and outside the axon. Of these, Na^+ and K^+ are the key players insofar as the nerve impulse is concerned. Rather than being impermeable, the axon membrane has many tubular protein pores of two kinds: channels that can open or close and let either Na^+ or K^+ ions through in a passive manner, like valves, and pumps that (using energy from the metabolism of glucose and oxygen) actively eject Na^+ ions (i.e., from inside the axon to outside) and bring in K^+ ions. Through the action of these active and passive pores, and the physical mechanisms of diffusion and the repulsion/attraction of like/unlike charges, a differential in charge, and hence potential (voltage), is established across the axon membrane which, in the resting state, is 70

millivolts, positive outside.

If the net excitation arriving from other cells sufficiently reduces that voltage, at the cell body end of the axon, then a sequence of opening and closing of pores is established, which results in a flow of Na^+ and K^+ ions and hence a voltage "blip," the action potential, proceeding down the axon. That wave is not like the flow of electrons in a copper wire, but rather like a water wave that results not from horizontal motion of the water, but from a differential up and down motion of water particles. This complicated process was pieced together by Alan Hodgkin and Andrew Huxley, in 1952, and is clearly discussed in the book by David H. Hubel.* Various electrical circuit analogs have been proposed, to model the nerve impluse, by Hodgkin and Huxley and others. They are all somewhat empirical and of the "fill and flush" type – that is, where a charge builds up and is then discharged through an electrical circuit, and they are described in the little book by F. C. Hoppensteadt.†

Of interest to us here is that the firing is repetitive (on the order of 100 impulses per second), and consists of a relaxation oscillation governed by the van der Pol equation (as discussed in Hoppensteadt). Further, it is known that as the excitation voltage is increased above the threshold, the magnitude of the action potential remains unchanged, but the firing frequency increases. If we plot the output (action potential) amplitude versus the input (excitation voltage) amplitude, the graph is as shown in Fig. 5. Since the graph of output amplitude versus input amplitude is not a straight line through the origin, the process must be nonlinear, which fact is also known through the governing equation being a van der Pol equation (or other such equation, depending upon the model adopted); indeed, any process where the output amplitude is zero until a critical threshold is reached is necessarily nonlinear. Since the individual neuron is a nonlinear device, surely the same is true for the entire central nervous system, and the natural and important question that asserts itself is "Why?". What is the functional purpose of that nonlinearity?

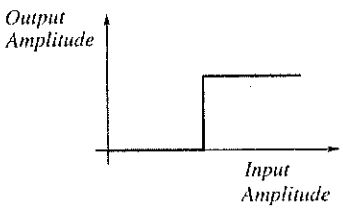


Figure 5. Input/output relation for a neuron.

Let us attempt an answer. We have seen that nonlinear systems are more complex than linear ones. Since our nervous system is responsible for carrying out complex tasks, it seems reasonable that the system chosen should be nonlinear. We can be more specific if we look at a single type of task, say visual perception, which is so complex that it occupies around a third of the million million neurons in the brain.

Perhaps the most striking revelation in studying visual perception is in discovering that one's visual perception is not a simple replica of the image that falls upon the retina but is an interpretation of that information, effected by visual processing that begins in the retina and continues up into the visual cortex of the brain. For instance, hold your two hands up, in front of your face, with one twice as far from your eyes as the other (about 8 and 16 inches). You should find that they look about the same size. Yet, if we replace your eyes with a camera, and take a picture, we see in the photo that one hand looks around twice as large as the other. Usually, we blame the camera for the "distortion," but the camera simply shows you the same

* *Eye, Brain, and Vision* (New York: W. H. Freeman and Company, 1988).

† *An Introduction to the Mathematics of Neurons* (Cambridge: Cambridge University Press, 1986).

information that is picked up by your retinas, the distortion is introduced by the brain as it interprets and reconstructs the data before presenting it to you as visual consciousness.

The latter is but one example of a principle of visual perception known as **size constancy**. The idea is that whereas the actual size of a physical object is invariant, the size of its retinal image varies dramatically as it is moved nearer or further from us. Size constancy is the processing, between the retina and visual consciousness, that compensates for such variation so as to stabilize our visual world. Thus, for instance, our hands look about the same size even when the retinal image of one is twice as large as that of the other. The functional advantage of that stabilization is to relieve our conscious mind of having to figure everything out; our brain does much of the figuring out and presents us with its findings so that our conscious attention can be directed to more pressing and singular matters.

Surely, size constancy requires a nonlinear perceptual system for if we take the retinal image size as the input amplitude and the perceived size as the output amplitude, then a linear system would show us the two hands just as a camera does. (Remember that for a linear system if we double the input we double the corresponding output if we triple the input we triple the output, and so on.)

In visual perception there are other constancy mechanisms as well, such as **brightness constancy** and **hue constancy**. To illustrate brightness constancy, consider the following simple experiment reported by Hubel in his book, cited above. We know from experience that a newspaper appears pretty much the same, whether we look at it in sunlight or in a dimly lit room: black print on white paper. Taking a newspaper and a light meter, Hubel found that the white paper reflected 20 times as much light outdoors as indoors, yet it looked about the same outdoors and indoors. If the perceptual system were linear, the white paper should have *looked* 20 times as bright outdoors compared to indoors. Even more striking, he found that the black letters actually reflected twice as much light outdoors as the white paper did indoors yet, whether indoors or outdoors, the black letters always looked black and white paper always looked white.

The point, then, is that these constancy mechanisms stabilize our perceived world and relieve our conscious mind from having to deal with newspapers that look 20 times brighter outdoors than indoors, hands that "grow" and "shrink" as they are moved to and fro, and so on, so that our conscious attention can be reserved for more singular matters such as not getting hit by a bus. These mechanisms are possible only by virtue of the nonlinearity of the central nervous system, which can be traced, in turn, to the highly nonlinear behavior of the basic building block of that system, the individual neuron.

You have no doubt heard about "the whole being greater than the sum of its parts." That idea expresses the essence of the Gestalt school of psychology which, by around 1920, supplanted the previously dominant molecular school of psychology, which had held that the whole *is* equal to the sum of the parts. To illustrate the Gestalt view, notice that the black dots in Fig. 6a are seen as a *group* of dots, not as a number of individual dots, and that the arrangement in Fig. 6b is seen as a triangle with sections removed, rather than as three bent lines. In fact, Max Wertheimer's

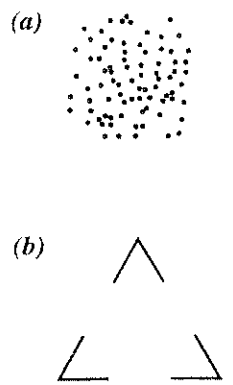


Figure 6. The whole is greater than the sum of its parts.

fundamental experiment, which launched the Gestalt concept in 1912, is as follows. If parallel lines of equal length are displayed on a screen successively, it is found that if the time interval and distance between them are sufficiently small, then they are perceived not as two separate lines but as a single line that *moves* laterally. (Today we recognize that idea as the basis of motion pictures!)

In mathematical terms, the molecular idea is reminiscent of the result for a linear differential equation $L[u] = f_1 + f_2 + \cdots + f_k$ that the response u to the combined input is simply the sum of the responses $u_1, u_2, \ldots, u_k$ to the individual inputs $f_1, f_2, \ldots, f_k$. We suggest here that, in effect, the contribution of the Gestaltists was to recognize the highly nonlinear nature of the perceptual system, even if they did not think of it or express it in those terms. We say more about the far-reaching effects of that nonlinearity upon human behavior in the next section.

Closure. The principal idea of this section is that of limit cycles, which occur only for nonlinear systems. The classic example of an equation with a limit cycle solution is the van der Pol equation, which we discuss, but that is by no means the only equation that exhibits a limit cycle. That limit cycle solution is said to be a **self-excited oscillation** because even the slightest disturbance from the equilibrium point at the origin results in an oscillation that grows and inevitably approaches the limit cycle as $t \to \infty$. The case of large ϵ is especially important in biological applications, and the corresponding limit cycle solution is a relaxation oscillation characterized by alternate t-intervals of slow and rapid change. Since the existence of a limit cycle is of great importance, there are numerous theorems available that help one to detect whether or not a limit cycle is present, but we include only the theorem of Levinson and Smith since it is helpful in our discussion of the van der Pol equation.

Finally, we discuss the action potential occuring during the firing of a neuron, as a biological illustration of a relaxation oscillation, and we use that example to point out the nonlinear nature of the neuron and central nervous system, and the profound implications of that nonlinearity.

EXERCISES 7.5

1. (a) Obtain computer results analogous to those presented in Fig. 2 and 3 for the case where $\epsilon = 0.1$. What value do you think the period approaches as $\epsilon \to 0$? Explain.
(b) Obtain computer results analogous to those presented in Fig. 2 and 3, for $\epsilon = 10$. NOTE: Be sure to make your t-integration step size small enough – namely, small compared to the time intervals of rapid change of $x(t)$.

2. Use Theorem 7.5.1 to show that the following equations admit limit cycles.

(a) $x'' - (1 - x^2)x' + x^3 = 0$
(b) $x'' + x^2(5x^2 - 3)x' + x = 0$

3. Identify and classify any singularities of the given equation in the x, y phase plane, where $x' = y$. Argue as convincingly as you can for or against the existence of any limit cycles, their shape, and their stability or instability. You should be able to tell a great deal from the equation itself, even in advance of any computer simulation.

(a) $x'' + (x^2 + x'^2 - 1)x' + x = 0$
(b) $x'' + (1 - x^2 - x'^2)x' + x = 0$

4. (*Hopf bifurcation*) (a) Show that the nonlinear system

$$x' = \epsilon x + y - x(x^2 + y^2), \tag{4.1a}$$

$$y' = -x + \epsilon y - y(x^2 + y^2) \tag{4.1b}$$

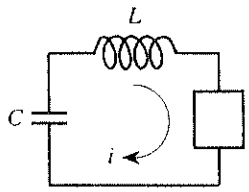
can be simplified to

$$r' = r(\epsilon - r^2), \tag{4.2a}$$

$$\theta' = -1 \tag{4.2b}$$

by the change of variables $x = r\cos\theta$, $y = r\sin\theta$ from the Cartesian x, y variables to the polar r, θ variables. HINT: Putting $x = r\cos\theta$, $y = r\sin\theta$ into (4.1) gives differential equations each of which contains both r' and θ' on the left-hand side. Suitable linear combinations of those equations give (4.2a,b), respectively. We suggest that you use the shorthand $\cos\theta \equiv c$ and $\sin\theta \equiv s$ for brevity.
(b) From (4.2) show that the origin in the x, y plane is a stable focus if $\epsilon < 0$ and an unstable focus if $\epsilon > 0$, and show that working from (4.1), instead, one obtains the same classification.
(c) Show from (4.2) that $r(t) = \sqrt{\epsilon}$ is a trajectory (if $\epsilon > 0$) and, in fact, a stable limit cycle. NOTE: Observe that zero is a bifurcation value of ϵ. As ϵ increases, a limit cycle is born as ϵ passes through the value zero, and its radius increases with ϵ. This is known as a **Hopf bifurcation**.
(d) Modify (4.1) so that it gives an *un*stable limit cycle at $r = \sqrt{\epsilon}$, instead.

5. The box in the circuit shown in the figure represents an "active element" such as a semiconductor or vacuum tube, the voltage drop across which is a known function $f(i)$ of the current i. Thus, Kirchhoff's voltage law gives $L\dfrac{di}{dt} + f(i) + \dfrac{1}{C}\displaystyle\int i\,dt = 0$.



(a) If f is of the form $f(i) = ai^3 - bi$, show that one obtains

$$Li'' + (3ai^2 - b)i' + \frac{1}{C}i = 0. \tag{5.1}$$

(b) Show that by a suitable scaling of both the independent and dependent variables one can obtain from (5.1) the van der Pol equation

$$I'' - \epsilon(1 - I^2)I' + I = 0, \tag{5.2}$$

where primes denote differentiation with respect to the new time variable τ, where $t = \alpha\tau$ and $i = \beta I$. That is, find α, β, and ϵ in terms of L, C, a, and b.

6. (*Rayleigh's equation and van der Pol relaxation oscillation*) (a) Show that if we set $x = z'$ in the van der Pol equation $x'' - \epsilon(1 - x^2)x' + x = 0$ and integrate once, we obtain $z'' - \epsilon(z' - z'^3/3) + z = C$, where C is a constant. Setting $z = u + C$, to absorb the C, obtain **Rayleigh's equation**

$$u'' - \epsilon\left(1 - \frac{u'^2}{3}\right)u' + u = 0 \tag{6.1}$$

on $u(t)$. The latter was studied by *Lord Rayleigh* (John William Strutt, 1842–1919) in connection with the vibration of a clarinet reed.
(b) Letting $u' = v$, reduce (6.1) to a system of two equations. Show that the only singular point of that system is $(0, 0)$, and classify its type.
(c) Choosing initial values for u and v, use computer software to obtain phase portraits and plots of $u(t)$ versus t, much as we have in Fig. 2 and 3, for $\epsilon = 0.2, 1$, and 5. For each of those ϵ's, estimate the amplitude and period of the limit cycle solution.
(d) To study the relaxation oscillation ($\epsilon \to \infty$) of the van der Pol equation it is more convenient to work with the Rayleigh equation (6.1), as we shall see. With $u' = v$, let us scale the u variable according to $u = \epsilon w$. Show that the resulting system is

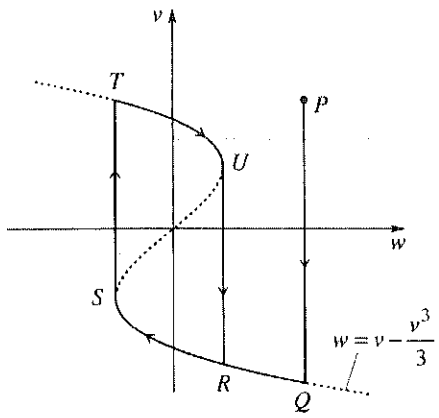
$$\epsilon w' = v, \tag{6.2a}$$

$$v' = \epsilon\left(v - \frac{v^3}{3}\right) - \epsilon w \tag{6.2b}$$

so

$$\frac{dv}{dw} = \epsilon^2\frac{(v - v^3/3) - w}{v}. \tag{6.3}$$

As $\epsilon \to \infty$ we see from (6.3) that $dv/dw \to \infty$ at all points in the v, w phase plane except on the curve $v - v^3/3 - w$. So for any initial point, say P, explain why the solution is as shown in the figure. For instance, why is the direction downward from P, and why does the trajectory jump from S to T and from U to R? The loop $RSTUR$ is traversed repeatedly and is the limit cycle. Finally, use the figure to sketch $v(t)$ and hence the limit cycle solution $x(t)$ of the van der Pol equation (for $\epsilon \to \infty$). The result should be similar to the $\epsilon = 5$ part of Fig. 3. (HINT: Recall the expression "$s' = a$" for the phase velocity in Exercise 8 of Section 7.2.) Finally, explain why it is convenient to work with the Rayleigh equation in order to find the relaxation oscillation of the van der Pol equation.



7. In connection with Fig. 1b we suggested that over the limit cycle the energy gain, while the representative point is in the strip $|x| < 1$, exactly balances the energy loss while the point is outside of that strip. Let us explore that idea.

(a) Multiplying (1) through by dx and integrating over one cycle, show that

$$\frac{1}{2}\, x'^2\Big|_i^f - \epsilon \int_i^f (1 - x^2)x'\, dx + \frac{1}{2}\, x^2\Big|_i^f = 0, \qquad (7.1)$$

where i and f simply denote the initial point and final point, respectively. Explain, further, why (7.1) reduces to

$$\int_i^f (1 - x^2)x'\, dx = 0, \qquad (7.2)$$

which expresses the balance stated above – namely, that the net work done over one cycle by the $\epsilon(1 - x^2)x'$ term in (1) is zero.

(b) For the case of small ϵ ($0 < \epsilon \ll 1$), seek the limit cycle solution in the form $x(t) \approx a \cos t$. Putting the latter into (7.2), show that one obtains $a = 2$ as the radius of the circular limit cycle, as claimed in the text. NOTE: Put differently, (7.1) is equivalent to

$$E\Big|_i^f = \epsilon \int_i^f (1 - x^2)x'\, dx, \qquad (7.3)$$

where $E = \frac{1}{2}x'^2 + \frac{1}{2}x^2$. That is, the change in the total energy E over one cycle is equal to the net work done by the force $\epsilon(1 - x^2)x'$. For a periodic motion $E\big|_i^f$ is zero which, once again, gives (7.2).

7.6 The Duffing Equation: Jumps and Chaos

7.6.1. Duffing equation and the jump phenomenon. Besides the van der Pol equation, also of great importance is the **Duffing equation**:

$$\boxed{mx'' + rx' + \alpha x + \beta x^3 = F_0 \cos \Omega t,} \qquad (1)$$

studied by *G. Duffing* around 1918. Whereas primary interest in the van der Pol equation is in the unforced equation and its self-excited limit cycle oscillation, most of the interest in the Duffing equation involves the various steady-state oscillations that can arise in response to a harmonic forcing function such as $F_0 \cos \Omega t$.

Physically, (1) arises in modeling the motion of a damped, forced, mechanical oscillator of mass m having a nonlinear spring. That is, the spring force is not kx but $\alpha x + \beta x^3$. We assume that $\alpha > 0$ but that β can be positive (for a "hard spring") or negative (for a "soft spring").

The linear version of (1), where $\beta = 0$, is discussed in Section 3.8, and an important result there consisted of the amplitude response curves – namely, the graphs of the amplitude of the steady-state vibration versus the driving frequency Ω for

various values of F_0. For the linear case we obtained two linearly independent homogeneous solutions and a particular solution and used linearity and superposition to form from them the general solution, which contained all possible solutions. Understand that because equation (1) is nonlinear, that approach is not applicable.

Consider first the undamped case ($r = 0$), and let $m = 1$ for simplicity, so (1) becomes

$$x'' + \alpha x + \beta x^3 = F_0 \cos \Omega t. \tag{2}$$

Further, suppose that β is small. Since (1) is nonlinear, we expect it to have a wealth of different sorts of solutions. Of these, particular interest attaches to the so-called **harmonic response**

$$x(t) \approx A \cos \Omega t \qquad \text{(for } \beta \text{ small)} \tag{3}$$

at the same frequency as the driving force. As shown in the exercises, pursuing an approximate solution of the form (3) yields the amplitude-frequency relation

$$\Omega^2 = \alpha + \frac{3}{4}\beta A^2 - \frac{F_0}{A}, \tag{4}$$

which gives the response curves shown in Fig. 1. For the linear case, where $\beta = 0$, (4) reduces to $A = F_0/(\alpha - \Omega^2)$, which result agrees with the amplitude-frequency relation found in Section 3.8. For $\beta > 0$ the curves bend to the right (shown), and for $\beta < 0$ they bend to the left (not shown). Thus, the effect of the nonlinear βx^3 term in (2) is to cause the response curves to bend to one side or the other. Recall from Section 3.8 that for the linear system ($\beta = 0$), the amplitude $|A|$ is infinite when the system is driven at its natural frequency $\sqrt{\alpha}$. [More precisely, we saw that the solution form $x(t) = A \cos \Omega t$ simply does not work for the system $x'' + \alpha x = F_0 \cos \Omega t$ if $\Omega = \sqrt{\alpha}$, but that the method of undetermined coefficients gives $x(t) = \dfrac{F_0}{2\Omega} t \sin \Omega t$, which growing oscillation is known as resonance.] However, because of the bending of the response curves, resonance cannot occur in the nonlinear case. That is, we see from Fig. 1 that if $\beta \neq 0$ then at each driving frequency Ω the response amplitude $|A|$ is finite.

What is the effect of including damping, of letting r be positive in (1) rather than zero? In that case we need to allow for a phase shift (as in Section 3.8), and seek $x(t) \approx A \cos(\Omega t + \Phi)$ in place of (3). The result would be a modified amplitude- frequency relation, in place of (4), and a "capping off" of the response curves as shown in Fig. 2a.

If $\Omega = \Omega_1$, for instance, then the response is at P_1 in Fig. 2b. Suppose we can vary the driving frequency Ω continuously by turning a control knob, like the volume knob on a radio. If we increase Ω slowly (remember that Ω is regarded as a constant in this analysis, so we need to increase it *very* slowly), then the representative point moves to the right along the response curve. But what happens when it reaches P_2, where the response curve has a vertical tangent? Numerical simulation reveals that the point jumps down to P_3, where it can then continue moving rightward on the response curve if Ω is increased further. That is, there is a transient

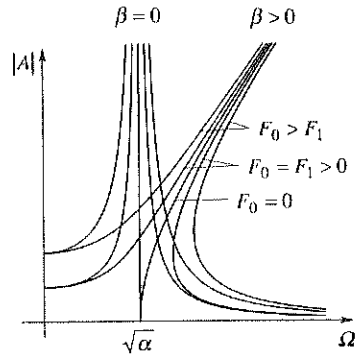


Figure 1. Amplitude response curves; undamped.

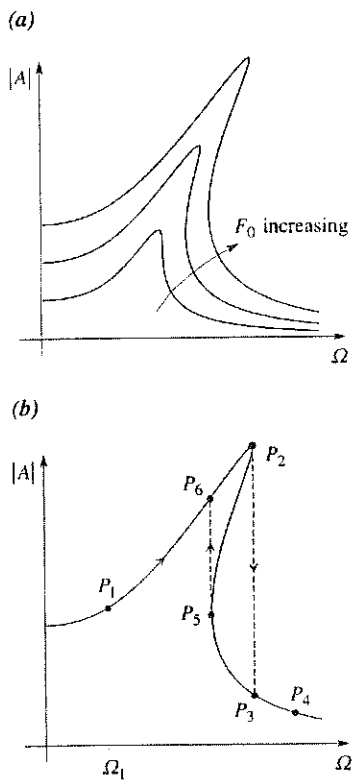


Figure 2. Amplitude response curves; damped.

period during which the oscillation changes from the large amplitude at P_2 to the small amplitude at P_3. Suppose that once the oscillation has settled down to the new amplitude we increase Ω further and stop at P_4. If we now *de*crease Ω the representative point does *not* jump up from P_3 to P_2. Rather, it continues to the point of vertical tangency at P_5, then jumps up to P_6 and continues, say to P_1. Thus, we meet the **jump phenomenon** associated with the Duffing equation, whereby a continuous variation of a system parameter (Ω) can lead to a discontinuous change in the output, namely, the jumps in amplitude from P_2 to P_3 and from P_5 to P_6.

Observe that the middle branch of the response curve, between P_5 and P_2, is inaccessible! That is, if we vary Ω so that the representative point moves from P_1 to P_4 and back again, it never moves along the middle branch.

What if we start the system at an Ω between the vertical dashed lines? Which of the three possible amplitudes will result? It can be shown that the solutions to (1) corresponding to points on the middle branch (P_5P_2) are unstable in essentially the same way that a hilltop is an unstable equilibrium point for a marble, while solutions corresponding to points on the upper and lower branches are stable in essentially the same way that the bottom of a valley is a stable equilibrium point for the marble. Thus, solutions corresponding to points on the middle branch will not be observed – either in the physical system or in its mathematical simulation. If we start the system at an Ω between the vertical dashed lines, and keep Ω fixed, the steady-state oscillation achieved will correspond to a point on the lower branch or on the upper branch. Given the initial conditions, it is possible to predict which of the two branches will "attract" the motion, but that story is well beyond our present scope.*

Both the limit cycle phenomenon exhibited by the van der Pol equation and the jump phenomenon exhibited by the Duffing equation are possible only for nonlinear equations; neither phenomenon could occur for a linear differential equation.

The jump from P_2 to P_3 in Fig. 2b is reminiscent of a Stock Market crash, such as occurred in 1929, yet there is no analog, in Stock Market behavior, of the *up*ward jump from P_5 to P_6. However, observe that in Fig. 2b there is only one control parameter, Ω. If there were two (or more), say Ω and Ψ, then instead of the curve in Fig. 2b we would have an $|A|$ surface above an Ω, Ψ plane. In that case it might be possible, by suitable variation of Ω and Ψ in the Ω, Ψ "control plane," to climb the mountain in a continuous manner, jump off the cliff at P_2, climb the mountain again, and so on, so that only downward jumps were obtained.

Finally, recall from Section 7.5 that the central nervous system is highly nonlinear. Thus, we might well be on the lookout for Duffing-like jumps in human (or animal) behavior due to continuous variation of one or more control parameters. Such jumps are indeed observed! For instance, if one bullies a dog into a corner it will retreat, but at some point it may "turn" on its tormentor, with its behavior jumping, instantaneously, from retreating to attacking. Similarly with human behavior. Two more illustrations: First, observe that our psychological moods likewise tend to change rather abruptly – compared to how long they last. Second, note that

*See Section 7.2 of D. W. Jordan and P. Smith, *Nonlinear Ordinary Differential Equations*, 2nd ed. (Oxford: Oxford University Press, 1987).

someone with an eating disorder might fast for a period of time, then jump almost instantaneously to binging, and vice versa. For a readable account of the modeling of such systems as these, see E. C. Zeeman's *Catastrophe Theory* (Reading, MA: Addison-Wesley, 1977).

7.6.2. Chaos. Consider the physical system shown in Fig. 3, a box containing a slender vertical steel reed cantilevered from the "ceiling" and two magnets attached to the "floor"; x is the horizontal deflection of the end of the reed. In equilibrium, the reed is stationary, with its tip at one magnet or the other. However, if we vibrate the box periodically in the horizontal direction, we can (if the magnets are not too strong) shake the reed loose from its equilibrium position and set it into motion. This experiment was carried out by F. Moon and P. Holmes,* to study chaos. They modeled the system by the Duffing equation

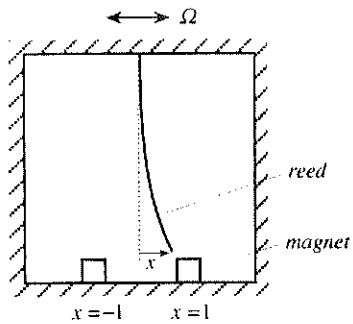


Figure 3. The "Moon beam" system of Moon and Holmes.

$$x'' + rx' - x + x^3 = F_0 \cos \Omega t, \tag{5}$$

where r is a (positive) damping coefficient and F_0 and Ω are the forcing function strength and frequency, respectively. The $-x + x^3$ terms approximate the force induced on the reed by the two competing magnets. It is zero at $x = 0, \pm 1$, which are therefore equilibrium points. To classify the equilibrium points, consider the unforced equation $x'' + rx' - x + x^3 = 0$ or, equivalently,

$$x' = y, \tag{6a}$$
$$y' = -ry + x - x^3. \tag{6b}$$

We leave it for the exercises for you to show that the origin $(x, y) = (0, 0)$ is an unstable equilibrium point, namely a saddle, and that $(\pm 1, 0)$ are stable equilibrium points (stable foci if $r < \sqrt{8}$ and stable nodes if $r > \sqrt{8}$). For the undriven system ($F_0 = 0$), imagine displacing the reed tip to $x = 0.5$, say, and releasing it from rest. Then it will undergo a damped motion about $x = 1$. If instead we release it from $x = -0.5$, say, it will undergo a damped motion about $x = -1$.

What will happen if we *force* the system ($F_0 > 0$)? We can imagine the reed undergoing a steady-state oscillation about $x = +1$ or $x = -1$, depending upon the initial conditions. To encourage physical insight, it is useful to consider the potential energy $V(x)$ associated with the magnetic force $F(x) = x - x^3$. Recalling from Section 7.4.2 that $F(x) = -V'(x)$, we have

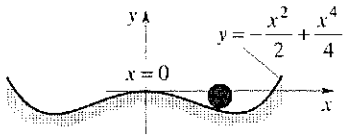


Figure 4. Double well.

$$V(x) = -\frac{x^2}{2} + \frac{x^4}{4}. \tag{7}$$

Thus, in place of a reed/magnet system we can conceptualize the system more intuitively as a mass in a double well as sketched in Fig. 4. That is, its gravitational potential energy is $V(x) = mgy = mg(-x^2/2 + x^4/4)$ which, except for the scale factor mg, is the same as $V(x)$ for the reed/magnet system, given in (7).

*F. C. Moon and P. J. Holmes, A Magnetoelastic Strange Attractor, *Journal of Sound and Vibration*, Vol. 65, pp. 275–296.

If the mass sits at the bottom of a well, then if we vibrate the system horizontally we expect the mass to oscillate within that well. If we vibrate so energetically that the mass jumps from one well to the other, then the situation becomes more complicated.

Rather than carry out the physical experiment, let us simulate it numerically by using computer software to solve (5). Let us fix $r = 0.3$ and $\Omega = 1.2$, and use the initial conditions $x(0) = 1$ and $x'(0) = 0.2$, say, so that if F_0 is not too large we expect an oscillation near $x = 1$ (i.e., in the right-hand well).* Our plan is to use successively larger F_0's and see what happens. The results are quite striking (like those obtained by Moon and Spencer). They are presented in Fig. 5, both in the x, y phase plane and as $x(t)$ versus t.

In Fig. 5a we set $F_0 = 0.2$ and find that after a brief transient period there results a steady-state oscillation near $x = 1$, as anticipated. That oscillation is of the same period as the forcing function (namely, $2\pi/\Omega = 2\pi/1.2 \approx 5.236$) so it is called a **period-1 oscillation**, or **harmonic oscillation**. Period-1 oscillations persist up to around $F_0 = 0.27$, but for $F_0 > 0.27$ different solution types arise.

For $F_0 = 0.28$ the solution is still periodic, but it takes two loops (in the x, y plane) to complete one period and the period is now doubled, namely, $4\pi/\Omega$. Thus, it is called a **period-2 oscillation**, or a **subharmonic oscillation of order 1/2**. In Fig. 5b–5f we omit the transient and display only the steady-state periodic solution so as not to obscure that display. Observe, in Fig. 5b, the point where the trajectory crosses itself. That crossing does not violate the existence and uniqueness theorem given in Section 7.3.1 because equation (5) is nonautonomous and the phase plane figure shows the *projection* of the non-self-intersecting curve in three-dimensional x, y, t space onto the x, y plane.

If we increase F_0 further, to 0.29, the forcing is sufficiently strong so that during the transient phase the mass (reed) is driven out of the right-hand potential well and ends up in a period-4 oscillation about the left-hand well. To observe this result we need to be patient and run the calculation to a sufficiently large time, namely, beyond $t \approx 400$. This period doubling continues as F_0 increases from 0.29 up to around 0.30. For $F_0 > 0.30$ a period-5 oscillation results that now encompasses both stable equilibrium points (Fig. 5d).

The regime $0.37 < F_0 < 0.65$ is found to be rather chaotic, with essentially random motions, as seen in Fig. 5e for the case $F_0 = 0.5$.

Reviewing these results, observe that as we increased F_0 the period of the motion increased until, when F_0 was increased above 0.37, periodicity was lost altogether and the motion became chaotic. (We can think of that motion as periodic but with infinite period.) It would be natural to expect that a further increase of F_0 would lead to even greater chaos (if one were to quantify degree of chaos), yet we find that if F_0 is increased to 0.65 we once again obtain a periodic solution, namely, the period-2 solution shown in Fig. 5f, and if F_0 is increased further to 0.73 then we obtain a period-1 solution (not shown).

In summary, we see that the forced Duffing equation admits a great variety

*These parameter values are the same as those chosen in Section 12.6 of D. W. Jordan and P. Smith (ibid). We refer you to that source for a more detailed discussion than we offer here.

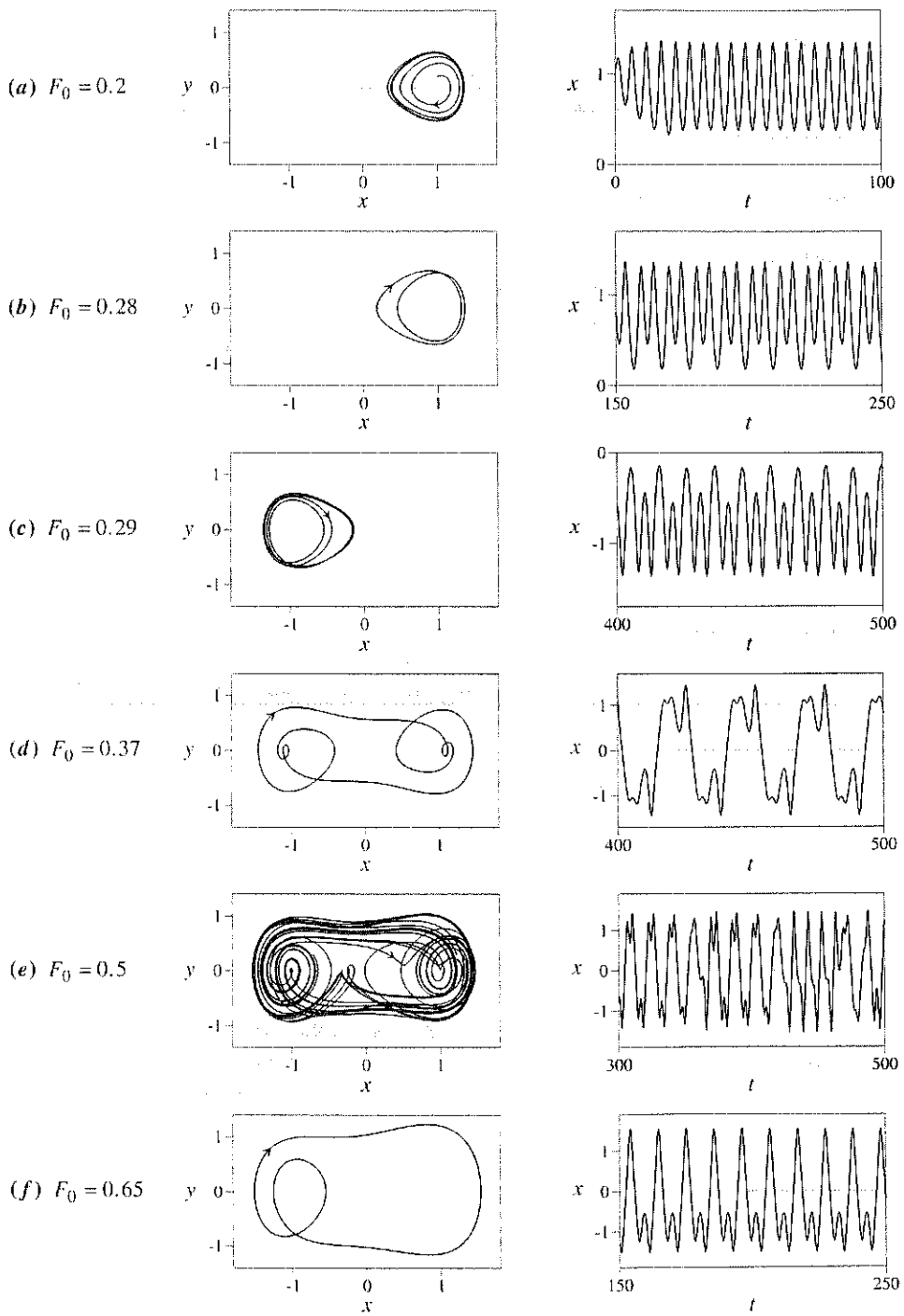


Figure 5. Response to (5): $r = 0.23$, $\Omega = 1.2$.

of periodic solutions and chaotic ones as well, and that these different regimes correspond to different intervals on an F_0 axis. (We chose to hold r and Ω fixed and to vary only F_0, but we could have varied r and/or Ω as well.) It is possible to predict analytically how the solution type varies with F_0, r, and Ω, but that analysis is beyond the scope of this introductory discussion.*

Having classified the response in Fig. 5e as chaotic, it behooves us to clarify what we mean by that. A reasonable working definition of **chaos** is *behavior in a deterministic system, over time, which is aperiodic and which is sensitive to initial conditions.*

By a system being **deterministic** we mean that the governing differential equation(s) and initial conditions imply the existence of a unique solution over subsequent time. Whether we are able to find that solution analytically, or whether we need to use computer simulation, is immaterial. For instance, for given values of r, F_0, Ω, the system consisting of equation (5), together with initial conditions x(0) and $x'(0)$, is deterministic. The choice $r = 0.3$, $F_0 = 0.37$, $\Omega = 1.2$, $x(0) = 1$, and $x'(0) = 0.2$, say, implies the unique response shown in Fig. 5d. If we rerun the numerical solution or solve the problem analytically (if we could), we obtain the same solution as shown in the figure. Likewise even for the chaotic response shown in Fig. 5e.

By the response being **aperiodic** we simply mean that it does not approach a periodic solution or a constant.

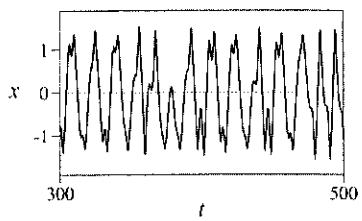


Figure 6. Sensitivity to initial conditions.

To illustrate what is meant by *sensitivity to initial conditions*, let us rerun the case corresponding to Fig. 5e, but with the initial conditions changed from $x(0) = 1$ and $x'(0) = 0.2$ to $x(0) = 1$ and $x'(0) = 0.2000000001$. Observe that the results (Fig. 6) bear virtually no resemblance to those in Fig. 5e. This circumstance is of great significance because if the initial conditions are known to only six significant figures, say, then the task of predicting the response is hopeless!

Another well known example of chaos is provided by the **Lorenz equations**

$$\begin{aligned} x' &= p(y - x), \\ y' &= (q - z)x - y, \\ z' &= xy - rz, \end{aligned} \tag{8}$$

where p, q, r are constants. This system was studied by the mathematical meteorologist *E. Lorenz* in 1963 in connection with the Bénard problem, whereby one seeks the effect of heating a horizontal fluid layer from below.† That problem is of fundamental interest in meteorology because the earth, having been heated by the sun during the day, radiates heat upward into the atmosphere in the evening, thus destabilizing the atmospheric layer above it. Lorenz's contribution was in discovering the chaotic nature of the solution for certain ranges of the physical parameters p, q, r, thereby suggesting the *impossibility* of meaningful long-range weather prediction. Some discussion of (8) is left for the exercises.

*See Section 12.6 in Jordan and Smith (ibid).

†E. Lorenz, Deterministic Nonperiodic Flows, *Journal of Atmospheric Sciences*, Vol. 20, pp. 130–141, 1963.

Perhaps the classic problem of chaos is that of turbulence, in fluid mechanics, be it in connection with the chaotic eddies and mixing behind a truck on a highway or the turbulent breakup of a rising filament of smoke.

To appreciate the revolution in physics that has resulted from recent work on chaos, one needs to understand the euphoria that greeted the birth of Newtonian mechanics and the calculus, according to which both the past and the future are contained in the system of differential equations and initial conditions at the present instant. In the words of Ivar Ekeland,* "Past and future are seen as equivalent, since both can be read from the present. Mathematics travels back in time as easily as a wanderer walks up a frozen river." As Ekeland points out, that statement is not quite true for, as we now know and as was understood by Poincaré even a century ago, deterministic nonlinear systems can turn out to be chaotic, in which case they are useless for long-term prediction.

Closure. The common thread in this section is the Duffing equation (1). For $\alpha > 0$ we study (1) in connection with the reed/magnet system of Moon and Holmes, shown in Fig. 3. Numerical solution of the governing equation (5), for a sequence of increasing F_0 values, leads to a variety of solution types: a harmonic response, various subharmonic responses, and even chaotic responses. The approach to chaos, as we increase F_0, is typical in that the onset of chaos is preceded by a sequence of period doublings.

We define chaos as behavior in a deterministic system (which must be nonlinear if it is to exhibit chaos) over time, which is aperiodic and so sensitive to initial conditions that accurate long-range predictions of the solution are not possible.

Computer software. To generate the responses shown in Fig. 5 and 6 we use the *Maple* phaseportrait command. However, it is worth mentioning how we obtain the graphs in Fig. 1 and 2, because (4) does not give A explicitly but only *im*plicitly as a function of Ω. For instance, suppose we wish to plot the graphs of y versus x, over $0 < x < 3$ and $0 < y < 4$, for the functions $y(x)$ given implicitly by the equations $y - y^3 = x$ and $4y - y^3 = x$. First, enter

with(plots):

and return, to access the subsequent plotting command. Then use the **implicitplot** command. Enter

implicitplot($\{y - y\hat{}3 = x,\ 4 * y - y\hat{}3 = x\},\ x = 0..3,\ y = 0..4,$
numpoints $= 1000$);

and return. Here, numpoints indicates the number of points to be used. To plot the single function $y - y^3 = x$, use $y - y\hat{}3 = x$ in place of implicitplot($\{y - y\hat{}3 = x,\ 4 * y - y\hat{}3 = x\}$.

**Mathematics and the Unexpected* (Chicago: Chicago University Press, 1988).

EXERCISES 7.6

1. (*Derivation of the Duffing amplitude-frequency relation*) We state in the text that if one seeks an approximate harmonic solution $x(t) \approx A\cos\Omega t$ to

$$x'' + \alpha x + \beta x^3 = F_0 \cos\Omega t, \tag{1.1}$$

then one obtains the ampitude-frequency relation

$$\Omega^2 = \alpha + \frac{3}{4}\beta A^2 - \frac{F_0}{A} \tag{1.2}$$

discovered by Duffing. A modern derivation of (1.2) would probably use a so-called *singular perturbation method* of *strained variables*, but here we will pursue a simpler iterative approach which is essentially that of Duffing; namely, we replace (1.1) by the iterative scheme

$$x''_{n+1} = -\alpha x_n - \beta x_n^3 + F_0 \cos\Omega t, \tag{1.3}$$

choose the initial iterate as $x(t) = A\cos\Omega t$, and then use (1.3) to find the successive iterates $x_1(t), x_2(t), \ldots$. It is surely not obvious whether that procedure will work, so it makes sense to try it out first for the simple *linear* case where $\beta = 0$, for which we know the exact solution.

(a) In that case ($\beta = 0$), show that if we seek a harmonic solution $x(t) = A\cos\Omega t$ of the Duffing equation (1.1) with $\beta = 0$, we obtain $A = F_0/(\alpha - \Omega^2)$, and hence the exact solution

$$x(t) = \frac{F_0}{\alpha\Omega^2}\cos\Omega t. \tag{1.4}$$

(b) Next, use (1.3) to generate $x_1(t), x_2(t), \ldots$, for $\beta = 0$, and show that

$$\begin{aligned} x_1(t) &= \frac{\alpha A - F_0}{\Omega^2}\cos\Omega t, \\ &\vdots \\ x_n(t) &= \left\{\left(\frac{\alpha}{\Omega^2}\right)^n A \right. \\ &\quad \left. -\frac{F_0}{\Omega^2}\left[1 + \left(\frac{\alpha}{\Omega^2}\right) + \cdots + \left(\frac{\alpha}{\Omega^2}\right)^{n-1}\right]\right\}\cos\Omega t. \end{aligned} \tag{1.5}$$

That is, put $x_0(t) = A\cos\Omega t$ into the right side of (1.3) and integrate twice to obtain x_1. Then put that x_1 into the right side of (1.3) and integrate twice to obtain x_2, and so on. By the time you reach x_3, the general result shown in (1.5) should be apparent.

(c) Recalling that the geometric series $1 + x + x^2 + x^3 + \cdots$ converges to $1/(1-x)$ if $|x| < 1$ and diverges otherwise, show that the $x_n(t)$ sequence given by (1.5) does indeed converge to the exact solution (1.4) as $n \to \infty$, provided that $|\alpha/\Omega^2| < 1$.

(d) In fact, show that if we equate the coefficients of $\cos\Omega t$ in $x_0(t) = A\cos\Omega t$ and $x_1(t) = \dfrac{\alpha A - F_0}{\Omega^2}\cos\Omega t$, we happen to obtain $A = F_0/(\alpha - \Omega^2)$, which agrees with the exact solution (1.4)!

(e) In view of the striking success in part (d), we are encouraged to expect good results even for the nonlinear case where $\beta \neq 0$. Thus, put $x_0(t) = A\cos\Omega t$ into the right side of (1.3) and integrate twice to obtain $x_1(t)$. Then, as in (d), equate the coefficients of $\cos\Omega t$ in $x_1(t)$ and $x_0(t)$, and show that the result is Duffing's relation (1.2). HINT: The identity $\cos^3\theta = (3\cos\theta + \cos 3\theta)/4$ should be helpful.

2. (*Computer problem regarding the Duffing jump phenomenon*) For the undamped case the amplitude-frequency relation is given by (4). For the damped case it is given by

$$F_0^2 = \left[(\alpha - \Omega^2)A + \frac{3}{4}\beta A^3\right]^2 + (r\Omega A)^2. \tag{2.1}$$

Throughout parts (a)–(g) let $F_0 = 2$, $\alpha = 1$, $\beta = 0.4$, and $r = 0.3$, for definiteness.

(a) Use (2.1) to generate a computer plot of $|A|$ versus Ω as we did in Fig. 2b. NOTE: Actually, there is no need to distinguish between A and $|A|$ since, unlike (4), (2.1) contains only even powers of A.

(b) For $\Omega = 1$ solve (2.1) for A. (HINT: Using *Maple*, for instance, use the fsolve command.) Next, use computer software such as the *Maple* phaseportrait command to solve

$$x'' + rx' + \alpha x + \beta x^3 = F_0 \cos\Omega t, \tag{2.2}$$

and plot $x(t)$ versus t over a sufficiently long time interval to obtain the steady-state response. Compare the amplitude of the resulting steady-state response with the value of A obtained from (2.1).

(c) Same as (b) but for an Ω point specified by your instructor, to the left of the first point of vertical tangency ($\Omega \approx 1.71$).

(d) Same as (b), but for an Ω point specified by your instructor, to the right of the second point of vertical tangency ($\Omega \approx 2.05$).

(e) Now consider an Ω between the two points of vertical tangency, say, $\Omega = 1.8$. Solve (2.1) for the three A values. Next, use computer software to solve (2.2) over a sufficiently long

time interval to obtain the steady-state response. Depending upon the initial conditions that you impose, you should obtain the smallest or largest of the three A values, but never the middle one. Keeping $x'(0) = 0$, determine the approximate value of $x(0)$ below which you obtain the small-amplitude response and above which you obtain the large-amplitude response.
(f) Same as (e) but for an Ω point specified by your instructor.
(g) Continuing to use the r, α, β, F_0 values given above, now let Ω be slowly varying according to $\Omega = 1.9 - 0.0005t$, and solve (2.2) over $0 < t < 800$ with the initial conditions $x(0) = x'(0) = 0$. Plot the resulting $x(t)$ versus t and discuss your results in terms of the ideas discussed in this section.

3. (a) Refer to (4) and Fig. 1. What is the asymptotic form of the graph of $|A|$ versus Ω as $|A| \to \infty$?
(b) In Fig. 1 we show several amplitude response curves for $\beta = 0$ and $\beta > 0$ and for several values of F_0. Obtain the analogous curves for the case where $\beta < 0$, either by a careful hand sketch or by computer plotting.

4. Determine the location and type of any singular points of (6).

5. (a)–(f) Obtain x, y and x, t plots for the cases depicted in Fig. 5. Your results should be the same as those in Fig. 5.

6. We stated that "period doubling continues as F_0 increases from 0.29 up to around 0.30." Find, by numerical experimentation, an F_0 that gives a period-8 oscillation (and, if possible, period-16) and obtain computer plots analogous to those in Fig. 5.

7. We stated that "if F_0 is increased further to 0.73, then we obtain a period-1 solution." Obtain computer plots for that case, analogous to those in Fig. 5.

8. We found an extreme sensitivity to initial conditions for the chaotic regime. Specifically, the plot in Fig. 6 bears little resemblance to the corresponding one in Fig. 5e, even though the initial conditions differed by only 10^{-10}. Show that that sensitivity is *not* found for the non-chaotic responses – namely, for the periodic responses. Specifically, rerun the cases reported in Fig. 5d and 5f, but with $x'(0) = 0.2$ changed to 0.20001, say. Do your results appear to reproduce those in Fig. 5?

9. The equation

$$x'' + 0.3x' + \sin x = F_0 \cos t \tag{9.1}$$

is similar to the one occurring in the Moon/Holmes experiment.

(a) Describe a physical problem that would have a governing equation of motion of that form. (We have assigned numerical values to all of the physical parameters except to F_0, which we leave for the purpose of numerical experimentation.)
(b) We leave this problem a bit more open ended than the foregoing ones, and simply ask you to carry out an analytical and experimental study of (9.1). For instance, you might investigate the singular points of the homogeneous version of (9.1), and also run computer solutions for a range of F_0 values, somewhat as done for equation (5).

Chapter 7 Review

In Sections 7.2–7.5 we study the autonomous system

$$x' = P(x, y), \qquad y' = Q(x, y) \tag{1}$$

mostly in the x, y phase plane. We focus considerable attention on the singular points of (1), the points (if any) where $P(x, y) = 0$ and $Q(x, y) = 0$. Linearizing equations (1) about each singular point, we obtain a simpler system of the form

$$\begin{aligned} X' &= aX + bY \\ Y' &= cX + dY, \end{aligned}$$

where $X = x - x_s$, $Y = y - y_s$, and (x_s, y_s) is the singular point. Considering only elementary singular points (that is, for which $ab - cd \neq 0$), we classify them as

centers, foci, nodes, and saddles. The Hartman–Grobman theorem assumes that the linearized system faithfully captures the nature of the local flow (except possibly for the borderline cases of proper nodes and centers, as explained in Section 7.4.1).

In Section 7.4 we study applications and introduced the idea of a bifurcation, whereby the behavior of the system changes qualitatively as a system parameter passes through a critical value. We illustrate the concept with an example of a saddle-node bifurcation from molecular biology.

In Section 7.5 we study the van der Pol equation

$$x'' - \epsilon(1 - x^2)x' + x = 0, \qquad (\epsilon > 0)$$

which introduce us to the concept of limit cycles and relaxation oscillations.

Finally, we study the forced Duffing equation

$$mx'' + rx' + \alpha x + \beta x^3 = F_0 \cos \Omega t$$

in two contexts. First, we consider it as modeling a mechanical oscillator, with nonlinear spring force $\alpha x + \beta x^3$, where $\alpha > 0$. Of the various possible solutions that can be obtained from different initial conditions, we study only the harmonic response – that is, the steady-state periodic response at the same frequency as the driving frequency Ω. The key feature that was revealed was the bending of the amplitude response curves and the resulting jump phenomenon, whereby the response amplitude jumps as Ω is increases or decreases slowly through a critical value.

We also consider it as modeling the "double well" reed/magnet system of Moon and Holmes. By numerical simulation, we find that if F_0 is not too large, then the oscillation is confined to one of the two wells. As F_0 is increased, there results a sequence of period doublings, giving so-called subharmonic responses, until F_0 becomes large enough to drive the response out of that well. Beyond a critical F_0 value, we then obtain a chaotic response involving both wells.

Chapter 8

Systems of Linear Algebraic Equations; Gauss Elimination

8.1 Introduction

There are many applications in science and engineering where application of the relevant physical law(s) immediately produces a set of linear algebraic equations. For instance, the application of Kirchoff's laws to a DC electrical circuit containing any number of resistors, batteries, and current loops immediately produces such a set of equations on the unknown currents. In other cases, the problem is stated in some other form such as one or more ordinary or partial differential equations, but the solution method eventually leads us to a system of linear algebraic equations. For instance, to find a particular solution to the differential equation

$$y''' - y'' = 3x^2 + 5\sin x \tag{1}$$

by the method of undetermined coefficients (Section 3.7.2), we seek it in the form

$$y_p(x) = Ax^4 + Bx^3 + Cx^2 + D\sin x + E\cos x. \tag{2}$$

Putting (2) into (1) and equating coefficients of like terms on both sides of the equation gives five linear algebraic equations on the unknown coefficients $A, B, \ldots, E$. Or, solving the so-called Laplace partial differential equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \tag{3}$$

on the rectangle $0 < x < 1$, $0 < y < 1$ by the method of finite differences (which is studied in Section 20.5), using a mesh size $\Delta x = \Delta y = 0.05$, gives $19^2 = 361$ linear algebraic equations on the unknown values of u at the 361 nodal points of the mesh. Our point here is not to get ahead of ourselves by plunging into partial differential equations, but to say that the solution of practical problems of interest in science and engineering *often* leads us to systems of linear algebraic equations.

Such systems often involve a great many unknowns. Thus, the question of existence (Does a solution exist?), which often sounds "too theoretical" to the practicing engineer, takes on great practical importance because a considerable computational effort is at stake.

The subject of linear algebra and matrices encompasses a great deal more than the theory and solution of systems of linear algebraic equations, but the latter is indeed a central topic and is foundational to others. Thus, we begin this sequence of five chapters (8–12) on **linear algebra** with an introduction to the theory of systems of linear algebraic equations, and their solution by the method of Gauss elimination. Results obtained here are used, and built upon, in Chapters 9–12.

Chapters 9 and 10 take us from vectors in 3-space to vectors in n-space and generalized vector space, to matrices and determinants. Linear systems of algebraic equations are considered again, in the second half of Chapter 10, in terms of rank, inverse matrix, LU decomposition, Cramer's rule, and linear transformation. Chapter 11 introduces the eigenvalue problem, diagonalization, and quadratic forms; areas of application include systems of ordinary differential equations, vibration theory, chemical kinetics, and buckling. Chapter 12 is optional and brief and provides an extension of results in Chapters 9–11 to complex spaces.

8.2 Preliminary Ideas and Geometrical Approach

The problem of finding solutions of equations of the form

$$f(x) = 0 \tag{1}$$

occupies a place of both practical and historical importance. Equation (1) is said to be **algebraic**, or **polynomial**, if $f(x)$ is expressible in the form $a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$, where $a_n \neq 0$ for definiteness [i.e., if $f(x)$ is a polynomial of finite degree n], and it is said to be **transcendental** otherwise.

EXAMPLE 1. The equations $6x - 5 = 0$ and $3x^4 - x^3 + 2x + 1 = 0$ are algebraic, whereas $x^3 + 2\sin x = 0$ and $e^x - 3 = 0$ are transcendental since $\sin x$ and e^x cannot be expressed as polynomials of finite degree. ∎

Besides the algebraic versus transcendental distinction, we classify (1) as **linear** if $f(x)$ is a first-degree polynomial,

$$a_1 x + a_0 = 0, \tag{2}$$

and **nonlinear** otherwise. Thus, the first equation in Example 1 is linear, and the other three are nonlinear.

While (1) is one equation in one unknown, we often encounter problems involving more than one equation and/or more than one unknown – that is, a **system**

of equations consisting of m equations in n unknowns, where $m \geq 1$ and $n \geq 1$,

$$\begin{aligned} f_1(x_1, \ldots, x_n) &= 0, \\ f_2(x_1, \ldots, x_n) &= 0, \\ &\vdots \\ f_m(x_1, \ldots, x_n) &= 0 \end{aligned} \tag{3}$$

such as

$$\begin{aligned} x_1 - \sin(x_1 + 7x_2) &= 0, \\ x_2^3 + x_2 - 5x_1 + 6 &= 0. \end{aligned} \tag{4}$$

In (4) it happens that $m = n$ (namely, $m = n = 2$) so that there are as many equations as unknowns. In general, however, m may be less than, equal to, or greater than n so we allow for $m \neq n$ in this discussion even though $m = n$ is the most important case.

In this chapter we consider only the case where (3) is **linear**, of the form

$$\boxed{\begin{array}{rcl} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &=& c_1, \quad \text{(eq.1)} \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &=& c_2, \quad \text{(eq.2)} \\ & \vdots & \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &=& c_m, \quad \text{(eq.m)} \end{array}} \tag{5}$$

and we restrict m and n to be finite, and the a_{ij}'s and c_j's to be real numbers. If all the c_j's are zero then (5) is **homogeneous**; if they are not all zero then (5) is **nonhomogeneous**.

The subscript notation adopted in (5) is not essential but is helpful in holding the nomenclature to a minimum, in rendering inherent patterns more visible, and in permitting a natural transition to matrix notation. The first subscript in a_{ij} indicates the equation, and the second indicates the x_j variable that it multiplies. For instance, a_{21} appears in the second equation and multiplies the x_1 variable. To avoid ambiguity we should write $a_{2,1}$ rather than a_{21} so that one does not mistakenly read the subscripts as twenty-one, but we will omit commas except when such ambiguity is not easily resolved from the context.

We say that a sequence of numbers $s_1, s_2, \ldots, s_n$ is a **solution** of (5) if and only if each of the m equations is satisfied numerically when we substitute s_1 for x_1, s_2 for x_2, and so on. If there exist *one or more* solutions to (5), we say that the system is **consistent**; if there is precisely *one* solution, that solution is **unique**; and if there is more than one, the solution is **nonunique**. If, on the other hand, there are *no* solutions to (5), the system is said to be **inconsistent**. The collection of all solutions to (5) is called its **solution set** so, by "solving (5)" we mean finding its solution set.

Let us begin with the simple case, where $m = n = 1$:

$$a_{11}x_1 = c_1. \tag{6}$$

In the generic case, $a_{11} \neq 0$ and (6) admits the unique solution $x_1 = c_1/a_{11}$, but if $a_{11} = 0$ there are two possibilities: if $c_1 \neq 0$ then there are no values of x_1 such that $0x_1 = c_1$ and (6) is inconsistent, but if $c_1 = 0$ then (6) becomes $0x_1 = 0$, and $x_1 = \alpha$ is a solution for *any* value of α; that is, the solution is nonunique.

Far from being too simple to be of interest, the case where $m = n = 1$ establishes a pattern that will hold in general, for any values of m and n. Specifically, the system (5) will admit a unique solution, no solution, or an infinity of solutions. For instance, it will never admit 4 solutions, 12 solutions, or 137 solutions.

Next, consider the case where $m = n = 2$:

$$a_{11}x_1 + a_{12}x_2 = c_1, \qquad \text{(eq.1)} \tag{7a}$$

$$a_{21}x_1 + a_{22}x_2 = c_2. \qquad \text{(eq.2)} \tag{7b}$$

If a_{11} and a_{12} are not both zero, then (eq.1) defines a straight line, say $L1$, in a Cartesian x_1, x_2 plane; that is the solution set of (eq.1) is the set of all points on that line. Similarly, if a_{21} and a_{22} are not both zero then the solution set of (eq.2) is the set of all points on a straight line $L2$. There exist exactly three possibilities, and these are illustrated in Fig. 1. First, the lines may intersect at a point, say P, in which case (7) admits the unique solution given by the coordinate pair x_1, x_2 of the point P (Fig. 1a). That is, any solution pair x_1, x_2 of (7) needs to be in the solution set of (eq.1) *and* in the solution set of (eq.2) hence at an intersection of $L1$ and $L2$. This is the generic case, and it occurs (Exercise 2) as long as

$$a_{11}a_{22} - a_{12}a_{21} \neq 0; \tag{8}$$

(8) is the analog of the $a_{11} \neq 0$ condition for the $m = n = 1$ case discussed above.

Second, the lines may be parallel and nonintersecting (Fig. 1b), in which case there is no solution. Then (7) is inconsistent, the solution set is empty.

Third, the lines may coincide (Fig. 1c), in which case the coordinate pair of each point on the line is a solution. Then (7) is consistent and there are an infinite number of solutions.

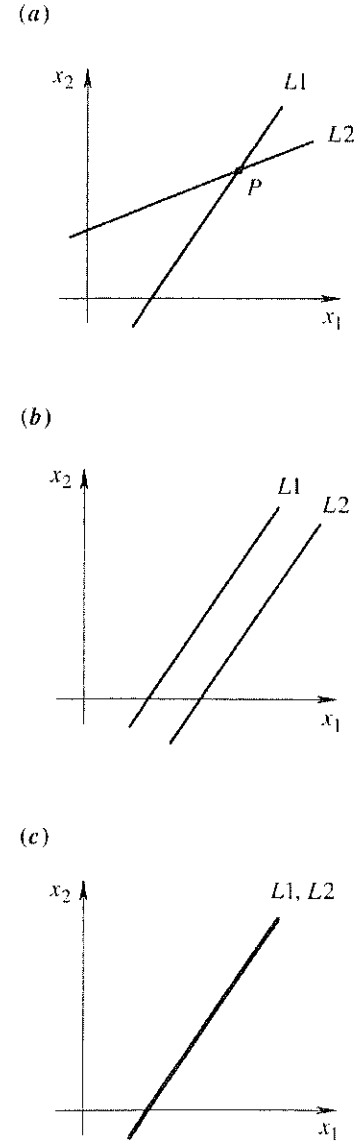


Figure 1. Existence and uniqueness for the system (7).

EXAMPLE 2.

$$\begin{aligned} 2x_1 - x_2 &= 5, \\ x_1 + 3x_2 &= -1, \end{aligned} \qquad \begin{aligned} x_1 + 3x_2 &= 1, \\ x_1 + 3x_2 &= 0, \end{aligned} \qquad \begin{aligned} x_1 + 3x_2 &= 1, \\ 2x_1 + 6x_2 &= 2, \end{aligned}$$

illustrate these three cases, respectively. ∎

Below (7) we said "If a_{11} and a_{12} are not both zero . . . ." What if they <u>are</u> both zero? Then if $c_1 \neq 0$ there is no solution of (7a), and hence there is no solution to the system (7). But if $c_1 = 0$, then (7a) reduces to $0 = 0$ and can be discarded, leaving just (7b). If a_{21} and a_{22} are not both zero, then (7b) gives a line of solutions, but if they are both zero then everything hinges on c_2. If $c_2 \neq 0$ there is no solution and (7) is inconsistent, but if $c_2 = 0$, so both (7a) and (7b) are simply $0 = 0$, then both x_1 and x_2 are arbitrary, and every point in the plane is a solution.

Next, consider the case where $m = n = 3$:

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = c_1, \qquad \text{(eq.1)} \qquad (9a)$$
$$a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = c_2, \qquad \text{(eq.2)} \qquad (9b)$$
$$a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = c_3. \qquad \text{(eq.3)} \qquad (9c)$$

Continuing the geometric approach exemplified by Fig. 1, observe that if a_{11}, a_{12}, a_{13} are not all zero then (eq.1) defines a plane, say $P1$, in Cartesian x_1, x_2, x_3 space, and similarly for (eq.2) and (eq.3). In the generic case, $P1$ and $P2$ intersect along a line L, and L pierces $P3$ at a point P. Then the x_1, x_2, x_3 coordinates of P give the unique solution of (9).

In the nongeneric case we can have no solution or an infinity of solutions in the following ways. There will be no solution if L is parallel to $P3$ and hence fails to pierce it, or if any two of the planes are parallel and not coincident. There will be an infinity of solutions if L lies in $P3$ (i.e., a line of solutions), if two planes are coincident and intersect the third (again, a line of solutions), or if all three planes are coincident (this time an entire plane of solutions).

The case where all of the a_{ij} coefficients are zero in one or more of equations (9) is left for the exercises.

An abstract extension of such geometrical reasoning can be continued even if $m = n \geq 4$. For instance, one speaks of $a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + a_{14}x_4 = c_1$ as defining a *hyperplane* in an abstract four-dimensional space. In fact, perhaps we should mention that even the familiar x_1, x_2 plane and x_1, x_2, x_3 space discussed here could be abstract as well. For instance, if x_1 and x_2 are unknown currents in two loops of an electrical circuit, then what physical meaning is there to an x_1, x_2 plane? None, but we can introduce it, create it, to assist our reasoning.

Closure. Most of this section is devoted to a geometrical discussion of the system (5) of linear algebraic equations. A great advantage of geometrical reasoning is that it brings our visual system into play. It is estimated that at least a third of the neurons in our brain are devoted to vision, hence our visual sense is extremely sophisticated. No wonder we say "Now I see what you mean; now I get the picture." The more geometry, pictures, visual images to aid our thinking, the better! We have not yet aimed at theorems, and have been content to lay the groundwork for the ideas of existence and uniqueness of solutions. In considering the cases where $m = n = 1$, $m = n = 2$, and $m = n = 3$, we have not meant to imply that we need to have $m = n$; *all* possibilities are considered in the next section. To proceed further, we need to consider the process of *finding* solutions, and that we do, in Section 8.3, by the method of Gauss elimination.

EXERCISES 8.2

1. True or false? If false, give a counterexample.

(a) An algebraic equation is necessarily linear.
(b) An algebraic equation is necessarily nonlinear.
(c) A transcendental equation is necessarily linear.
(d) A transcendental equation is necessarily nonlinear.
(e) A linear equation is necessarily algebraic.
(f) A nonlinear equation is necessarily algebraic.
(g) A linear equation is necessarily transcendental.
(h) A nonlinear equation is necessarily transcendental.

2. Derive the condition (8) as the necessary and sufficient condition for (7) to admit a unique solution.

3. (a) Discuss all possibilities of the existence and uniqueness of solutions of (9) from a geometrical point of view, in the event that $a_{11} = a_{12} = a_{13} = 0$, but a_{21}, a_{22}, a_{23} and a_{31}, a_{32}, a_{33} are not all zero.
(b) Same as (a), but with $a_{21} = a_{22} = a_{23} = 0$ as well.
(c) Same as (a), but with $a_{21} = a_{22} = a_{23} = a_{31} = a_{32} = a_{33} = 0$ as well.

8.3 Solution by Gauss Elimination

8.3.1. Motivation. In this section we continue to consider the system of m linear algebraic equations

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= c_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= c_2, \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= c_m, \end{aligned} \tag{1}$$

in the n unknowns $x_1, \ldots, x_n$, and develop the solution technique known as **Gauss elimination**. To motivate the ideas, we begin with an example.

EXAMPLE 1. Determine the solution set of the system

$$\begin{aligned} x_1 + x_2 - x_3 &= 1, \\ 3x_1 + x_2 + x_3 &= 9, \\ x_1 - x_2 + 4x_3 &= 8. \end{aligned} \tag{2}$$

Keep the first equation intact, and add -3 times the first equation to the second (as a replacement for the second equation), and add -1 times the first equation to the third (as a replacement for the third equation). These steps yield the new "indented" system

$$\begin{aligned} x_1 + x_2 - x_3 &= 1, \\ -2x_2 + 4x_3 &= 6, \\ -2x_2 + 5x_3 &= 7. \end{aligned} \tag{3}$$

Next, keep the first two of these intact, and add -1 times the second equation to the third, and obtain

$$\begin{aligned} x_1 + x_2 - x_3 &= 1, \\ -2x_2 + 4x_3 &= 6, \\ x_3 &= 1. \end{aligned} \tag{4}$$

Finally, multiplying the second of these by $-1/2$ to normalize the leading coefficient (to unity), gives

$$\begin{aligned} x_1 + x_2 - x_3 &= 1, \quad \text{(eq.1)} \\ x_2 - 2x_3 &= -3, \quad \text{(eq.2)} \\ x_3 &= 1. \quad \text{(eq.3)} \end{aligned} \tag{5}$$

It is helpful to think of the original system (2) as a tangle of string that we wish to unravel. The first step is to find a loose end and that is, in effect, what the foregoing process of successive indentations has done for us. Specifically, (eq.3) in (5) is the "loose end," and with that in hand we may unravel (5) just as we would unravel a tangle: putting $x_3 = 1$ into (eq.2) gives $x_2 = -1$, and then putting $x_3 = 1$ and $x_2 = -1$ into (eq.1) gives $x_1 = 3$. Thus, we obtain the unique solution

$$x_3 = 1, \quad x_2 = -1, \quad x_1 = 3. \tag{6}$$

COMMENT 1. From a mathematical point of view, the system (2) was a "tangle" because the equations were **coupled**; that is, each equation contained more than one unknown. Actually, the final system (5) is coupled too, since (eq.1) contains all three unknowns and (eq.2) contains two of them. However, the coupling in (5) is not as debilitating because (5) is in what we call **triangular form**. Thus, we were able to solve (eq.3) for x_3, put that value into (eq.2) and solve for x_2, and then put these values into (eq.1) and solve for x_1, which steps are known as **back substitution**.

COMMENT 2. However, the process begs this question: Is it obvious that the systems (2)–(5) all have the same solution sets so that when we solve (5) we are actually solving (2)? That is, is it not conceivable that in applying the arithmetic steps that carried us from (2) to (5) we might, inadvertently, have altered the solution set? For example, $x - 1 = 4$ has the unique solution $x = 5$, but if we innocently square both sides, the resulting equation $(x-1)^2 = 16$ admits the *two* solutions $x = 5$ and $x = -3$. ∎

The question just raised applies to linear systems in general. It is answered in Theorem 8.3.1 that follows, but first we define two terms: "equivalent systems" and "elementary equation operations."

Two linear systems in n unknowns, x_1 through x_n, are said to be **equivalent** if their solution sets are identical.

The following operations on linear systems are known as **elementary equation operations**:

1. Addition of a multiple of one equation to another
   Symbolically: (eq.j) $\to$ (eq.j) + α (eq.k)

2. Multiplication of an equation by a nonzero constant
   Symbolically: (eq.j) $\to$ α (eq.j)

3. Interchange of two equations
   Symbolically: (eq.j) $\leftrightarrow$ (eq.k)

Then we can state the following result.

THEOREM 8.3.1 *Equivalent Systems*
If one linear system is obtained from another by a finite number of elementary equation operations, then the two systems are equivalent.

Outline of Proof: The truth of this claim for elementary equation operations of types 2 and 3 should be evident, so we confine our remarks to operations of type 1. It suffices to look at the effect of one such operation. Thus, suppose that a given linear system A is altered by replacing its jth equation by its jth plus α times its kth, its other equations being kept intact. Let us call the new system A'. Surely, every solution of A will also be a solution A' since we have merely added equal quantities to equal quantities. That is, *if A' results from A by the application of an elementary equation operation of type 1, then every solution of A is also a solution of A'*. Further, we can convert A' back to A by an elementary equation operation of type 1, namely, by replacing the jth equation of A' by the jth equation of A' plus $-\alpha$ times the kth equation of A'. Consequently, it follows from the italicized result (two sentences back) that every solution of A' is also a solution of A. Then A and A' are equivalent, as claimed. ■

In Example 1, we saw that each step is an elementary equation operation: Three elementary equation operations of type 1 took us from (2) to (4), and one of type 2 took us from (4) to (5); finally, the back substitution amounted to several operations of type 1. Thus, according to Theorem 8.3.1, equivalence was maintained throughout so we can be sure that (6) is the solution set of the original system (2) (as can be verified by direct substitution).

The system in Example 1 admitted a unique solution. To see how the method of successive elimination works out when there is no solution, or a nonunique solution, let us work two more examples.

EXAMPLE 2. *Inconsistent System.* Consider the system

$$\begin{aligned} 2x_1 + 3x_2 - 2x_3 &= 4, \\ x_1 - 2x_2 + x_3 &= 3, \\ 7x_1 \qquad - x_3 &= 2. \end{aligned} \tag{7}$$

Keep the first equation intact, add $-\frac{1}{2}$ times the first equation to the second (eq.2 $\to$ eq.2 $-\frac{1}{2}$ eq.1), and add $-\frac{7}{2}$ times the first to the third (eq.3 $\to$ eq.3 $-\frac{7}{2}$ eq.1):

$$\begin{aligned} 2x_1 + 3x_2 - 2x_3 &= 4, \\ -\tfrac{7}{2}x_2 + 2x_3 &= 1, \\ -\tfrac{21}{2}x_2 + 6x_3 &= -12. \end{aligned} \tag{8}$$

Keep the first two equations intact, and add -3 times the second equation to the third (eq.3

$\rightarrow$ eq.3 -3 eq.2):

$$\begin{aligned} 2x_1 + 3x_2 - 2x_3 &= 4, \\ -\tfrac{7}{2}x_2 + 2x_3 &= 1, \\ 0 &= -15. \end{aligned} \tag{9}$$

Any solution x_1, x_2, x_3 of (9) must satisfy each of the three equations, but there are no values of x_1, x_2, x_3 that can satisfy $0 = -15$. Thus, (9) is inconsistent (has no solution), and therefore (7) is as well.

COMMENT. The source of the inconsistency is the fact that whereas the left-hand side of the third equation is 2 times the left-hand side of the first equation plus 3 times the left-hand side of the second, the right-hand sides do not bear that relationship: $2(4)+3(3) = 17 \neq 2$. [While that built-in contradiction is not obvious from (7), it eventually comes to light in the third equation in (9).] If we modify the system (7) by changing the final 2 in (7) to 17, then the final -12 in (8) becomes a 3, and the final -15 in (9) becomes a zero:

$$\begin{aligned} 2x_1 + 3x_2 - 2x_3 &= 4, \\ -\tfrac{7}{2}x_2 + 2x_3 &= 1, \\ 0 &= 0 \end{aligned} \tag{10}$$

or, multiplying the first by $\frac{1}{2}$ and the second by $-\frac{2}{7}$,

$$\begin{aligned} x_1 + \tfrac{3}{2}x_2 - x_3 &= 2, \\ x_2 - \tfrac{4}{7}x_3 &= -\tfrac{2}{7}, \end{aligned} \tag{11a,b}$$

where we have discarded the identity $0 = 0$. Thus, by changing the c_j's so as to be "compatible," the system now admits an infinity of solutions rather than none. Specifically, we can let x_3 (or x_2, it doesn't matter which) in (11b) be *any* value, say α, where α is arbitrary. Then (11b) gives $x_2 = -\frac{2}{7} + \frac{4}{7}\alpha$, and putting these into (11a), $x_1 = \frac{17}{7} + \frac{1}{7}\alpha$. Thus, we have the infinity of solutions

$$x_3 = \alpha, \quad x_2 = -\frac{2}{7} + \frac{4}{7}\alpha, \quad x_1 = \frac{17}{7} + \frac{1}{7}\alpha \tag{12}$$

for any α. Evidently, two of the three planes intersect, giving a line that lies in the third plane, and equations (12) are parametric equations of that line! ∎

EXAMPLE 3. *Nonunique Solution.* Consider the system of four equations in six unknowns ($m = 4, n = 6$)

$$\begin{aligned} 2x_2 + x_3 + 4x_4 + 3x_5 + x_6 &= 2, \\ x_1 - x_2 + x_3 \qquad\qquad + 2x_6 &= 0, \\ x_1 + x_2 + 2x_3 + 4x_4 + x_5 + 2x_6 &= 3, \\ x_1 - 3x_2 \qquad - 4x_4 - 2x_5 + x_6 &= 0. \end{aligned} \tag{13}$$

Wanting the top equation to begin with x_1 and subsequent equations to indent at the left,

let us first move the top equation to the bottom (eq.1 $\leftrightarrow$ eq.4):

$$\begin{aligned} x_1 - 3x_2 \qquad\quad - 4x_4 - 2x_5 + x_6 &= 0, \\ x_1 - x_2 + x_3 \qquad\qquad\quad + 2x_6 &= 0, \\ x_1 + x_2 + 2x_3 + 4x_4 + x_5 + 2x_6 &= 3, \\ 2x_2 + x_3 + 4x_4 + 3x_5 + x_6 &= 2. \end{aligned} \tag{14}$$

Add -1 times the first equation to the second (eq.2 $\to$ eq.2 -1 eq.1) and third (eq.3 $\to$ eq.3 -1 eq.1) equations:

$$\begin{aligned} x_1 - 3x_2 \qquad\quad - 4x_4 - 2x_5 + x_6 &= 0, \\ 2x_2 + x_3 + 4x_4 + 2x_5 + x_6 &= 0, \\ 4x_2 + 2x_3 + 8x_4 + 3x_5 + x_6 &= 3, \\ 2x_2 + x_3 + 4x_4 + 3x_5 + x_6 &= 2. \end{aligned} \tag{15}$$

Add -2 times the second to the third (eq.3 $\to$ eq.3 -2 eq.2) and -1 times the second to the fourth (eq.4 $\to$ eq.4 -1 eq.2):

$$\begin{aligned} x_1 - 3x_2 \qquad\quad - 4x_4 - 2x_5 + x_6 &= 0, \\ 2x_2 + x_3 + 4x_4 + 2x_5 + x_6 &= 0, \\ -x_5 - x_6 &= 3, \\ x_5 &= 2. \end{aligned} \tag{16}$$

Add the third to the fourth (eq.4 $\to$ eq.4 + eq.3):

$$\begin{aligned} x_1 - 3x_2 \qquad\quad - 4x_4 - 2x_5 + x_6 &= 0, \\ 2x_2 + x_3 + 4x_4 + 2x_5 + x_6 &= 0, \\ -x_5 - x_6 &= 3, \\ -x_6 &= 5. \end{aligned} \tag{17}$$

Finally, multiply the second, third, and fourth by $\frac{1}{2}$, -1, and -1, respectively, to normalize the leading coefficients (eq.2 $\to \frac{1}{2}$ eq.2, eq.3 $\to -1$ eq.3, eq.4 $\to -1$ eq.4):

$$\begin{aligned} x_1 - 3x_2 \qquad\quad - 4x_4 - 2x_5 + x_6 &= 0, \\ x_2 + \tfrac{1}{2}x_3 + 2x_4 + x_5 + \tfrac{1}{2}x_6 &= 0, \\ x_5 + x_6 &= -3, \\ x_6 &= -5. \end{aligned} \tag{18}$$

The last two equations give $x_6 = -5$ and $x_5 = 2$, and these values can be substituted back into the second equation. In that equation we can let x_4 be arbitrary, say α_1, and we can also let x_3 be arbitrary, say α_2. Then that equation gives x_2 and, again by back substitution, the first equation gives x_1. The result is the infinity of solutions

$$\begin{aligned} x_6 = -5, \quad x_5 = 2, \quad x_4 = \alpha_1, \quad x_3 = \alpha_2, \\ x_2 = \frac{1}{2} - 2\alpha_1 - \frac{1}{2}\alpha_2, \quad x_1 = \frac{21}{2} - 2\alpha_1 - \frac{3}{2}\alpha_2, \end{aligned} \tag{19}$$

where α_1 and α_2 are arbitrary. ∎

If a solution set contains p independent arbitrary parameters $(\alpha_1, \ldots, \alpha_p)$, we call it (in this text) a **p-parameter family of solutions**. Thus, (12) and (19) are

one- and two-parameter families of solutions, respectively. Each choice of values for $\alpha_1, \ldots, \alpha_p$ yields a **particular solution**. In (19), for instance, the choice $\alpha_1 = 1$ and $\alpha_2 = 0$ yields the particular solution $x_1 = \frac{17}{2}$, $x_2 = -\frac{3}{2}$, $x_3 = 0$, $x_4 = 1$, $x_5 = 2$, and $x_6 = -5$.

8.3.2. Gauss elimination. The method of Gauss elimination,* illustrated in Examples 1–3, can be applied to *any* linear system (1), whether or not the system is consistent, and whether or not the solution is unique. Though hard to tell from the foregoing hard calculations, the method is efficient and is commonly available in computer systems.

Observe that the end result of the Gauss elimination process enables us to determine, merely from the pattern of the final equations, whether or not a solution exists and is unique. For instance, we can see from the pattern of (5) that there is a unique solution, from the bottom equation in (9) that there no solution, and from the extra double indentation in (18) that there is a two-parameter family of solutions.

As representative of the case where $m < n$, let $m = 3$ and $n = 5$. There are four possible final patterns, and these are shown schematically in Fig. 1. For instance, the third equation in Fig. 1a could be $x_3 - 6x_4 + 2x_5 = 0$ or $x_3 + 2x_4 + 0x_5 = 4$, and the given third equation in Fig. 1b could be $0 = 6$ or $0 = 0$. It may seem foolish to include the case shown in Fig. 1d because there are no x_j's (all of the a_{ij} coefficients being zero), but it is *possible* so we have included it. From these patterns we draw these conclusions: (a) there exists a two-parameter family of solutions; (b) there is no solution (the system is inconsistent) if the right-hand member of the third equation is nonzero, and a three-parameter family of solutions if the latter is zero; (c) there is no solution if either of the right-hand members of the second and third equations is nonzero, and a four-parameter family of solutions if each of them is zero; (d) there is no solution if any of the right-hand members is nonzero, and a five-parameter family of solutions if each of them is zero.

It may appear that Fig. 1 does not cover all possible cases. For instance, what about the case shown in Fig. 2? That case can be converted to the case shown in Fig. 1a simply by renaming the unknowns: let x_3 become x_2 and let x_5 become x_3. Specifically, let $x_1 \to x_1$, $x_3 \to x_2$, $x_5 \to x_3$, $x_4 \to x_4$, and $x_2 \to x_5$.

The case where $m \geq n$ can be studied in a similar manner, and we can draw the following general conclusions.

```
(a)  • • • • • = •
       • • • • = •
         • • • = •

(b)  • • • • • = •
       • • • • = •
               0 = •

(c)  • • • • • = •
               0 = •
               0 = •

(d)            0 = •
               0 = •
               0 = •
```

Figure 1. The final pattern; $m = 3$, $n = 5$.

```
• • • • • = •
      • • • = •
            • = •
```

Figure 2. Was this case not covered?

THEOREM 8.3.2 *Existence / Uniqueness for Linear Systems*
If $m < n$, the system (1) can be consistent or inconsistent. If it is consistent it cannot have a unique solution; it will have a p-parameter family of solutions, where $n - m \leq p \leq n$. If $m \geq n$, (1) can be consistent or inconsistent. If it is

*The method is attributed to *Karl Friedrich Gauss* (1777–1855), who is generally regarded as the foremost mathematician of the nineteenth century and often referred to as the "prince of mathematicians."

consistent it can have a unique solution or a p-parameter family of solutions, where $1 \leq p \leq n$.

The next theorem follows immediately from Theorem 8.3.2, but we state it separately for emphasis.

THEOREM 8.3.3 *Existence/Uniqueness for Linear Systems*
Every system (1) necessarily admits no solution, a unique solution, or an infinity of solutions.

Observe that a system (1) is inconsistent only if, in its Gauss-eliminated form, one or more of the equations is of the form zero equal to a nonzero number. But that can never happen if every c_j in (1) is zero, that is, if (1) is **homogeneous**.

THEOREM 8.3.4 *Existence/Uniqueness for Homogeneous Systems*
Every homogeneous linear system of m equations in n unknowns is consistent. Either it admits the unique trivial solution or else it admits an infinity of nontrivial solutions in addition to the trivial solution. If $m < n$, then there is an infinity of nontrivial solutions in addition to the trivial solution.

In summary, not only did the method of Gauss elimination provide us with an efficient and systematic solution procedure, it also led us to important results regarding the existence and uniqueness of solutions.

8.3.3. Matrix notation. In applying Gauss elimination, we quickly discover that writing the variables $x_1, \ldots, x_n$ over and over is inefficient, and even tends to upstage the more central role of the a_{ij}'s and c_j's. It is therefore preferable to omit the x_j's altogether and to work directly with the rectangular array

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} & c_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & c_2 \\ \vdots & \vdots & & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} & c_m \end{bmatrix}, \tag{20}$$

known as the **augmented matrix** of the system (1), that is, the **coefficient matrix**

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \tag{21}$$

augmented by the column of c_j's. By *matrix* we simply mean a rectangular array of numbers, called **elements**; it is customary to enclose the elements between parentheses to emphasize that the entire matrix is regarded as a single entity. A horizontal line of elements is called a **row**, and a vertical line is called a **column**. Counting rows from the top, and columns from the left,

$$a_{21} \quad a_{22} \quad \cdots \quad a_{2n} \quad c_2 \qquad \text{and} \qquad \begin{matrix} c_1 \\ c_2 \\ \vdots \\ c_m \end{matrix},$$

say, are the second row and $(n+1)$th column, respectively, of the augmented matrix (20).

In terms of the abbreviated matrix notation, the calculation in Example 1 would look like this.

Original system:

$$\begin{bmatrix} 1 & 1 & -1 & 1 \\ 3 & 1 & 1 & 9 \\ 1 & -1 & 4 & 8 \end{bmatrix}.$$

Add -3 times first row to second row, and add -1 times first row to third row:

$$\begin{bmatrix} 1 & 1 & -1 & 1 \\ 0 & -2 & 4 & 6 \\ 0 & -2 & 5 & 7 \end{bmatrix}.$$

Add -1 times second row to third row, and multiply second row by $-\frac{1}{2}$:

$$\begin{bmatrix} 1 & 1 & -1 & 1 \\ 0 & 1 & -2 & -3 \\ 0 & 0 & 1 & 1 \end{bmatrix}. \tag{22}$$

Thus, corresponding to the so-called elementary equation operations on members of a system of linear equations there are **elementary row operations** on the augmented matrix, as follows:

1. Addition of a multiple of one row to another:
   Symbolically: (jth row) $\to$ (jth row) $+ \alpha$(kth row)

2. Multiplication of a row by a nonzero constant:
   Symbolically: (jth row) $\to \alpha$(jth row)

3. Interchange of two rows:
   Symbolically: (jth row) $\leftrightarrow$ (kth row)

And we say that two matrices are **row equivalent** if one can be obtained from the other by finitely many elementary row operations.

8.3.4. Gauss–Jordan reduction. With the Gauss elimination completed, the remaining steps consist of back substitution. In fact, those steps are elementary row operations as well. The difference is that whereas in the Gauss elimination we proceed from the top down, in the back substitution we proceed from the bottom up.

EXAMPLE 4. To illustrate, let us return to Example 1 and pick up at the end of the Gauss elimination, with (5), and complete the back substitution steps using elementary row operations. In matrix format, we begin with

$$\begin{bmatrix} 1 & 1 & -1 & 1 \\ 0 & 1 & -2 & -3 \\ 0 & 0 & 1 & 1 \end{bmatrix}. \tag{23}$$

Keeping the bottom row intact, add 2 times that row to the second, and add 1 times that row to the first:

$$\begin{bmatrix} 1 & 1 & 0 & 2 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 1 \end{bmatrix}. \tag{24}$$

Now keeping the bottom two rows intact, add -1 times the second row to the first:

$$\begin{bmatrix} 1 & 0 & 0 & 3 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 1 \end{bmatrix}, \tag{25}$$

which is the solution: $x_1 = 3$, $x_2 = -1$, $x_3 = 1$ as obtained in Example 1. ∎

The entire process, of Gauss elimination plus back substitution, is known as **Gauss–Jordan reduction**, after Gauss and *Wilhelm Jordan* (1842–1899). The final result is an augmented matrix in **reduced row-echelon form**. That is:

1. In each row not made up entirely of zeros, the first nonzero element is a 1, a so-called **leading 1**.
2. In any two consecutive rows not made up entirely of zeros, the leading 1 in the lower row is to the right of the leading 1 in the upper row.
3. If a column contains a leading 1, every other element in that column is a zero.
4. All rows made up entirely of zeros are grouped together at the bottom of the matrix.

For instance, (25) is in reduced row-echelon form, as is the final matrix in the next example.

EXAMPLE 5. Let us return to Example 3 and finish the Gauss–Jordan reduction, beginning with (18):

$$\begin{bmatrix} 1 & -3 & 0 & -4 & -2 & 1 & 0 \\ 0 & 1 & \frac{1}{2} & 2 & 1 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & -3 \\ 0 & 0 & 0 & 0 & 0 & 1 & -5 \end{bmatrix} \to \begin{bmatrix} 1 & 0 & \frac{3}{2} & 2 & 1 & \frac{5}{2} & 0 \\ 0 & 1 & \frac{1}{2} & 2 & 1 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & -3 \\ 0 & 0 & 0 & 0 & 0 & 1 & -5 \end{bmatrix} \to$$

$$\begin{bmatrix} 1 & 0 & \frac{3}{2} & 2 & 0 & \frac{3}{2} & 3 \\ 0 & 1 & \frac{1}{2} & 2 & 0 & -\frac{1}{2} & 3 \\ 0 & 0 & 0 & 0 & 1 & 1 & -3 \\ 0 & 0 & 0 & 0 & 0 & 1 & -5 \end{bmatrix} \to \begin{bmatrix} \mathbf{1} & 0 & \frac{3}{2} & 2 & 0 & 0 & \frac{21}{2} \\ 0 & \mathbf{1} & \frac{1}{2} & 2 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \mathbf{1} & 0 & 2 \\ 0 & 0 & 0 & 0 & 0 & \mathbf{1} & -5 \end{bmatrix}.$$

The last augmented matrix is in reduced row-echelon form. The four leading 1's are displayed in **bold type**, and we see that, as a result of the back substitution steps, only 0's are to be found above each leading 1. The final augmented matrix once again gives the solution (19). ∎

8.3.5. Pivoting. Recall that the first step in the Gauss elimination of the system

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= c_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= c_2, \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= c_m, \end{aligned} \tag{26}$$

is to subtract a_{21}/a_{11} times the first equation from the second, a_{31}/a_{11} times the first equation from the third, and so on, while keeping the first equation intact. The first equation is called the **pivot equation** (or, the first row is the pivot row if one is using the matrix format), and a_{11} is called the **pivot**. That step produces an indented system of the form

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= c_1, \\ a'_{22}x_2 + \cdots + a'_{2n}x_n &= c'_2, \\ &\vdots \\ a'_{m2}x_2 + \cdots + a'_{mn}x_n &= c'_m. \end{aligned} \tag{27}$$

Next, we keep the first *two* equations intact and use the second equation as the new pivot equation to indent the third through mth equations, and so on.

Naturally, we need each pivot to be nonzero. For instance, we need $a_{11} \neq 0$ for $a_{21}/a_{11}, a_{31}/a_{11}, \ldots$ to be defined. If a pivot is zero, interchange that equation

with any one below it, such as the next equation or last equation (as we did in Example 3), until a nonzero pivot is available. Such interchange of equations is called **partial pivoting**. If a pivot is zero we have no choice but to use partial partial pivoting, but in practice even a nonzero pivot should be rejected if it is "very small," since the smaller it is the more susceptible is the calculation to the adverse effect of machine roundoff error (see Exercise 13). To be as safe as possible, one can choose the pivot equation as the one with the largest leading coefficient (relative to the other coefficients in the equation).

Closure. Beginning with a system of coupled linear algebraic equations, one can use a sequence of elementary operations to minimize the coupling between the equations while leaving the solution set intact. Besides putting forward the important method of Gauss elimination, which is used heavily in the following chapters, we used that method to establish several important theoretical results regarding the existence and uniqueness of solutions.

The Gauss elimination and Gauss–Jordan reduction discussions lead naturally to a convenient, and equivalent, formulation in matrix notation. We will return to the concept of matrices in Chapter 10, and develop it in detail.

Computer software. Chapters 8–12 cover the domain known as linear algebra. A great many calculations in linear algebra can be carried out using computer algebra systems. In *Maple*, for instance, a great many commands ("functions") are contained within the **linalg** package. A listing of these commands can be obtained by entering ?linalg. That list includes the **linsolve** command, which can be used to solve a system of m linear algebraic equations in n unknowns. To access linsolve (or any other command within the linalg package), first enter with(linalg). Then, linsolve(A,b) solves (1) for $x_1, \ldots, x_n$, where A is the coefficient matrix and b is the column of c_j's. For instance, the system

$$\begin{aligned} x_1 - x_2 + 2x_3 - 3x_4 &= 4, \\ x_1 + 2x_2 - x_3 + 3x_4 &= 1 \end{aligned} \tag{28}$$

admits the two-parameter family of solutions

$$x_4 = \alpha_1, \quad x_3 = \alpha_2, \quad x_2 = -1 - 2\alpha_1 + \alpha_2, \quad x_1 = 3 + \alpha_1 - \alpha_2, \tag{29}$$

where α_1, α_2 are arbitrary. To solve (28) using *Maple*, enter

with(linalg):

then return and enter

linsolve(array([[1, −1, 2, −3], [1, 2, −1, 3]]), array([4, 1]));

and return. The output is

$$[-_t_1 + _t_2 + 3, \ _t_1 - 2_t_2 - 1, \ _t_1, \ _t_2]$$

where the entries are $x_1, \ldots, x_4$ and where $_t_1$ and $_t_2$ are arbitrary constants. With $_t_1 = \alpha_2$ and $_t_2 = \alpha_1$, this result is the same as (29). If you prefer, you could use the sequence

```
with(linalg):
A := array([[1, -1, 2, -3], [1, 2, -1, 3]]):
b := array([4, 1]):
linsolve(A, b);
```

instead. If the system is inconsistent, then either the output will be NULL, or there will be no output.

EXERCISES 8.3

1. Derive the solution set for each of the following systems using Gauss elimination and augmented matrix format. Document each step (e.g., 2nd row→ 2nd row + 5 times 1st row), and classify the result (e.g., unique solution, the system is inconsistent, 3-parameter family of solutions, etc.).

(a) $\begin{aligned} 2x - 3y &= 1 \\ 5x + y &= 2 \end{aligned}$

(b) $\begin{aligned} 2x + y &= 0 \\ 3x - 2y &= 0 \end{aligned}$

(c) $x + 2y = 4$

(d) $\begin{aligned} x - y + z &= 1 \\ 2x - y - z &= 8 \end{aligned}$

(e) $2x_1 - x_2 - x_3 - 5x_4 = 6$

(f) $\begin{aligned} 2x_1 - x_2 - x_3 - 3x_4 &= 0 \\ x_1 - x_2 + 4x_3 &= 2 \end{aligned}$

(g) $\begin{aligned} x + 2y + 3z &= 4 \\ 5x + 6y + 7z &= 8 \\ 9x + 10y + 11z &= 12 \end{aligned}$

(h) $\begin{aligned} x_1 + x_2 - 2x_3 &= 3 \\ x_1 - x_2 - 3x_3 &= 1 \\ x_1 - 3x_2 - 4x_3 &= -1 \end{aligned}$

(i) $\begin{aligned} 2x_1 - x_2 &= 6 \\ 3x_1 + 2x_2 &= 4 \\ x_1 + 10x_2 &= -12 \\ 6x_1 + 11x_2 &= -2 \end{aligned}$

(j) $\begin{aligned} x_1 - x_2 + 2x_3 + x_4 &= -1 \\ 2x_1 + x_2 + x_3 - x_4 &= 4 \\ x_1 + 2x_2 - x_3 - 2x_4 &= 5 \\ x_1 + x_3 &= 1 \end{aligned}$

(k) $\begin{aligned} x_1 + x_3 &= 1 \\ x_1 + 2x_2 - x_3 - 2x_4 &= 5 \\ x_1 - x_2 + 2x_3 + x_4 &= 0 \\ 2x_1 + x_2 + x_3 - x_4 &= 4 \end{aligned}$

(l) $\begin{aligned} x_3 + x_4 &= 2 \\ 4x_2 - x_3 + x_4 &= 0 \\ x_1 - x_2 + 2x_3 + x_4 &= 4 \end{aligned}$

(m) $\begin{aligned} x + 2y + 3z &= 5 \\ 2x + 3y + 4z &= 8 \\ 3x + 4y + 5z &= c \\ x + y &= 2 \end{aligned}$

for $c = 10$, and again, for $c = 11$

(n) $\begin{aligned} 2x + y + z &= 10 \\ 3x + y - z &= 6 \\ x - 2y - 4z &= -10 \end{aligned}$

(o) $\begin{aligned} 2x_1 + x_2 &= 1 \\ x_1 + 2x_2 + x_3 &= 1 \\ x_2 + 2x_3 + x_4 &= 1 \\ x_3 + 2x_4 &= 1 \end{aligned}$

(p) $\begin{aligned} 2x_1 + x_2 &= 0 \\ x_1 + 2x_2 + x_3 &= -1 \\ x_2 + 2x_3 &= -4 \end{aligned}$

(q) $\begin{aligned} 2x_1 + x_2 + x_4 + 2x_5 &= 0 \\ x_1 + x_2 - x_3 &= 0 \\ x_1 + x_2 + x_3 - 3x_4 + 2x_5 &= 0 \\ 2x_1 + 2x_2 - x_3 + x_5 &= 0 \end{aligned}$

2. (a)–(q) Same as Exercise 1 but using Gauss–Jordan reduction instead of Gauss elimination.

3. (a)–(q) Same as Exercise 1 but using computer software such as the *Maple* linsolve command.

4. Can 20 linear algebraic equations in 14 unknowns have a unique solution? Be inconsistent? Have a two-parameter family of solutions? Have a 14-parameter family of solutions? Have a 16-parameter family of solutions? Explain.

5. Let

$$\begin{aligned} a_1x_1 + a_2x_2 + a_3x_3 &= 0, \\ b_1x_1 + b_2x_2 + b_3x_3 &= 0 \end{aligned}$$

represent any two planes through the origin in a Cartesian x_1, x_2, x_3 space. For the case where the planes intersect along a line, show whether or not that line necessarily passes through the origin.

6. If possible, adapt the methods of this section to solve the following *nonlinear* systems. If it is *not* possible, say so.

(a) $\begin{aligned} x_1^2 + 2x_2^2 - x_3^2 &= 29 \\ x_1^2 + x_2^2 + x_3^2 &= 19 \\ 3x_1^2 + 4x_2^2 \qquad &= 67 \end{aligned}$

(b) $\begin{aligned} x + 3y &= 13 \\ \sin x + 2y &= 5 \end{aligned}$

(c) $\begin{aligned} \sin x + \sin y \qquad &= 1 \\ \sin x - \sin y + 4\cos z &= 1.2 \\ \sin x + \sin y + 2\cos z &= 1.6 \end{aligned}$

where $-\pi/2 \le x \le \pi/2$, $-\pi/2 \le y \le \pi/2$, and $0 \le z \le 2\pi$.

7. For what values of the parameter λ do the following homogeneous (do you agree that they are homogeneous?) systems admit *non*trivial solutions? Find the nontrivial solutions corresponding to each such λ.

(a) $\begin{aligned} 2x + y &= \lambda x \\ x + 2y &= \lambda y \end{aligned}$

(b) $\begin{aligned} 2x - y &= \lambda x \\ -x + 2y &= \lambda y \end{aligned}$

(c) $\begin{aligned} x - 2y &= \lambda x \\ 4x - 8y &= \lambda y \end{aligned}$

(d) $\begin{aligned} z &= \lambda x \\ z &= \lambda y \\ x + y + z &= \lambda z \end{aligned}$

(e) $\begin{aligned} x + y + z &= \lambda x \\ y + z &= \lambda y \\ 2z &= \lambda z \end{aligned}$

(f) $\begin{aligned} 2x + y + z &= \lambda x \\ x + 2y + z &= \lambda y \\ x + y + 2z &= \lambda z \end{aligned}$

8. Evaluate these excerpts from examination papers.

(a) "Given the system

$$\begin{aligned} x_1 - 2x_2 &= 0, \\ 2x_1 - 4x_2 &= 0, \end{aligned}$$

add -2 times the first equation to the second and add $-\frac{1}{2}$ times the second equation to the first. By these Gauss elimination steps we obtain the equivalent system $0 = 0$ and $0 = 0$, and hence the two-parameter family of solutions $x_1 = \alpha_1$ (arbitrary), $x_2 = \alpha_2$ (arbitrary)."

(b) "Given the system

$$\begin{aligned} x_1 + x_2 - 4x_3 &= 0, \\ 2x_1 - x_2 + x_3 &= 0, \end{aligned}$$

since both left-hand sides equal zero, they must equal each other. Hence we have the equation

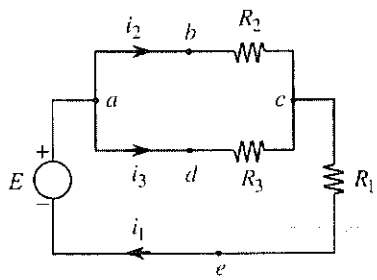
$$x_1 + x_2 - 4x_3 = 2x_1 - x_2 + x_3,$$

which equation is equivalent to the original system."

9. Make up an example of an inconsistent linear algebraic system of equations, with

(a) $m = 2, n = 4$ (b) $m = 1, n = 4$

10. *(Physical example of nonexistence and nonuniqueness; DC circuit)* Kirchoff's current and voltage laws were given in Section 2.3.1. If we apply those laws to the DC circuit shown,



we obtain the equations

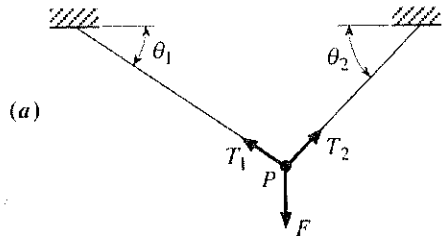
$$\begin{aligned} i_1 - i_2 - i_3 &= 0, && \text{(junction } a) \\ i_1 - i_2 - i_3 &= 0, && \text{(junction } c) \\ R_2i_2 - R_3i_3 &= 0, && \text{(loop } abcda) \\ R_1i_1 + R_2i_2 &= E, && \text{(loop } abcea) \\ R_1i_1 + R_3i_3 &= E, && \text{(loop } adcea) \end{aligned} \tag{10.1}$$

where i_1, i_2, i_3 are the three currents (measured as positive in the direction assumed in the figure). R_1, R_2, R_3 are the resistances of the three resistors, and E is the voltage rise (from e to a) induced by the battery or other source. [Evidently, we did not need to apply the current law to both junctions since the resulting equations are identical. Similarly, it may be that not all of the loop equations are needed. But rather than try to decide which of equations (10.1) to keep and which to discard, let us merely keep them all.] We now state the problem: Obtain the solution set of equations (10.1) by Gauss elimination. If there is no solution, or if there is a nonunique solution, explain that result in physical terms. Take

(a) $R_1 = R_2 = R_3 \equiv R \quad (R \neq 0)$
(b) $R_1 = R_2 \equiv R,\ R_2 = 2R \quad (R \neq 0)$
(c) $R_1 \equiv R,\ R_2 = R_3 = 2R \quad (R \neq 0)$

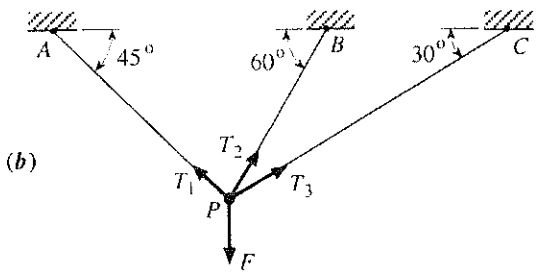
(d) $R_1 \equiv R,\ R_2 = 4R,\ R_3 = 6R \quad (R \neq 0)$
(e) $R_2 \equiv R,\ R_1 = R_3 = 0$
(f) $R_1 \equiv R,\ R_2 = R_3 = 0 \quad (R \neq 0)$
(g) $R_1 = R_2 = R_3 = 0$

11. *(Physical example of nonexistence and nonuniqueness: statically indeterminate structures)* (a) Consider the static equilibrium of the system shown, consisting of two weightless



cables connected at P, at which point a vertical load F is applied. Requiring an equilibrium of vertical force components, and horizontal force components too, derive two linear algebraic equations on the unknown tensions T_1 and T_2. Are there any combinations of angles θ_1 and θ_2 (where $0 \leq \theta_1 \leq \frac{\pi}{2}$ and $0 \leq \theta_2 \leq \frac{\pi}{2}$) such that there is either no solution or a nonunique solution? Explain.

(b) This time let there be three cables at angles of 45°, 60°, and 30° as shown. Again, requiring an equilibrium of vertical and



horizontal forces at P, derive two linear algebraic equations on the unknown tensions T_1, T_2, T_3. Show that the equations are consistent so there is a nonunique solution. NOTE: We say that such a structure is **statically indeterminate** because the forces in it cannot be determined from the laws of statics alone. What information needs to be added if we are to complete the evaluation of T_1, T_2, T_3? What is needed is information about the relative stiffness of the cables. We pursue this to a conclusion in (c), below.

(c) [*Completion of part (b)*] Before the load F is applied, locate an x, y Cartesian coordinate system at P. Let P be 1 foot below the "ceiling" so the coordinates of A, B, C are $(-1, 1)$, $(1/\sqrt{3}, 1)$, and $(\sqrt{3}, 1)$, respectively. Now apply the load F. The point P will move to a point (x, y), and we assume that the cables are stiff enough so that x and y are very small: $|x| \ll 1$ and $|y| \ll 1$. Let the cables obey Hooke's law: $T_1 = k_1\delta_1$, $T_2 = k_2\delta_2$, and $T_3 = k_3\delta_3$, where δ_j is the increase in length of the jth cable due to the tension T_j. Since P moves to (x, y), it follows that

$$\begin{aligned}
\delta_1 &= \sqrt{(x+1)^2 + (y-1)^2} - \sqrt{2} \\
&= \sqrt{2 + 2(x-y) + (x^2+y^2)} - \sqrt{2} \\
&\approx \sqrt{2 + 2(x-y)} - \sqrt{2} \\
&= \sqrt{2}\,[1 + (x-y)]^{1/2} - \sqrt{2} \\
&\approx \sqrt{2}\left[1 + \frac{1}{2}(x-y)\right] - \sqrt{2} = \frac{1}{\sqrt{2}}(x-y).
\end{aligned} \tag{11.1}$$

Explain each step in (11.1), and show, similarly, that

$$\delta_2 \approx -\frac{1}{2}x - \frac{\sqrt{3}}{2}y, \tag{11.2}$$

$$\delta_3 \approx -\frac{\sqrt{3}}{2}x - \frac{1}{2}y. \tag{11.3}$$

Thus,

$$\begin{aligned}
T_1 &= k_1\delta_1 \approx \frac{k_1}{\sqrt{2}}(x-y), \\
T_2 &= k_2\delta_2 \approx -\frac{k_2}{2}(x + \sqrt{3}y), \\
T_3 &= k_3\delta_3 \approx -\frac{k_3}{2}(\sqrt{3}x + y).
\end{aligned} \tag{11.4}$$

Putting (11.4) into the two equilibrium equations obtained in (b) then gives two equations in the unknown displacements x, y. Show that that system can be solved uniquely for x and y, and thus complete the solution for T_1, T_2, T_3.

12. *(Roundoff error difficulty due to small pivots)* To illustrate how small pivots can accentuate the effects of roundoff error, consider the system

$$\begin{aligned}
0.005x_1 + 1.47x_2 &= 1.49, \\
0.975x_1 + 2.32x_2 &= 6.22
\end{aligned} \tag{12.1}$$

with exact solution $x_1 = 4$ and $x_2 = 1$. Suppose that our computer carries three significant figures and then rounds off. Using the first equation as our pivot equation, Gauss elimination gives

$$\begin{bmatrix} 0.005 & 1.47 & 1.49 \\ 0.975 & 2.32 & 6.22 \end{bmatrix} \to \begin{bmatrix} 0.005 & 1.47 & 1.49 \\ 0 & -285 & -284 \end{bmatrix}$$

so $x_2 = 284/285 = 0.996$ and $x_1 = [1.49 - (1.47)(0.996)]/0.005 = (1.49 - 1.46)/0.005 = 6$. Show that if we use partial pivoting and then use the first equation of the system

$$\begin{aligned} 0.975x_1 + 2.32x_2 &= 6.22, \\ 0.005x_1 + 1.47x_2 &= 1.49 \end{aligned} \tag{12.2}$$

as our pivot equation, we obtain the result $x_1 = 4.00$ and $x_2 = 1.00$ (which happens to be exactly correct).

13. (*Ill-conditioned systems*) Practically speaking, our numerical calculations are normally carried out on computers, be they hand-held calculators or large digital computers. Such machines carry only a finite number of significant figures and thus introduce *roundoff error* into most calculations. One might expect (or hope) that such slight deviations will lead to answers that are only slightly in error. For example, the solution of

$$\begin{aligned} x + \quad y &= 2, \\ x - 1.014y &= 0 \end{aligned} \tag{13.1}$$

is $x \approx 1.007$, $y \approx 0.993$, whereas the solution of the rounded-off version

$$\begin{aligned} x + \quad y &= 2, \\ x - 1.01y &= 0 \end{aligned}$$

is very much the same, namely $x \approx 1.005$, $y \approx 0.995$. In sharp contrast, the solutions of

$$\begin{aligned} x + \quad y &= 2, \\ x + 1.014y &= 0 \end{aligned} \tag{13.2}$$

and the rounded-off version

$$\begin{aligned} x + \quad y &= 2, \\ x + 1.01y &= 0, \end{aligned}$$

$x \approx 144.9$, $y \approx -142.9$ and $x = 202$, $y = -200$, respectively; (13.2) is an example of a so-called **ill-conditioned** system (ill-conditioned in the sense that small changes in the coefficients lead to large changes in the solution). Here, we ask the following: Explain why (13.2) is much more sensitive to roundoff than (13.1) by exploring the two cases graphically, that is, in the x, y plane.

Chapter 8 Review

This chapter deals with systems of linear algebraic equations, m equations in n unknowns, insofar as the existence and uniqueness of solutions and solution technique. We find that there are three possibilities: a unique solution, no solution, and an infinity of solutions. If one or more solutions exist then the system is said to be consistent, if there are no solutions then it is inconsistent.

The key, in assessing existence/uniqueness as well as in finding solutions, is provided by elementary operations because they enable us to manipulate the system so as to reduce the coupling to a minimum, while at the same time keeping the solution set intact.

The method of Gauss elimination is introduced, as a systematic solution procedure based upon the three elementary operations, and it is shown that the subsequent back substitution steps amount to elementary operations as well. The entire process, Gauss elimination followed by the back substitution, is known as Gauss–Jordan reduction. Realize that the latter is a solution method, or algorithm, not a formula for the solution. Explicit solution formulas are developed, but not until Chapter 10.

We find that the process of Gauss elimination and Gauss–Jordan reduction are expressed most conveniently in matrix notation, although that notation is not essential to the method. In subsequent chapters the matrix approach is developed

more fully.

The most important results of this chapter are contained in Theorems 8.3.1–3. Finally, we also stress the value of geometrical and visual reasoning, and suggest that you keep that idea in mind as we proceed.

Chapter 9

Vector Space

9.1 Introduction

Normally, one meets vectors for the first time within some physical context – in studying mechanics, electric and magnetic fields, and so on. There, the vectors exist within two- or three-dimensional space and correspond to force, velocity, position, magnetic field, and so on. They have both *magnitude* and *direction*; they can be *scaled* by multiplicative factors, *added* according to the parallelogram law; *dot* and *cross product* operations are defined between vectors; the angle between two vectors is defined; vectors can be *expanded* as linear combinations of *base vectors*; and so on.

Alternatively, there exists a highly formalized axiomatic approach to vectors known as *linear vector space* or *abstract vector space*. Although this generalized vector concept is essentially an outgrowth of the more primitive system of "arrow vectors" in 2-space and 3-space, described above, it extends well beyond that system in scope and applicability.

For pedagogical reasons, we break the transition from 2-space and 3-space to abstract vector space into two steps: in Sections 9.4 and 9.5 we introduce a generalization to "n-space," and in Section 9.6 we complete the extension to general vector space, including *function spaces* where the vectors are functions! However, we do not return to function spaces until Chapter 17, in connection with Fourier series and the Sturm–Liouville theory; in Chapters 9–12 our chief interest is in n-space.

9.2 Vectors; Geometrical Representation

Some quantities that we encounter may be completely defined by a single real number, or magnitude; the mass or kinetic energy of a given particle, and the temperature or salinity at some point in the ocean, are examples. Others are not defined solely by a magnitude but rather by a magnitude and a direction, examples being force, velocity, momentum, and acceleration. Such quantities are called

vectors.

The defining features of a vector being magnitude and direction suggests the geometric representation of a vector as a directed line segment, or "arrow," where the length of the arrow is scaled according to the magnitude of the vector. For example, if the wind is blowing at 8 meters/sec from the northeast, that defines a wind-velocity vector $\mathbf{v}$, where we adopt **boldface type** to signify that the quantity is a vector; alternative notations include the use of an overhead arrow as in $\vec{v}$. Choosing, according to convenience, a scale of 5 meters/sec per centimeter, say, the geometric representation of $\mathbf{v}$ is as shown in Fig. 1. Denoting the magnitude, or **norm**, of any vector $\mathbf{v}$ as $\|\mathbf{v}\|$, we have $\|\mathbf{v}\| = 8$ for the $\mathbf{v}$ vector in Fig. 1.

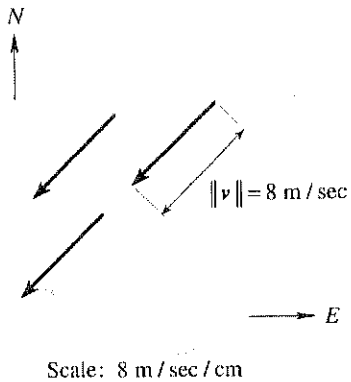


Figure 1. Geometric representation of $\mathbf{v}$.

Observe that the *location* of a vector is not specified, only its magnitude and direction. Thus, the two unlabeled arrows in Fig. 1 are equally valid alternative representations of $\mathbf{v}$. That is not to say that the physical *effect* of the vector will be entirely independent of its position. For example, it should be apparent that the motion of the body $\mathcal{B}$ induced by a force $\mathbf{F}$ (Fig. 2) will certainly depend on the point of application of $\mathbf{F}$* as will the stress field induced in $\mathcal{B}$. Nevertheless, the two vectors in Fig. 2 are still regarded as equal, as are the three in Fig. 1.

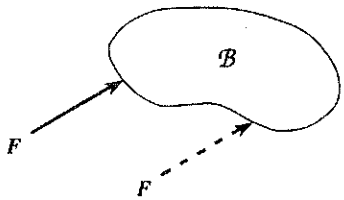


Figure 2. Position of a vector.

Like numbers, vectors do not become useful until we introduce rules for their manipulation, that is, a vector algebra. Having elected the arrow representation of vectors, the vector algebra that we now introduce will, likewise, be geometric.

First, we say that two vectors are **equal** if and only if their lengths are identical and if their directions are identical as well.

Next, we define a process of **addition** between any two vectors $\mathbf{u}$ and $\mathbf{v}$. The first step is to move $\mathbf{v}$ (if necessary), parallel to itself, so that its tail coincides with the head of $\mathbf{u}$. Then the sum, or *resultant*, $\mathbf{u} + \mathbf{v}$ is defined as the arrow from the tail of $\mathbf{u}$ to the head of $\mathbf{v}$, as in Fig. 3a. Reversing the order, $\mathbf{v} + \mathbf{u}$ is as shown in Fig. 3b. Equivalently, we may place $\mathbf{u}$ and $\mathbf{v}$ tail to tail, as in Fig. 3c. Comparing Fig. 3c with Fig. 3a and b, we see that the diagonal of the parallelogram (Fig. 3c), is both $\mathbf{u} + \mathbf{v}$ and $\mathbf{v} + \mathbf{u}$. Thus,

$$\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}, \tag{1}$$

so addition is commutative. One may show (Exercise 3) that it is associative as well,

$$(\mathbf{u} + \mathbf{v}) + \mathbf{w} = \mathbf{u} + (\mathbf{v} + \mathbf{w}). \tag{2}$$

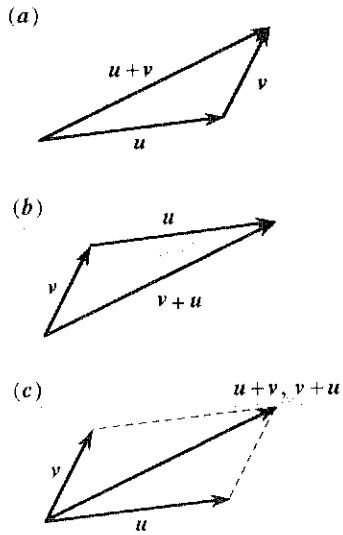


Figure 3. Vector addition.

Next, we define any vector of zero length to be a **zero vector**, denoted as $\mathbf{0}$. Its length being zero, its direction is immaterial; any direction may be assigned if desired. From the definition of addition above, it follows that

$$\mathbf{u} + \mathbf{0} = \mathbf{0} + \mathbf{u} = \mathbf{u} \tag{3}$$

for each vector $\mathbf{u}$.

Corresponding to $\mathbf{u}$ we define a **negative inverse** "$-\mathbf{u}$" such that if $\mathbf{u}$ is any nonzero vector, then $-\mathbf{u}$ is determined uniquely, as shown in Fig. 4a; that is, it is

*Students of mechanics know that the point of application of $\mathbf{F}$ affects the *rotational* part of the motion but not the *translational* part.

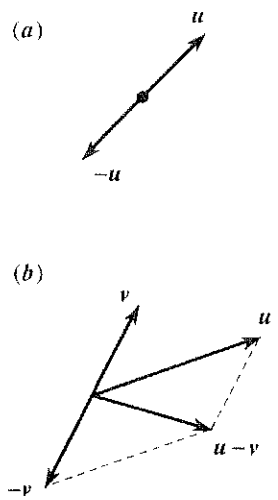


Figure 4. $-\mathbf{u}$ and vector subtraction.

of the same length as $\mathbf{u}$ but is directed in the opposite direction (again, $\mathbf{u}$ and $-\mathbf{u}$ have the same length, the length of $-\mathbf{u}$ is *not* negative). For the zero vector we have $-\mathbf{0} = \mathbf{0}$. We denote $\mathbf{u} + (-\mathbf{v})$ as $\mathbf{u} - \mathbf{v}$ ("$\mathbf{u}$ minus $\mathbf{v}$") but emphasize that it is really the addition of $\mathbf{u}$ and $-\mathbf{v}$, as in Fig. 4b.

Finally, we introduce another operation, called **scalar multiplication**, between any vector $\mathbf{u}$ and any scalar (i.e., a real number) α: If $\alpha \neq 0$ and $\mathbf{u} \neq \mathbf{0}$, then $\alpha\mathbf{u}$ is a vector whose length is $|\alpha|$ times the length of $\mathbf{u}$ and whose direction is the same as that of $\mathbf{u}$ if $\alpha > 0$, and the opposite if $\alpha < 0$; if $\alpha = 0$ and/or $\mathbf{u} = \mathbf{0}$, then $\alpha\mathbf{u} = \mathbf{0}$. This definition is illustrated in Fig. 5. It follows from this definition that scalar multiplication has the following algebraic properties:

$$\alpha(\beta\mathbf{u}) = (\alpha\beta)\mathbf{u}, \tag{4a}$$

$$(\alpha + \beta)\mathbf{u} = \alpha\mathbf{u} + \beta\mathbf{u}, \tag{4b}$$

$$\alpha(\mathbf{u} + \mathbf{v}) = \alpha\mathbf{u} + \alpha\mathbf{v}, \tag{4c}$$

$$1\mathbf{u} = \mathbf{u}, \tag{4d}$$

where α, β are any scalars and $\mathbf{u}, \mathbf{v}$ are any vectors.

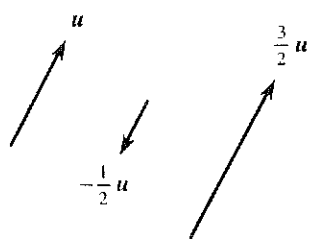


Figure 5. Scalar multiplication.

Observe that the parallelogram rule of vector addition is a definition so it does not need to be proved. Nevertheless, definitions are not necessarily fruitful so it is worthwhile to reflect for a moment on why the parallelogram rule has proved important and useful. Basically, if we say that "the sum of $\mathbf{u}$ and $\mathbf{v}$ is $\mathbf{w}$," and thereby pass from the two vectors $\mathbf{u}, \mathbf{v}$ to the single vector $\mathbf{w}$, it seems fair to expect some sort of equivalence to exist between the action of $\mathbf{w}$ and the joint action of $\mathbf{u}$ and $\mathbf{v}$. For example, if $\mathbf{F}_1$ and $\mathbf{F}_2$ are two forces acting on a body $\mathcal{B}$, as shown in Fig. 6, it is known from fundamental principles of mechanics that their combined effect will be the same as that due to the single force $\mathbf{F}$, so it seems reasonable and natural to say that $\mathbf{F}$ is the sum of $\mathbf{F}_1$ and $\mathbf{F}_2$. This concept goes back at least as far as *Aristotle* (384–322 B.C.). Thus, while the algebra of vectors is developed here as an essentially mathematical matter, it is important to appreciate the role of physics and physical motivation.

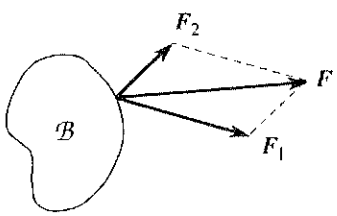


Figure 6. Physical motivation for parallelogram addition.

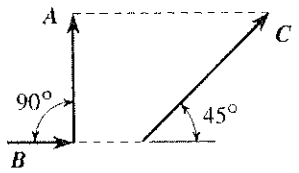
In closing this section, let us remark that our foregoing discussion should not be construed to imply that objects of physical interest are necessarily vectors (as are force and velocity) or scalars [as are temperature, mass, and speed (i.e., the *magnitude of the velocity vector*)]. For example, in the study of mechanics one finds that more than a magnitude and a direction are needed to fully define the state of stress at a point; in fact, a "second-order tensor" is needed – a quantity that is more exotic than a vector in much the same way that a vector is more exotic than a scalar.*

*For an introduction to tensors, we recommend to the interested reader the 68-page book *Tensor Analysis* by H. D. Block (Columbus, OH: Charles E. Merrill, 1962).

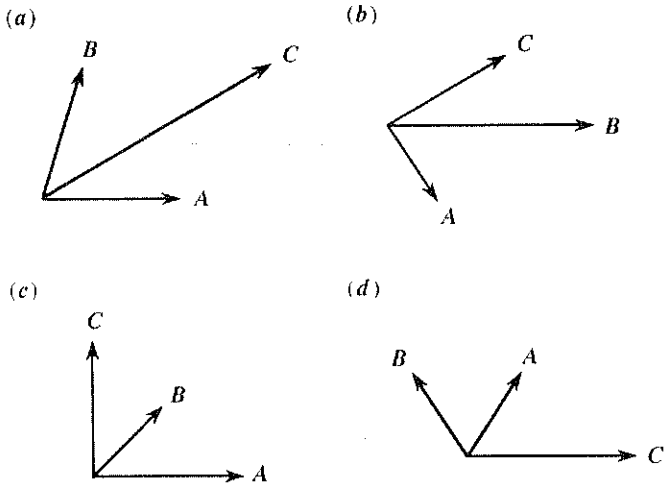
EXERCISES 9.2

1. Trace the vectors **A**, **B**, **C**, shown where **A** is twice as long as **B**. Then determine each of the following by graphical means.

(a) $\mathbf{A} + \mathbf{B} + \mathbf{C}$ (b) $\mathbf{B} - \mathbf{A}$
(c) $\mathbf{A} - \mathbf{C} + 3\mathbf{B}$ (d) $2(\mathbf{B} - \mathbf{A}) + 6\mathbf{0}$
(e) $\mathbf{A} + (4\mathbf{B} - \mathbf{C})$ (f) $\mathbf{A} + 2\mathbf{B} - 2\mathbf{C}$



2. In each case, **C** can be expressed as a linear combination of **A** and **B**, that is, as $\mathbf{C} = \alpha\mathbf{A} + \beta\mathbf{B}$. Trace the three vectors and by graphical means determine α and β.



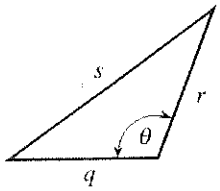
3. Show that the associative property (2) follows from the graphical definition of vector addition.

4. Derive the following from the definitions of vector addition and scalar multiplication:

(a) property (4a) (b) property (4b)
(c) property (4c) (d) property (4d)

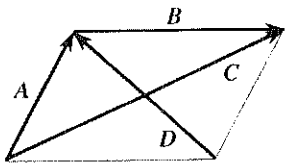
5. (a) If $\|\mathbf{A}\| = 1$, $\|\mathbf{B}\| = 2$, and $\|\mathbf{C}\| = 5$, can $\mathbf{A} + \mathbf{B} + \mathbf{C} = \mathbf{0}$?

HINT: Use the *law of cosines* $s^2 = q^2 + r^2 - 2qr\cos\theta$ (see the accompanying figure) or the Euclidean proposition that the length of any one side of a triangle cannot exceed the sum of the lengths of the other two sides.
(b) Repeat part (a), with "$\|\mathbf{A}\| = 1$" changed to $\|\mathbf{A}\| = 4$.

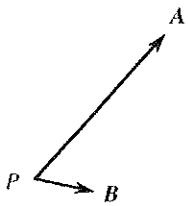


6. Use the definitions and properties given in the reading to show that $\mathbf{A} + \mathbf{B} = \mathbf{C}$ implies that $\mathbf{A} = \mathbf{C} - \mathbf{B}$.

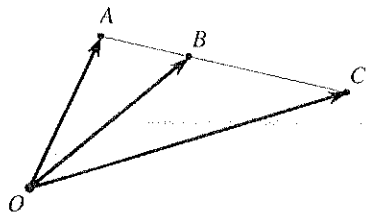
7. (a) Show that if $\mathbf{A} + \mathbf{B} = \mathbf{0}$ and **A** and **B** are not parallel, then each of **A** and **B** must be **0**.
(b) Vectors are often of help in deriving geometrical relationships. For example, to show that the *diagonals of a parallelogram bisect each other* one may proceed as follows. From the accompanying figure $\mathbf{A} + \mathbf{B} = \mathbf{C}$, $\mathbf{A} - \alpha\mathbf{D} = \beta\mathbf{C}$, and $\mathbf{A} = \mathbf{B} + \mathbf{D}$. Eliminating **A** and **B**, we obtain $(2\beta - 1)\mathbf{C} = (1 - 2\alpha)\mathbf{D}$, and since **C** and **D** are not parallel, it must be true [per part (a)] that $2\beta - 1 = 1 - 2\alpha = 0$ (i.e., $\alpha = \beta = \frac{1}{2}$), which completes the proof. We now state the problem: Use this sort of procedure to show that *a line from one vertex of a parallelogram to the midpoint of a nonadjacent side trisects a diagonal.*



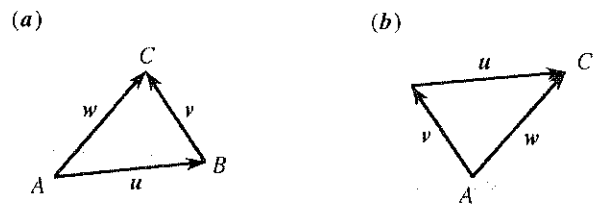
8. If (see the accompanying figure) the vector $\mathbf{A} + \alpha\mathbf{B}$ is placed with its tail at point P, show the line generated by its head as α varies between $-\infty$ and $+\infty$.



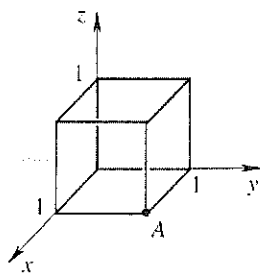
9. If (see the accompanying figure) $\|\mathbf{AB}\| / \|\mathbf{AC}\| = \alpha$, show that $\mathbf{OB} = \alpha\,\mathbf{OC} + (1 - \alpha)\mathbf{OA}$.



10. One may express *linear displacement* as a vector: If a particle moves from point A to point B, the displacement vector is the directed line segment, say $\mathbf{u}$, from A to B. For example, observe that a displacement $\mathbf{u}$ from A to B, followed by a displacement $\mathbf{v}$ from B to C, is equivalent to a single displacement $\mathbf{w}$ from A to C: $\mathbf{u}+\mathbf{v} = \mathbf{w}$ [part (a) in the accompanying figure]. Reversing the order, displacements $\mathbf{v}$ and then $\mathbf{u}$ also



carry us from A to C: $\mathbf{v}+\mathbf{u} = \mathbf{w}$ [part (b) in the figure]. Thus, $\mathbf{u}+\mathbf{v} = \mathbf{v}+\mathbf{u}$ so that the commutativity axiom (1) is indeed satisfied. How about *angular* displacements? Suppose that we express the angular displacement of a rigid body about an axis as $\boldsymbol{\theta}$, where the magnitude of $\boldsymbol{\theta}$ is equal to the angle of rotation, and the orientation of $\boldsymbol{\theta}$ is along the axis of rotation, in the direction specified by the "right-hand rule." That is, if we curl the fingers of our right hand about the axis of rotation, in the direction of rotation, then the direction $\boldsymbol{\theta}$ along the axis of rotation is the direction in which our thumb points. The problem is to show that $\boldsymbol{\theta}$, defined in this way, is *not* a proper vector quantity. HINT: Considering the unit cube shown below, say, show (by keeping track of the coordinates of the corner A) that the orientation that results from a rotation of $\pi/2$ about the x axis, followed by a rotation of $\pi/2$ about the y axis, is not the same as that which results when the order of the rotations is reversed. NOTE: If you have encountered angular velocity vectors (usually denoted as $\boldsymbol{\omega}$ or $\boldsymbol{\Omega}$), in mechanics, it may seem strange to you that *finite rotations* (assigned a vector direction by the right-hand rule) are not true vectors. The idea is that angular velocity involves *infinitesimal* rotations, and infinitesimal rotations (assigned a vector direction by the right-hand rule) *are* true vectors. This subtle point is discussed in many sources (e.g., Robert R. Long, *Engineering Science Mechanics*, Englewood Cliffs, NJ: Prentice Hall, 1963, pp. 31–36).



9.3 Introduction of Angle and Dot Product

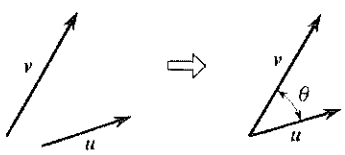


Figure 1. The angle θ between $\mathbf{u}$ and $\mathbf{v}$.

Continuing our discussion, we define here the angle between two vectors and a "dot product" operation between two vectors. The angle θ between two nonzero vectors $\mathbf{u}$ and $\mathbf{v}$ will be understood to mean the ordinary angle between the two vectors when they are arranged tail to tail as in Fig. 1. (We will not attempt to define θ if one or both of the vectors is $\mathbf{0}$.) Of course, this definition of θ is ambiguous in that there are *two* such angles, an interior angle ($\leq \pi$) and an exterior angle ($\geq \pi$); for definiteness, we choose θ to be the interior angle,

$$0 \leq \theta \leq \pi, \tag{1}$$

as in Fig. 1. Unless explicitly stated otherwise, angular measure will be understood to be in *radians*.

Next, we define the so-called **dot product**, $\mathbf{u} \cdot \mathbf{v}$, between two vectors $\mathbf{u}$ and $\mathbf{v}$ as

$$\boxed{\mathbf{u} \cdot \mathbf{v} \equiv \begin{cases} \|\mathbf{u}\| \, \|\mathbf{v}\| \cos\theta & \text{if } \mathbf{u}, \mathbf{v} \neq \mathbf{0}, \\ 0 & \text{if } \mathbf{u} = \mathbf{0} \text{ or } \mathbf{v} = 0; \end{cases}} \tag{2a,b}$$

$\|\mathbf{u}\|$, $\|\mathbf{v}\|$, and $\cos\theta$ are scalars so $\mathbf{u} \cdot \mathbf{v}$ is a scalar, too.*

By way of geometrical interpretation, observe (Fig. 2a) that $\|\mathbf{u}\| \cos\theta$ is the length of the orthogonal projection of $\mathbf{u}$ on the line of action of $\mathbf{v}$ so that $\mathbf{u} \cdot \mathbf{v} = \|\mathbf{u}\| \, \|\mathbf{v}\| \cos\theta = (\|\mathbf{v}\|)(\|\mathbf{u}\| \cos\theta)$ is the length of $\mathbf{v}$ times the length of the orthogonal projection of $\mathbf{u}$ on the line of action of $\mathbf{v}$.† Actually, that statement holds if $0 \leq \theta \leq \pi/2$; if $\pi/2 < \theta \leq \pi$, the cosine is negative, and $\mathbf{u} \cdot \mathbf{v} = \|\mathbf{u}\| \, \|\mathbf{v}\| \cos\theta$ is the negative of the length of $\mathbf{v}$ times the length of the orthogonal projection of $\mathbf{u}$ on the line of action of $\mathbf{v}$.

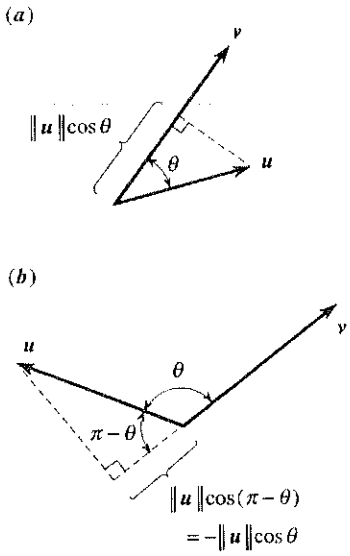


Figure 2. Projection of $\mathbf{u}$ on $\mathbf{v}$.

EXAMPLE 1. *Work Done by a Force.* In mechanics the *work* W done when a body undergoes a linear displacement from an initial point A to a final point B, under the action of a constant force $\mathbf{F}$ (Fig. 3), is defined as the length of the orthogonal projection of $\mathbf{F}$ on the line of displacement, positive if $\mathbf{F}$ is "assisting" the motion (i.e., if $0 \leq \theta < \pi/2$, as in Fig. 3a) and negative if $\mathbf{F}$ is "opposing" the motion (i.e., if $\pi/2 < \theta < \pi$, as in Fig. 3b), times the displacement. By the displacement we mean the length of the vector $\mathbf{AB}$ with head at B and tail at A. But that product is precisely the dot product of $\mathbf{F}$ with $\mathbf{AB}$,

$$W = \mathbf{F} \cdot \mathbf{AB}. \tag{3}$$

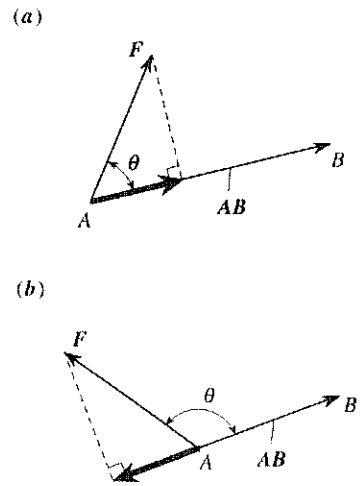


Figure 3. Work done by $\mathbf{F}$.

An important special case of the dot product occurs when $\theta = \pi/2$. Then $\mathbf{u}$ and $\mathbf{v}$ are perpendicular, and

$$\mathbf{u} \cdot \mathbf{v} = \|\mathbf{u}\| \, \|\mathbf{v}\| \cos\frac{\pi}{2} = 0. \tag{4}$$

Also of importance is the case where $\mathbf{u} = \mathbf{v}$. Then, according to (2),

$$\mathbf{u} \cdot \mathbf{u} \equiv \begin{cases} \|\mathbf{u}\| \, \|\mathbf{u}\| \cos 0 = \|\mathbf{u}\|^2 & \text{if } \mathbf{u} \neq \mathbf{0}, \\ 0 & \text{if } \mathbf{u} = \mathbf{0} \end{cases} \tag{5}$$

so that we have

$$\boxed{\|\mathbf{u}\| = \sqrt{\mathbf{u} \cdot \mathbf{u}}} \tag{6}$$

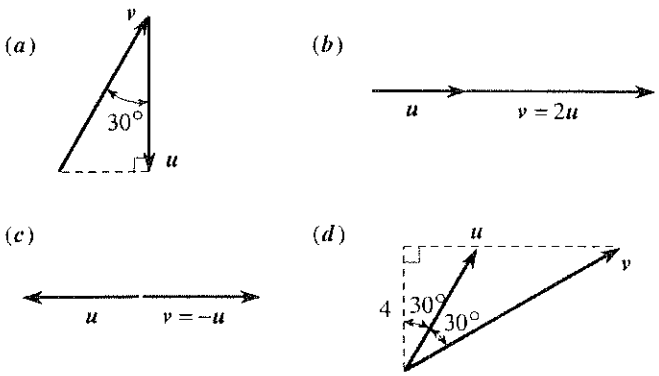
*You may wonder why (2b) is needed since if $\mathbf{u} = \mathbf{0}$, say, then $\|\mathbf{u}\| = 0$ and $\|\mathbf{u}\| \, \|\mathbf{v}\| \cos\theta$ is apparently zero anyway. But the point is that if $\mathbf{u}$ and/or $\mathbf{v}$ are $\mathbf{0}$, then θ is undefined; hence $\cos\theta$ (and even zero times $\cos\theta$) is undefined, too.

†Alternatively, we could decompose $\mathbf{u} \cdot \mathbf{v} = \|\mathbf{u}\| \, \|\mathbf{v}\| \cos\theta = (\|\mathbf{u}\|)(\|\mathbf{v}\| \cos\theta)$; that is, as the length of $\mathbf{u}$ times the length of the orthogonal projection of $\mathbf{v}$ on the line of action of $\mathbf{u}$.

whether $\mathbf{u} \neq \mathbf{0}$ or not. The relationship (6) between the dot product and the norm will be useful in subsequent sections.

EXERCISES 9.3

1. Evaluate $\mathbf{u} \cdot \mathbf{v}$ in each case. In (a) $\|\mathbf{u}\| = 5$, in (b) $\|\mathbf{u}\| = 3$, and in (c) $\|\mathbf{u}\| = 6$.



2. (*Properties of the dot product*) Prove each of the following properties of the dot product, where α, β are any scalars, and $\mathbf{u}, \mathbf{v}, \mathbf{w}$ are any vectors.

(a) $\mathbf{u} \cdot \mathbf{v} = \mathbf{v} \cdot \mathbf{u}$ (commutativity)

(b) $\mathbf{u} \cdot \mathbf{u} > 0$ for all $\mathbf{u} \neq \mathbf{0}$
$\quad\quad\quad = 0$ for all $\mathbf{u} = \mathbf{0}$ (nonnegativeness)

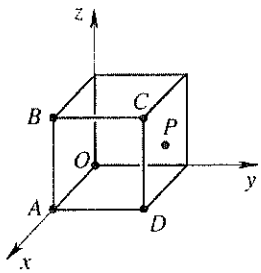
(c) $(\alpha\mathbf{u} + \beta\mathbf{v}) \cdot \mathbf{w} = \alpha(\mathbf{u} \cdot \mathbf{w}) + \beta(\mathbf{v} \cdot \mathbf{w})$ (linearity)

HINT: In proving part (c), you may wish to show, first, that part (c) is equivalent to the two conditions $(\mathbf{u} + \mathbf{v}) \cdot \mathbf{w} = \mathbf{u} \cdot \mathbf{w} + \mathbf{v} \cdot \mathbf{w}$ and $(\alpha\mathbf{u}) \cdot \mathbf{v} = \alpha(\mathbf{u} \cdot \mathbf{v})$.

3. Using the properties given in Exercise 2, show that

$$(\mathbf{u} + \mathbf{v}) \cdot (\mathbf{w} + \mathbf{x}) = \mathbf{u} \cdot \mathbf{w} + \mathbf{u} \cdot \mathbf{x} + \mathbf{v} \cdot \mathbf{w} + \mathbf{v} \cdot \mathbf{x}. \quad (3.1)$$

4. Consider the unit cube shown, where P is the midpoint of the right-hand face. Evaluate each of the following using the definition (2), and (3.1) in Exercise 3. HINT: To evaluate $\mathbf{AC} \cdot \mathbf{OP}$, for instance, write $\mathbf{AC} \cdot \mathbf{OP} = (\mathbf{AD} + \mathbf{DC}) \cdot (\mathbf{OD} + \mathbf{DP})$ and then use (3.1).



(a) $\mathbf{OC} \cdot \mathbf{AB}$ (b) $\mathbf{BA} \cdot \mathbf{OP}$ (c) $\mathbf{AC} \cdot \mathbf{OP}$ (d) $\mathbf{OC} \cdot \mathbf{CP}$
(e) $\mathbf{OC} \cdot \mathbf{OP}$ (f) $\mathbf{BC} \cdot \mathbf{OP}$ (g) $\mathbf{AO} \cdot \mathbf{OP}$ (h) $\mathbf{CP} \cdot \mathbf{DP}$
(i) $\mathbf{BP} \cdot \mathbf{DB}$ (j) $\mathbf{PB} \cdot \mathbf{CO}$ (k) $\mathbf{AP} \cdot \mathbf{PB}$ (l) $\mathbf{AO} \cdot \mathbf{PA}$

5. Referring to the figure in Exercise 4, use the dot product to compute the following angles. (See the hint in Exercise 4.) You may use (2), (6), and (3.1).

(a) APO (b) APB (c) APC (d) APD
(e) ABP (f) ACP (g) BPO (h) BPC
(i) BPD (j) BOP (k) CPO (l) DPO

6. If $\mathbf{u}$ and $\mathbf{v}$ are nonzero, show that $\mathbf{w} = \|\mathbf{v}\|\,\mathbf{u} + \|\mathbf{u}\|\,\mathbf{v}$ bisects the angle between $\mathbf{u}$ and $\mathbf{v}$. (You may use any of the properties given in Exercise 2.)

9.4 n-Space

Here we move away from our dependence on the arrow representation of vectors, in 2-space and 3-space, by introducing an alternative representation in terms of 2-tuples and 3-tuples. This step will lead us to a more general notion of vectors in "n-space."

The idea is simple and is based on the familiar representation of *points* in Cartesian 1-, 2-, and 3-space as 1-, 2-, and 3-tuples of real numbers. For example, the 2-tuple (a_1, a_2) denotes the point P indicated in Fig. 1a, where a_1, a_2 are the x, y coordinates, respectively. But it can also serve to denote the vector $\mathbf{OP}$ in Fig. 1b or, indeed, any equivalent vector $\mathbf{QR}$.

Thus the vector is now represented as the 2-tuple (a_1, a_2) rather than as an arrow, and while pictures may still be drawn, as in Fig. 1b, they are no longer essential and can be discarded if we wish – at least once the algebra of 2-tuples is established (in the next paragraph). The set of all such real 2-tuple vectors will be called **2-space** and will be denoted by the symbol $\mathbb{R}^2$; that is,

$$\mathbb{R}^2 = \{(a_1, a_2) \mid a_1, a_2 \text{ real numbers}\}. \tag{1}$$

Vectors $\mathbf{u} = (u_1, u_2)$ and $\mathbf{v} = (v_1, v_2)$ in $\mathbb{R}^2$ are defined to be *equal* if $u_1 = v_1$ and $u_2 = v_2$; their *sum* is defined as*

$$\mathbf{u} + \mathbf{v} \equiv (u_1 + v_1, u_2 + v_2) \tag{2}$$

as can be seen from Fig. 2; the *scalar multiple* $\alpha\mathbf{u}$ is defined, for any scalar α, as

$$\alpha\mathbf{u} \equiv (\alpha u_1, \alpha u_2); \tag{3}$$

the *zero vector* is

$$\mathbf{0} \equiv (0, 0); \tag{4}$$

and the *negative of* $\mathbf{u}$ is

$$-\mathbf{u} \equiv (-u_1, -u_2). \tag{5}$$

Similarly, for $\mathbb{R}^3$:

$$\mathbb{R}^3 = \{(a_1, a_2, a_3) \mid a_1, a_2, a_3 \text{ real numbers}\}. \tag{6}$$

$$\mathbf{u} + \mathbf{v} \equiv (u_1 + v_1, u_2 + v_2, u_3 + v_3), \tag{7}$$

and so on.†

It may not be evident that we have gained much since the arrow and n-tuple representations are essentially equivalent. But, in fact, the n-tuple format begins to "open doors." For example, the instantaneous state of the electrical circuit (consisting of a battery and two resistors) shown in Fig. 3 may be defined by the two currents i_1 and i_2 or, equivalently, by the single 2-tuple vector (i_1, i_2). Thus, even though "magnitudes," "directions," and "arrow vectors" may not leap to mind in describing the system shown in Fig. 3, a vector representation *is* quite natural within the n-tuple framework, and that puts us in a position, in dealing with that electrical system, to make use of whatever vector theorems and techniques are available, as developed in subsequent sections and chapters.

*We use the $\equiv$ equal sign to mean *equal to by definition*.

†The space $\mathbb{R}^1$ of 1-tuples will not be of interest here.

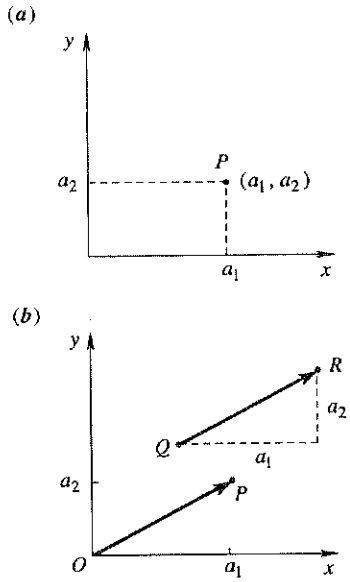


Figure 1. 2-tuple representation.

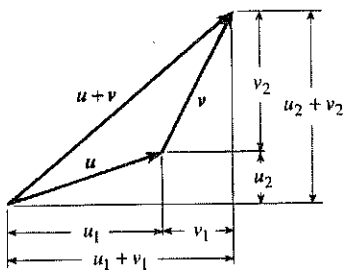


Figure 2. Establishing (2).

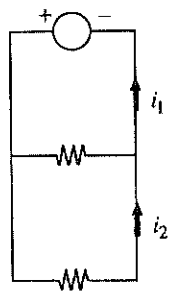


Figure 3. Electrical system.

Indeed, why stop at 3-tuples? One may introduce the set of all ordered real **n-tuple** vectors, even if n is greater than 3. We call this **n-space**, and denote it as $\mathbb{R}^n$, that is,

$$\mathbb{R}^n = \{(a_1, \ldots, a_n) \mid a_1, \ldots, a_n \text{ real numbers }\}. \tag{8}$$

Consider two vectors, $\mathbf{u} = (u_1, \ldots, u_n)$ and $\mathbf{v} = (v_1, \ldots, v_n)$, in $\mathbb{R}^n$. The scalars $u_1, \ldots, u_n$ and $v_1, \ldots, v_n$ are called the **components** of $\mathbf{u}$ and $\mathbf{v}$. As you may well expect, based on our foregoing discussion of $\mathbb{R}^2$ and $\mathbb{R}^3$, $\mathbf{u}$ and $\mathbf{v}$ are said to be *equal* if $u_1 = v_1, \ldots, u_n = v_n$, and we define

$$\begin{aligned}
\mathbf{u} + \mathbf{v} &\equiv (u_1 + v_1, \ldots, u_n + v_n), && \text{(addition)} && (9a)\\
\alpha\mathbf{u} &\equiv (\alpha u_1, \ldots, \alpha u_n), && \text{(scalar multiplication)} && (9b)\\
\mathbf{0} &\equiv (0, \ldots, 0), && \text{(zero vector)} && (9c)\\
-\mathbf{u} &\equiv (-1)\mathbf{u}, && \text{(negative inverse)} && (9d)\\
\mathbf{u} - \mathbf{v} &\equiv \mathbf{u} + (-\mathbf{v}). && && (9e)
\end{aligned}$$

From these definitions we may deduce the following properties:

$$\begin{aligned}
\mathbf{u} + \mathbf{v} &= \mathbf{v} + \mathbf{u}, && \text{(commutativity)} && (10a)\\
(\mathbf{u} + \mathbf{v}) + \mathbf{w} &= \mathbf{u} + (\mathbf{v} + \mathbf{w}), && \text{(associativity)} && (10b)\\
\mathbf{u} + \mathbf{0} &= \mathbf{u}, && && (10c)\\
\mathbf{u} + (-\mathbf{u}) &= \mathbf{0}, && && (10d)\\
\alpha(\beta\mathbf{u}) &= (\alpha\beta)\mathbf{u}, && \text{(associativity)} && (10e)\\
(\alpha + \beta)\mathbf{u} &= \alpha\mathbf{u} + \beta\mathbf{u}, && \text{(distributivity)} && (10f)\\
\alpha(\mathbf{u} + \mathbf{v}) &= \alpha\mathbf{u} + \alpha\mathbf{v}, && \text{(distributivity)} && (10g)\\
1\mathbf{u} &= \mathbf{u}, && && (10h)\\
0\mathbf{u} &= \mathbf{0}, && && (10i)\\
(-1)\mathbf{u} &= -\mathbf{u}, && && (10j)\\
\alpha\mathbf{0} &= \mathbf{0}. && && (10k)
\end{aligned}$$

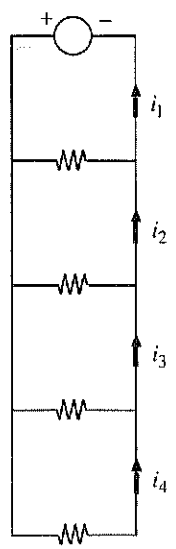


Figure 4. Another electrical circuit.

To illustrate how such n-tuples might arise, observe that the state of the electrical system shown in Fig. 4, may be defined at any instant by the four currents i_1, i_2, i_3, i_4, and that these may be regarded as the components of a single vector $\mathbf{i} = (i_1, i_2, i_3, i_4)$ in $\mathbb{R}^4$.

Of course, the notation of $(u_1, \ldots, u_n)$ as a *point* or *arrow* in an "n-dimensional space" can be realized graphically only if $n \leq 3$; if $n > 3$, the interpretation is valid only in an abstract, or schematic, sense. However, our inability to carry out traditional Cartesian graphical constructions for $n > 3$ will be no hindrance. Indeed, part of the idea here is to move *away* from a dependence on graphical constructions.

Having extended the vector concept to $\mathbb{R}^n$, you may well wonder if further extension is possible. Such extension is not only possible, it constitutes an important step in modern mathematics; more about this in Section 9.6.

EXERCISES 9.4

1. If $\mathbf{t} = (5,0,1,2)$, $\mathbf{u} = (2,-1,3,4)$, $\mathbf{v} = (4,-5,1)$, $\mathbf{w} = (-1,-2,5,6)$, evaluate each of the following (as a single vector); if the operation is undefined (i.e., has not been defined here), state that. At each step cite the equation number of the definition or property being used.

(a) $2\mathbf{t} + 7\mathbf{u}$ (b) $3\mathbf{t} - 5\mathbf{u}$
(c) $4[\mathbf{u} + 5(\mathbf{w} - 2\mathbf{u})]$ (d) $4\mathbf{t}\mathbf{u} + \mathbf{w}$
(e) $-\mathbf{w} + \mathbf{t}$ (f) $2\mathbf{t}/\mathbf{u}$
(g) $\mathbf{t} + 2\mathbf{u} + 3\mathbf{w}$ (h) $\mathbf{t} - 2\mathbf{u} - 4\mathbf{v}$
(i) $\mathbf{u}(3\mathbf{t} + \mathbf{w})$ (j) $\mathbf{u}^2 + 2\mathbf{t}$
(k) $2\mathbf{t} + 7\mathbf{u} - 4$ (l) $\mathbf{u} + \mathbf{w}\mathbf{t}$
(m) $\sin \mathbf{u}$ (n) $\mathbf{w} + \mathbf{t} - 2\mathbf{u}$

2. Let $\mathbf{u} = (1,3,0,-2)$, $\mathbf{v} = (2,0,-5,0)$, and $\mathbf{w} = (4,3,2,-1)$.

(a) If $3\mathbf{u} - \mathbf{x} = 4(\mathbf{v} + 2\mathbf{x})$, solve for $\mathbf{x}$ (i.e., find its components).
(b) If $\mathbf{x} + \mathbf{u} + \mathbf{v} + \mathbf{w} = \mathbf{0}$, solve for $\mathbf{x}$.

3. Let $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$ be as given in Exercise 2. Citing the definition or property used, at each step, solve each of the following for $\mathbf{x}$. NOTE: Besides the definitions and properties stated in this section, it should be clear that if $\mathbf{x} = \mathbf{y}$, then $\mathbf{x}+\mathbf{z} = \mathbf{y}+\mathbf{z}$ for any $\mathbf{z}$, and $\alpha\mathbf{x} = \alpha\mathbf{y}$ for any α (adding and multiplying equals by equals.)

(a) $3\mathbf{x} + 2(\mathbf{u} - 5\mathbf{v}) = \mathbf{w}$
(b) $3\mathbf{x} = 4\mathbf{0} + (1,0,0,0)$
(c) $\mathbf{u} - 4\mathbf{x} = \mathbf{0}$
(d) $\mathbf{u} + \mathbf{v} - 2\mathbf{x} = \mathbf{w}$

4. If $\mathbf{t} = (2,1,3)$, $\mathbf{u} = (1,2,-4)$, $\mathbf{v} = (0,1,1)$, $\mathbf{w} = (-2,1,-1)$, solve each of the following for the scalars $\alpha_1, \alpha_2, \alpha_3$. If no such scalars exist, state that.

(a) $\alpha_1\mathbf{t} + \alpha_2\mathbf{u} + \alpha_3\mathbf{v} = \mathbf{0}$
(b) $\alpha_1\mathbf{t} + \alpha_2\mathbf{v} + \alpha_3\mathbf{w} = \mathbf{0}$
(c) $\alpha_1\mathbf{t} + \alpha_2\mathbf{u} + \alpha_3\mathbf{w} = (1,3,2)$
(d) $\alpha_1\mathbf{t} + \alpha_2\mathbf{v} + \alpha_3\mathbf{w} = (2,0,-1)$
(e) $\alpha_1\mathbf{u} + \alpha_2\mathbf{v} = \mathbf{0}$
(f) $\alpha_1\mathbf{u} + \alpha_2\mathbf{v} = \alpha_3\mathbf{w} - (2,0,0)$

5. (a) If $\mathbf{u}$ and $\mathbf{v}$ are given 4-tuples and $\mathbf{0} = (0,0,0,0)$, does the vector equation $\alpha_1\mathbf{u}+\alpha_2\mathbf{v} = \mathbf{0}$ necessarily have nontrivial solutions for the scalars α_1 and α_2? Explain. (If the answer is "no," a counterexample will suffice.)
(b) Repeat part (a), but where $\mathbf{u}$, $\mathbf{v}$ are 3-tuples and $\mathbf{0} = (0,0,0)$.
(c) Repeat part (a), but where $\mathbf{u}$, $\mathbf{v}$ are 2-tuples and $\mathbf{0} = (0,0)$.

9.5 Dot Product, Norm, and Angle for n-Space

9.5.1. Dot product, norm, and angle. We wish to define the norm of an n-tuple vector, and the dot product and angle between two n-tuple vectors, just as we did for "arrow vectors." These definitions should be expressed in terms of the *components* of the n-tuples since the graphical and geometrical arguments used for arrow vectors will not be possible here for $n > 3$. Thus, if $\mathbf{u} = (u_1, \ldots, u_n)$, we wish to define the *norm* or "length" of $\mathbf{u}$, denoted as $\|\mathbf{u}\|$, in terms of the components $u_1, \ldots, u_n$ of $\mathbf{u}$; and given another vector $\mathbf{v} = (v_1, \ldots, v_n)$, we wish to define the *angle* θ between $\mathbf{u}$ and $\mathbf{v}$, and the *dot product* $\mathbf{u} \cdot \mathbf{v}$, in terms of $u_1, \ldots, u_n$ and $v_1, \ldots, v_n$.

Furthermore, we would like these definitions to reduce to the definitions given in Sections 9.2 and 9.3 in the event that $n = 2$ or 3.

Let us begin with the dot product. Our plan is to return to the arrow vector formula

$$\mathbf{u} \cdot \mathbf{v} = \|\mathbf{u}\| \, \|\mathbf{v}\| \cos\theta, \tag{1}$$

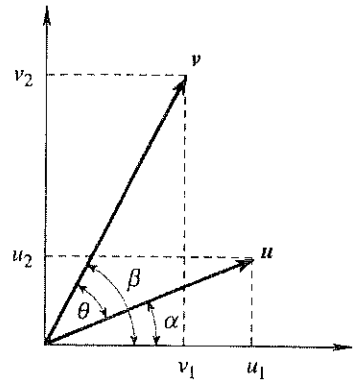


Figure 1. $\mathbf{u} \cdot \mathbf{v}$ in terms of components.

to re-express it in terms of vector components for $\mathbb{R}^2$ and $\mathbb{R}^3$, and then to generalize those forms to $\mathbb{R}^n$.

If $\mathbf{u}$ and $\mathbf{v}$ are vectors in $\mathbb{R}^2$ as shown in Fig. 1, formula (1) may be expressed in terms of the components of $\mathbf{u}$ and $\mathbf{v}$ as follows:

$$\begin{aligned}\mathbf{u} \cdot \mathbf{v} &= \|\mathbf{u}\| \, \|\mathbf{v}\| \cos\theta \\ &= \|\mathbf{u}\| \, \|\mathbf{v}\| \cos(\beta - \alpha) \\ &= \|\mathbf{u}\| \, \|\mathbf{v}\| (\cos\beta \cos\alpha + \sin\beta \sin\alpha) \\ &= (\|\mathbf{u}\| \cos\alpha)(\|\mathbf{v}\| (\cos\beta) + (\|\mathbf{u}\| \sin\alpha)(\|\mathbf{v}\| (\sin\beta) \\ &= u_1 v_1 + u_2 v_2. \end{aligned} \tag{2}$$

We state, without derivation, that the analogous result for $\mathbb{R}^3$ is

$$\mathbf{u} \cdot \mathbf{v} = u_1 v_1 + u_2 v_2 + u_3 v_3. \tag{3}$$

Generalizing (2) and (3) to $\mathbb{R}^n$, it is eminently reasonable to define the (scalar-valued) dot product of two n-tuple vectors $\mathbf{u} = (u_1, \ldots, u_n)$ and $\mathbf{v} = (v_1, \ldots, v_n)$ as

$$\boxed{\mathbf{u} \cdot \mathbf{v} \equiv u_1 v_1 + u_2 v_2 + \cdots + u_n v_n = \sum_{j=1}^{n} u_j v_j.} \tag{4}$$

Observe that we have not proved (4); it is a definition.

Defining the dot product is the key, for now $\|\mathbf{u}\|$ and θ follow readily. Specifically, we define

$$\boxed{\|\mathbf{u}\| \equiv \sqrt{\mathbf{u} \cdot \mathbf{u}} = \sqrt{\sum_{j=1}^{n} u_j^2}} \tag{5}$$

in accordance with equation (6) in Section 9.3, and

$$\boxed{\theta \equiv \cos^{-1}\left(\frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \, \|\mathbf{v}\|}\right),} \tag{6}$$

from (1), where the inverse cosine is understood to be in the interval $[0, \pi]$.* Notice the generalized Pythagorean-theorem nature of (5).

Other dot products and norms are sometimes defined for n-space, but we choose to use (4) and (5), which are known as the **Euclidean dot product** and **Euclidian norm**, respectively. To signify that the Euclidean dot product and norm have been adopted, we henceforth refer to the space as **Euclidean n-space**, rather

*By the "interval $[a, b]$ on a real x axis," we mean the points $a \le x \le b$. Such an interval is said to be **closed** since it includes the two endpoints. To denote the **open** interval $a < x < b$, we write (a, b). Similarly, $[a, b)$ means $a \le x < b$, and $(a, b]$ means $a < x \le b$. Implicit in the closed-interval notation $[a, b]$ is the finiteness of a and b.

than just n-space. We will still denote it by the symbol $\mathbb{R}^n$ (although some authors prefer the notation $\mathbb{E}^n$).

EXAMPLE 1. Let $\mathbf{u} = (1, 0)$ and $\mathbf{v} = (2, -2)$. Then

$$\begin{aligned} \mathbf{u} \cdot \mathbf{v} &= (1)(2) + (0)(-2) = 2, \\ \|\mathbf{u}\| &= \sqrt{(1)^2 + (0)^2} = 1, \\ \|\mathbf{v}\| &= \sqrt{(2)^2 + (-2)^2} = 2\sqrt{2}, \\ \theta &= \cos^{-1}\left(\frac{2}{2\sqrt{2}}\right) = \frac{\pi}{4} \quad (\text{or } 45^\circ), \end{aligned}$$

as is readily verified if we sketch $\mathbf{u}$ and $\mathbf{v}$ as arrow vectors in a Cartesian plane. ∎

EXAMPLE 2. Let $\mathbf{u} = (2, -2, 4, -1)$ and $\mathbf{v} = (5, 9, -1, 0)$. Then,

$$\mathbf{u} \cdot \mathbf{v} = (2)(5) + (-2)(9) + (4)(-1) + (-1)(0) = -12, \tag{7}$$

$$\|\mathbf{u}\| = \sqrt{(2)^2 + (-2)^2 + (4)^2 + (-1)^2} = 5, \tag{8}$$

$$\|\mathbf{v}\| = \sqrt{(5)^2 + (9)^2 + (-1)^2 + (0)^2} = \sqrt{107}, \tag{9}$$

$$\theta = \cos^{-1}\left(\frac{-12}{5\sqrt{107}}\right) \approx \cos^{-1}(-0.232) \approx 1.805 \quad (\text{or } 103.4^\circ). \tag{10}$$

In this case, n $(= 4)$ is greater than 3 so (7) through (10) are not to be understood in any physical or graphical sense, but merely in terms of the definitions (4) to (6).

COMMENT. The dot product of $\mathbf{u} = (2, -2, 4)$ and $\mathbf{v} = (5, 9, -1, 0)$, on the other hand, is *not defined* since here $\mathbf{u}$ and $\mathbf{v}$ are members of different spaces, $\mathbb{R}^3$ and $\mathbb{R}^4$, respectively. It is *not* legitimate to augment $\mathbf{u}$ to the form $(2, -2, 4, 0)$ on the grounds that "surely adding a zero can't hurt." ∎

There is one catch that you may have noticed: (6) serves to define a (real) θ only if the argument of the inverse cosine is less than or equal to unity in magnitude. That this is indeed true is not so obvious. Nevertheless, that

$$-1 \le \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \, \|\mathbf{v}\|} \le 1 \qquad \text{or} \qquad |\mathbf{u} \cdot \mathbf{v}| \le \|\mathbf{u}\| \, \|\mathbf{v}\| \tag{11}$$

does necessarily hold will be proved in a moment. Whereas double braces denote vector norm, the single braces in (11) denote the absolute value of the scalar $\mathbf{u} \cdot \mathbf{v}$.

9.5.2. Properties of the dot product. The dot product defined by (4) possesses the following important properties:

$$\textit{Commutative:} \qquad \mathbf{u} \cdot \mathbf{v} = \mathbf{v} \cdot \mathbf{u}, \tag{12a}$$

$$\textit{Nonnegative:} \qquad \begin{aligned} \mathbf{u} \cdot \mathbf{u} &> 0 \quad \text{for all } \mathbf{u} \ne \mathbf{0} \\ &= 0 \quad \text{for } \mathbf{u} = \mathbf{0}, \end{aligned} \tag{12b}$$

$$\textit{Linear:} \qquad (\alpha \mathbf{u} + \beta \mathbf{v}) \cdot \mathbf{w} = \alpha(\mathbf{u} \cdot \mathbf{w}) + \beta(\mathbf{v} \cdot \mathbf{w}), \tag{12c}$$

for any scalars α, β and any vectors $\mathbf{u}, \mathbf{v}, \mathbf{w}$. The linearity condition (12c) is equivalent to the two conditions $(\mathbf{u}+\mathbf{v}) \cdot \mathbf{w} = (\mathbf{u} \cdot \mathbf{w}) + (\mathbf{v} \cdot \mathbf{w})$ and $(\alpha \mathbf{u}) \cdot \mathbf{v} = \alpha(\mathbf{u} \cdot \mathbf{v})$. Verification of these claims is left for the exercises.

EXAMPLE 3. Expand the dot product $(6\mathbf{t} - 2\mathbf{u}) \cdot (\mathbf{v} + 4\mathbf{w})$. Using (12), we obtain

$$\begin{aligned}
(6\mathbf{t} - 2\mathbf{u}) \cdot (\mathbf{v} + 4\mathbf{w}) &= 6[\mathbf{t} \cdot (\mathbf{v} + 4\mathbf{w})] - 2[\mathbf{u} \cdot (\mathbf{v} + 4\mathbf{w})] && \text{by (12c)} \\
&= 6[(\mathbf{v} + 4\mathbf{w}) \cdot \mathbf{t}] - 2[(\mathbf{v} + 4\mathbf{w}) \cdot \mathbf{u}] && \text{by (12a)} \\
&= 6(\mathbf{v} \cdot \mathbf{t}) + 24(\mathbf{w} \cdot \mathbf{t}) - 2(\mathbf{v} \cdot \mathbf{u}) - 8(\mathbf{w} \cdot \mathbf{u}) && \text{by (12c)}
\end{aligned}$$

in much the same way that we obtain $(a - b)(c + d) = ac + ad - bc - bd$ in scalar arithmetic. ∎

As a consequence of (12) we are in a position to prove the promised inequality (11), namely, the **Schwarz inequality***

$$\boxed{|\mathbf{u} \cdot \mathbf{v}| \leq \|\mathbf{u}\| \, \|\mathbf{v}\| \, .} \tag{13}$$

To derive this result, we start with the inequality

$$(\mathbf{u} + \alpha \mathbf{v}) \cdot (\mathbf{u} + \alpha \mathbf{v}) \geq 0, \tag{14}$$

which is guaranteed by (12b), for any scalar α and any vectors $\mathbf{u}$ and $\mathbf{v}$. Expanding the left-hand side and noting that $\mathbf{u} \cdot \mathbf{u} = \|\mathbf{u}\|^2$ and $\mathbf{v} \cdot \mathbf{v} = \|\mathbf{v}\|^2$, (14) becomes†

$$\|\mathbf{u}\|^2 + 2\alpha \mathbf{u} \cdot \mathbf{v} + \alpha^2 \|\mathbf{v}\|^2 \geq 0. \tag{15}$$

Regarding $\mathbf{u}$ and $\mathbf{v}$ as fixed and α as variable, the left-hand side is then a quadratic function of α. If we choose α so as to minimize the left-hand side, then (15) will be as close to an equality as possible and hence as informative as possible. Thus, setting $d(\text{left-hand side})/d\alpha = 0$, we obtain

$$2\mathbf{u} \cdot \mathbf{v} + 2\alpha \|\mathbf{v}\|^2 = 0 \qquad \text{or} \qquad \alpha = -\frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{v}\|^2}.$$

Putting this optimal value of α back into (15) gives us

$$\begin{aligned}
\|\mathbf{u}\|^2 - 2\frac{(\mathbf{u} \cdot \mathbf{v})^2}{\|\mathbf{v}\|^2} + \frac{(\mathbf{u} \cdot \mathbf{v})^2}{\|\mathbf{v}\|^2} &\geq 0, \\
\|\mathbf{u}\|^2 \|\mathbf{v}\|^2 - 2(\mathbf{u} \cdot \mathbf{v})^2 + (\mathbf{u} \cdot \mathbf{v})^2 &\geq 0, \\
\|\mathbf{u}\|^2 \|\mathbf{v}\|^2 &\geq (\mathbf{u} \cdot \mathbf{v})^2,
\end{aligned}$$

*After *Hermann Amandus Schwarz* (1843–1921). The names *Cauchy* and *Bunyakovsky* are also associated with this well-known inequality.

†Does a term such as $\alpha \mathbf{u} \cdot \mathbf{v}$ in (15) mean $(\alpha \mathbf{u}) \cdot \mathbf{v}$, or $\alpha(\mathbf{u} \cdot \mathbf{v})$? It does not matter; by virtue of (12c) (with $\beta = 0$ and $\mathbf{w}$ changed to $\mathbf{v}$), $(\alpha \mathbf{u}) \cdot \mathbf{v} = \alpha(\mathbf{u} \cdot \mathbf{v})$, so the parentheses are not needed.

and taking square roots of both sides yields the Schwarz inequality (13).*

Thus, it was not merely a matter of luck that the arguments of the inverse cosines were smaller than unity in magnitude in Examples 1 and 2, it was guaranteed in advance by the Schwarz inequality (13).

9.5.3. Properties of the norm. Since the norm is related to the dot product according to

$$\|\mathbf{u}\| = \sqrt{\mathbf{u}\cdot\mathbf{u}}, \tag{16}$$

the properties (12) of the dot product should imply certain corresponding properties of the norm. These properties are as follows:

$$\textit{Scaling}: \qquad \|\alpha\mathbf{u}\| = |\alpha|\,\|\mathbf{u}\|, \tag{17a}$$

$$\textit{Nonnegative}: \qquad \begin{aligned} \|\mathbf{u}\| &> 0 \quad \text{for all } \mathbf{u} \neq \mathbf{0} \\ &= 0 \quad \text{for } \mathbf{u} = \mathbf{0}, \end{aligned} \tag{17b}$$

$$\textit{Triangle Inequality}: \qquad \|\mathbf{u}+\mathbf{v}\| \le \|\mathbf{u}\| + \|\mathbf{v}\|. \tag{17c}$$

Equation (17a) simply says that $\alpha\mathbf{u}$ is $|\alpha|$ times as long as $\mathbf{u}$, and for arrow representations of 2-tuples or 3-tuples the triangle inequality (17c) amounts to the Euclidean proposition that the length of any one side of a triangle cannot exceed the sum of the lengths of the other two sides (Fig. 2). Less obvious, however, is the fact that (17c) holds for n-tuples for n's > 3.

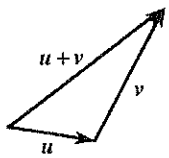


Figure 2. Triangle inequality.

Let us prove only (17a) and (17c) since (17b) follows readily from (16) and (12b). First, (17a):

$$\begin{aligned}
\|\alpha\mathbf{u}\| &= \sqrt{(\alpha\mathbf{u})\cdot(\alpha\mathbf{u})} && \text{by (16)} \\
&= \sqrt{\alpha\mathbf{u}\cdot(\alpha\mathbf{u})} && \text{by (12c) with } \beta = 0 \text{ and } \mathbf{w} = \alpha\mathbf{u} \\
&= \sqrt{\alpha(\alpha\mathbf{u})\cdot\mathbf{u}} && \text{by (12a)} \\
&= \sqrt{\alpha^2\mathbf{u}\cdot\mathbf{u}} && \text{by (12c) with } \beta = 0 \text{ and } \mathbf{w} = \mathbf{u} \\
&= |\alpha|\sqrt{\mathbf{u}\cdot\mathbf{u}} = |\alpha|\,\|\mathbf{u}\|.
\end{aligned}$$

Turning to (17c), we find that

$$\begin{aligned}
\|\mathbf{u}+\mathbf{v}\|^2 &= (\mathbf{u}+\mathbf{v})\cdot(\mathbf{u}+\mathbf{v}) && \text{by (16)} \\
&= \mathbf{u}\cdot\mathbf{u} + \mathbf{v}\cdot\mathbf{u} + \mathbf{u}\cdot\mathbf{v} + \mathbf{v}\cdot\mathbf{v} \\
&= \|\mathbf{u}\|^2 + 2\mathbf{u}\cdot\mathbf{v} + \|\mathbf{v}\|^2 \\
&\le \|\mathbf{u}\|^2 + 2\,|\mathbf{u}\cdot\mathbf{v}| + \|\mathbf{v}\|^2 \\
&\le \|\mathbf{u}\|^2 + 2\,\|\mathbf{u}\|\,\|\mathbf{v}\| + \|\mathbf{v}\|^2 && \text{by (13)} \\
&= (\|\mathbf{u}\| + \|\mathbf{v}\|)^2
\end{aligned}$$

*That the choice $\alpha = -\mathbf{u}\cdot\mathbf{v}/\|\mathbf{v}\|^2$ *min*imizes the left-hand side of (15) follows from the fact that $d^2(\text{left-hand side})/d\alpha^2 = 2\,\|\mathbf{v}\|^2 > 0$.

so that

$$\|\mathbf{u}+\mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|,$$

as claimed. A key step was the use of the Schwarz inequality (13), but we also used the simple inequality $\mathbf{u}\cdot\mathbf{v} \leq |\mathbf{u}\cdot\mathbf{v}|$, which holds since $\mathbf{u}\cdot\mathbf{v}$ is a (positive, zero, or negative) real number; that is, if $\mathbf{u}\cdot\mathbf{v}$ is negative, then the $<$ holds, and if $\mathbf{u}\cdot\mathbf{v}$ is zero or positive, then the $=$ holds.

EXAMPLE 4. Let us verify the triangle inequality for a specific example, say the vectors $\mathbf{u} = (2, 1, 3, -1)$ and $\mathbf{v} = (0, 4, 2, 1)$. Then $\mathbf{u}+\mathbf{v} = (2, 5, 5, 0)$ so (17c) becomes

$$\sqrt{54} \leq \sqrt{15} + \sqrt{21}$$

or $7.348 \leq 3.873 + 4.583$, which is indeed true. ∎

9.5.4. Orthogonality. If $\mathbf{u}$ and $\mathbf{v}$ are nonzero vectors such that $\mathbf{u}\cdot\mathbf{v} = 0$, then

$$\theta = \cos^{-1}\left(\frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|\,\|\mathbf{v}\|}\right) = \cos^{-1}\left(\frac{0}{\|\mathbf{u}\|\,\|\mathbf{v}\|}\right) = \cos^{-1}(0) = \frac{\pi}{2}, \tag{18}$$

and we say that $\mathbf{u}$ and $\mathbf{v}$ are *perpendicular*. [Here we have used the nonzeroness of $\mathbf{u}$ and $\mathbf{v}$ in the third equality in (18); if $\mathbf{u}$ and/or $\mathbf{v}$ *were* $\mathbf{0}$, we would have had $\cos^{-1}(0/\|\mathbf{u}\|\,\|\mathbf{v}\|) = \cos^{-1}(0/0)$, which is not defined.]

But to equate the condition $\mathbf{u}\cdot\mathbf{v} = 0$ to perpendicularity ($\theta = \pi/2$) would not be correct since $\mathbf{u}\cdot\mathbf{v}$ will also be zero in the event that $\mathbf{u}$ and/or $\mathbf{v}$ are $\mathbf{0}$, in which case θ is not defined. Let us therefore make a distinction between perpendicularity and "orthogonality." We will say that $\mathbf{u}$ and $\mathbf{v}$ are **orthogonal** if

$$\mathbf{u}\cdot\mathbf{v} = 0. \tag{19}$$

Only if $\mathbf{u}$ and $\mathbf{v}$ are both nonzero does their orthogonality imply their perpendicularity (i.e., $\theta = \pi/2$). With this definition, we see that the zero vector $\mathbf{0}$ is orthogonal to *every* vector including itself (Exercise 14).

Finally, we say that a *set* of vectors, say $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$, is an **orthogonal set** if every vector in the set is orthogonal to every other one:

$$\mathbf{u}_i \cdot \mathbf{u}_j = 0 \quad \text{if } i \neq j. \tag{20}$$

EXAMPLE 5. $\mathbf{u}_1 = (2, 3, -1, 0)$, $\mathbf{u}_2 = (1, 2, 8, 3)$, $\mathbf{u}_3 = (9, -6, 0, 1)$ is an orthogonal set because $\mathbf{u}_1\cdot\mathbf{u}_2 = \mathbf{u}_1\cdot\mathbf{u}_3 = \mathbf{u}_2\cdot\mathbf{u}_3 = 0$. ∎

EXAMPLE 6. $\mathbf{u}_1 = (1, 3)$, $\mathbf{u}_2 = (0, 0)$ is an orthogonal set because $\mathbf{u}_1\cdot\mathbf{u}_2 = 0$. ∎

9.5.5. Normalization. Any nonzero vector $\mathbf{u}$ can be scaled to have unit length by multiplying it by $1/\|\mathbf{u}\|$ so we say that the vector

$$\hat{\mathbf{u}} = \frac{1}{\|\mathbf{u}\|}\mathbf{u} \tag{21}$$

has been "normalized." That $\hat{\mathbf{u}}$ has unit length is readily verified:

$$\begin{aligned} \|\hat{\mathbf{u}}\| = \left\|\frac{1}{\|\mathbf{u}\|}\mathbf{u}\right\| &= \left|\frac{1}{\|\mathbf{u}\|}\right| \|\mathbf{u}\| && \text{by (17a)} \\ &= \frac{1}{\|\mathbf{u}\|}\|\mathbf{u}\| && \text{by (17b)} \\ &= 1. \end{aligned}$$

A vector of unit length is called a **unit vector**. We will often use the caret notation $\hat{\mathbf{u}}$ for unit vectors.

EXAMPLE 7. Normalize $\mathbf{u} = (1, -1, 0, 2)$. Since $\|\mathbf{u}\| = \sqrt{\mathbf{u}\cdot\mathbf{u}} = \sqrt{6}$, we have

$$\hat{\mathbf{u}} = \frac{1}{\|\mathbf{u}\|}\mathbf{u} = \frac{1}{\sqrt{6}}(1, -1, 0, 2) = \left(\frac{1}{\sqrt{6}}, -\frac{1}{\sqrt{6}}, 0, \frac{2}{\sqrt{6}}\right). \quad \blacksquare$$

A set of vectors is said to be **orthonormal** if it is orthogonal and if each vector is normalized (i.e., is a unit vector). We will use that term so frequently that it will be useful to abbreviate it as ON, but be aware that that abbreviation is not standard. Thus, $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ is ON if and only if $\mathbf{u}_i \cdot \mathbf{u}_j = 0$ whenever $i \neq j$ (for orthogonality), and $\mathbf{u}_j \cdot \mathbf{u}_j = 1$ for each j (so $\|\mathbf{u}_j\| = 1$, so the set is normalized).

The symbol

$$\delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \tag{22}$$

will be useful, and is known as the **Kronecker delta**, after *Leopold Kronecker* (1823–1891). Thus, $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ is ON if and only if

$$\boxed{\mathbf{u}_i \cdot \mathbf{u}_j = \delta_{ij}} \tag{23}$$

for $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, k$.

EXAMPLE 8. Let

$$\mathbf{u}_1 = (1, 0, 0, 0), \quad \mathbf{u}_2 = \left(0, \frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}}\right), \quad \mathbf{u}_3 = \left(0, \frac{1}{\sqrt{2}}, 0, -\frac{1}{\sqrt{2}}\right).$$

Then $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ is ON because $\|\mathbf{u}_1\| = \|\mathbf{u}_2\| = \|\mathbf{u}_3\| = 1$ and $\mathbf{u}_1 \cdot \mathbf{u}_2 = \mathbf{u}_1 \cdot \mathbf{u}_3 = \mathbf{u}_2 \cdot \mathbf{u}_3 = 0$. ∎

Closure. In this section we introduce a dot product $\mathbf{u} \cdot \mathbf{v}$, a norm $\|\mathbf{u}\|$, and an angle θ between two vectors for n-space. Their introduction is not a matter of derivation but, rather, a matter of definition. The definitions are designed as extensions of the definitions for the familiar "arrow" vectors of 2- and 3-space, somewhat as the upper floors of a home are built upon the foundation rather than being placed on an adjacent lot. Those extensions become apparent once we express the dot product, norm, and angle for arrow vectors in n-tuple notation. The key is the definition

$$\mathbf{u} \cdot \mathbf{v} = u_1 v_1 + u_2 v_2 + \cdots + u_n v_n,$$

because then the arrow vector formula $\|\mathbf{u}\| = \sqrt{\mathbf{u} \cdot \mathbf{u}}$ gives us a norm, and the arrow vector formula $\theta = \cos^{-1}(\mathbf{u} \cdot \mathbf{v} / \|\mathbf{u}\| \, \|\mathbf{v}\|)$ gives us an angle θ between $\mathbf{u}$ and $\mathbf{v}$.

From these definitions we then derive the properties (12a,b,c) of the dot product and (17a,b,c) of the norm. In 2- and 3-space the triangle inequality amounts to the familiar Euclidean proposition that the length of any one side of a triangle cannot exceed the sum of the lengths of the other two sides, but in n-space, for $n > 3$, it amounts to an abstract generalization of that notion and does not have such a realizable physical or geometrical interpretation.

EXERCISES 9.5

1. Given the following vectors $\mathbf{u}$ and $\mathbf{v}$, determine $\|\mathbf{u}\|$, $\|\mathbf{v}\|$ and θ (in radians and degrees). If $\mathbf{u}$ and $\mathbf{v}$ are orthogonal, state that.

(a) $\mathbf{u} = (4, 3)$, $\mathbf{v} = (2, -1)$
(b) $\mathbf{u} = (1, 2, 3, 4)$, $\mathbf{v} = (-4, -3, -2, -1)$
(c) $\mathbf{u} = (3, 0, 1)$, $\mathbf{v} = (-2, 3, 6)$
(d) $\mathbf{u} = (2, 2, 2)$, $\mathbf{v} = (-4, -5, -6)$
(e) $\mathbf{u} = (2, 5)$, $\mathbf{v} = (10, -4)$
(f) $\mathbf{u} = (1, 2, 3, 4)$, $\mathbf{v} = (4, 3, 2, 1)$
(g) $\mathbf{u} = (3, 2, 0, -1, 1)$, $\mathbf{v} = (-5, 0, 0, 2, 4)$

2. State whether or not each of the following expressions is defined.

(a) $\|\mathbf{u}\| \, \mathbf{u}$
(b) $\mathbf{u} \cdot (\mathbf{v} \cdot \mathbf{w})$
(c) $\|(\mathbf{u} \cdot \mathbf{v})\mathbf{v}\|$
(d) $(\mathbf{u} + \mathbf{v}) \cdot \mathbf{w}$
(e) $(\mathbf{u} + \mathbf{v}) \cdot (\mathbf{u} - \mathbf{v})$
(f) $\mathbf{u} + 6(\mathbf{v} \cdot \mathbf{w})$
(g) $\cos^{-1}(2\mathbf{u} + \mathbf{v})$
(h) $\mathbf{u} / \|\mathbf{u}\|^2$
(i) $(7\mathbf{u}) \cdot (2\mathbf{v})$
(j) $\|\mathbf{u} + 3\mathbf{u}^2\|$

3. Let us denote, as points in 2- and 3-space, $A = (2, 0)$, $B = (3, -1)$, $C = (5, 0)$, $D = (4, 2)$, $E = (2, 2)$, $F = (1, 3, -2)$, $G = (2, 0, 4)$, $H = (5, 4, 3)$, $I = (-3, -1, 0)$, $J = (0, 0, 0)$. Determine, by vector methods, all interior angles and their sum, in degrees, for each of the following polygons.

(a) $ABCA$	(b) $ABCDA$	(c) $ABCDEA$
(d) $BCDB$	(e) $BCDEB$	(f) $FGHF$
(g) $FGIF$	(h) $GHIG$	(i) $FGJF$
(j) $GHJG$	(k) $HIJH$	(l) $FIJF$

4. (a)–(g) Normalize each pair of $\mathbf{u}$, $\mathbf{v}$ vectors in Exercise 1; that is, obtain $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$.

5. If vectors $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, represented as arrows, form a triangle such that $\mathbf{A} = \mathbf{B} + \mathbf{C}$, derive the *law of cosines* $C^2 = A^2 + B^2 - 2AB\cos\alpha$, where α is the interior angle between $\mathbf{A}$ and $\mathbf{B}$, and where A, B, C are the lengths of $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, respectively, by starting with the identity $\mathbf{C} \cdot \mathbf{C} = (\mathbf{A} - \mathbf{B}) \cdot (\mathbf{A} - \mathbf{B})$.

6. (*Orthogonalization*) In each of following, find scalars α, β, γ and vectors $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$ such that $\mathbf{u}_1 = \mathbf{u}$, $\mathbf{u}_2 = \mathbf{u} + \alpha\mathbf{v}$, $\mathbf{u}_3 = \mathbf{u} + \beta\mathbf{v} + \gamma\mathbf{w}$ is a nonzero orthogonal set, that is,

$\mathbf{u}_1 \cdot \mathbf{u}_2 = 0$, $\mathbf{u}_1 \cdot \mathbf{u}_3 = 0$, and $\mathbf{u}_2 \cdot \mathbf{u}_3 = 0$, where $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3 \neq \mathbf{0}$. If this is *not* possible, state that.

(a) $\mathbf{u} = (1,3,0)$, $\mathbf{v} = (2,3,0)$, $\mathbf{w} = (2,1,-3)$
(b) $\mathbf{u} = (2,0,-1)$, $\mathbf{v} = (1,2,3)$, $\mathbf{w} = (3,-2,-5)$
(c) $\mathbf{u} = (1,0,0)$, $\mathbf{v} = (2,1,0)$, $\mathbf{w} = (3,2,-1)$
(d) $\mathbf{u} = (1,2,0,1)$, $\mathbf{v} = (1,0,1,1)$, $\mathbf{w} = (2,-1,1,1)$
(e) $\mathbf{u} = (1,0,0,0)$, $\mathbf{v} = (1,1,0,0)$, $\mathbf{w} = (1,1,1,0)$
(f) $\mathbf{u} = (1,-1,1,-1)$, $\mathbf{v} = (1,2,0,1)$, $\mathbf{w} = (0,2,1,0)$
(g) $\mathbf{u} = (1,2)$, $\mathbf{v} = (0,2)$, $\mathbf{w} = (1,-1)$
(h) $\mathbf{u} = (3,0)$, $\mathbf{v} = (1,1)$, $\mathbf{w} = (-1,2)$

7. If $\mathbf{u} = (1,3,-4,2)$ and $\mathbf{v} = (2,0,0,3)$, evaluate the following.

(a) $\|\mathbf{u} - \mathbf{v}\|$ (b) $\|3\mathbf{u} - 2\mathbf{v}\| + \|-\mathbf{v}\|$
(c) $\left\| \frac{1}{\|\mathbf{u}\|}\mathbf{u} \right\|$ (d) $\|\mathbf{u}\| + \|\mathbf{v}\|$

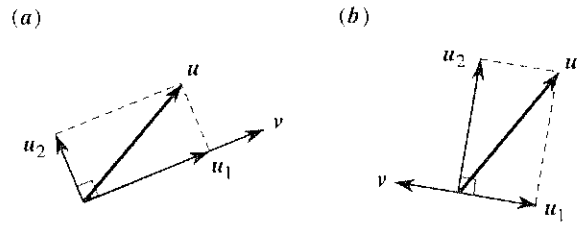
8. Derive the following identities.

(a) $\|\mathbf{u}+\mathbf{v}\|^2 + \|\mathbf{u}-\mathbf{v}\|^2 = 2\|\mathbf{u}\|^2 + 2\|\mathbf{v}\|^2$
(b) $\|\mathbf{u}+\mathbf{v}\|^2 - \|\mathbf{u}-\mathbf{v}\|^2 = 4\mathbf{u}\cdot\mathbf{v}$
(c) Verify parts (a) and (b) for the case $\mathbf{u} = (2,0,1,1)$ and $\mathbf{v} = (1,-3,0,2)$.

9. Find all nonzero vectors (if any) orthogonal to the following vectors.

(a) $(3,0,-1)$
(b) $(2,1,1)$ and $(1,2,3)$
(c) $(1,1,0,-1)$
(d) $(1,3,4,0)$ and $(2,-1,0,5)$
(e) $(6,-1,2,2)$, $(1,4,3,0)$, and $(4,-9,-4,2)$
(f) $(6,-1,2,2)$, $(1,4,3,0)$, and $(4,-5,-4,2)$
(g) $(1,-2,0)$, $(2,3,1)$, and $(7,0,2)$
(h) $(2,1,-1)$, $(1,1,1)$, and $(3,2,1)$

10. (*Orthogonal separation*) It is sometimes desired to separate a given nonzero vector $\mathbf{u}$ into the sum of two orthogonal vectors, one parallel to and one perpendicular to some other nonzero vector $\mathbf{v}$, as sketched in part (a) of the accompanying figure. That is, $\mathbf{u} = \mathbf{u}_1 + \mathbf{u}_2$, where $\mathbf{u}_1$ is of the form $\alpha\mathbf{v}$, and $\mathbf{u}_2 \cdot \mathbf{u}_1 = 0$. We call $\mathbf{u}_1$ the *orthogonal projection of* $\mathbf{u}$ *on* $\mathbf{v}$, and we call $\mathbf{u}_2$ the *component of* $\mathbf{u}$ *orthogonal to* $\mathbf{v}$.



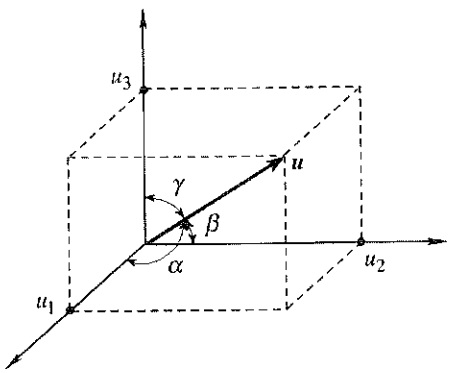
(a) Show that $\mathbf{u}_1$ and $\mathbf{u}_2$ can be found, in terms of $\mathbf{u}$ and $\mathbf{v}$, as

$$\begin{aligned} \mathbf{u}_1 &= (\mathbf{u}\cdot\hat{\mathbf{v}})\hat{\mathbf{v}} \qquad \text{where} \quad \hat{\mathbf{v}} = \mathbf{v}/\|\mathbf{v}\|, \\ \mathbf{u}_2 &= \mathbf{u} - \mathbf{u}_1. \end{aligned} \tag{10.1}$$

Is (10.1) valid only for 2- and 3-space, or does it hold, without modification, for n-space as well? Explain.
(b) Use (10.1) to carry out the separation for the cases where $\mathbf{u} = (1,3)$ and $\mathbf{v} = (2,4)$, and where $\mathbf{u} = (1,3)$ and $\mathbf{v} = (-1,-2)$. Interpret your results graphically for each of these cases.
(c) Use (10.1) to carry out the separation for $\mathbf{u} = (2,3,1)$ and $\mathbf{v} = (0,2,3)$.
(d) Repeat part (c), for $\mathbf{u} = (1,2,-1)$, $\mathbf{v} = (3,-1,1)$.
(e) Repeat part (c), for $\mathbf{u} = (3,0,5,6)$, $\mathbf{v} = (1,-2,0,4)$.
(f) Repeat part (c), for $\mathbf{u} = (2,1,0,0,3)$, $\mathbf{v} = (0,0,1,-2,1)$.

11. (a) Prove the associative property $(\alpha\mathbf{u})\cdot\mathbf{v} = \alpha(\mathbf{u}\cdot\mathbf{v})$.
(b) Prove the distributive property $(\mathbf{u}+\mathbf{v})\cdot\mathbf{w} = \mathbf{u}\cdot\mathbf{w}+\mathbf{v}\cdot\mathbf{w}$.
(c) Prove that the linearity property (12c) is equivalent to the two properties given in parts (a) and (b).

12. (*Direction cosines*) The **direction cosines** of a vector $\mathbf{u} = (u_1, u_2, u_3)$ in 3-space are defined as $l_1 \equiv \cos\alpha$, $l_2 \equiv \cos\beta$, $l_3 \equiv \cos\gamma$, where α, β, γ are the angles between $\mathbf{u}$ and the positive coordinate axes, as shown.



(a) Obtain general expressions for l_1, l_2, l_3 in terms of the components u_1, u_2, u_3.
(b) Evaluate l_1, l_2, l_3 for $\mathbf{u} = (2,-1,5)$.
(c) Evaluate l_1, l_2, l_3 for $\mathbf{u} = (2,-4,1)$.
(d) Evaluate l_1, l_2, l_3 for $\mathbf{u} = (4,0,-3)$.
(e) Show that $l_1^2 + l_2^2 + l_3^2 = 1$.

13. If $\mathbf{u}\cdot\mathbf{v} = 0$ and $\mathbf{v}\cdot\mathbf{w} = 0$, does that imply that $\mathbf{u}\cdot\mathbf{w} = 0$? Prove or disprove. HINT: If a claim is true, it needs to be proved in general, that is, for all possible cases. But if it is false, it can be disproved merely by putting forward a single counterexample.

14. Determine whether or not each set of vectors is orthogonal. (a) $(1,3)$, $(-6,2)$, $(0,0)$
(b) $(2,3,0)$, $(-3,2,1)$, $(1,1,1)$, $(1,-3,1)$
(c) $(1,0,0,0)$, $(0,1,0,0)$, $(0,0,1,0)$, $(0,0,0,1)$
(d) $(1,1,1,1)$, $(1,-1,1,-1)$, $(0,1,0,-1)$, $(2,0,-2,0)$
(e) $(2,1,-1,1)$, $(1,1,3,0)$, $(1,-1,0,-1)$, $(2,1,1,1)$

15. Determine a unit vector along the line of intersection of the following two planes in $\mathbb{R}^3$. NOTE: Do not use "cross products" since this topic has not yet been discussed here.

(a) $\begin{aligned} x_1 + 2x_2 - x_3 &= 8 \\ x_1 - x_2 + x_3 &= 0 \end{aligned}$

(b) $\begin{aligned} x_1 + x_2 \qquad &= 0 \\ x_1 - x_2 + 2x_3 &= 0 \end{aligned}$

(c) $\begin{aligned} x_1 - x_2 - 5x_3 &= 0 \\ x_2 + 4x_3 &= 6 \end{aligned}$

(d) $\begin{aligned} 2x_1 \qquad + x_3 &= 1 \\ x_1 + 4x_2 \qquad &= 1 \end{aligned}$

(e) $\begin{aligned} x_1 - 5x_2 + x_3 &= 4 \\ 2x_1 - x_2 - x_3 &= 3 \end{aligned}$

(f) $\begin{aligned} x_1 + x_2 + 12x_3 &= 0 \\ x_1 + 2x_2 + 12x_3 &= 5 \end{aligned}$

(g) $\begin{aligned} x_1 - x_2 - x_3 &= 2 \\ x_1 - x_2 - 2x_3 &= 5 \end{aligned}$

16. (*Schwarz inequality*) To make (15) as close to an equality as possible, and hence as informative as possible, we minimized the left-hand side by setting $d(\text{left-hand side})/d\alpha = 0$. That step gave $\alpha = -\mathbf{u}\cdot\mathbf{v}/\|\mathbf{v}\|^2$, and putting that result back into (15) gave the Schwarz inequality. That proof is valid for $\mathbb{R}^n$ for any n (≥ 1). For the special case of $\mathbb{R}^2$, show that the optimal α is $-\mathbf{u}\cdot\mathbf{v}/\|\mathbf{v}\|^2$ by using a graphical approach; that is, using a suitable sketch. HINT: Given $\mathbf{u}$ and $\mathbf{v}$, make $\mathbf{u}+\alpha\mathbf{v}$ as short as possible.

9.6 Generalized Vector Space

9.6.1. Vector space. In Section 9.5 we generalize our vector concept from the familiar arrow vectors of 2- and 3-space to n-tuple vectors in abstract n-space, and it is n-space that is used in the remainder of this chapter and in Chapters 10–12. Yet, it is interesting to wonder if further generalization is possible. The answer is yes, and we will complete that story in this section. Far from being just a mathematical curiosity, the results will be essential in later chapters, when we study Fourier series, Sturm–Liouville theory, and partial differential equations.

The idea is as follows. In preceding sections we introduced the vectors and arithmetic rules for their manipulation, and *then* derived the various properties, such as $\mathbf{u}+\mathbf{v} = \mathbf{v}+\mathbf{u}$, $\mathbf{u}+\mathbf{0} = \mathbf{u}$, $\alpha(\beta\mathbf{u}) = (\alpha\beta)\mathbf{u}$, and so on. In generalizing, the essential idea is to reverse the cart and the horse. Specifically, we elevate the derived properties to axioms, or requirements, and regard the vectors as "objects," the nature of which is not restricted in advance. They may be chosen to be n-tuples or whatever; all that we ask is that a plus (+) operation, a zero vector, a negative inverse, and scalar multiplication be defined such that all of the vector space axioms are satisfied. Thus:

DEFINITION 9.6.1 *Vector Space*
We call a (nonempty) set $\mathcal{S}$ of "objects," which are denoted by **boldface type** and referred to as *vectors*, a **vector space** if the following requirements are met:

(i) An operation, which will be called vector *addition* and denoted as $+$, is defined between any two vectors in $\mathcal{S}$ in such a way that if $\mathbf{u}$ and $\mathbf{v}$ are in $\mathcal{S}$, then $\mathbf{u}+\mathbf{v}$ is too (i.e., $\mathcal{S}$ is *closed under addition*). Furthermore,

$$\mathbf{u}+\mathbf{v}=\mathbf{v}+\mathbf{u}, \qquad \text{(commutative)} \tag{1}$$

$$(\mathbf{u}+\mathbf{v})+\mathbf{w}=\mathbf{u}+(\mathbf{v}+\mathbf{w}). \qquad \text{(associative)} \tag{2}$$

(ii) $\mathcal{S}$ contains a unique *zero vector* $\mathbf{0}$ such that

$$\mathbf{u}+\mathbf{0}=\mathbf{u} \tag{3}$$

for each $\mathbf{u}$ in $\mathcal{S}$.

(iii) For each $\mathbf{u}$ in $\mathcal{S}$ there is a unique vector "$-\mathbf{u}$" in $\mathcal{S}$, called the *negative inverse of* $\mathbf{u}$, such that

$$\mathbf{u}+(-\mathbf{u})=\mathbf{0}. \tag{4}$$

We denote $\mathbf{u}+(-\mathbf{v})$ as $\mathbf{u}-\mathbf{v}$ for brevity, but emphasize that it is actually the $+$ operation between $\mathbf{u}$ and $-\mathbf{v}$.

(iv) Another operation, called *scalar multiplication*, is defined such that if $\mathbf{u}$ is any vector in $\mathcal{S}$ and α is any scalar,* then the scalar multiple $\alpha\mathbf{u}$ is in $\mathcal{S}$, too (i.e., $\mathcal{S}$ is *closed under scalar multiplication*). Further, we require that

$$\alpha(\beta\mathbf{u})=(\alpha\beta)\mathbf{u}, \qquad \text{(associative)} \tag{5}$$

$$(\alpha+\beta)\mathbf{u}=\alpha\mathbf{u}+\beta\mathbf{u}, \qquad \text{(distributive)} \tag{6}$$

$$\alpha(\mathbf{u}+\mathbf{v})=\alpha\mathbf{u}+\alpha\mathbf{v}, \qquad \text{(distributive)} \tag{7}$$

$$1\mathbf{u}=\mathbf{u}, \tag{8}$$

if the vectors $\mathbf{u}$, $\mathbf{v}$ are in $\mathcal{S}$, and α, β are scalars.

Observe that if we write $\mathbf{u}+\mathbf{v}+\mathbf{w}$, it is not clear whether we mean $(\mathbf{u}+\mathbf{v})+\mathbf{w}$ (i.e., first add $\mathbf{u}$ and $\mathbf{v}$, and then add the result to $\mathbf{w}$) or $\mathbf{u}+(\mathbf{v}+\mathbf{w})$. However, the associative property (2) guarantees that it does not matter, so the parentheses can be omitted without ambiguity. Similarly, $\alpha\beta\mathbf{u}$ is unambiguous by virtue of (5).

EXAMPLE 1. $\mathbb{R}^n$*-Space.* Surely, the n-space $\mathbb{R}^n$, defined earlier, does constitute a vector space; after all, the axioms listed in Definition 9.6.1 come from the properties of $\mathbb{R}^n$ listed in Section 9.4. Thus, there is no need to check to see if those axioms are satisfied.

Instead, and for heuristic purposes, let us modify our addition operation from

$$\mathbf{u}+\mathbf{v}\equiv(u_1+v_1,\ldots,u_n+v_n) \tag{9}$$

*We continue to restrict all scalars to be (finite) real numbers. Hence, we call the vector space a **real vector space**.

to

$$\mathbf{u} + \mathbf{v} \equiv (u_1 + 2v_1, \ldots, u_n + 2v_n), \tag{10}$$

and see if (10) works; that is, let us see if the vector space axioms listed under (i) in Definition 9.6.1 are still satisfied if we use (10) as our addition operation instead of (9). According to (10),

$$\mathbf{v} + \mathbf{u} \equiv (v_1 + 2u_1, \ldots, v_n + 2u_n) \tag{11}$$

so a comparison of (10) and (11) shows that the commutativity axiom (1) is satisfied only if $u_j + 2v_j = v_j + 2u_j$ $(j = 1, \ldots, n)$, hence only if $v_j = u_j$, hence only if $\mathbf{v} = \mathbf{u}$. Since (1) does not hold for any chosen vectors $\mathbf{u}$ and $\mathbf{v}$, but only for vectors $\mathbf{u}$ and $\mathbf{v}$ that are equal, we conclude that if $\mathbf{u} + \mathbf{v}$ is defined by (10), then we do *not* have a vector space. Of course, it is possible that (10) violates other axioms besides (1), but one failure is sufficient to show that the set is not a legitimate vector space.

COMMENT. Observe that we have not shown that $\mathbf{u} + \mathbf{v}$ *must* be defined as in (9); conceivably,

$$\mathbf{u} + \mathbf{v} \equiv (u_1^2 + v_1, \ldots, u_n^2 + v_n) \tag{12}$$

or

$$\mathbf{u} + \mathbf{v} \equiv (u_1 - v_1, \ldots, u_n - v_n) \tag{13}$$

might work; that is, might satisfy the requirements listed under (i). Thus, understand that the plus signs on the left- and right-hand sides of (9) are not the same. The ones on the right denote the usual addition of real numbers (e.g., $2 + 5 = 7$), whereas the one on the left is more exotic; it denotes a certain operation between vectors $\mathbf{u}$ and $\mathbf{v}$, which is being defined by (9), or (10), or (12), or (13). To emphasize that point we could use a different notation such as $\mathbf{u} * \mathbf{v}$, in place of $\mathbf{u} + \mathbf{v}$, as some authors do. However, having made that point let us continue to use $\mathbf{u} + \mathbf{v}$. ∎

$\mathbb{R}^n$ is but one example of a vector space. Many other useful spaces can be introduced by using objects other than n-tuples as the vectors. For example, the vectors may be functions, matrices, or whatever, provided that vector addition, a zero vector, a negative inverse, and scalar multiplication are defined such that all of the vector space axioms are satisfied. For nowhere in Definition 9.6.1 is the *nature* of the vectors specified or in any way restricted.

EXAMPLE 2. *A Function Space.* This time, let the vectors be functions. Specifically, let $\mathbf{u} = u(x)$ be any continuous function defined on $0 \le x \le 1$, say. For the addition operation let

$$\mathbf{u} + \mathbf{v} \equiv u(x) + v(x); \tag{14a}$$

that is, let $\mathbf{u} + \mathbf{v}$ be the function whose values are the ordinary sum $u(x) + v(x)$. For scalar multiplication let

$$\alpha\mathbf{u} \equiv \alpha u(x); \tag{14b}$$

for the zero vector choose the zero function

$$\mathbf{0} \equiv 0; \tag{14c}$$

and for the negative of $\mathbf{u}$ define

$$-\mathbf{u} \equiv -u(x); \tag{14d}$$

that is, the function whose values are $-u(x)$.

With these definitions, we can verify that all of the vector space requirements are satisfied, so that the set $\mathcal{S}$ of such vectors is a bona fide vector space. For instance, if $\mathbf{u} = u(x)$ and $\mathbf{v} = v(x)$ are continuous on $0 \le x \le 1$, then so is $\mathbf{u} + \mathbf{v} = u(x) + v(x)$ so $\mathcal{S}$ is closed under addition. Further, $\mathbf{v} + \mathbf{u} = v(x) + u(x) = u(x) + v(x) = \mathbf{u} + \mathbf{v}$,* so addition satisfies the commutative property (1), and so on.

This $\mathcal{S}$ is but one example of a **function space**, a space in which the vectors are functions. ∎

The following theorem is useful, and its proof illustrates the axiomatic approach.

THEOREM 9.6.1 *Properties of Scalar Multiplication*
If $\mathbf{u}$ is any vector in a vector space $\mathcal{S}$ and α is any scalar, then

$$0\mathbf{u} = \mathbf{0}, \tag{15a}$$

$$(-1)\mathbf{u} = -\mathbf{u}, \tag{15b}$$

$$\alpha\mathbf{0} = \mathbf{0}. \tag{15c}$$

Proof: These results follow from our definition of vector space. To prove (15a), one line of approach is as follows:

$$\begin{aligned} 0\mathbf{u} + \mathbf{u} &= 0\mathbf{u} + 1\mathbf{u} && \text{by (8)} \\ &= (0+1)\mathbf{u} && \text{by (6)} \\ &= 1\mathbf{u} \\ &= \mathbf{u} && \text{by (8).} \end{aligned}$$

Then

$$\begin{aligned} 0\mathbf{u} + \mathbf{u} + (-\mathbf{u}) &= \mathbf{u} + (-\mathbf{u}) \\ 0\mathbf{u} + \mathbf{0} &= \mathbf{0} && \text{by (4),} \\ 0\mathbf{u} &= \mathbf{0} && \text{by (3).} \end{aligned}$$

The remaining two, (15b) and (15c), are left for the exercises. ∎

9.6.2. Inclusion of inner product and/or norm. Observe that there is no mention of a dot product or a norm either in Definition 9.6.1 or in Examples 1 or 2. Indeed, a vector space $\mathcal{S}$ need not *have* a dot product (also called an **inner product**) or a norm defined for it. If it does have an inner product it is called an *inner product*

*The second equality holds because $v(x) + u(x)$ is the ordinary sum of two real numbers; e.g., $4 + 3 = 3 + 4$.

space; if it has a norm it is called a *normed vector space*; and if it has both it is called a *normed inner product space*.

If we do choose to introduce an inner product for S, how is it to be defined? Do you remember the idea of reversing the cart and the horse? That is how we do it. Equations (12a,b,c) in Section 9.5.2 were shown to be properties of the inner product $\mathbf{u} \cdot \mathbf{v} = u_1 v_1 + \cdots + u_n v_n$. We now take those properties and elevate them to axioms, or requirements, that are to be satisfied by any inner product of any vector space.

Similarly, we take the properties (17a,b,c) of the norm, in Section 9.5.3, and elevate them to axioms, or requirements, that are to be satisfied by any norm of any vector space.

Let us tabulate them here:

REQUIREMENTS OF INNER PRODUCT

Commutative:	$\mathbf{u} \cdot \mathbf{v} = \mathbf{v} \cdot \mathbf{u},$		(16a)
Nonnegative:	$\mathbf{u} \cdot \mathbf{u} > 0$	for all $\mathbf{u} \neq \mathbf{0}$,	
	$= 0$	for $\mathbf{u} = \mathbf{0}$,	(16b)
Linear:	$(\alpha \mathbf{u} + \beta \mathbf{v}) \cdot \mathbf{w} = \alpha(\mathbf{u} \cdot \mathbf{w}) + \beta(\mathbf{v} \cdot \mathbf{w}),$		(16c)

and

REQUIREMENTS OF NORM

Scaling:	$\|\alpha \mathbf{u}\| = \|\alpha\| \, \|\mathbf{u}\|,$		(17a)
Nonnegative:	$\|\mathbf{u}\| > 0$	for all $\mathbf{u} \neq \mathbf{0}$,	
	$= 0$	for $\mathbf{u} = \mathbf{0}$,	(17b)
Triangle Inequality:	$\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|.$		(17c)

Let us illustrate.

EXAMPLE 3. $\mathbb{R}^n$*-Space.* If we wish to add an inner product to the vector space $\mathbb{R}^n$, we can use the choice

$$\mathbf{u} \cdot \mathbf{v} \equiv u_1 v_1 + \cdots + u_n v_n = \sum_{j=1}^{n} u_j v_j. \tag{18a}$$

We know that (18a) satisfies the requirements (16) because the latter were deduced, in Section 9.5.2, as properties that *follow* from (18a). A variation of (18a) that still satisfies (16) is (Exercise 6)

$$\mathbf{u}\cdot\mathbf{v} \equiv w_1u_1v_1 + \cdots + w_nu_nv_n = \sum_{j=1}^{n} w_ju_jv_j, \tag{18b}$$

where the w_j's are fixed positive constants known as "weights" because they attach more or less weight to the different components of $\mathbf{u}$ and $\mathbf{v}$. For instance, consider $\mathbb{R}^2$ and let $w_1 = 5$ and $w_2 = 3$. Then if $\mathbf{u} = (2, -4)$ and $\mathbf{v} = (1, 6)$ we have $\mathbf{u}\cdot\mathbf{v} = 5(2)(1) + 3(-4)(6) = -62$.

Note that for (18b) to be a legitimate inner product we must have $w_j > 0$ for each j. For suppose, still in $\mathbb{R}^2$, that $w_1 = 3$ and $w_2 = -2$. Then, for $\mathbf{u} = (1, 5)$, say, we have $\mathbf{u}\cdot\mathbf{u} = 3(1)(1) - 2(5)(5) = -47 < 0$, in violation of (16b). Or, suppose that $w_1 = 3$ and $w_2 = 0$. Then, for $\mathbf{u} = (0, 4)$, say, we have $\mathbf{u}\cdot\mathbf{u} = 3(0)(0) + 0(4)(4) = 0$ even though $\mathbf{u} \neq \mathbf{0}$, again in violation of (16b).

Now, suppose that we wish to add a norm. If for any vector space $\mathcal{S}$ we already have an inner product, then a legitimate norm can always be obtained from that inner product as $\|\mathbf{u}\| = \sqrt{\mathbf{u}\cdot\mathbf{u}}$, and that choice is called the *natural norm*. Thus, the natural norms corresponding to (18a) and (18b) are

$$\|\mathbf{u}\| \equiv \sqrt{\sum_{1}^{n} u_j^2} \quad \text{and} \quad \|\mathbf{u}\| \equiv \sqrt{\sum_{1}^{n} w_ju_j^2}, \tag{19a,b}$$

respectively.

However, we do not *have* to choose the natural norm. For instance, we could use (18a) as our inner product, and choose

$$\|\mathbf{u}\| \equiv |u_1| + \cdots + |u_n| = \sum_{j=1}^{n} |u_j| \tag{20}$$

as our norm (Exercise 8). The latter is used by Struble in his book on differential equations,* probably because it is algebraically simpler than the *Euclidean norm* (19a) or the *modified Euclidean norm* (19b). Furthermore, he defines no inner product whatsoever. Struble calls (20) the *taxicab norm* since a taxicab driver judges the distance from the corner of 5th Avenue and 34th Street to the corner of 2nd Avenue and 49th Street as 18 blocks, not $\sqrt{234}$ blocks. ∎

EXAMPLE 4. *The Function Space of Example 2.* How might we choose an inner product for the *function* space $\mathcal{S}$ defined in Example 2? To motivate our choice, let us imagine approximating any given function (i.e., vector) $u(x)$ in $\mathcal{S}$ in a piecewise-constant manner as depicted in Fig. 1. That is, divide the x interval $(0 \leq x \leq 1)$ into n equal parts and define the approximating piecewise-constant function, over each subinterval as

*R. A. Struble, *Nonlinear Differential Equations* (New York: McGraw-Hill, 1962).

the value of $u(x)$ at the left endpoint of that subinterval. If we represent the piecewise-constant function as the n-tuple $(u_1, \ldots, u_n)$, then we have, in a heuristic sense,

$$u(x) \approx (u_1, \ldots, u_n). \tag{21}$$

Similarly, for any other function $v(x)$ in $\mathcal{S}$,

$$v(x) \approx (v_1, \ldots, v_n). \tag{22}$$

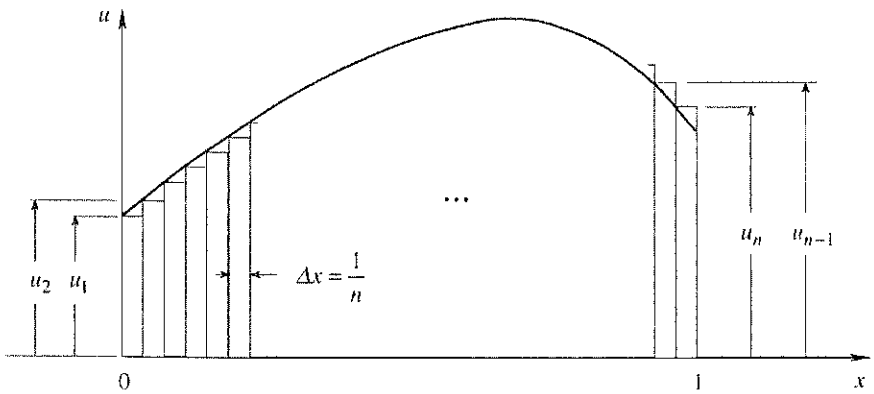


Figure 1. Staircase approximation of $u(x)$.

The n-tuple vectors on the right-hand sides of (21) and (22) are members of $\mathbb{R}^n$. For that space, let us adopt the inner product

$$(u_1, \ldots, u_n) \cdot (v_1, \ldots, v_n) = \sum_{j=1}^{n} u_j v_j \Delta x, \tag{23}$$

that is, (18b) with all of the w_j weights the same, namely, the subinterval width Δx. If we let $n \to \infty$, the "staircase approximations" approach $u(x)$ and $v(x)$, and the sum in (23) tends to the integral $\int_0^1 u(x)v(x)\,dx$.

This heuristic reasoning suggests the inner product

$$\boxed{\Big\langle u(x), v(x) \Big\rangle \equiv \int_0^1 u(x)v(x)\,dx.} \tag{24a}$$

We can denote it as $\mathbf{u} \cdot \mathbf{v}$ and call it the dot product, or we can denote it as $< u(x), v(x) >$ and call it the inner product. For function spaces, the latter notation is somewhat standard, and is our choice in this text.

COMMENT 1. By no means do we claim our staircase idea to be a rigorous derivation of (24a). In fact, it is neither rigorous nor a derivation; it is *heuristic motivation* for the *definition* (24a). We leave it for the exercises to verify that (24a) does satisfy the requirements (16).

COMMENT 2. Just as (18b) is a legitimate generalization of (18a), (if $w_j > 0$ for $1 \le j \le n$), we expect that

$$\boxed{\Big\langle u(x), v(x) \Big\rangle \equiv \int_0^1 u(x)v(x)w(x)\,dx} \tag{24b}$$

is a legitimate generalization of (24a) [if $w(x) > 0$ for $0 \le x \le 1$], proof of which claim is left for the exercises. The inner product (24b) is prominent when we study Fourier series and the Sturm–Liouville theory in Chapter 17.

COMMENT 3. Naturally, if we wish to define a norm as well, we could use a natural norm based on (24a) or (24b), for instance

$$\|\mathbf{u}\| \equiv \sqrt{< u(x), u(x) >} = \sqrt{\int_0^1 u^2(x)w(x)\,dx} \tag{25}$$

based on (24b).

COMMENT 4. Notice carefully that the concept of the *dimension* of a vector space has not yet been introduced, although it is in Section 9.10. There, we define dimension and find that $\mathbb{R}^n$ is n-dimensional (which claim is probably not a great shock). Since the staircase approximation (21) becomes exact only as $n \to \infty$, it appears that our function space S is infinite dimensional!

COMMENT 5. A bit of notation: the set of functions that are defined and continuous on $[0, 1]$ (i.e., $0 \le x \le 1$) is usually denoted as $C^0[0, 1]$. If not only are the functions continuous but also all derivatives through order k, then the set is denoted as $C^k[0, 1]$. ∎

Closure. Using n-space as a ladder, we complete our generalization of vector space by taking the properties of $\mathbb{R}^n$ (such as $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$) and turning them into the axioms, or requirements, to be met by *any* vector space. Thus, attention shifted from the objects, the vectors, to those requirements. There is no restriction on the nature of the vectors, which can be arrows, n-tuples, matrices, functions, or oranges. For us, the most important vector spaces are $\mathbb{R}^n$ and various function spaces; $\mathbb{R}^n$ is used in the remainder of this chapter and Chapters 10–12, and function spaces are used in Chapter 17 when we study Fourier series and Sturm-Liouville theory.

To illustrate the power of the axiomatic approach, recall the Schwarz inequality $|\mathbf{u} \cdot \mathbf{v}| \le \|\mathbf{u}\|\,\|\mathbf{v}\|$, proved in Section 9.5.2 for $\mathbb{R}^n$. That result holds for *any* normed inner product space with natural norm $\|\mathbf{u}\| = \sqrt{\mathbf{u} \cdot \mathbf{u}}$ for it followed from properties of $\mathbb{R}^n$, which properties are subsequently elevated to axioms for general vector space. Thus, it represents many properties rolled into one. For example, in $\mathbb{R}^n$, with the dot product (18a) it says

$$\left|\sum_{j=1}^{n} u_j v_j\right| \le \sqrt{\sum_1^n u_j^2}\,\sqrt{\sum_1^n v_j^2}, \tag{26}$$

in the function space of Examples 2 and 4; with the inner product (24b) and norm (25) it says

$$\left|\int_0^1 u(x)v(x)w(x)\,dx\right| \le \sqrt{\int_0^1 u^2(x)w(x)\,dx}\,\sqrt{\int_0^1 v^2(x)w(x)\,dx}, \tag{27}$$

and so on.

EXERCISES 9.6

1. Recall that $\mathbb{R}^n$ is the vector space ("real" vector space since all scalars are to be real numbers) in which the vectors are n-tuples $\mathbf{u} = (u_1, \ldots, u_n)$, with the definitions

$$\begin{aligned} \mathbf{u} + \mathbf{v} &= (u_1, \ldots, u_n) + (v_1, \ldots, v_n) \\ &\equiv (u_1 + v_1, \ldots, u_n + v_n), && (1.1) \\ \mathbf{0} &\equiv (0, \ldots, 0), && (1.2) \\ -\mathbf{u} &\equiv (-u_1, \ldots, -u_n), && (1.3) \\ \alpha\mathbf{u} &\equiv (\alpha u_1, \ldots, \alpha u_n). && (1.4) \end{aligned}$$

If we make the following modifications, do we still have a vector space? If not, specify all requirements within Definition 9.6.1 that fail to be met.

(a) only vectors of the form $\mathbf{u} = (u, u, \ldots, u)$ admitted, where $-\infty < u < \infty$
(b) only vectors of the form $\mathbf{u} = (u, 2u, 3u, \ldots, nu)$ admitted, where $-\infty < u < \infty$
(c) only the vector $(0, \ldots, 0)$ admitted (this is an example of a **zero vector space**, a vector space containing only the zero vector)
(d) $\mathbf{u} + \mathbf{v} \equiv (u_1 - v_1, \ldots, u_n - v_n)$, in place of (1.1)
(e) $\mathbf{u} + \mathbf{v} \equiv (0, \ldots, 0)$ for all $\mathbf{u}$'s and $\mathbf{v}$'s, in place of (1.1)
(f) $\alpha\mathbf{u} \equiv (\alpha^2 u_1, \ldots, \alpha^2 u_n)$, in place of (1.4)

2. We noted in Example 1 that the definition (10) of vector addition violates axiom (1). Does it violate any others as well? Explain.

3. Prove (15b), that $(-1)\mathbf{u} = -\mathbf{u}$.

4. Prove (15c), that $\alpha\mathbf{0} = \mathbf{0}$.

5. Prove that if $\alpha\mathbf{u} = \mathbf{0}$ then $\alpha = 0$ and/or $\mathbf{u} = \mathbf{0}$.

6. Show that the inner product (18b) does satisfy the requirements (16).

7. We stated in Example 3 that if for any vector space $\mathcal{S}$ we already have an inner product, then a legitimate norm can always be obtained from that inner product as $\|\mathbf{u}\| = \sqrt{\mathbf{u} \cdot \mathbf{u}}$, which choice is called the natural norm. Prove that claim.

8. Show that the "taxicab norm" (20) is a legitimate norm – that is, that it satisfies the requirements (17).

9. (a) Does the choice $\|\mathbf{u}\| = \max\limits_{1 \le j \le n} |u_j|$, for $\mathbb{R}^n$, satisfy the requirements (17)? Explain.
(b) How about $\|\mathbf{u}\| = \min\limits_{1 \le j \le n} |u_j|$, for $\mathbb{R}^n$?

10. Let $\mathcal{S}$ be the set of real-valued polynomial functions, of degree n, defined on $a \le x \le b$. If $\mathbf{u} = a_0 + a_1 x + \cdots + a_n x^n$ and $\mathbf{v} = b_0 + b_1 x + \cdots + b_n x^n$ are any two such functions, and α is any (real) scalar, define the sum $\mathbf{u} + \mathbf{v}$ and the scalar multiple $\alpha\mathbf{u}$ as

$$\begin{aligned} (\mathbf{u} + \mathbf{v})(x) &= (a_0 + b_0) + (a_1 + b_1)x + \cdots + (a_n + b_n)x^n, \\ (\alpha\mathbf{u})(x) &= \alpha a_0 + \alpha a_1 x + \cdots + \alpha a_n x^n, \end{aligned}$$

respectively. Further, let $\mathbf{0}$ be the function $0 + 0x + \cdots + 0x^n$, and let $-\mathbf{u}$ be the function $-a_0 - a_1 x + \cdots - a_n x^n$. Show that $\mathcal{S}$ is a vector space.

11. Show that the inner product (24b) does satisfy the requirements (16).

12. (*Schwarz inequality*) We derive the Schwarz inequality

$$|\mathbf{u} \cdot \mathbf{v}| \le \|\mathbf{u}\| \, \|\mathbf{v}\| \qquad (12.1)$$

for $\mathbb{R}^n$ space in Section 9.5.2. The latter holds not only for $\mathbb{R}^n$ but for *any* normed inner product space with the natural norm $\|\mathbf{u}\| = \sqrt{\mathbf{u} \cdot \mathbf{u}}$. In this exercise we simply ask you to verify (12.1) by working out the left- and right-hand sides for these specific cases:

(a) $\mathbf{u} = (3, 1, -1, 0)$ and $\mathbf{v} = (1, 2, 5, -4)$ in $\mathbb{R}^4$, with the inner product (18a)
(b) $\mathbf{u} = (1, 2, 4, -3)$ and $\mathbf{v} = (0, 4, 1, 1)$ in $\mathbb{R}^4$, with $\mathbf{u} \cdot \mathbf{v} = u_1 v_1 + 5u_2 v_2 + 3u_3 v_3 + 2u_4 v_4$
(c) $\mathbf{u} = (1, 1, 1, 1, 1)$ and $\mathbf{v} = (2, 2, 2, 2, 2)$ in $\mathbb{R}^5$, with $\mathbf{u} \cdot \mathbf{v} = u_1 v_1 + 2u_2 v_2 + 3u_3 v_3 + 4u_4 v_4 + 5u_5 v_5$
(d) $\mathbf{u} = 2 + x$ and $\mathbf{v} = 3x^2$ in the function space of Example 4, with the inner product $\mathbf{u} \cdot \mathbf{v} = \langle u(x), v(x) \rangle = \int_0^1 u(x)v(x)\,dx$
(e) Same as (d), but with $\langle u(x), v(x) \rangle = \int_0^1 u(x)v(x)(2 + 5x)\,dx$

13. (*Solution space*) (a) Consider a set of m linear homogeneous algebraic equations in the n unknowns $x_1, \ldots, x_n$, and denote each solution of the system as an n-tuple vector $\mathbf{x} = (x_1, \ldots, x_n)$ in $\mathbb{R}^n$. Show that the set of all such vectors, with the usual definitions [$\mathbf{u} + \mathbf{v} = (u_1 + v_1, \ldots, u_n + v_n)$, $\alpha\mathbf{u} = (\alpha u_1, \ldots, \alpha u_n)$, $-\mathbf{u} = (-u_1, \ldots, -u_n)$, $\mathbf{0} = (0, \ldots, 0)$], is a vector space. That space is called the **solution space** of the system.

(b) If the system is *non*homogeneous, is the set of solutions still a vector space? Explain.

14. (*Solution space*) Show that the solutions of a linear homogeneous differential equation (with the same definitions of $\mathbf{u}+\mathbf{v}$, $\alpha\mathbf{u}$, $-\mathbf{u}$, and $\mathbf{0}$ as in Example 2) constitute a vector space, the so-called **solution space** of that differential equation.

9.7 Span and Subspace

Here, we begin a sequence of closely related ideas: span, linear dependence, basis, expansion, and dimension. The concepts, definitions, and theorems hold for any vector space, but our illustrative examples are restricted to the n-space $\mathbb{R}^n$, this being the case of most interest in Chapters 9–12.

We begin with the idea of the "span" of a set of vectors.

DEFINITION 9.7.1 *Span*
If $\mathbf{u}_1, \ldots, \mathbf{u}_k$ are vectors in a vector space S, then the set of all linear combinations of these vectors, that is, all vectors of the form

$$\mathbf{u} = \alpha_1\mathbf{u}_1 + \cdots + \alpha_k\mathbf{u}_k, \tag{1}$$

where $\alpha_1, \ldots, \alpha_k$ are scalars is called the **span** of $\mathbf{u}_1, \ldots, \mathbf{u}_k$ and is denoted as span $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$.

The set $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ is called the **generating set** of span $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$.

Let us illustrate with some vector sets in $\mathbb{R}^2$ and $\mathbb{R}^3$ so we can support the discussion with diagrams.

EXAMPLE 1. Determine the span of the single vector

$$\mathbf{u}_1 = (4, 2) \tag{2}$$

in $\mathbb{R}^2$. Then span $\{\mathbf{u}_1\}$ is the set of all vectors that are scalar multiples of $\mathbf{u}_1$. Hence, span $\{\mathbf{u}_1\}$ is the set of all vectors on the line L in Fig. 1, such as $\mathbf{u} = 2\mathbf{u}_1 = (8, 4)$, $\mathbf{v} = -\frac{1}{2}\mathbf{u}_1 = (-2, -1)$, and $\mathbf{0} = 0\mathbf{u}_1 = (0, 0)$. We say that $\mathbf{u}_1$ *generates* the line L. ∎

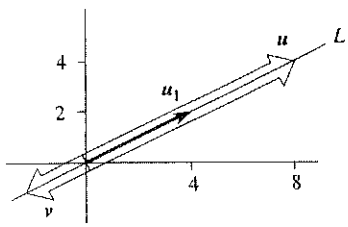


Figure 1. Span $\{\mathbf{u}_1\}$.

EXAMPLE 2. Determine the span of the two vectors

$$\mathbf{u}_1 = (4, 2), \quad \mathbf{u}_2 = (-8, -4). \tag{3}$$

Span $\{\mathbf{u}_1, \mathbf{u}_2\}$ is, once again, the line L in Fig. 1 (i.e., the set of all vectors on L), for both $\mathbf{u}_1$ and $\mathbf{u}_2$ lie along L, so any linear combination of them, $\alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2$, does too. Similarly, span $\{(4,2), (-8,-4), (18,9), (0,0)\}$ is the line L. ∎

Observe that the line L, in Examples 1 and 2, is only a subset of the vector space $\mathbb{R}^2$. Observe that that subset of $\mathbb{R}^2$ is itself a vector space, a so-called "subspace" of $\mathbb{R}^2$. For if $\mathbf{u}$ and $\mathbf{v}$ are any two vectors on L, then $\mathbf{u} + \mathbf{v}$ is on L, too, so the set *is* closed under addition; similarly, if $\mathbf{u}$ is on L, so is $\alpha\mathbf{u}$, for any scalar α, so the set *is* closed under scalar multiplication; L *does* contain the zero vector [since we can set all the α's in (1) equal to zero]; and for each $\mathbf{u}$ on L there is a (unique) vector $-\mathbf{u}$ on L such that $\mathbf{u} + (-\mathbf{u}) = \mathbf{0}$.

DEFINITION 9.7.2 *Subspace*
If a subset $\mathcal{T}$ of a vector space $\mathcal{S}$ is itself a vector space (with the same definitions as $\mathcal{S}$ for vector addition $\mathbf{u}+\mathbf{v}$, scalar multiplication $\alpha\mathbf{u}$, zero vector $\mathbf{0}$, and negative vector $-\mathbf{u}$), then $\mathcal{T}$ is a **subspace** of $\mathcal{S}$.

Usually, a subspace of $\mathcal{S}$ is only a part of $\mathcal{S}$, as the line L is only a part of $\mathbb{R}^2$, but since a subset of a set can be all of that set, a subspace of $\mathcal{S}$ can be all of $\mathcal{S}$. For instance, $\mathbb{R}^2$ is a subspace of $\mathbb{R}^2$.

THEOREM 9.7.1 *Span as Subspace*
If $\mathbf{u}_1, \ldots, \mathbf{u}_k$ are vectors in a vector space $\mathcal{S}$, then span $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ is itself a vector space, a subspace of $\mathcal{S}$.

For instance, the line L in Fig. 1 is a subspace of $\mathbb{R}^2$. Proof of Theorem 9.7.1 is left for the exercises.

EXAMPLE 3. Is the span of

$$\mathbf{u}_1 = (5,1), \quad \mathbf{u}_2 = (1,3) \tag{4}$$

all of $\mathbb{R}^2$ or only a part of $\mathbb{R}^2$? To determine the extent of span $\{\mathbf{u}_1, \mathbf{u}_2\}$, let $\mathbf{v} = (v_1, v_2)$ be any given vector in $\mathbb{R}^2$, and try to express

$$\mathbf{v} = \alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2. \tag{5}$$

That is,

$$\begin{aligned}(v_1, v_2) &= \alpha_1(5,1) + \alpha_2(1,3)\\ &= (5\alpha_1, \alpha_1) + (\alpha_2, 3\alpha_2).\\ &= (5\alpha_1 + \alpha_2, \alpha_1 + 3\alpha_2).\end{aligned} \tag{6}$$

Equating components, we obtain the linear equations

$$\begin{aligned} 5\alpha_1 + \alpha_2 &= v_1, \\ \alpha_1 + 3\alpha_2 &= v_2 \end{aligned} \tag{7}$$

in α_1, α_2. Applying Gauss elimination, (7) becomes

$$\begin{aligned} \alpha_1 + \tfrac{1}{5}\alpha_2 &= \tfrac{1}{5}v_1, \\ \alpha_2 &= \tfrac{5}{14}v_2 - \tfrac{1}{14}v_1. \end{aligned} \tag{8}$$

It is clear from the Gauss-reduced form (8) that the system is *consistent* (solvable for α_1, α_2) for *every* vector $\mathbf{v}$ in $\mathbb{R}^2$. Hence, we may conclude that span $\{\mathbf{u}_1, \mathbf{u}_2\}$ is all of $\mathbb{R}^2$; we say that $\{\mathbf{u}_1, \mathbf{u}_2\}$ spans $\mathbb{R}^2$. (Here we use "span" as a verb; in Definition 9.7.1 it is introduced as a noun.)

Thus, every $\mathbf{v}$ in $\mathbb{R}^2$ can be expressed as a linear combination of vector $\mathbf{u}_1$ and $\mathbf{u}_2$. As representative, let $\mathbf{v} = (6, -4)$ so $v_1 = 6$ and $v_2 = -4$. Then (8) gives $\alpha_2 = -\frac{13}{7}$ and $\alpha_1 = \frac{11}{7}$, so that (5) becomes

$$\mathbf{v} = \tfrac{11}{7}\mathbf{u}_1 - \tfrac{13}{7}\mathbf{u}_2. \tag{9}$$

To see this in graphical terms, observe from Fig. 2 that $\mathbf{v} = \mathbf{OA} + \mathbf{OB}$, where (with the aid of a scale) $\mathbf{OA} \approx 1.6\mathbf{u}_1$ and $\mathbf{OB} \approx -1.9\mathbf{u}_2$. Thus, $\mathbf{v} \approx 1.6\mathbf{u}_1 - 1.9\mathbf{u}_2$, in agreement with (9).

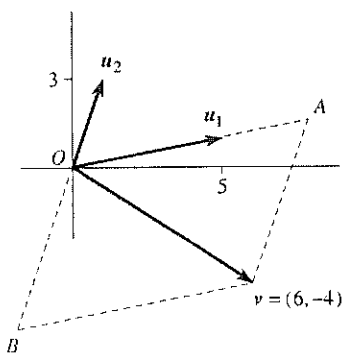


Figure 2. Representation of $\mathbf{v}$.

COMMENT. Suppose that we add $\mathbf{u}_3 = (2, 2)$ to the set. It should be evident that span $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ is all of $\mathbb{R}^2$, again, since $\{\mathbf{u}_1, \mathbf{u}_2\}$ spanned $\mathbb{R}^2$ even "without any help" from $\mathbf{u}_3$. But in case this is not clear, let us go through steps analogous to steps (5) to (8):

$$\mathbf{v} = \alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2 + \alpha_3\mathbf{u}_3 \tag{10}$$

so $(v_1, v_2) = (5\alpha_1 + \alpha_2 + 2\alpha_3, \alpha_1 + 3\alpha_2 + 2\alpha_3)$. Thus,

$$\begin{aligned} 5\alpha_1 + \alpha_2 + 2\alpha_3 &= v_1, \\ \alpha_1 + 3\alpha_2 + 2\alpha_3 &= v_2, \end{aligned}$$

or

$$\begin{aligned} \alpha_1 + \tfrac{1}{5}\alpha_2 + \tfrac{2}{5}\alpha_3 &= \tfrac{1}{5}v_1, \\ \alpha_2 + \tfrac{4}{7}\alpha_3 &= \tfrac{5}{14}v_2 - \tfrac{1}{14}v_1. \end{aligned} \tag{11}$$

Like (8), (11) is *consistent* for every $\mathbf{v}$ in $\mathbb{R}^2$ so $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ spans $\mathbb{R}^2$, as claimed. Whereas (8) had a unique solution so that the representation (5) was unique, (11) happens to have an infinity of solutions so that the representation (10) is *not* unique. ∎

EXAMPLE 4. As a final example, consider the span of

$$\mathbf{u}_1 = (1, 2, 2), \quad \mathbf{u}_2 = (-1, 0, 2) \tag{12}$$

in $\mathbb{R}^3$. Setting

$$\mathbf{v} = \alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2, \tag{13}$$

we have

$$\begin{aligned} \alpha_1 - \alpha_2 &= v_1, \\ 2\alpha_1 \qquad &= v_2, \\ 2\alpha_1 + 2\alpha_2 &= v_3, \end{aligned}$$

or, after Gauss elimination,

$$\begin{aligned} \alpha_1 - \alpha_2 &= v_1, \\ \alpha_2 &= \tfrac{1}{2}v_2 - v_1, \\ 0 &= v_3 - 2v_2 + 2v_1. \end{aligned} \tag{14}$$

Now, span $\{\mathbf{u}_1, \mathbf{u}_2\}$ is the set of all possible vectors $\mathbf{v}$ given by (13), i.e., all vectors $\mathbf{v}$ for which the system (14) is consistent, i.e., all vectors $\mathbf{v} = (v_1, v_2, v_3)$ such that

$$2v_1 - 2v_2 + v_3 = 0 \tag{15}$$

[so that the last of equations (14) is $0 = 0$ rather than a contradiction].

In geometrical terms, on the other hand, span $\{\mathbf{u}_1, \mathbf{u}_2\}$ should be the subset of $\mathbb{R}^3$ consisting of the *plane* that passes through $\mathbf{u}_1$ and $\mathbf{u}_2$ ($\mathbf{u}_1$ and $\mathbf{u}_2$ are shown in Fig. 3). How does that fact correlate with (15)? As a matter of fact, (15) *is* the equation of a plane in 3-space, and that plane does pass through the origin, through the tip of $\mathbf{u}_1$ [i.e., the point $(1, 2, 2)$], and through the tip of $\mathbf{u}_2$ [the point $(-1, 0, 2)$]. Hence, it *is* the plane through $\mathbf{u}_1$ and $\mathbf{u}_2$ so the analytical approach, namely, steps (13) to (15) and our geometrical interpretation are in agreement.

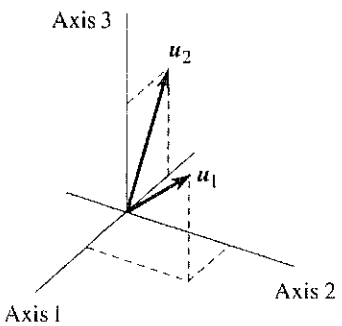


Figure 3. $\mathbf{u}_1$ and $\mathbf{u}_2$.

We conclude that span $\{\mathbf{u}_1, \mathbf{u}_2\}$ is not all of $\mathbb{R}^3$; it is only the subspace of $\mathbb{R}^3$ consisting of the plane (i.e., all vectors in the plane) containing the given vectors $\mathbf{u}_1$ and $\mathbf{u}_2$.

COMMENT. Since span $\{\mathbf{u}_1, \mathbf{u}_2\}$ is a plane, would it be correct to say that span $\{\mathbf{u}_1, \mathbf{u}_2\}$ is $\mathbb{R}^2$? No, that would be incorrect; $\mathbb{R}^2$ is made up of *two*-tuples, while the vectors in the above-mentioned plane are *three*-tuples. Thus, $\mathbb{R}^2$ space is not relevant in this problem. All that can be said here is that span $\{\mathbf{u}_1, \mathbf{u}_2\}$ is the subspace of $\mathbb{R}^3$ consisting of the plane containing the vectors $\mathbf{u}_1$ and $\mathbf{u}_2$, that is, the plane defined by (15). ∎

Closure. In leading up to the concept of bases and expansions, the two key ideas are span and linear independence. In this section we introduce the idea of span; in the next section we introduce linear dependence and linear independence. Although the concept of span holds for any vector space, such as $\mathbb{R}^6$, we suggest that you focus on the foregoing examples in two- and three-spaces, so that you can use the two- and three-dimensional drawings to promote understanding.

EXERCISES 9.7

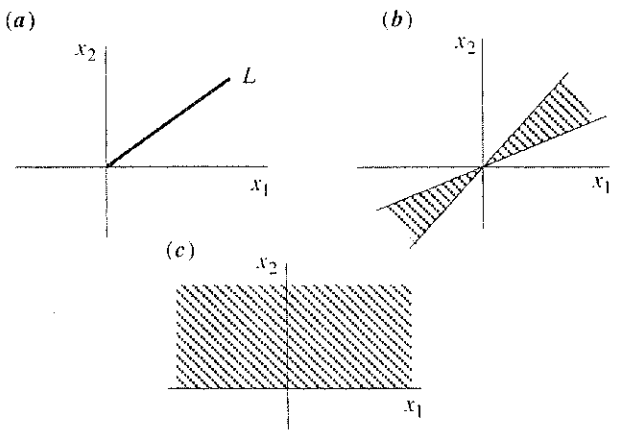
1. Show whether the vectors

(a) $(1,0,\ldots,0), (0,1,0,\ldots,0), \ldots, (0,\ldots,0,1)$ span $\mathbb{R}^n$
(b) $(0,0,0,1), (0,0,1,1), (0,1,1,1), (1,1,1,1)$ span $\mathbb{R}^4$
(c) $(1,2,0,4), (2,3,1,-1), (0,1,0,1), (0,0,0,0), (1,1,2,3)$ span $\mathbb{R}^4$
(d) $(1,3,2,2), (5,7,1,0), (-1,-2,-4,3)$ span $\mathbb{R}^4$
(e) $(1,0,1), (2,1,-1), (1,2,-5)$ span $\mathbb{R}^3$
(f) $(1,1,2), (0,0,0), (2,1,0), (-1,0,3)$ span $\mathbb{R}^3$
(g) $(2,0,3), (-1,2,4), (-5,2,-2)$ span $\mathbb{R}^3$
(h) $(1,3,0), (2,-1,1), (1,1,4)$ span $\mathbb{R}^3$
(i) $(-1,2,4), (-5,2,-2), (2,0,3), (1,2,3)$ span $\mathbb{R}^3$
(j) $(0,0,0), (2,1,4), (-1,3,5)$ span $\mathbb{R}^3$
(k) $(2,1,3), (1,-1,2)$ span $\mathbb{R}^3$
(l) $(2,1,-1), (1,3,1), (5,5,-1), (0,5,3)$ span $\mathbb{R}^3$
(m) $(-4,1,0), (2,2,2), (1,2,3)$ span $\mathbb{R}^3$
(n) $(-3,1,0), (1,1,1), (-1,7,5)$ span $\mathbb{R}^3$
(o) $(1,2), (2,1)$ span $\mathbb{R}^2$
(p) $(1,2), (2,1), (4,5)$ span $\mathbb{R}^2$
(q) $(1,2), (2,1), (2,3), (2,-4)$ span $\mathbb{R}^2$

2. (a) Sketch any two vectors that span the space of all vectors in the plane of the paper.
(b) Sketch any three such vectors.
(c) Sketch any four such vectors.

3. Are the following vector sets subspaces of $\mathbb{R}^2$? (See accompanying figure.) Explain.

(a) the straight line L that extends from the origin to infinity
(b) the wedge-shaped region (including its boundary lines) that extends to infinity in both directions
(c) the upper half plane $x_2 \geq 0$



4. (*Solution space*) First, review Exercise 13a in Section 9.6. That solution space is a subspace of $\mathbb{R}^n$. To illustrate, consider the simple system $x_1 + 3x_2 = 0$; that is, $m = 1$, $n = 2$, $a_{11} = 1$, and $a_{12} = 3$. The solution is $x_2 = \alpha$ (arbitrary), $x_1 = -3\alpha$, or $\mathbf{x} = (x_1, x_2) = \alpha(-3, 1)$ so the solution space is the span of the vector $(-3, 1)$, that is, span $\{(-3, 1)\}$. In this manner, determine the solution space for each of the following examples.

(a) $x_1 - x_2 + 4x_3 = 0$ in $\mathbb{R}^3$
(b) $x_1 + x_2 + x_3 - x_4 = 0$ in $\mathbb{R}^4$
(c) $\begin{aligned} x_1 - x_2 + x_3 &= 0 \\ x_1 + x_2 + x_3 &= 0 \end{aligned}$ in $\mathbb{R}^3$
(d) $\begin{aligned} x_1 + 3x_2 - x_3 + x_4 &= 0 \\ x_1 + 2x_3 + x_4 &= 0 \end{aligned}$ in $\mathbb{R}^4$
(e) $\begin{aligned} x_1 - x_2 + x_3 - 2x_4 &= 0 \\ x_1 - x_2 + x_4 + 2x_5 &= 0 \end{aligned}$ in $\mathbb{R}^5$
(f) $\begin{aligned} x_1 + x_2 - x_3 + x_4 &= 0 \\ x_1 + 2x_2 - x_4 &= 0 \\ x_2 + x_3 &= 0 \end{aligned}$ in $\mathbb{R}^4$
(g) $\begin{aligned} x_1 + x_2 + 2x_3 - 2x_4 &= 0 \\ x_1 + x_2 + 2x_3 + x_5 &= 0 \\ 2x_4 + x_5 &= 0 \end{aligned}$ in $\mathbb{R}^5$

5. Find any two vectors in $\mathbb{R}^3$ that span the plane

(a) $x_1 - 2x_2 + 4x_3 = 0$ (b) $2x_1 + x_2 - 6x_3 = 0$
(c) $x_1 + 5x_3 = 0$ (d) $x_1 + 4x_2 + x_3 = 0$
(e) $x_2 + 2x_3 = 0$ (f) $3x_1 - x_2 - x_3 = 0$

6. Show whether the given sets are identical. Explain.

(a) span $\{(2,-1,-1), (3,1,0)\}$ and span $\{(2,-1,-1), (5,5,2)\}$
(b) span $\{(1,2,3), (2,-1,1)\}$ and span $\{(1,2,3), (3,1,5)\}$
(c) span $\{(4,1,0), (1,1,1)\}$ and span $\{(1,1,1), (2,-1,-2)\}$
(d) span $\{(1,2,-1), (-3,0,0)\}$ and span $\{(1,0,0), (1,3,0)\}$
(e) span $\{(1,0,1,2), (-1,1,1,0)\}$ and span $\{(0,1,2,2), (1,1,3,4)\}$
(f) span $\{(1,0,1,2), (1,1,1,1), (1,2,3,4)\}$ and span $\{(2,0,-1,0), (0,-1,2,3), (4,3,2,1)\}$
(g) span $\{(1,0,1,1), (2,1,1,0), (1,2,2,1)\}$ and span $\{(2,-1,0,0), (1,-2,0,1), (3,5,4,1)\}$
(h) span $\{(1,2,3,0), (0,1,0,2), (2,3,0,1)\}$ and span $\{(1,0,-3,-1), (-1,1,3,3), (1,2,1,1)\}$

7. Find any two ON (orthonormal) vectors in

(a) span $\{(1,2), (6,-1)\}$
(b) span $\{(1,2,4), (2,-1,3)\}$
(c) span $\{(1,-1,0), (1,2,3)\}$

(d) span $\{(2,1,0),(0,1,2)\}$
(e) span $\{(1,1,0,1),(0,2,-1,1)\}$
(f) span $\{(-2,3,1,1),(0,2,-1,1)\}$

8. Prove Theorem 9.7.1.

9.8 Linear Dependence

The definition of the linear dependence or independence of a set of vectors is essentially identical to Definition 3.2.1 for a set of functions, with the word "functions" changed to "vectors:"

DEFINITION 9.8.1 *Linear Dependence and Linear Independence*
A set of vectors $\{\mathbf{u}_1,\ldots,\mathbf{u}_k\}$ is said to be **linearly dependent** if at least one of them can be expressed as a linear combination of the others. If none can be so expressed, then the set is **linearly independent**.

Thus, we urge you to review Section 3.2 in conjunction with your study of this section. As in Chapter 3, we frequently use the abbreviations **LD** and **LI** to stand for linearly dependent and linearly independent, respectively.

EXAMPLE 1. Let $\mathbf{u}_1=(1,0)$, $\mathbf{u}_2=(1,1)$, and $\mathbf{u}_3=(5,4)$. These are LD since, by inspection, we can express $\mathbf{u}_3$ as a linear combination of $\mathbf{u}_1$ and $\mathbf{u}_2$: $\mathbf{u}_3=\mathbf{u}_1+4\mathbf{u}_2$. (Alternatively, we could express $\mathbf{u}_2=\frac{1}{4}\mathbf{u}_3-\frac{1}{4}\mathbf{u}_1$, or $\mathbf{u}_1=-4\mathbf{u}_2+\mathbf{u}_3$). ∎

EXAMPLE 2. Let $\mathbf{u}_1=(1,0)$ and $\mathbf{u}_2=(1,1)$. These are LI since $\mathbf{u}_1$ cannot be expressed as a "linear combination of the others," namely, as a scalar multiple of $\mathbf{u}_2$, nor can $\mathbf{u}_2$ be expressed as a scalar multiple of $\mathbf{u}_1$. ∎

EXAMPLE 3. Let $\mathbf{u}_1=(2,-1)$, $\mathbf{u}_2=(0,0)$, and $\mathbf{u}_3=(0,1)$. These are LD since we can express $\mathbf{u}_2=0\mathbf{u}_1+0\mathbf{u}_3$. (The fact that we *cannot* express $\mathbf{u}_1$ as a linear combination of $\mathbf{u}_2$ and $\mathbf{u}_3$, nor $\mathbf{u}_3$ as a linear combination of $\mathbf{u}_1$ and $\mathbf{u}_2$ does not alter our conclusion, for recall the words "at least one" in the definition.) ∎

It is implicit in Definition 9.8.1 that $\mathbf{u}_1,\ldots,\mathbf{u}_k$ are all members of the same vector space; in Examples 1 to 3 that space was $\mathbb{R}^2$. Thus, it would make no sense to ask whether $\mathbf{u}_1=(2,5)$ and $\mathbf{u}_2=(4,3,0,1)$ are linearly dependent or not since $\mathbf{u}_1$ is a member of $\mathbb{R}^2$ while $\mathbf{u}_2$ is a member of $\mathbb{R}^4$.

The preceding examples are simple enough to be worked by inspection. In more complicated cases, the following theorem provides a systematic approach for

determining whether a given vector set is linearly dependent or linearly independent.

THEOREM 9.8.1 *Test for Linear Dependence / Independence*
A finite set of vectors $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ is LD if and only if there exist scalars α_j, not all zero, such that

$$\alpha_1 \mathbf{u}_1 + \cdots + \alpha_k \mathbf{u}_k = \mathbf{0}; \tag{1}$$

if (1) holds only if all the α_j's are zero, then the set is LI.

Proof is essentially the same as for Theorem 3.2.1.

EXAMPLE 4. Consider the 4-tuples

$$\mathbf{u}_1 = (2, 0, 1, -3), \quad \mathbf{u}_2 = (0, 1, 1, 1), \quad \mathbf{u}_3 = (2, 2, 3, 0). \tag{2}$$

To see if these vectors are LI or LD, appeal directly to (1):

$$\alpha_1(2, 0, 1, -3) + \alpha_2(0, 1, 1, 1) + \alpha_3(2, 2, 3, 0) = (0, 0, 0, 0), \tag{3}$$

or $(2\alpha_1 + 2\alpha_3,\ \alpha_2 + 2\alpha_3,\ \alpha_1 + \alpha_2 + 3\alpha_3,\ -3\alpha_1 + \alpha_2) = (0, 0, 0, 0)$. Thus,

$$\begin{aligned} 2\alpha_1 \qquad\quad + 2\alpha_3 &= 0, \\ \alpha_2 + 2\alpha_3 &= 0, \\ \alpha_1 + \alpha_2 + 3\alpha_3 &= 0, \\ -3\alpha_1 + \alpha_2 \qquad\quad &= 0. \end{aligned} \tag{4}$$

Applying Gauss elimination yields

$$\begin{aligned} 2\alpha_1 \qquad\quad + 2\alpha_3 &= 0, \\ \alpha_2 + 2\alpha_3 &= 0, \\ \alpha_3 &= 0, \\ 0 &= 0. \end{aligned} \tag{5}$$

This system admits only the trivial solution, $\alpha_1 = \alpha_2 = \alpha_3 = 0$ so $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$ are LI. ∎

EXAMPLE 5. Consider the 3-tuples

$$\mathbf{u}_1 = (1, 0, 1), \quad \mathbf{u}_2 = (1, 1, 1), \quad \mathbf{u}_3 = (1, 1, 2), \quad \mathbf{u}_4 = (1, 2, 1). \tag{6}$$

Working from (1), as in Example 4, we have

$$\begin{aligned} \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 &= 0, \\ \alpha_2 + \alpha_3 + 2\alpha_4 &= 0, \\ \alpha_1 + \alpha_2 + 2\alpha_3 + \alpha_4 &= 0, \end{aligned} \tag{7}$$

or, after Gauss elimination,

$$\begin{aligned} \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 &= 0, \\ \alpha_2 + \alpha_3 + 2\alpha_4 &= 0, \\ \alpha_3 \qquad &= 0. \end{aligned} \tag{8}$$

This time, there exist nontrivial solutions for the α_j's so the vectors $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$ are LD. [Specifically, (8) gives $\alpha_3 = 0$, $\alpha_4 = \alpha$, $\alpha_2 = -2\alpha$, $\alpha_1 = \alpha$ where α is arbitrary. With $\alpha = 1$, say, (1) becomes $\mathbf{u}_1 - 2\mathbf{u}_2 + 0\mathbf{u}_3 + \mathbf{u}_4 = \mathbf{0}$.] ∎

We conclude this section with four modest theorems, the first three being essentially the same as Theorems 3.2.4–3.2.6 for functions.

THEOREM 9.8.2 *Linear Dependence / Independence of Two Vectors*
A set of two vectors $\{\mathbf{u}_1, \mathbf{u}_2\}$ is LD if and only if one is expressible as a scalar multiple of the other.

THEOREM 9.8.3 *Linear Dependence of Sets Containing the Zero Vector*
A set containing the zero vector is LD.

THEOREM 9.8.4 *Equating Coefficients*
Let $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ be LI. Then, for

$$a_1\mathbf{u}_1 + \cdots + a_k\mathbf{u}_k = b_1\mathbf{u}_1 + \cdots + b_k\mathbf{u}_k$$

to hold, it is necessary and sufficient that $a_j = b_j$ for each $j = 1, \ldots, k$. That is, the coefficients of corresponding vectors on the left- and right-hand sides must match.

THEOREM 9.8.5 *Orthogonal Sets*
Every finite orthogonal set of (nonzero) vectors is LI.

Proof of Theorem 9.8.5: Dot $\mathbf{u}_1$ into both sides of

$$\alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2 + \cdots + \alpha_k\mathbf{u}_k = \mathbf{0}. \tag{9}$$

In other words,

$$\begin{aligned} \mathbf{u}_1 \cdot (\alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2 + \cdots + \alpha_k\mathbf{u}_k) &= \mathbf{u}_1 \cdot \mathbf{0}, \\ \alpha_1\mathbf{u}_1 \cdot \mathbf{u}_1 + \alpha_2\mathbf{u}_1 \cdot \mathbf{u}_2 + \cdots + \alpha_k\mathbf{u}_1 \cdot \mathbf{u}_k &= 0, \\ \alpha_1 \|\mathbf{u}_1\|^2 + 0 + \cdots + 0 &= 0. \end{aligned} \tag{10}$$

Now $\mathbf{u}_1 \neq \mathbf{0}$ implies that $\|\mathbf{u}_1\| \neq 0$ so it follows from (10) that $\alpha_1 = 0$. Similarly, dotting $\mathbf{u}_2$ into (9) gives $\alpha_2 = 0$, and so on. Since $\alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$, the $\mathbf{u}_j$'s must be LI, as claimed. ■

EXAMPLE 6. The set $\{(2,1),(1,5)\}$ in $\mathbb{R}^2$ is LI because neither vector can be expressed as a scalar multiple of the other. ■

EXAMPLE 7. Let

$$\mathbf{u}_1 = (4,-1,1,2), \quad \mathbf{u}_2 = (3,0,2,5), \quad \mathbf{u}_3 = (0,0,0,0)$$

in $\mathbb{R}^4$. The set is LD, according to Theorem 9.8.3 because it contains the zero vector $\mathbf{u}_3 = \mathbf{0}$. That is, $\mathbf{u}_3$ can be expressed as a linear combination of $\mathbf{u}_1$ and $\mathbf{u}_2$: $\mathbf{u}_3 = 0\mathbf{u}_1 + 0\mathbf{u}_2$. If the preceding sentence is not clear, rewrite the equation as $0\mathbf{u}_1 + 0\mathbf{u}_2 - 1\mathbf{u}_3 = \mathbf{0}$ and observe that the α_j coefficients (0, 0, and -1) are not all zero. ■

Closure. The foregoing discussion of the linear dependence / independence of vectors is essentially the same as the discussion of the linear dependence / independence of functions in Section 3.2, except that the Wronskian determinant test did not carry over.

EXERCISES 9.8

1. (a) Can a set be neither LD nor LI? Explain.
(b) Can a set be both LD and LI? Explain.

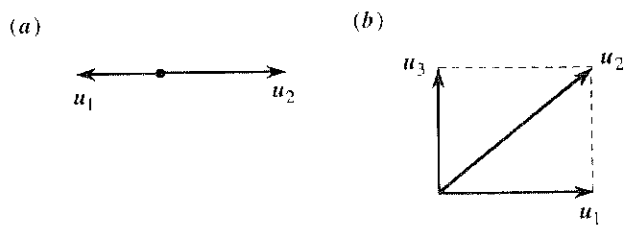
2. Show that the following sets are LD by expressing one of the vectors as a linear combination of the others.

(a) $\{(1,1),(1,2),(3,4)\}$
(b) $\{(1,4),(2,8),(3,-1)\}$
(c) $\{(1,-1),(4,2),(-3,3)\}$
(d) $\{(1,2,3),(3,2,1),(5,5,5)\}$
(e) $\{(1,0,0),(0,1,0),(3,3,0),(2,-7,9)\}$

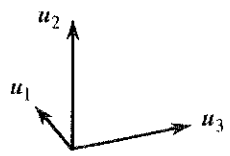
3. Determine whether the following set is LI or LD. If it is LD, then give a linear relation among the vectors.

(a) $(1,3),(2,0),(-1,3),(7,3)$
(b) $(1,3),(2,0),(1,2),(-1,5)$
(c) $(2,3,0),(1,-2,3)$
(d) $(2,3,0),(1,-2,4),(1,1,0),(1,1,1)$
(e) $(0,0,2),(0,0,3),(2,-1,5),(1,2,4),(7,9,1),(2,0,-4)$
(f) $(2,3,0,0),(1,-5,0,2),(3,1,2,2)$
(g) $(1,3,2,0),(4,1,-2,-2),(0,2,0,3),(4,7,1,2)$
(h) $(2,0,1,-1,0),(1,2,0,3,1),(4,-4,3,-9,-2)$
(i) $(1,3,0),(0,1,-1),(0,0,0)$
(j) $(1,1,0,0),(1,-1,0,0),(0,0,-2,2),(0,0,1,1)$
(k) $(1,-3,0,2,1),(-2,6,0,-4,-2)$
(l) $(5,4,1,1),(0,0,0,0),(1,9,-7,2)$
(m) $(1,2,3,4),(2,3,4,5)$
(n) $(2,1,-1),(1,4,2),(3,-2,-4)$
(o) $(7,1,0),(-1,1,4),(2,3,5)$
(p) $(1,2,-1),(1,0,1),(3,-2,5)$
(q) $(3,1,0,0),(1,-2,4,1),(2,1,6,5)$
(r) $(2,4,0,1),(1,0,1,2),(1,-3,1,2),(1,1,-1-1)$

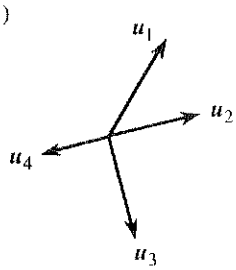
4. Show, by graphical means, that the vector sets shown below, and lying in the plane of the paper, are LD. (The emphasis here is on the method and ideas, not on graphical precision.)



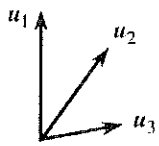
(c)



(d)



(e)



5. If $\mathbf{u}_1$ and $\mathbf{u}_2$ are LI, $\mathbf{u}_1$ and $\mathbf{u}_3$ are LI and $\mathbf{u}_2$ and $\mathbf{u}_3$ are LI, does it follow that $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ is LI? Prove or disprove.

6. Prove or disprove:

(a) $\mathbf{v}$ is in span $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ if $\{\mathbf{v}, \mathbf{u}_1, \ldots, \mathbf{u}_k\}$ is LD.
(b) $\mathbf{v}$ is not in span $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ if $\{\mathbf{v}, \mathbf{u}_1, \ldots, \mathbf{u}_k\}$ is LI.
(c) $\mathbf{v}$ is not in span $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ if and only if $\{\mathbf{v}, \mathbf{u}_1, \ldots, \mathbf{u}_k\}$ is LI.

7. (a) Prove Theorem 9.8.2.
(b) Prove Theorem 9.8.3.
(c) Prove Theorem 9.8.4.

9.9 Bases, Expansions, Dimension

9.9.1. Bases and expansions. In the calculus we learn that a given function $f(x)$ can be "expanded" as a linear combination of powers of x (namely $1, x, x^2, \ldots$),

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots. \tag{1}$$

We call $a_0, a_1, a_2, \ldots$ the "expansion coefficients," and these can be computed from $f(x)$ as $a_j = f^{(j)}(0)/j!$. Such representation of a given function is important, and examples such as $e^x = 1 + x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \cdots$ and $\sin x = x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 - \cdots$ are familiar to us.

Likewise useful, in Chapters 9–12, are the expansion of a given vector $\mathbf{u}$ in terms of a set of "base vectors" $\mathbf{e}_1, \ldots, \mathbf{e}_k$:

$$\mathbf{u} = \alpha_1 \mathbf{e}_1 + \cdots + \alpha_k \mathbf{e}_k. \tag{2}$$

How do we come up with such sets of base vectors and, once we know the $\mathbf{e}_j$'s and the given $\mathbf{u}$, how do we compute the expansion coefficients α_j? The story is simpler than for the power series of functions because whereas (1) is an infinite series and one needs to deal with the sophisticated issue of convergence, our vector expansions in Chapters 9–12 entail only a *finite* number of terms.

Beginning simply, consider the vector space $\mathbb{R}^2$, the set of all vectors in the plane of the paper. In particular, consider the vectors $\mathbf{e}_1$ and $\mathbf{e}_2$ shown in Fig. 1a. It should be evident (Theorem 9.8.2) that $\mathbf{e}_1$ and $\mathbf{e}_2$ are LI and that they span the space so that any given vector, such as $\mathbf{u}$ in Fig. 1b and $\mathbf{v}$ in Fig. 1c, can be expressed as a linear combination of them.

For the vector $\mathbf{u}$, for example, $\mathbf{u} = \mathbf{OA} + \mathbf{OB}$; with the aid of a scale, $\mathbf{OA} = 1.6\mathbf{e}_1$ and $\mathbf{OB} = 2\mathbf{e}_2$, so that

$$\mathbf{u} = 1.6\mathbf{e}_1 + 2\mathbf{e}_2. \tag{3}$$

Similarly (Fig. 1c),

$$\mathbf{v} = 2\mathbf{e}_1 - 2.5\mathbf{e}_2, \tag{4}$$

and so on, for any given vector in the plane. Of course, the zero vector is simply $\mathbf{0} = 0\mathbf{e}_1 + 0\mathbf{e}_2$.

The formulas (3) and (4) are examples of the *expansion* of a given vector [$\mathbf{u}$ in (3), $\mathbf{v}$ in (4)] in terms of a set of *base vectors* [the set $\{\mathbf{e}_1, \mathbf{e}_2\}$].

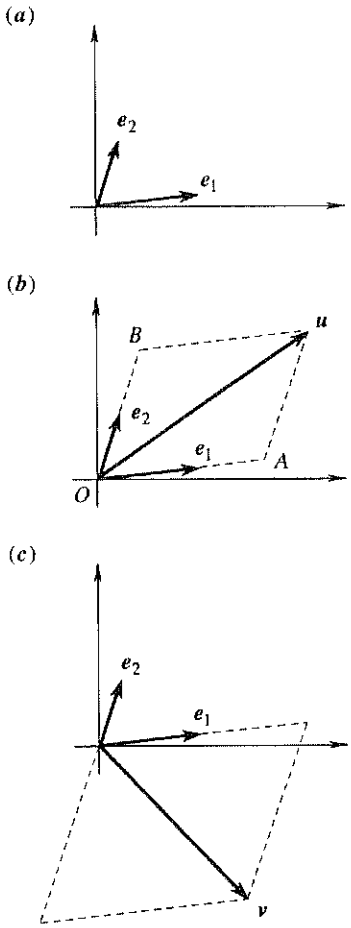


Figure 1. Vector expansion in $\mathbb{R}^2$.

DEFINITION 9.9.1 *Basis*
A finite set of vectors $\{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$ in a vector space $\mathcal{S}$ is a **basis** for $\mathcal{S}$ if each vector $\mathbf{u}$ in $\mathcal{S}$ can be expressed (i.e., "expanded") uniquely in the form

$$\mathbf{u} = \alpha_1\mathbf{e}_1 + \cdots + \alpha_k\mathbf{e}_k = \sum_{j=1}^{k} \alpha_j\mathbf{e}_j. \tag{5}$$

By the expansion (5) being unique, we mean that the α_j *expansion coefficients* are uniquely determined.

THEOREM 9.9.1 *Test for Basis*
A finite set $\{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$ in a vector space $\mathcal{S}$ is a basis for $\mathcal{S}$ if and only if it spans $\mathcal{S}$ and is LI.

Proof: First, it follows from the definition of the verb span that every vector $\mathbf{u}$ in $\mathcal{S}$ can be expanded as in (5) if and only if the set $\{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$ spans $\mathcal{S}$. Turning to the question of the uniqueness of the expansion, suppose that both expansions

$$\mathbf{u} = \alpha_1\mathbf{e}_1 + \cdots + \alpha_k\mathbf{e}_k, \tag{6}$$
$$\mathbf{u} = \beta_1\mathbf{e}_1 + \cdots + \beta_k\mathbf{e}_k \tag{7}$$

hold for any given vector $\mathbf{u}$ in $\mathcal{S}$. Subtracting (7) from (6) gives

$$(\alpha_1 - \beta_1)\mathbf{e}_1 + \cdots + (\alpha_k - \beta_k)\mathbf{e}_k = \mathbf{0}. \tag{8}$$

Now, each of the coefficients $(\alpha_1 - \beta_1), \ldots, (\alpha_k - \beta_k)$ in (8) must be zero, in which case $\alpha_1 = \beta_1, \ldots, \alpha_k = \beta_k$ and expansions (6) and (7) are identical if and only if

the set $\{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$ is LI. Hence, the expansion (5) is unique if and only if the set is LI, and this completes the proof. ■

The key idea revealed in the foregoing proof is that a basis needs to contain *enough* vectors but *not too many*: enough so that the set spans the space and can therefore be used to expand any given vector in the space, but not too many, in order that such expansions will be unique.

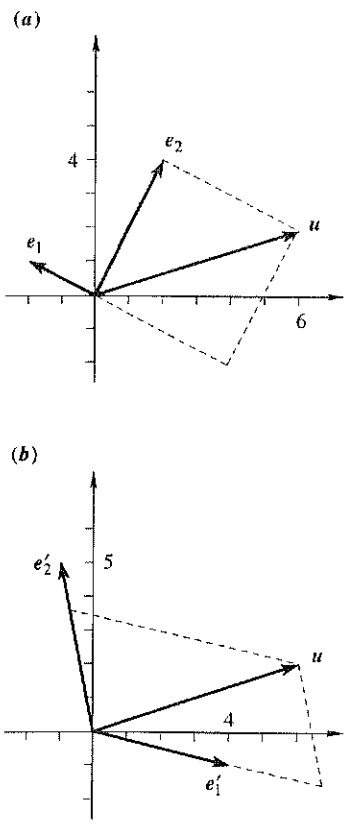


Figure 2. Two bases for $\mathbb{R}^2$.

EXAMPLE 1. Consider the vectors

$$\mathbf{e}_1 = (-2, 1), \quad \mathbf{e}_2 = (2, 4). \tag{9}$$

As may be verified, the set (9) is LI and spans $\mathbb{R}^2$ and is therefore a basis for $\mathbb{R}^2$.

Using that set to expand the vector $\mathbf{u} = (6, 2)$, say, we express

$$\mathbf{u} = \alpha_1\mathbf{e}_1 + \alpha_2\mathbf{e}_2, \tag{10}$$

or $(6, 2) = (-2\alpha_1, \alpha_1) + (2\alpha_2, 4\alpha_2)$. Hence,

$$\begin{aligned} -2\alpha_1 + 2\alpha_2 &= 6, \\ \alpha_1 + 4\alpha_2 &= 2. \end{aligned} \tag{11}$$

Solving (11), $\alpha_1 = -2$ and $\alpha_2 = 1$ so the expansion (10) is

$$\mathbf{u} = -2\mathbf{e}_1 + \mathbf{e}_2, \tag{12}$$

as displayed in Fig. 2a.

It is to be emphasized that the basis (9) shown in Fig. 2a is by no means the *only* basis for $\mathbb{R}^2$; there are slews of them. For example, it is readily verified that another is

$$\mathbf{e}_1' = (4, -1), \quad \mathbf{e}_2' = (-1, 5), \tag{13}$$

and in this case the expansion of $\mathbf{u} = (6, 2)$ is found to be

$$\mathbf{u} = \frac{32}{19}\mathbf{e}_1' + \frac{14}{19}\mathbf{e}_2', \tag{14}$$

as depicted in Fig. 2b.

COMMENT. The difference between the expansions (12) and (13) is *not* at odds with the notion of uniqueness since the two expansions are with respect to different bases. In other words, (12) *is* the unique expansion of $\mathbf{u}$ in terms of the $\mathbf{e}_1, \mathbf{e}_2$ basis, and (14) *is* the unique expansion of $\mathbf{u}$ in terms of the $\mathbf{e}_1', \mathbf{e}_2'$ basis. ■

9.9.2. Dimension. If we always worked in 2-space or 3-space, the concept of dimension would hardly need elaboration; for example, 3-space is three-dimensional, a plane within it is two-dimensional, and a line within it is one-dimensional. However, having generalized our vector concept beyond 3-space, we need to clarify the idea of dimension.

DEFINITION 9.9.2 *Dimension*

If the greatest number of LI vectors that can be found in a vector space S is k,

where $1 \leq k < \infty$, then $\mathcal{S}$ is **k-dimensional**, and we write

$$\dim \mathcal{S} = k.$$

If $\mathcal{S}$ is the zero vector space (i.e., if it contains only the zero vector), we define $\dim \mathcal{S} = 0$. If an arbitrarily large number of LI vectors can be found in $\mathcal{S}$, we say that $\mathcal{S}$ is **infinite-dimensional.***

To determine the dimension of a given vector space, it may be more convenient to use the following theorem than to work directly from Definition 9.9.2.

THEOREM 9.9.2 *Test for Dimension*
If a vector space $\mathcal{S}$ admits a basis consisting of k vectors, then $\mathcal{S}$ is k-dimensional.

Proof: Let $\{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$ be a basis for $\mathcal{S}$. Because these vectors form a basis, they must be LI. Hence, we have *at least* k LI vectors in $\mathcal{S}$, and it remains to show that *no more than* k LI vectors can be found in $\mathcal{S}$. Suppose that vectors $\mathbf{e}'_1, \ldots, \mathbf{e}'_{k+1}$ in $\mathcal{S}$ are LI. Each of these can be expanded in terms of the given base vectors, as

$$\begin{aligned} \mathbf{e}'_1 &= a_{11}\mathbf{e}_1 + \cdots + a_{1k}\mathbf{e}_k, \\ &\vdots \\ \mathbf{e}'_{k+1} &= a_{k+1,1}\mathbf{e}_1 + \cdots + a_{k+1,k}\mathbf{e}_k, \end{aligned} \tag{15}$$

say. Putting these expressions into the equation

$$\alpha_1\mathbf{e}'_1 + \alpha_2\mathbf{e}'_2 + \cdots + \alpha_{k+1}\mathbf{e}'_k = \mathbf{0} \tag{16}$$

and grouping terms gives

$$(\alpha_1 a_{11} + \cdots + \alpha_{k+1}a_{k+1,1})\,\mathbf{e}_1 + \cdots + (\alpha_1 a_{1k} + \cdots + \alpha_{k+1}a_{k+1,k})\,\mathbf{e}_k = \mathbf{0}.$$

But the set $\{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$ is LI since it is a basis, so each coefficient in the preceding equation must be zero:

$$\begin{aligned} a_{11}\alpha_1 + \cdots + a_{k+1,1}\alpha_{k+1} &= 0, \\ &\vdots \\ a_{1k}\alpha_1 + \cdots + a_{k+1,k}\alpha_{k+1} &= 0. \end{aligned} \tag{17}$$

These are k linear homogeneous equations in the $k+1$ unknowns α_1 through α_{k+1}, and such a system necessarily admits nontrivial solutions (Theorem 8.3.4). Thus, the α's in (17) are not all necessarily zero so the vectors $\mathbf{e}'_1, \ldots, \mathbf{e}'_{k+1}$ could

*Infinite-dimensional function spaces will be studied in Chapter 17.

not have been LI after all. Hence, it is not possible to find more than k LI vectors in $\mathcal{S}$, and this completes the proof. ■

The spaces of chief concern in Chapters 9–12 are the n-tuple spaces $\mathbb{R}^n$ and subspaces thereof. For $\mathbb{R}^n$ we can say the following.

THEOREM 9.9.3 *Dimension of* $\mathbb{R}^n$
The dimension of $\mathbb{R}^n$ is n: $\dim \mathbb{R}^n = n$.

Proof: The vectors

$$\begin{aligned} \mathbf{e}_1 &= (1, 0, \ldots, 0), \\ \mathbf{e}_2 &= (0, 1, 0, \ldots, 0), \\ &\vdots \\ \mathbf{e}_n &= (0, \ldots, 0, 1) \end{aligned} \tag{18}$$

constitute a basis for $\mathbb{R}^n$ because any vector $\mathbf{u} = (u_1, \ldots, u_n)$ in $\mathbb{R}^n$ can be expanded uniquely as $\mathbf{u} = u_1\mathbf{e}_1 + \cdots + u_n\mathbf{e}_n$. Since this basis contains n vectors, it follows from Theorem 9.9.2 that $\mathbb{R}^n$ is n-dimensional. ■

Indeed, we might well have questioned the reasonableness of our definition of dimension if $\mathbb{R}^n$ had turned out to be other than n-dimensional! The ON basis (18) is called the **standard basis** for $\mathbb{R}^n$ (and is the n-space generalization of the "$\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}$" ON basis that might be known to you from other courses).

Finally, what about the dimension of a *subspace*, for example, the subspace of $\mathbb{R}^3$ that is spanned by two given vectors?

THEOREM 9.9.4 *Dimension of Span* $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$
The dimension of span $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$, where the $\mathbf{u}_j$'s are not all zero, denoted as dim [span $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$], is equal to the greatest number of LI vectors within the generating set $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$.

Proof: Denote the generating set $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ as U. Let the greatest number of LI vectors in U be N, where $1 \leq N \leq k$. It may be assumed, without loss of generality, that the members of U have been numbered so that $\mathbf{u}_1, \ldots, \mathbf{u}_N$ are LI. Then each of the remaining members of U, namely $\mathbf{u}_{N+1}, \ldots, \mathbf{u}_k$, can be expressed as a linear combination of $\mathbf{u}_1, \ldots, \mathbf{u}_N$. Surely, then, each vector in span U can similarly be expressed as a linear combination of $\mathbf{u}_1, \ldots, \mathbf{u}_N$. Now $\{\mathbf{u}_1, \ldots, \mathbf{u}_N\}$ is LI and spans span U. According to Theorem 9.9.3, then, the dimension of span U is N; that is, it is the same as the greatest number of LI vectors in U, as was to be proved. ■

EXAMPLE 2. Let

$$\mathbf{u}_1 = (3, -1, 2, 1), \quad \mathbf{u}_2 = (1, 1, 0, -1), \quad \mathbf{u}_3 = (4, 0, 2, 0).$$

These vectors are, of course, members of $\mathbb{R}^4$. But since $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$ are only three vectors, dim [span $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$] is *at most* three. In fact, it is *not* three since we see that $\mathbf{u}_3 = \mathbf{u}_1 + \mathbf{u}_2$. But $\mathbf{u}_1$ and $\mathbf{u}_2$, say, are LI since neither is a scalar multiple of the other. Thus, there are only two LI vectors within the generating set so dim [span $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$] $= 2$. ∎

In Example 2 we determined that the greatest number of LI vectors in the generating set was 2 by inspection. What if we wish to determine dim [span $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$] where the $\mathbf{u}_j$'s are members of $\mathbb{R}^8$, and $k = 6$, say? For such a large problem we cannot expect "inspection" to work. Yet, what are we to do, test the $\mathbf{u}_j$'s for linear independence one at a time, two at a time, three at a time, and so on, until we determine the greatest number of LI vectors in $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$? That would be quite tedious. No, we will see later, in Chapter 10, that the best way to determine the greatest number of LI vectors in a given set is to determine the "rank" of a certain matrix, and that can be done by the extremely efficient method of elementary row operations. Meanwhile, in the present section, we "get by" by keeping the examples and exercises simple enough so that we can rely on inspection.

Let us return, now, to our discussion of bases and expansions.

9.9.3. Orthogonal bases. If, as in Example 1, there are many bases for a given space, then how do we decide which one to select? We will find that in most applications the most convenient basis to use is dictated by the context, so let us not worry about that now. This point is addressed in Chapter 11 as well as in the chapters on PDEs.

However, we do wish to show, here, that *orthogonal* bases are to be preferred whenever possible. For observe from Example 1 that to expand $\mathbf{u}$ (that is, to compute the α_j expansion coefficients) we needed to solve the system (11) of two equations in two unknowns. Similarly, if we seek to expand a given vector in $\mathbb{R}^8$, then there will be eight base vectors (because $\mathbb{R}^8$ is eight-dimensional) and eight α_j expansion coefficients, and these will be found by solving a system [analogous to (11)] of eight equations in the eight unknown α_j's. Thus, the expansion process can be quite laborious.

On the other hand, suppose that $\{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$ is an *orthogonal* basis for $\mathcal{S}$; that is, it is not only a basis but also happens to be an orthogonal set:

$$\mathbf{e}_i \cdot \mathbf{e}_j = 0 \quad \text{if} \quad i \neq j. \tag{19}$$

Suppose that we wish to expand a given vector $\mathbf{u}$ in $\mathcal{S}$ in terms of that basis; that is, we wish to determine the coefficients $\alpha_1, \ldots, \alpha_k$ in the expansion

$$\mathbf{u} = \alpha_1 \mathbf{e}_1 + \alpha_2 \mathbf{e}_2 + \cdots + \alpha_k \mathbf{e}_k. \tag{20}$$

To accomplish this, dot (20) with $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_k$, in turn. Doing so, and using

(19), we obtain the linear system

$$\begin{aligned}
\mathbf{u}\cdot\mathbf{e}_1 &= (\mathbf{e}_1\cdot\mathbf{e}_1)\alpha_1 + 0\alpha_2 + \cdots + 0\alpha_k,\\
\mathbf{u}\cdot\mathbf{e}_2 &= 0\alpha_1 + (\mathbf{e}_2\cdot\mathbf{e}_2)\alpha_2 + 0\alpha_3 + \cdots + 0\alpha_k,\\
&\vdots\\
\mathbf{u}\cdot\mathbf{e}_k &= 0\alpha_1 + \cdots + 0\alpha_{k-1} + (\mathbf{e}_k\cdot\mathbf{e}_k)\alpha_k,
\end{aligned} \tag{21}$$

where all of the quantities $\mathbf{u}\cdot\mathbf{e}_1, \ldots, \mathbf{u}\cdot\mathbf{e}_k, \mathbf{e}_1\cdot\mathbf{e}_1, \ldots, \mathbf{e}_k\cdot\mathbf{e}_k$ are computable since $\mathbf{u}, \mathbf{e}_1, \ldots, \mathbf{e}_k$ are known. The crucial point is that even though (21) is still k equations in the k unknown α_j's, the system is *uncoupled* (i.e., the only unknown in the first equation is α_1, the only one in the second is α_2, and so on) and readily gives

$$\alpha_1 = \frac{\mathbf{u}\cdot\mathbf{e}_1}{\mathbf{e}_1\cdot\mathbf{e}_1}, \quad \alpha_2 = \frac{\mathbf{u}\cdot\mathbf{e}_2}{\mathbf{e}_2\cdot\mathbf{e}_2}, \quad \ldots, \quad \alpha_k = \frac{\mathbf{u}\cdot\mathbf{e}_k}{\mathbf{e}_k\cdot\mathbf{e}_k}, \tag{22}$$

provided, of course, that none of the denominators vanish. But these quantities cannot vanish because $\mathbf{e}_j\cdot\mathbf{e}_j = \|\mathbf{e}_j\|^2$, which is zero if and only if $\mathbf{e}_j = \mathbf{0}$, and this cannot be because if any $\mathbf{e}_j$ were $\mathbf{0}$, then the set $\{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$ would be LD (Theorem 9.8.3), and hence not a basis.

Thus, *if the* $\{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$ *basis is orthogonal, the expansion of any given* $\mathbf{u}$ *is simply*

$$\boxed{\mathbf{u} = \left(\frac{\mathbf{u}\cdot\mathbf{e}_1}{\mathbf{e}_1\cdot\mathbf{e}_1}\right)\mathbf{e}_1 + \cdots + \left(\frac{\mathbf{u}\cdot\mathbf{e}_k}{\mathbf{e}_k\cdot\mathbf{e}_k}\right)\mathbf{e}_k = \sum_{j=1}^{k}\left(\frac{\mathbf{u}\cdot\mathbf{e}_j}{\mathbf{e}_j\cdot\mathbf{e}_j}\right)\mathbf{e}_j.} \tag{23}$$

If, besides being orthogonal, the $\mathbf{e}_j$'s are normalized ($\|\mathbf{e}_j\| = 1$) so that they constitute an ON (orthonormal) basis, then (23) simplifies slightly to

$$\boxed{\mathbf{u} = (\mathbf{u}\cdot\hat{\mathbf{e}}_1)\,\hat{\mathbf{e}}_1 + \cdots + (\mathbf{u}\cdot\hat{\mathbf{e}}_k)\,\hat{\mathbf{e}}_k = \sum_{j=1}^{k}(\mathbf{u}\cdot\hat{\mathbf{e}}_j)\,\hat{\mathbf{e}}_j,} \tag{24}$$

where we recall that carets denote unit vectors.

EXAMPLE 3. Expand $\mathbf{u} = (4, 3, -3, 6)$ in terms of the orthogonal base vectors $\mathbf{e}_1 = (1, 0, 2, 0)$, $\mathbf{e}_2 = (0, 1, 0, 0)$, $\mathbf{e}_3 = (-2, 0, 1, 5)$, $\mathbf{e}_4 = (-2, 0, 1, -1)$ of $\mathbb{R}^4$. This basis is orthogonal but not ON so we use (23) rather than (24). Computing $\mathbf{u}\cdot\mathbf{e}_1 = -2$, $\mathbf{e}_1\cdot\mathbf{e}_1 = 5$, and so on, (23) gives

$$\mathbf{u} = -\frac{2}{5}\mathbf{e}_1 + 3\mathbf{e}_2 + \frac{19}{30}\mathbf{e}_3 - \frac{17}{6}\mathbf{e}_4. \tag{25}$$

Alternatively, we could have inferred, from $\mathbf{u} = \alpha_1\mathbf{e}_1 + \cdots + \alpha_4\mathbf{e}_4$, the four equations

$$\begin{aligned} \alpha_1 \quad\quad - 2\alpha_3 - 2\alpha_4 &= 4, \\ \alpha_2 \quad\quad\quad\quad\quad &= 3, \\ 2\alpha_1 \quad\quad + \alpha_3 + \alpha_4 &= -3, \\ 5\alpha_3 - \alpha_4 &= 6 \end{aligned} \tag{26}$$

on the four unknown α_j's, and solved these by Gauss elimination, but it is much easier to "cash in" on the orthogonality of the basis and to use (23). If we choose to work with an ON basis, we can scale the $\mathbf{e}_j$'s as $\hat{\mathbf{e}}_1 = \frac{1}{\sqrt{5}}(1,0,2,0)$, $\hat{\mathbf{e}}_2 = (0,1,0,0)$, $\hat{\mathbf{e}}_3 = \frac{1}{\sqrt{30}}(-2,0,1,5)$, $\hat{\mathbf{e}}_4 = \frac{1}{\sqrt{6}}(-2,0,1,-1)$. Then (24) gives

$$\mathbf{u} = -\frac{2}{\sqrt{5}}\hat{\mathbf{e}}_1 + 3\hat{\mathbf{e}}_2 + \frac{19}{\sqrt{30}}\hat{\mathbf{e}}_3 - \frac{17}{\sqrt{6}}\hat{\mathbf{e}}_4, \tag{27}$$

which result is equivalent to (25). ∎

Given a *non*orthogonal basis there are three possibilities. First, one can use it and face up to the tedious expansion process. Second, one can "trade the nonorthogonal basis in" for an orthogonal basis using the **Gram–Schmidt orthogonalization procedure**, which procedure is introduced briefly in the exercises and discussed in detail in the next section. Third, one can retain the nonorthogonal basis but streamline the expansion process by computing and utilizing a set of **dual**, or **reciprocal, vectors** corresponding to the given basis, as described in the exercises.

Closure. This section is about the **expansion** of vectors, in a given vector space $\mathcal{S}$, in terms of a set of *base vectors*. A set of vectors $\{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$ in $\mathcal{S}$ is a basis for $\mathcal{S}$ if each vector $\mathbf{u}$ in $\mathcal{S}$ *can* be expanded as a *unique* linear combination of the $\mathbf{e}_j$'s. We showed (Theorem 9.9.1) that $\{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$ is indeed a basis for $\mathcal{S}$ if and only if it spans $\mathcal{S}$ (so each $\mathbf{u}$ *can* be expanded) and is LI (so the expansion is *unique*). The number of vectors in any basis for $\mathcal{S}$ is called the *dimension* of $\mathcal{S}$. For instance, $\mathbb{R}^n$ admits the *standard basis* (18), comprised of n vectors, so $\mathbb{R}^n$ is n-dimensional. And the greatest number of LI vectors in a set $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ is the dimension of their span.

We found that the expansion process (i.e., the determination of the expansion coefficients) can be quite laborious if there are many base vectors, but is extremely simple if the basis is orthogonal, or ON, in which case the expansions are given by (23) or (24), respectively. You should remember those two formulas and be able to derive them as well.

EXERCISES 9.9

1. Show whether the following is a basis.

(a) $(1,0),(1,1),(1,2)$ for $\mathbb{R}^2$
(b) $(3,2),(-1,-5)$ for $\mathbb{R}^2$
(c) $(1,1)$ for $\mathbb{R}^2$
(d) $(2,0,1),(5,-1,2),(1,-1,0)$ for $\mathbb{R}^3$
(e) $(5,-1,2),(2,0,1),(1,-1,1)$ for $\mathbb{R}^3$
(f) $(2,1,0,6),(7,-1,-2,3),(4,3,2,1)$ for $\mathbb{R}^4$
(g) $(4,3,-2,1),(5,0,0,0),(2,1,-3,0),(1,2,4,5)$ for $\mathbb{R}^4$
(h) $(4,2,0,0),(1,2,3,0),(5,-2,3,1),(0,-6,0,1)$ for $\mathbb{R}^4$
(i) $(1,0,0,0),(1,1,0,0),(1,1,1,0),(1,1,1,1)$ for $\mathbb{R}^4$
(j) $(3,0,0,1),(2,0,0,1),(1,3,5,6),(4,-2,1,3)$ for $\mathbb{R}^4$
(k) $(1,3,-1,2),(1,2,4,3),(2,5,3,5),(3,7,7,8)$ for $\mathbb{R}^4$
(l) $(2,3,5,0),(1,-1,2,3),(4,1,2,3),(5,4,1,0),(1,2,4,6)$ for $\mathbb{R}^4$
(m) $(1,0,0,0),(0,1,0,0),(0,0,1,0),(0,0,0,1),(0,0,0,0)$ for $\mathbb{R}^4$
(n) $(1,0,0,0),(0,1,0,0),(0,0,1,0),(0,0,0,0)$ for $\mathbb{R}^4$
(o) $(1,1,2),(4,-2,-1)$ for span $\{(2,-4,-5),(1,-1,-1)\}$
(p) $(1,1,2),(4,-2,-1)$ for span $\{(3,-5,-6),(1,2,1)\}$
(q) $(1,1,1),(1,-1,2)$ for span $\{(2,4,1),(1,7,-2)\}$
(r) $(1,2,3),(1,0,4)$ for span $\{(3,2,0),(1,1,-1)\}$

2. Expand each vector $\mathbf{u}$ in terms of the orthogonal basis $\{\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3\}$ of $\mathbb{R}^3$, where $\mathbf{e}_1=(2,1,3)$, $\mathbf{e}_2=(1,-2,0)$, $\mathbf{e}_3=(6,3,-5)$.

(a) $\mathbf{u}=(9,-2,4)$ (b) $\mathbf{u}=(1,0,0)$
(c) $\mathbf{u}=(0,1,5)$ (d) $\mathbf{u}=(3,-1,1)$
(e) $\mathbf{u}=(0,5,0)$ (f) $\mathbf{u}=(1,2,3)$

3. (a)–(f) Expand each of the $\mathbf{u}$ vectors in Exercise 2 in terms of the ON basis $\{\hat{\mathbf{e}}_1,\hat{\mathbf{e}}_2,\hat{\mathbf{e}}_3\}$ of $\mathbb{R}^3$, where $\hat{\mathbf{e}}_1,\hat{\mathbf{e}}_2,\hat{\mathbf{e}}_3$ are normalized versions of $\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3$ given in Exercise 2.

4. Expand each vector $\mathbf{u}$ in terms of the orthogonal basis $\{\mathbf{e}_1,\ldots,\mathbf{e}_4\}$ of $\mathbb{R}^4$, where $\mathbf{e}_1=(2,0,-1,-5)$, $\mathbf{e}_2=(2,0,-1,1)$, $\mathbf{e}_3=(0,1,0,0)$, $\mathbf{e}_4=(1,0,2,0)$.

(a) $\mathbf{u}=(1,0,0,0)$ (b) $\mathbf{u}=(0,6,0,0)$
(c) $\mathbf{u}=(2,5,1,-3)$ (d) $\mathbf{u}=(4,3,-2,0)$
(e) $\mathbf{u}=(1,2,0,5)$ (f) $\mathbf{u}=(2,-7,4,1)$
(g) $\mathbf{u}=(0,0,0,9)$ (h) $\mathbf{u}=(2,3,-2,1)$
(i) $\mathbf{u}=(0,0,5,0)$ (j) $\mathbf{u}=(1,1,1,1)$

5. Verify that the $\{\mathbf{e}_1,\ldots,\mathbf{e}_4\}$ vectors given in Example 3 are a basis for $\mathbb{R}^4$. Also, solve (26) by Gauss elimination and verify that the α_j's thus obtained agree with those given in (25).

6. If $\{\mathbf{e}_1,\ldots,\mathbf{e}_k\}$ is an orthogonal set in a vector space $\mathcal{S}$, is it a basis

(a) for $\mathcal{S}$? (b) for span $\{\mathbf{e}_1,\ldots,\mathbf{e}_k\}$?

7. (*Zero vector space*) Show that a **zero vector space** (i.e., a vector space consisting of the zero vector alone) has no basis.

8. Let $\mathbf{u}_1=(1,0,0)$, $\mathbf{u}_2=(0,1,0)$, $\mathbf{u}_3=(0,0,1)$, $\mathbf{u}_4=(1,1,0)$, $\mathbf{u}_5=(0,1,1)$, $\mathbf{u}_6=(1,1,1)$, and $\mathbf{u}_7=(0$ Evaluate each of the following.

(a) dim [span $\{\mathbf{u}_1\}$]
(b) dim [span $\{\mathbf{u}_1,\mathbf{u}_2\}$]
(c) dim [span $\{\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_3\}$]
(d) dim [span $\{\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_3,\mathbf{u}_4\}$]
(e) dim [span $\{\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_4\}$]
(f) dim [span $\{\mathbf{u}_1,\mathbf{u}_4,\mathbf{u}_5\}$]
(g) dim [span $\{\mathbf{u}_5,\mathbf{u}_6,\mathbf{u}_7\}$]
(h) dim [span $\{\mathbf{u}_1,\mathbf{u}_5,\mathbf{u}_6,\mathbf{u}_7\}$]

9. Let $\mathbf{u}_1=(1,0,0,0)$, $\mathbf{u}_2=(1,1,0,0)$, $\mathbf{u}_3=(1,1,1,0)$, $\mathbf{u}_4=(1,1,1,1)$, $\mathbf{u}_5=(0,0,0,1)$, $\mathbf{u}_6=(3,3,3,3)$. Evaluate each of the following.

(a) dim [span $\{\mathbf{u}_1,\mathbf{u}_3,\mathbf{u}_5\}$]
(b) dim [span $\{\mathbf{u}_1,\mathbf{u}_4,\mathbf{u}_6\}$]
(c) dim [span $\{\mathbf{u}_2,\mathbf{u}_4,\mathbf{u}_6\}$]
(d) dim [span $\{\mathbf{u}_5\}$]
(e) dim [span $\{\mathbf{u}_3,\mathbf{u}_4\}$]
(f) dim [span $\{\mathbf{u}_3,\mathbf{u}_4,\mathbf{u}_5,\mathbf{u}_6\}$]
(g) dim [span $\{\mathbf{u}_4,\mathbf{u}_6\}$]
(h) dim [span $\{\mathbf{u}_2,\mathbf{u}_3,\mathbf{u}_4,\mathbf{u}_5,\mathbf{u}_6\}$]

10. (a)–(f) Determine the dimension of the solution space in Exercise 4 of Section 9.7.

11. (*Gram–Schmidt orthogonalization process*) Given k LI vectors $\mathbf{v}_1,\ldots,\mathbf{v}_k$, it is possible to obtain from them k ON vectors, say $\hat{\mathbf{e}}_1,\ldots,\hat{\mathbf{e}}_k$, in span $\{\mathbf{v}_1,\ldots,\mathbf{v}_k\}$ by the **Gram–Schmidt process**, after *Jörgen P. Gram* (1850–1916) and *Erhardt Schmidt* (1876–1959), by taking $\mathbf{e}_1$ equal to $\mathbf{v}_1$, taking $\mathbf{e}_2$ equal to a suitable linear combination of $\mathbf{v}_1$, $\mathbf{v}_2$, taking $\mathbf{e}_3$ equal to a suitable linear combination of $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_3$, and so on, and then normalizing the results. The resulting ON set is as

follows:

$$\begin{aligned}
\hat{\mathbf{e}}_1 &= \frac{\mathbf{v}_1}{\|\mathbf{v}_1\|}, \\
\hat{\mathbf{e}}_2 &= \frac{\mathbf{v}_2 - (\mathbf{v}_2 \cdot \hat{\mathbf{e}}_1)\hat{\mathbf{e}}_1}{\|\mathbf{v}_2 - (\mathbf{v}_2 \cdot \hat{\mathbf{e}}_1)\hat{\mathbf{e}}_1\|}, \\
&\vdots \\
\hat{\mathbf{e}}_j &= \frac{\mathbf{v}_j - \sum_{i=1}^{j-1}(\mathbf{v}_j \cdot \hat{\mathbf{e}}_i)\hat{\mathbf{e}}_i}{\left\|\mathbf{v}_j - \sum_{i=1}^{j-1}(\mathbf{v}_j \cdot \hat{\mathbf{e}}_i)\hat{\mathbf{e}}_i\right\|} \quad \text{through } j = k.
\end{aligned} \tag{11.1}$$

We now state the problem: Verify that each $\hat{\mathbf{e}}_j$ defined by (11.1) is a linear combination of $\mathbf{v}_1, \ldots, \mathbf{v}_j$, and that the $\hat{\mathbf{e}}_j$'s are ON. [In verifying that $\|\hat{\mathbf{e}}_j\| = 1$, be sure to show that each denominator in (11.1) is nonzero.]

12. In each case use the Gram–Schmidt formula (11.1) in Exercise 11 to obtain an ON set from the given LI set.

(a) $(4,0),(2,1)$
(b) $(1,-2),(3,4)$
(c) $(1,0,0),(1,1,0),(1,1,1)$
(d) $(1,1,0),(2,-1,1),(1,0,3)$
(e) $(1,1,1),(2,0,-1)$
(f) $(1,1,1),(1,0,1),(1,1,0)$
(g) $(1,2,1),(1,-1,2),(-1,3,1)$
(h) $(2,0,1),(1,1,1),(-2,0,3)$
(i) $(2,1,1,0),(1,5,-1,2)$
(j) $(6,-1,1,2,1),(2,3,-1,1,4)$

13. (*The dual or reciprocal vectors*) For definiteness, consider our vector space $\mathcal{S}$ to be $\mathbb{R}^n$.

(a) If $\{\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_n\}$ is an ON basis for $\mathbb{R}^n$, and $\mathbf{u}$ is in $\mathbb{R}^n$, then by dotting $\hat{\mathbf{e}}_i$, into both sides of the equation $\mathbf{u} = \sum_{j=1}^n \alpha_j \hat{\mathbf{e}}_j$, we find that $\alpha_i = \mathbf{u} \cdot \hat{\mathbf{e}}_i$ so that the expansion of $\mathbf{u}$ in terms of the given basis is

$$\mathbf{u} = \sum_{j=1}^n (\mathbf{u} \cdot \hat{\mathbf{e}}_j)\hat{\mathbf{e}}_j. \tag{13.1}$$

If, instead, we have a basis $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ which is *not* ON, then, as noted in the text, the expansion process is not so simple. However, suppose that we can find a set $\{\mathbf{e}_1^*, \ldots, \mathbf{e}_n^*\}$ such that

$$\mathbf{e}_i \cdot \mathbf{e}_j^* = \begin{cases} 1, & i = j \\ 0, & i \neq j. \end{cases} \tag{13.2}$$

Then show that dotting $\mathbf{e}_i^*$ into $\mathbf{u} = \sum_{j=1}^n \alpha_j \mathbf{e}_j$ gives $\alpha_i = \mathbf{u} \cdot \mathbf{e}_i^*$ so that

$$\mathbf{u} = \sum_{j=1}^n (\mathbf{u} \cdot \mathbf{e}_j^*)\mathbf{e}_j. \tag{13.3}$$

The set $\{\mathbf{e}_1^*, \ldots, \mathbf{e}_n^*\}$ is called the **dual**, or **reciprocal**, **set** corresponding to the original set $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$. (We will see in the last exercise in Section 10.6 that the dual set exists, is unique, and is itself a basis for $\mathbb{R}^n$, the so-called **dual** or **reciprocal basis**.)

(b) Given the basis $\mathbf{e}_1 = (1,0), \mathbf{e}_2 = (1,1)$ for $\mathbb{R}^2$, use equation (13.2) to determine the dual vectors $\mathbf{e}_1^*, \mathbf{e}_2^*$. Then use equation (13.3) to expand $\mathbf{u} = (3,1)$. Sketch $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_1^*, \mathbf{e}_2^*, \mathbf{u}$ to scale, and verify the expansion graphically, that is, by means of the parallelogram rule of vector addition.

(c) Repeat part (b), for $\mathbf{e}_1 = (2,1), \mathbf{e}_2 = (0,2), \mathbf{u} = (-3,2)$,

(d) Repeat part (b), for $\mathbf{e}_1 = (-1,1), \mathbf{e}_2 = (2,1), \mathbf{u} = (0,4)$,

(e) Given the basis $\mathbf{e}_1 = (1,0,0), \mathbf{e}_2 = (1,1,0), \mathbf{e}_3 = (1,1,1)$ for $\mathbb{R}^3$, use equation (13.2) to determine the dual vectors $\mathbf{e}_1^*, \mathbf{e}_2^*, \mathbf{e}_3^*$. Then use equation (13.3) to expand each of the vectors $\mathbf{u} = (4,-1,5)$, $\mathbf{v} = (0,0,2)$, $\mathbf{w} = (5,-2,3)$. Be sure to see that the dual vectors get computed once and for all, for a given basis $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$; once we have got them, expansions of the form (13.3) are simple.

(f) Repeat part (e) for $\mathbf{e}_1 = (2,0,1), \mathbf{e}_2 = (1,1,0), \mathbf{e}_3 = (1,-1,3)$, and $\mathbf{u} = (6,1,0)$, $\mathbf{v} = (1,2,4)$, $\mathbf{w} = (0,3,0)$.

(g) Show that if the $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ basis does happen to be ON, then the dual vectors coalesce with the $\mathbf{e}_j$'s, i.e., $\mathbf{e}_j^* = \mathbf{e}_j$ for $j = 1, 2, \ldots, n$.

9.10 Best Approximation

Let $\mathcal{S}$ be a normed inner product vector space (i.e., a vector space with both a norm and an inner, or dot, product defined), and let the norm be the "natural norm" $\|\mathbf{u}\| = \sqrt{\mathbf{u}\cdot\mathbf{u}}$. We know that if $\{\mathbf{e}_1, \ldots, \mathbf{e}_N\}$ is a basis for $\mathcal{S}$, then any vector $\mathbf{u}$ in $\mathcal{S}$ can be (uniquely) expanded in the form $\mathbf{u} = \sum_{j=1}^{N} c_j\mathbf{e}_j$. If the basis is orthogonal, then the expansion process is easy, with the c_j's computed, from the given vector $\mathbf{u}$ and the base vectors $\mathbf{e}_j$, as $c_j = (\mathbf{u}\cdot\mathbf{e}_j)/(\mathbf{e}_j\cdot\mathbf{e}_j)$. And if the basis is not only orthogonal but ON, then $\mathbf{u} = \sum_{j=1}^{N} c_j\hat{\mathbf{e}}_j$, where $c_j = \mathbf{u}\cdot\hat{\mathbf{e}}_j$.

However, what if we do not have a "full deck?" That is, what if $\{\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_N\}$ is ON, but falls short of being a basis for $\mathcal{S}$ (i.e., $N < \dim \mathcal{S}$)? If $\mathbf{u}$ happens to fall within span $\{\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_N\}$, which subspace of $\mathcal{S}$ we denote as $\mathcal{T}$, then it can still be expanded in terms of $\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_N$, but if it is not in $\mathcal{T}$, then it cannot be so expanded.

In the latter case the question arises, what is the best *approximation* of $\mathbf{u}$ in terms of $\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_N$? In this section we answer that question in general, and illustrate the results for the case where $\mathcal{S}$ is $\mathbb{R}^n$. Later in this book, when we study Fourier series and partial differential equations, our interest will be in function spaces instead.

9.10.1. Best approximation and orthogonal projection. The *best approximation problem*, which we address is this: given a vector $\mathbf{u}$ in $\mathcal{S}$, and an ON set $\{\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_N\}$ in $\mathcal{S}$, what is the best approximation

$$\mathbf{u} \approx c_1\hat{\mathbf{e}}_1 + \cdots + c_N\hat{\mathbf{e}}_N = \sum_{j=1}^{N} c_j\hat{\mathbf{e}}_j? \tag{1}$$

That is, how do we compute the c_j coefficients so as to render the error vector $\mathbf{E} = \mathbf{u} - \sum_{j=1}^{N} c_j\hat{\mathbf{e}}_j$ as small as possible? In other words, how do we choose the c_j's so as to minimize the norm of the error vector $\|\mathbf{E}\|$? If $\|\mathbf{E}\|$ is a minimum, then so is $\|\mathbf{E}\|^2$, so let us minimize $\|\mathbf{E}\|^2$ (to avoid square roots), where

$$\begin{aligned}\|\mathbf{E}\|^2 = \mathbf{E}\cdot\mathbf{E} &= \left(\mathbf{u} - \sum_{j=1}^{N} c_j\hat{\mathbf{e}}_j\right)\cdot\left(\mathbf{u} - \sum_{j=1}^{N} c_j\hat{\mathbf{e}}_j\right)\\ &= \mathbf{u}\cdot\mathbf{u} - 2\sum_{j=1}^{N} c_j\,(\mathbf{u}\cdot\hat{\mathbf{e}}_j) + \sum_{j=1}^{N} c_j^2,\end{aligned} \tag{2}$$

and where the step

$$\begin{aligned}\left(\sum_{1}^{N} c_j\hat{\mathbf{e}}_j\right)\cdot\left(\sum_{1}^{N} c_j\hat{\mathbf{e}}_j\right) &= (c_1\hat{\mathbf{e}}_1 + \cdots + c_N\hat{\mathbf{e}}_N)\cdot(c_1\hat{\mathbf{e}}_1 + \cdots + c_N\hat{\mathbf{e}}_N)\\ &= c_1^2 + \cdots + c_N^2 = \sum_{1}^{N} c_j^2\end{aligned} \tag{3}$$

follows from the orthonormality of the $\hat{\mathbf{e}}_j$'s.

Defining $\mathbf{u} \cdot \hat{\mathbf{e}}_j \equiv \alpha_j$ and noting that $\mathbf{u} \cdot \mathbf{u} = \|\mathbf{u}\|^2$, we may express (2) as

$$\|\mathbf{E}\|^2 = \sum_{j=1}^{N} c_j^2 - 2\sum_{j=1}^{N} \alpha_j c_j + \|\mathbf{u}\|^2,$$

or, completing the square, as

$$\|\mathbf{E}\|^2 = \sum_{j=1}^{N} (c_j - \alpha_j)^2 + \|\mathbf{u}\|^2 - \sum_{j=1}^{N} \alpha_j^2. \tag{4}$$

Observe that $\mathbf{u}$ and the ON set $\{\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_N\}$ are given so that $\|\mathbf{u}\|$ and the α_j's in (4) are fixed computable quantities: $\|\mathbf{u}\| = \sqrt{\mathbf{u} \cdot \mathbf{u}}$ and $\alpha_j = \mathbf{u} \cdot \hat{\mathbf{e}}_j$ for $j = 1, 2, \ldots, N$. Thus, in seeking to minimize the right-hand side of (4), the only control we exercise is in our choice of the c_j's. The right-hand side of (4) is greater than or equal to zero,* and so is the $\sum_{j=1}^{N} (c_j - \alpha_j)^2$ term containing the c_j's. Thus, the best that we can do is to set $c_j = \alpha_j$ $(j = 1, 2, \ldots, N)$. With that choice, our best approximation (1) becomes

$$\boxed{\mathbf{u} \approx \sum_{j=1}^{N} (\mathbf{u} \cdot \hat{\mathbf{e}}_j)\, \hat{\mathbf{e}}_j.} \tag{5}$$

Let us summarize these results.

THEOREM 9.10.1 *Best Approximation*
Let $\mathbf{u}$ be any vector in a normed inner product vector space $\mathcal{S}$ with natural norm ($\|\mathbf{u}\| = \sqrt{\mathbf{u} \cdot \mathbf{u}}$), and let $\{\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_N\}$ be an ON set in $\mathcal{S}$. Then the best approximation (1) is obtained when the c_j's are given by $c_j = \mathbf{u} \cdot \hat{\mathbf{e}}_j$, as indicated in (5).

EXAMPLE 1. Let $\mathcal{S}$ be $\mathbb{R}^2$, $N = 1$, $\hat{\mathbf{e}}_1 = \frac{1}{13}(12, 5)$, and $\mathbf{u} = (1, 1)$, as shown in Fig. 1. Find the best approximation $\mathbf{u} \approx c_1 \hat{\mathbf{e}}_1$, that is, the best approximation of $\mathbf{u}$ in span $\{\hat{\mathbf{e}}_1\}$ (which is the line L). Theorem 9.10.1 gives $c_1 = \mathbf{u} \cdot \hat{\mathbf{e}}_1 = 17/13$, and hence the best approximation

$$\mathbf{u} \approx \frac{17}{13} \hat{\mathbf{e}}_1, \tag{6}$$

which is the vector $\mathbf{OA}$ in Fig. 1.

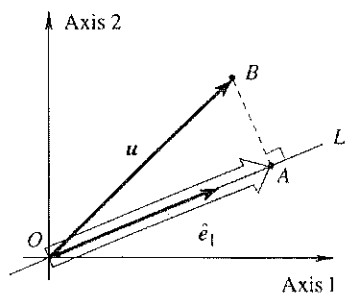


Figure 1. Best approximation of $\mathbf{u}$ in span $\{\hat{\mathbf{e}}_1\}$.

COMMENT. Observe from the figure that the best approximation $\mathbf{OA}$ is the *orthogonal projection* of $\mathbf{u}$ onto span $\{\mathbf{e}_1\}$, which orthogonality is verified by the calculation

*This fact may not be obvious due to the minus sign in front of the last summation. But remember that the right-hand side of (4) is equal to $\|\mathbf{E}\|^2$, and surely $\|\mathbf{E}\|^2 \geq 0$.

$\mathbf{AB}\cdot\hat{\mathbf{e}}_1 = (\mathbf{u}-\mathbf{OA})\cdot\hat{\mathbf{e}}_1 = (\mathbf{u}-\frac{17}{13}\hat{\mathbf{e}}_1)\cdot\hat{\mathbf{e}}_1 = \frac{17}{13}-\frac{17}{13} = 0$. That result makes perfect sense since if $c_1\hat{\mathbf{e}}_1$ is to be the best approximation to $\mathbf{u}$, then the distance from the tip of $\mathbf{u}$ to the tip of $c_1\hat{\mathbf{e}}_1$ (which is some point on L) should be as small as possible. That shortest distance is the perpendicular distance from the tip of $\mathbf{u}$ to the line L. ∎

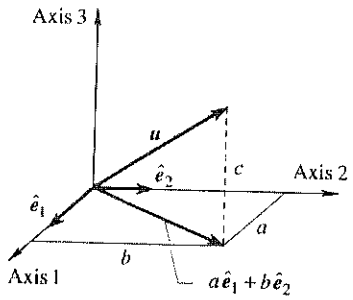


Figure 2. Best approximation of $\mathbf{u}$ in span $\{\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2\}$.

EXAMPLE 2. Let $\mathcal{S}$ be $\mathbb{R}^3$, let $N = 2$ with $\hat{\mathbf{e}}_1 = (1,0,0)$ and $\hat{\mathbf{e}}_2 = (0,1,0)$, and let $\mathbf{u} = (a,b,c)$, as shown in Fig. 2. Computing the coefficients in (5) as $\mathbf{u}\cdot\hat{\mathbf{e}}_1 = a$ and $\mathbf{u}\cdot\hat{\mathbf{e}}_2 = b$, (5) becomes

$$\mathbf{u} \approx a\hat{\mathbf{e}}_1 + b\hat{\mathbf{e}}_2. \tag{7}$$

The latter is an equality if $c = 0$. That is, (7) is an equality if $\mathbf{u}$ happens to lie in span $\{\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2\}$, but if $c \neq 0$ then the best approximation $a\hat{\mathbf{e}}_1 + b\hat{\mathbf{e}}_2$ to $\mathbf{u}$ is the orthogonal projection of $\mathbf{u}$ onto span $\{\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2\}$. ∎

In Examples 1 and 2, $\mathcal{S}$ was $\mathbb{R}^2$ and $\mathbb{R}^3$, respectively, so we were able to draw useful pictures. In each case we discovered that the best approximation of $\mathbf{u}$ on the subspace $\mathcal{T}$ of $\mathcal{S}$ spanned by $\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_N$ was the orthogonal projection of $\mathbf{u}$ onto $\mathcal{T}$. Is that result true in all cases? That is, is the error vector $\mathbf{E}$ necessarily orthogonal to $\mathcal{T}$? Since the error vector is

$$\mathbf{E} = \mathbf{u} - \sum_{j=1}^{N} (\mathbf{u}\cdot\hat{\mathbf{e}}_j)\,\hat{\mathbf{e}}_j, \tag{8}$$

we have

$$\begin{aligned}\mathbf{E}\cdot\hat{\mathbf{e}}_k &= \left[\mathbf{u} - \sum_{j=1}^{N} (\mathbf{u}\cdot\hat{\mathbf{e}}_j)\,\hat{\mathbf{e}}_j\right]\cdot\hat{\mathbf{e}}_k \\ &= \mathbf{u}\cdot\hat{\mathbf{e}}_k - (\mathbf{u}\cdot\hat{\mathbf{e}}_k)(1) = 0\end{aligned} \tag{9}$$

for each $k = 1, 2, \ldots, N$, where the second equality follows from the fact that $\hat{\mathbf{e}}_j\cdot\hat{\mathbf{e}}_k = 0$ if $j \neq k$ and 1 if $j = k$.

Since $\mathbf{E}$ is orthogonal to every one of the $\hat{\mathbf{e}}_j$'s, it is therefore orthogonal to every vector in $\mathcal{T}$. In that sense we say that the right-hand side of (5) is the **orthogonal projection of $\mathbf{u}$ onto $\mathcal{T}$**, and denote it as $\text{proj}_{\mathcal{T}}\,\mathbf{u}$:

$$\text{proj}_{\mathcal{T}}\,\mathbf{u} \equiv \sum_{j=1}^{N} (\mathbf{u}\cdot\hat{\mathbf{e}}_j)\,\hat{\mathbf{e}}_j. \tag{10}$$

The idea that the best approximation of $\mathbf{u}$ in $\mathcal{T}$ is the orthogonal projection of $\mathbf{u}$ onto $\mathcal{T}$ lends a welcome geometrical interpretation to the problem of best approximation. In fact, let us rephrase Theorem 9.10.1 in terms of orthogonal projection.

THEOREM 9.10.1′ *Best Approximation by Orthogonal Projection*
Let $\mathbf{u}$ be any vector in a normed inner product vector space $\mathcal{S}$ with natural norm

($\|\mathbf{u}\| = \sqrt{\mathbf{u}\cdot\mathbf{u}}$), and let $\{\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_N\}$ be an ON set in $\mathcal{S}$. Denote the subspace span $\{\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_N\}$ of $\mathcal{S}$ as $\mathcal{T}$. Then the best approximation of $\mathbf{u}$ in $\mathcal{T}$ (i.e., of the form $c_1\hat{\mathbf{e}}_1 + \cdots + c_N\hat{\mathbf{e}}_N$) is given by the orthogonal projection of $\mathbf{u}$ onto $\mathcal{T}$, namely, by $\text{proj}_{\mathcal{T}}\,\mathbf{u}$.

9.10.2. Kronecker delta. When working with ON sets it is convenient to use the **Kronecker delta** symbol δ_{jk}, defined as

$$\delta_{jk} \equiv \begin{cases} 1, & j = k \\ 0, & j \neq k \end{cases} \tag{11}$$

and named after *Leopold Kronecker* (1823–1891), who contributed to algebra and the theory of equations. The subscripted j and k are usually positive integers. Clearly, δ_{jk} is symmetric in its indices j and k:

$$\delta_{jk} = \delta_{kj}. \tag{12}$$

To illustrate the use of the Kronecker delta, suppose that $\{\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_N\}$ is an ON basis for some space $\mathcal{S}$, and that we wish to expand a given $\mathbf{u}$ in $\mathcal{S}$ as

$$\mathbf{u} = \sum_{j=1}^{N} c_j \hat{\mathbf{e}}_j. \tag{13}$$

To determine the c_j's, dot $\hat{\mathbf{e}}_k$ into both sides, where k is any integer such that $1 \leq k \leq N$, and use the fact that $\hat{\mathbf{e}}_j \cdot \hat{\mathbf{e}}_k = \delta_{jk}$ (because the $\hat{\mathbf{e}}_j$'s are ON):

$$\begin{aligned} \mathbf{u}\cdot\hat{\mathbf{e}}_k &= \left(\sum_{j=1}^{N} c_j \hat{\mathbf{e}}_j\right) \cdot \hat{\mathbf{e}}_k = \sum_{j=1}^{N} c_j \left(\hat{\mathbf{e}}_j \cdot \hat{\mathbf{e}}_k\right) \\ &= \sum_{j=1}^{N} c_j \delta_{jk} = c_k. \end{aligned} \tag{14}$$

Thus, $c_k = \mathbf{u}\cdot\hat{\mathbf{e}}_k$ for each $k = 1, 2, \ldots, N$ so (13) becomes

$$\mathbf{u} = \sum_{j=1}^{N} \left(\mathbf{u}\cdot\hat{\mathbf{e}}_j\right) \hat{\mathbf{e}}_j. \tag{15}$$

Closure. Principal interest, in this brief section, is in the best approximation of a given vector $\mathbf{u}$ in a normed inner product vector space $\mathcal{S}$ in terms of an ON set $\{\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_N\}$ which falls short of being a basis for $\mathcal{S}$ inasmuch as $N < \dim \mathcal{S}$. Of course, if $N = \dim \mathcal{S}$ so the set is a basis, then we have the equality (15), but if $N < \dim \mathcal{S}$, then the best approximation of $\mathbf{u}$ is given by (5), best in the vector sense; that is, the norm of the error vector [i.e., the norm of the difference between

the left- and right-hand sides of (15)] is minimized. From a geometric point of view, (15) says that the best approximation of $\mathbf{u}$ is the orthogonal projection of $\mathbf{u}$ onto the span of $\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_N$, which concept is explained in Examples 1 and 2, and which should be understood. In those examples we use the usual inner product for $\mathbb{R}^n$ (namely, $\mathbf{u} \cdot \mathbf{v} = u_1 v_1 + \cdots + u_n v_n$) and the corresponding natural norm (the Euclidean norm), but it should be understood that the results of this section (Theorems 9.10.1 and 9.10.1′) hold for any choice of the norm and inner product, provided that we use the "natural norm" $\|\mathbf{u}\| = \sqrt{\mathbf{u} \cdot \mathbf{u}}$.

EXERCISES 9.10

1. We concluded from (4) that the best choice for the c_j's is $c_j = \alpha_j = \mathbf{u} \cdot \hat{\mathbf{e}}_j$. Show that this same result is obtained from (2) by setting $\partial \|\mathbf{E}\|^2 / \partial c_j = 0$, and verify that the extremum thus obtained is a minimum.

2. Let $\mathcal{S}$ be $\mathbb{R}^5$, and let $N = 3$ with $\hat{\mathbf{e}}_1 = \frac{1}{\sqrt{5}}(1,0,2,0,0)$, $\hat{\mathbf{e}}_2 = \frac{1}{\sqrt{6}}(2,0,-1,0,1)$, $\hat{\mathbf{e}}_3 = (0,0,0,1,0)$. Find the best approximation to the given $\mathbf{u}$ vector within span $\{\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \hat{\mathbf{e}}_3\}$, and the norm of the error vector.

(a) $(3,-2,0,0,5)$ (b) $(0,0,0,2,1)$ (c) $(3,0,1,4,1)$
(d) $(1,1,0,1,1)$ (e) $(0,2,0,0,0)$ (f) $(1,0,-3,3,1)$
(g) $(0,7,0,3,0)$ (h) $(1,2,3,4,5)$ (i) $(5,4,3,2,1)$

3. Let $\mathcal{S}$ be $\mathbb{R}^4$, and let

$$\hat{\mathbf{e}}_1 = \tfrac{1}{\sqrt{3}}(1,1,0,-1), \quad \hat{\mathbf{e}}_2 = \tfrac{1}{\sqrt{3}}(1,-1,-1,0),$$
$$\hat{\mathbf{e}}_3 = \tfrac{1}{\sqrt{3}}(1,0,1,1), \quad \hat{\mathbf{e}}_4 = \tfrac{1}{\sqrt{3}}(0,1,-1,1).$$

Find the best approximation to $\mathbf{u} = (4,-2,1,6)$ within span $\{\hat{\mathbf{e}}_1\}$, span $\{\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2\}$, span $\{\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \hat{\mathbf{e}}_3\}$, and span $\{\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \hat{\mathbf{e}}_3, \hat{\mathbf{e}}_4\}$, and in each case compute the norm of the error vector, $\|\mathbf{E}\|$.

4. Same as Exercise 3, but for the given $\mathbf{u}$ vector.

(a) $(4,1,0,-1)$ (b) $(3,-1,1,2)$ (c) $(0,0,2,5)$
(d) $(1,2,4,4)$ (e) $(0,5,3,-1)$ (f) $(2,0,-1,-1)$

5. (*Bessel inequality*) Beginning with (4), derive the **Bessel inequality**

$$\boxed{\sum_{j=1}^{N} (\mathbf{u} \cdot \hat{\mathbf{e}}_j)^2 \le \|\mathbf{u}\|^2.} \tag{5.1}$$

Notice that if $\mathbf{u}$ happens to be in span $\{\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_k\}$, or if $\dim \mathcal{S} = N$, then (5.1) becomes an equality. In two and three dimensions that equality is actually the Pythagorean theorem, and in more than three dimensions it amounts to an abstract extension of that theorem.

6. (*A different inner product*) In Examples 1 and 2 we use the "usual" inner product for $\mathbb{R}^n$, $\mathbf{u} \cdot \mathbf{v} = u_1 v_1 + \cdots + u_n v_n$, but that is not the only acceptable one. In Example 3 of Section 9.6 we see that another acceptable inner product is

$$\mathbf{u} \cdot \mathbf{v} = w_1 u_1 v_1 + \cdots + w_n u_n v_n, \tag{6.1}$$

where the w_j's are fixed positive constants, or "weights."
(a) Rework Example 1 using the modified inner product $\mathbf{u} \cdot \mathbf{v} = 2u_1 v_1 + u_2 v_2$ and its corresponding natural norm $\|\mathbf{u}\| = \sqrt{2u_1^2 + u_2^2}$. Show that the resulting best approximation $\frac{29}{313}(12,5)$, which is not the same as the best approximation $\frac{17}{169}(12,5)$ given by (6). HINT: You will need to rescale $\mathbf{e}_1$ so as to be a unit vector according to the new norm.
(b) Whereas the error vector $\mathbf{AB}$ (Fig. 1) is orthogonal *and* perpendicular to span $\{\hat{\mathbf{e}}_1\}$, show that in this exercise the error vector is indeed orthogonal to span $\{\hat{\mathbf{e}}_1\}$, as promised in the text, but *not* perpendicular to it. To explain this "paradox," show that for the modified inner product the orthogonality of two nonzero vectors does not imply their perpendicularity.

7. Verify the last step in (14), that $\sum_{j=1}^{N} c_j \delta_{jk} = c_k$.

8. Verify the following, where (i,j,k,l) run from 1 to N.

(a) $\sum_{j=1} \delta_{ij} = 1$ (b) $\sum_j \delta_{ij}\delta_{jk} = \delta_{ik}$
(c) $\sum_j \sum_k \delta_{ij}\delta_{jk}\delta_{kl} = \delta_{il}$

Chapter 9 Review

We begin with the two- and three-dimensional "arrow vector" concept that is probably already familiar to you from an introductory course in physics, where the vectors denoted forces, velocities, and so on. For such vectors, vector addition $\mathbf{u}+\mathbf{v}$, scalar multiplication $(\alpha\mathbf{u})$, a zero vector $(\mathbf{0})$, a negative inverse $[-\mathbf{u}=(-1)\mathbf{u}]$, a norm $(\|\mathbf{u}\|)$, a dot product

$$\mathbf{u}\cdot\mathbf{v} = \|\mathbf{u}\|\,\|\mathbf{v}\|\cos\theta, \tag{1}$$

and the angle $\theta = \cos^{-1}\left(\dfrac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|\,\|\mathbf{v}\|}\right)$ between $\mathbf{u}$ and $\mathbf{v}$ are all defined.

From there, we generalize to abstract n-space, where $\mathbf{u}=(u_1,\ldots,u_n)$, by defining vector addition, and so on, in such a way that they agree with the corresponding arrow vector definitions when $n=2$ and $n=3$. For instance,

$$\mathbf{u}\cdot\mathbf{v} = \sum_{j=1}^{n} u_j v_j, \tag{2}$$

$$\|\mathbf{u}\| = \sqrt{\mathbf{u}\cdot\mathbf{u}} = \sqrt{\sum_{j=1}^{n} u_j^2}, \tag{3}$$

and

$$\theta = \cos^{-1}\frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|\,\|\mathbf{v}\|}. \tag{4}$$

From these definitions, we derived various *properties* such as

$$\mathbf{u}+\mathbf{v} = \mathbf{v}+\mathbf{u}, \qquad \text{(commutative)} \tag{5}$$

$$(\mathbf{u}+\mathbf{v})+\mathbf{w} = \mathbf{u}+(\mathbf{v}+\mathbf{w}), \qquad \text{(associative)} \tag{6}$$

and so on, along with the following properties of the dot product and norm.

Dot Product

Commutative: $$\mathbf{u}\cdot\mathbf{v} = \mathbf{v}\cdot\mathbf{u}, \tag{7a}$$

Nonnegative: $$\begin{aligned}\mathbf{u}\cdot\mathbf{u} &> 0 \quad \text{for all } \mathbf{u}\neq\mathbf{0}\\ &= 0 \quad \text{for } \mathbf{u}=\mathbf{0},\end{aligned} \tag{7b}$$

Linear: $$(\alpha\mathbf{u}+\beta\mathbf{v})\cdot\mathbf{w} = \alpha(\mathbf{u}\cdot\mathbf{w})+\beta(\mathbf{v}\cdot\mathbf{w}), \tag{7c}$$

Norm

Scaling: $$\|\alpha\mathbf{u}\| = |\alpha|\,\|\mathbf{u}\|, \tag{8a}$$

Nonnegative: $$\begin{aligned}\|\mathbf{u}\| &> 0 \quad \text{for all } \mathbf{u}\neq\mathbf{0}\\ &= 0 \quad \text{for } \mathbf{u}=\mathbf{0},\end{aligned} \tag{8b}$$

Triangular Inequality: $$\|\mathbf{u}+\mathbf{v}\| \le \|\mathbf{u}\|+\|\mathbf{v}\|. \tag{8c}$$

To complete the extension to generalized vector space, we reverse the cart and the horse by elevating these various properties to the level of axioms, or requirements. That is, we let the fundamental objects, the vectors, be whatever we choose them to be, and then define addition and scalar multiplication operations, a zero vector, a negative inverse, a dot or "inner" product (if we wish), and a norm (if we wish), so that those axioms are satisfied. Our chief interest, in introducing generalized vector space, is in function spaces, but we will not work with function spaces until Chapter 17, when we study Fourier series and the Sturm–Liouville theory.

Next, we introduce the concept of *span* and *linear dependence*, primarily so that we can develop the idea of the *expansion* of a given vector in a vector space $\mathcal{S}$ in terms of a set of *base vectors* for $\mathcal{S}$. We define a set of vectors $\{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$ to be a *basis* for $\mathcal{S}$ if each vector $\mathbf{u}$ in $\mathcal{S}$ can be expressed ("expanded") uniquely in the form $\mathbf{u} = \alpha_1\mathbf{e}_1 + \cdots + \alpha_k\mathbf{e}_k$, and prove that a set $\{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$ is a basis for $\mathcal{S}$ if and only if it spans $\mathcal{S}$ and is LI (linearly independent). In particular, orthogonal bases are especially convenient because of the ease with which one can compute the expansion coefficients α_j. The result is

$$\mathbf{u} = \left(\frac{\mathbf{u}\cdot\mathbf{e}_1}{\mathbf{e}_1\cdot\mathbf{e}_1}\right)\mathbf{e}_1 + \cdots + \left(\frac{\mathbf{u}\cdot\mathbf{e}_k}{\mathbf{e}_k\cdot\mathbf{e}_k}\right)\mathbf{e}_k \tag{9}$$

if the basis is orthogonal, and

$$\mathbf{u} = (\mathbf{u}\cdot\hat{\mathbf{e}}_1)\,\hat{\mathbf{e}}_1 + \cdots + (\mathbf{u}\cdot\hat{\mathbf{e}}_k)\,\hat{\mathbf{e}}_k \tag{10}$$

if it is ON (orthonormal); (9) and (10) should be understood and remembered.

Finally, we study the question of the best *approximation* of a given vector $\mathbf{u}$ in a vector space $\mathcal{S}$ in terms of an ON set $\{\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_N\}$ which falls short of being a basis for $\mathcal{S}$. We show that the best approximation (i.e., the one that minimizes the norm of the error vector) is

$$\mathbf{u} \approx \sum_{j=1}^{N} (\mathbf{u}\cdot\hat{\mathbf{e}}_j)\,\hat{\mathbf{e}}_j \tag{11}$$

which, in geometrical language, is the *orthogonal projection of* $\mathbf{u}$ *onto the span of* $\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_N$.

Chapter 10

Matrices and Linear Equations

10.1 Introduction

We have already met matrices in Section 8.3.3, but they were introduced there only as a notational convenience for the implementation of Gauss elimination and Gauss–Jordan reduction. In the present chapter we focus on matrix theory itself, which theory will enable us to obtain additional important results regarding the solution of systems of linear algebraic equations.

One way to view matrix theory is to think in terms of a parallel with function theory. In our mathematical training, we first study numbers – the points on a real number axis. Then we study functions, which are mappings, or transformations, from one real axis to another. For instance, $f(x) = x^2$ maps the point $x = 3$, say, on an x axis to the point $f = 9$ on an f axis. Just as functions act upon numbers, we shall see that matrices act upon vectors and are mappings from one vector space to another. Having studied vectors, in Chapter 9, we can now turn our attention to matrices.

Historically, matrix theory did not become a part of undergraduate engineering science curricula until around 1960, when digital computers became widely available in academia.

10.2 Matrices and Matrix Algebra

A **matrix** is a rectangular array of quantities that are called the **elements** of the matrix. Normally, the elements will be real numbers, although they may occasionally be other objects such as differential operators or even matrices. Some of these cases will be met as we go along; for the present, however, let us consider the elements to be *real numbers*. The complex case is studied in Chapter 12.

Specifically, any matrix $\mathbf{A}$ may be expressed as

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \tag{1}$$

where the brackets (or, in some texts, parentheses) are used to emphasize that the entire array is to be regarded as a single entity. A horizontal line of elements is called a **row**, and a vertical line is called a **column**. Counting rows from the top and columns from the left, then

$$a_{21} \quad a_{22} \quad \cdots \quad a_{2n} \qquad \text{and} \qquad \begin{matrix} a_{13} \\ a_{23} \\ \vdots \\ a_{m3}, \end{matrix}$$

say, are the second row and third column, respectively. Thus, we call the first subscript on a_{ij} the *row index*, and the second subscript the *column index*.

We usually use boldface capital letters to denote matrices and lightface lowercase letters to denote their elements. The matrix $\mathbf{A}$ in (1) is seen to have m rows and n columns and is therefore said to be $m \times n$ (read "m by n"); we shall refer to this as the *form* of $\mathbf{A}$. In some applications m and/or n may be infinite, but here we shall consider only matrices of finite size: $1 \leq m < \infty$, $1 \leq n < \infty$. Furthermore, m and n may, but need not, be equal. If $\mathbf{A}$ is small we may wish to dispense with the subscript notation for the elements. For example, if $m = n = 2$, we may prefer

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \quad \text{to} \quad \mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \tag{2}$$

but if $\mathbf{A}$ is large this becomes inconvenient. The double-subscript notation employed in (1) is especially convenient for digital computer calculations.

In view of the subscript notation in (1), one also writes

$$\mathbf{A} = \{a_{ij}\} \tag{3}$$

for short, where a_{ij} is called the *ij element* and $i = 1, \ldots, m$ and $j = 1, \ldots, n$. Some authors write $a_{i,j}$ in place of a_{ij} to avoid ambiguity – for example, to prevent us from reading a_{21} as a-sub-twenty-one, but we will omit the commas, except when such ambiguity is not easily resolved from the context.

EXAMPLE 1. The matrices

$$\mathbf{A} = \begin{bmatrix} 3 & -1 \\ 0 & 2 \\ 7 & 5 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 8 & -6 & 0 \\ 1 & 3 & 27 \\ 2 & 9 & 4 \end{bmatrix}, \quad \text{and} \quad \mathbf{C} = [8, -7, 0, 4, 3]$$

are 3×2, 3×3, and 1×5, respectively. If we denote $\mathbf{B} = \{b_{ij}\}$, then $b_{11} = 8$, $b_{12} = -6$, $b_{32} = 9$, and so on. $\mathbf{C}$ happens to be a single row and it seems best to separate the elements by commas, but the commas are not essential. ∎

Two matrices are said to be **equal** if they are of the same form and if their corresponding elements are equal. For instance, none of the matrices above are equal, but if $\mathbf{D} = [8, -7, 0, 4, 3]$ then, $\mathbf{D} = \mathbf{C}$.

One may be tempted to identify $\mathbf{C}$, above, as a 5-tuple *vector* rather than as a 1×5 matrix. That would be a bit premature since vectors are not merely objects; they have rules for vector addition and scalar multiplication defined, whereas our matrices are, thus far, just mathematical "objects." In fact, our next step is to define some arithmetic operations for matrices so that they may be manipulated in useful ways. For vectors we defined two arithmetic operations, vector addition and scalar multiplication; for matrices we define three: matrix addition, scalar multiplication, *and* the multiplication of matrices.

Matrix addition. If $\mathbf{A} = \{a_{ij}\}$ and $\mathbf{B} = \{b_{ij}\}$ are any two matrices of the same form, say $m \times n$, then their sum $\mathbf{A} + \mathbf{B}$ is defined as

$$\boxed{\mathbf{A} + \mathbf{B} \equiv \{a_{ij} + b_{ij}\}} \tag{4}$$

and is itself an $m \times n$ matrix. If $\mathbf{A}$ and $\mathbf{B}$ are of the same form, they are said to be *conformable for addition*; if they are not of the same form, then $\mathbf{A} + \mathbf{B}$ is *not defined.*

EXAMPLE 2. If

$$\mathbf{A} = \begin{bmatrix} 2 & 0 & -6 \\ 1 & 3 & 4 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} -1 & 2 & 0 \\ 15 & 6 & 3 \end{bmatrix}, \quad \text{and} \quad \mathbf{C} = \begin{bmatrix} 4 & 2 \\ -1 & 0 \end{bmatrix}, \tag{5}$$

then,

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} 1 & 2 & -6 \\ 16 & 9 & 7 \end{bmatrix}, \tag{6}$$

but $\mathbf{A} + \mathbf{C}$ and $\mathbf{B} + \mathbf{C}$ are not defined since $\mathbf{A}$ and $\mathbf{B}$ are 2×3 while $\mathbf{C}$ is 2×2. ∎

Scalar multiplication. If $\mathbf{A} = \{a_{ij}\}$ is any $m \times n$ matrix and c is any scalar, their product is defined as

$$\boxed{c\mathbf{A} \equiv \{ca_{ij}\},} \tag{7}$$

and is itself an $m \times n$ matrix; we do not distinguish between $c\mathbf{A}$ and $\mathbf{A}c$. Furthermore, we denote

$$-\mathbf{A} \equiv (-1)\mathbf{A}. \tag{8}$$

In place of $\mathbf{A} + (-\mathbf{B})$, we simply write $\mathbf{A} - \mathbf{B}$, and call it the *difference* of $\mathbf{A}$ and $\mathbf{B}$, or $\mathbf{A}$ *minus* $\mathbf{B}$.

EXAMPLE 3. If $\mathbf{A}$ and $\mathbf{C}$ are the matrices in Example 1, then

$$3\mathbf{A} = \begin{bmatrix} 9 & -3 \\ 0 & 6 \\ 21 & 15 \end{bmatrix} \quad \text{and} \quad -\mathbf{C} = [-8, 7, 0, -4, -3]. \quad \blacksquare$$

We shall list the important properties of matrix addition and scalar multiplication in a moment, but first let us define the so-called **zero matrix 0** to be any $m \times n$ matrix all the elements of which are zero. For example,

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \mathbf{0} \quad \text{and} \quad \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = \mathbf{0},$$

the first being 2×3, the second being 3×1.

THEOREM 10.2.1 *Properties of Matrix Addition and Scalar Multiplication*
If $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are $m \times n$ matrices, $\mathbf{0}$ is an $m \times n$ zero matrix, and α, β are any scalars, then

$$\begin{aligned}
\mathbf{A} + \mathbf{B} &= \mathbf{B} + \mathbf{A}, && \text{(commutativity)} && (9a)\\
(\mathbf{A} + \mathbf{B}) + \mathbf{C} &= \mathbf{A} + (\mathbf{B} + \mathbf{C}), && \text{(associativity)} && (9b)\\
\mathbf{A} + \mathbf{0} &= \mathbf{A}, && && (9c)\\
\mathbf{A} + (-\mathbf{A}) &= \mathbf{0}, && && (9d)\\
\alpha(\beta\mathbf{A}) &= (\alpha\beta)\mathbf{A}, && \text{(associativity)} && (9e)\\
(\alpha + \beta)\mathbf{A} &= \alpha\mathbf{A} + \beta\mathbf{A}, && \text{(distributivity)} && (9f)\\
\alpha(\mathbf{A} + \mathbf{B}) &= \alpha\mathbf{A} + \alpha\mathbf{B}, && \text{(distributivity)} && (9g)\\
1\mathbf{A} &= \mathbf{A}, && && (9h)\\
0\mathbf{A} &= \mathbf{0}, && && (9i)\\
\alpha\mathbf{0} &= \mathbf{0}. && && (9j)
\end{aligned}$$

The proof follows from the foregoing definitions and is left for the exercises.

Observe that there are no surprises in (9); the usual rules of arithmetic are seen to apply. For the special case where $\mathbf{A}$ consists of a single row (or column), we see that the definitions of addition and scalar multiplication above are identical to those introduced in Section 9.4 for n-tuple vectors. Thus, we may properly refer to the matrices

$$\mathbf{A} = [a_{11}, \ldots, a_{1n}] \quad \text{and} \quad \mathbf{A} = \begin{bmatrix} a_{11} \\ \vdots \\ a_{n1} \end{bmatrix} \tag{10}$$

as n-dimensional **row** and **column vectors**, respectively.

Matrix multiplication. Judging from the rather natural way in which matrix addition and scalar multiplication are defined, by (4) and (7), one might well expect the multiplication of two matrices $\mathbf{A} = \{a_{ij}\}$ and $\mathbf{B} = \{b_{ij}\}$ to be defined only if $\mathbf{A}$ and $\mathbf{B}$ are of the same form, with the definition $\mathbf{AB} \equiv \{a_{ij}b_{ij}\}$. In fact, this is *not* the case. Instead, the standard definition of matrix multiplication is the one suggested by Cayley.* Called the *Cayley product*, it is as follows: if $\mathbf{A} = \{a_{ij}\}$ is any $m \times n$ matrix and $\mathbf{B} = \{b_{ij}\}$ is any $n \times p$ matrix (so that the number of columns of $\mathbf{A}$ is equal to the number of rows of $\mathbf{B}$), then the product $\mathbf{AB}$ is defined as

$$\boxed{\mathbf{AB} \equiv \left\{ \sum_{k=1}^{n} a_{ik}b_{kj} \right\}; \qquad (1 \le i \le m, \ 1 \le j \le p)} \tag{11}$$

that is, if we denote $\mathbf{AB} = \mathbf{C} = \{c_{ij}\}$, then

$$c_{ij} = \sum_{k=1}^{n} a_{ik}b_{kj}. \tag{12}$$

If the number of columns of $\mathbf{A}$ is equal to the number of rows of $\mathbf{B}$, then $\mathbf{A}$ and $\mathbf{B}$ are said to be *conformable for multiplication*; if not, the product $\mathbf{AB}$ is *not defined*. NOTE: The relative forms of $\mathbf{A}$, $\mathbf{B}$, and their product $\mathbf{C}$ are important and, as stated above, are as follows:

$$\begin{array}{ccccc} \mathbf{A} & \text{times} & \mathbf{B} & = & \mathbf{C}. \\ m \times n & & n \times p & & m \times p \\ & \nwarrow \text{equal} \nearrow & & & \end{array} \tag{13}$$

EXAMPLE 4. Suppose that

$$\mathbf{A} = \begin{bmatrix} 2 & 0 & -5 \\ 1 & 3 & 2 \\ 4 & 1 & -1 \\ 0 & 2 & 7 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} 5 & 1 \\ -2 & 3 \\ 1 & 0 \end{bmatrix}.$$

*Arthur Cayley (1821–1895) produced around 200 papers in a 15-year period during which he was engaged in the practice of law. In 1863, he accepted a professorship of mathematics at Cambridge.

Then $\mathbf{A}$ is 4×3 and $\mathbf{B}$ is 3×2. Since the number of columns of $\mathbf{A}$ (namely, 3) is the same as the number of rows of $\mathbf{B}$, the product $\mathbf{AB}$ is defined and, according to (13), will be 4×2. According to the definition (12),

$$\mathbf{AB} = \underset{4\times 3}{\begin{bmatrix} 2 & 0 & -5 \\ 1 & 3 & 2 \\ \boxed{4} & \boxed{1} & \boxed{-1} \\ 0 & 2 & 7 \end{bmatrix}} \underset{3\times 2}{\begin{bmatrix} 5 & \boxed{1} \\ -2 & \boxed{3} \\ 1 & \boxed{0} \end{bmatrix}} = \underset{4\times 2}{\begin{bmatrix} 5 & 2 \\ 1 & 10 \\ 17 & \boxed{7} \\ 3 & 6 \end{bmatrix}} = \mathbf{C}. \tag{14}$$

To compute c_{32}, for example, (12) gives

$$c_{32} = \sum_{k=1}^{3} \underset{\substack{\text{3rd row}\\ \text{of } \mathbf{A}}}{a_{3k}}\, \underset{\substack{\text{2nd column}\\ \text{of } \mathbf{B}}}{b_{k2}} = (4)(1) + (1)(3) + (-1)(0) = 7,$$

as indicated by the arrows in (14). One more:

$$c_{11} = \sum_{k=1}^{3} a_{1k}b_{k1} = (2)(5) + (0)(-2) + (-5)(1) = 5.$$

We move across the rows of the first matrix and down the columns of the second.

COMMENT. Observe that c_{32} is the *dot product* of the third row of $\mathbf{A}$, considered as a 3-tuple vector, with the second column of $\mathbf{B}$. More generally, if $\mathbf{AB} = \mathbf{C} = \{c_{ij}\}$, then c_{ij} is the dot product of the ith row of $\mathbf{A}$ with the jth column of $\mathbf{B}$. Thus, the number of elements in the rows of $\mathbf{A}$ (namely, the number of columns in $\mathbf{A}$) must equal the number of elements in the columns of $\mathbf{B}$ (namely, the number of rows of $\mathbf{B}$). ∎

EXAMPLE 5. Two more examples:

$$\underset{4\times 3}{\begin{bmatrix} 5 & 0 & -1 \\ 2 & 3 & 4 \\ 1 & 0 & 6 \\ 0 & 0 & 1 \end{bmatrix}} \underset{3\times 1}{\begin{bmatrix} 3 \\ 2 \\ 5 \end{bmatrix}} = \underset{4\times 1}{\begin{bmatrix} (5)(3)+(0)(2)+(-1)(5) \\ (2)(3)+(3)(2)+(4)(5) \\ (1)(3)+(0)(2)+(6)(5) \\ (0)(3)+(0)(2)+(1)(5) \end{bmatrix}} = \begin{bmatrix} 10 \\ 32 \\ 33 \\ 5 \end{bmatrix},$$

and

$$\underset{2\times 2}{\begin{bmatrix} -3 & 1 \\ 10 & 2 \end{bmatrix}} \underset{2\times 2}{\begin{bmatrix} 1 & 0 \\ 2 & 4 \end{bmatrix}} = \underset{2\times 2}{\begin{bmatrix} (-3)(1)+(1)(2) & (-3)(0)+(1)(4) \\ (10)(1)+(2)(2) & (10)(0)+(2)(4) \end{bmatrix}}$$

$$= \begin{bmatrix} -1 & 4 \\ 14 & 8 \end{bmatrix}. \quad ∎$$

It is extremely important to see that *matrix multiplication is not, in general, commutative; that is,*

$$\mathbf{AB} \neq \mathbf{BA}, \tag{15}$$

except in exceptional cases. For suppose that $\mathbf{A} = m \times n$ and $\mathbf{B} = n \times p$ (i.e., $\mathbf{A}$ is $m \times n$ and $\mathbf{B}$ is $n \times p$) so that $\mathbf{AB}$ is at least defined. However, $\mathbf{BA} = (n \times p)(m \times n)$ is not even defined, let alone equal to $\mathbf{AB}$, unless $p = m$. Assuming that that is the case,

$$\mathbf{BA} = (n \times m)(m \times n) = n \times n, \tag{16a}$$

whereas

$$\mathbf{AB} = (m \times n)(n \times m) = m \times m. \tag{16b}$$

Comparing (16a) with (16b), we see that we must also have $m = n$ if $\mathbf{AB}$ and $\mathbf{BA}$ are to be of the same form and hence *possibly* equal. Thus, a necessary condition for $\mathbf{AB}$ to equal $\mathbf{BA}$ (i.e., for $\mathbf{A}$ and $\mathbf{B}$ to commute under multiplication) is that $\mathbf{A}$ and $\mathbf{B}$ both be $n \times n$.

EXAMPLE 6. If

$$\mathbf{A} = \begin{bmatrix} 2 & 3 \\ 3 & 0 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} -2 & 3 \\ 3 & -4 \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} 2 & 3 \\ 1 & 0 \end{bmatrix},$$

we find that $\mathbf{A}$ and $\mathbf{B}$ commute ($\mathbf{AB} = \mathbf{BA}$), but $\mathbf{A}$ and $\mathbf{C}$ do not ($\mathbf{AC} \neq \mathbf{CA}$), nor do $\mathbf{B}$ and $\mathbf{C}$. ∎

Thus, the condition that $\mathbf{A}$ and $\mathbf{B}$ must both be $n \times n$, for $\mathbf{A}$ and $\mathbf{B}$ to commute, is *necessary but not sufficient.* In view of the importance of which factor is first and which is second in a matrix product $\mathbf{AB}$, we sometimes say that $\mathbf{B}$ is **pre-multiplied** by $\mathbf{A}$, and $\mathbf{A}$ is **post-multiplied** by $\mathbf{B}$ so as to leave no doubt as to which factor is first and which is second.

The lack of commutativity, in general, is a major setback so we must wonder why Cayley's definition has been adopted rather than the simpler one that comes to mind, $\mathbf{AB} \equiv \{a_{ij}b_{ij}\}$, which would surely yield commutativity since $\mathbf{BA} = \{b_{ij}a_{ij}\} = \{a_{ij}b_{ij}\} = \mathbf{AB}$ (the second equality following from the commutativity of the multiplication of ordinary numbers). A sufficiently compelling reason to use Cayley's definition involves the application of matrix notation to systems of linear algebraic equations for it turns out that, with Cayley's definition of multiplication, the system of m linear algebraic equations

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= c_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= c_2, \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= c_m \end{aligned} \tag{17}$$

in the n unknowns $x_1, \ldots, x_n$ is equivalent to the single compact matrix equation

$$\mathbf{Ax} = \mathbf{c}, \tag{18}$$

where

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \text{and} \quad \mathbf{c} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix}. \tag{19}$$

$\mathbf{A}$ is called the **coefficient matrix**, and $\mathbf{x}$ is the unknown; that is, its components are the unknowns $x_1, \ldots, x_n$. To verify the claimed equivalence, work out the product $\mathbf{Ax}$, and set the result equal to $\mathbf{c}$. That step gives

$$\underset{m \times 1}{\begin{bmatrix} a_{11}x_1 + \cdots + a_{1n}x_n \\ a_{21}x_1 + \cdots + a_{2n}x_n \\ \vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n \end{bmatrix}} = \underset{m \times 1}{\begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_m \end{bmatrix}}. \tag{20}$$

These two $m \times 1$ matrices (or m-dimensional column vectors) will be equal if and only if each of their corresponding m elements (or components) are equal. Thus, (18) is equivalent to the m scalar equations (17), so (17) and (18) are equivalent, as claimed. This important result is of course a consequence of (11) and provides strong support for adopting that definition of matrix multiplication.

Any $n \times n$ matrix $\mathbf{A} = \{a_{ij}\}$ is said to be **square**, and of **order** n,* and the elements $a_{11}, a_{22}, \ldots, a_{nn}$ are said to lie on the **main diagonal** of $\mathbf{A}$ – that is, the diagonal from the upper left corner to the lower right corner. Notice that to be able to multiply any matrix $\mathbf{A}$ with itself, $\mathbf{A}$ needs to be square. For suppose that $\mathbf{A}$ is $m \times n$; then we have

$$\underset{m \times \underline{n}}{\mathbf{A}} \qquad \underset{\underline{m} \times n}{\mathbf{A}}$$

and we need n (the number of columns in the first matrix) to equal m (the number of rows in the second) for the multiplication to be defined. If $\mathbf{A}$ *is* square and p is any positive integer, we define

$$\underbrace{\mathbf{AA} \cdots \mathbf{A}}_{p \text{ factors}} \equiv \mathbf{A}^p. \tag{21}$$

The familiar laws of exponents,

$$\mathbf{A}^p\mathbf{A}^q = \mathbf{A}^{p+q}, \qquad (\mathbf{A}^p)^q = \mathbf{A}^{pq} \tag{22}$$

follow for any positive integers p and q.

*Thus, we distinguish between form and order: the **form** of an $m \times n$ matrix is $m \times n$, the **order** of an $n \times n$ matrix is n.

If, in particular, the only nonzero elements of a square matrix lie on the main diagonal, **A** is said to be a **diagonal matrix**. For example,

$$\mathbf{A} = \begin{bmatrix} 3 & 0 \\ 0 & -5 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} 7 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

are diagonal, and any diagonal matrix of order n can be denoted as

$$\mathbf{D} = \begin{bmatrix} d_{11} & 0 & \cdots & 0 \\ 0 & d_{22} & & \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & d_{nn} \end{bmatrix}, \tag{23}$$

where not all of the d_{jj}'s are zero,* and all of the off-diagonal elements are zero; that is, $d_{ij} = 0$ if $i \neq j$. It is left for the exercises to show that

$$\mathbf{D}^p = \begin{bmatrix} d_{11}^p & 0 & \cdots & 0 \\ 0 & d_{22}^p & & \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & d_{nn}^p \end{bmatrix}, \tag{24}$$

for any positive integer p.

If, furthermore, $d_{11} = d_{22} = \cdots = d_{nn} = 1$, then **D** is called the **identity matrix** **I**. Thus,

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & 1 \end{bmatrix} = \{\delta_{ij}\}, \tag{25}$$

where δ_{ij} is the **Kronecker delta** symbol defined in Section 9.10.2, namely,

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases} \tag{26}$$

It is sometimes convenient to include a subscript n to indicate the order of **I**. For example,

$$\mathbf{I}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{I}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

*If all of the diagonal elements were zero as well, then *all* the elements would be zero, and it would be more reasonable to describe **D** as a zero matrix, **0**.

The key property of the identity matrix is that if $\mathbf{A}$ is any square matrix of the same order as $\mathbf{I}$, then

$$\mathbf{IA} = \mathbf{AI} = \mathbf{A}, \tag{27}$$

proof of which is left for the exercises. In other words, $\mathbf{I}$ is the matrix analog of the number 1 in scalar arithmetic! From the first equality in (27), we see that one case in which commutativity *does* hold is when one of the matrices is the identity matrix $\mathbf{I}$.

Finally, it is convenient to extend our definition of $\mathbf{A}^p$ [recall (21)] to the case where $p = 0$. If $\mathbf{A}$ is any $n \times n$ matrix, we define

$$\mathbf{A}^0 \equiv \mathbf{I}, \tag{28}$$

where $\mathbf{I}$ is an $n \times n$ identity matrix.

Perhaps we should take a moment to mention that whereas the identity matrix $\mathbf{I}$ is necessarily square, the zero matrix $\mathbf{0}$ is simply $m \times n$, not necessarily square. It is readily verified that

$$\mathbf{0A} = \mathbf{0} \quad \text{and} \quad \mathbf{A0} = \mathbf{0} \tag{29}$$

for any matrix $\mathbf{A}$. In the first of these equations, suppose that the $\mathbf{0}$ on the left is $m \times n$ and that $\mathbf{A}$ is $n \times p$. Then the $\mathbf{0}$ on the right is $m \times p$; that is, it is not necessarily of the same form as the one on the left.

In view of the general failure of commutativity, as stated in (15), we may well wonder if any other familiar arithmetic rules fail to hold for the multiplication of matrices. The answer is yes; the following rules for real numbers (a, b, c) do *not* carry over to matrices:

1. $ab = ba$ (commutativity).
2. If $ab = ac$ and $a \neq 0$, then $b = c$ (cancellation rule).
3. If $ab = 0$, then $a = 0$ and/or $b = 0$.
4. If $a^2 = 1$, then $a = +1$ or -1.

To add emphasis, we state these difficulties as a theorem.

THEOREM 10.2.2 *"Exceptional" Properties of Matrix Multiplication*

(i) $\mathbf{AB} \neq \mathbf{BA}$ in general.
(ii) Even if $\mathbf{A} \neq \mathbf{0}$, $\mathbf{AB} = \mathbf{AC}$ does not imply that $\mathbf{B} = \mathbf{C}$.
(iii) $\mathbf{AB} = \mathbf{0}$ does not imply that $\mathbf{A} = \mathbf{0}$ and/or $\mathbf{B} = \mathbf{0}$.
(iv) $\mathbf{A}^2 = \mathbf{I}$ does not imply that $\mathbf{A} = +\mathbf{I}$ or $-\mathbf{I}$.

The first of these has already been discussed, and the others are discussed in the exercises. Theorem 10.2.2 notwithstanding, several important properties do carry over from the multiplication of real numbers to the multiplication of matrices:

THEOREM 10.2.3 *"Ordinary" Properties of Matrix Multiplication*
If α, β are scalars, and the matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ are suitably conformable, then

$$\begin{aligned}
(\alpha\mathbf{A})\mathbf{B} &= \mathbf{A}(\alpha\mathbf{B}) = \alpha(\mathbf{AB}), && \text{(associativity)} && (30a)\\
\mathbf{A}(\mathbf{BC}) &= (\mathbf{AB})\mathbf{C}, && \text{(associativity)} && (30b)\\
(\mathbf{A}+\mathbf{B})\mathbf{C} &= \mathbf{AC}+\mathbf{BC}, && \text{(distributivity)} && (30c)\\
\mathbf{C}(\mathbf{A}+\mathbf{B}) &= \mathbf{CA}+\mathbf{CB}, && \text{(distributivity)} && (30d)\\
\mathbf{A}(\alpha\mathbf{B}+\beta\mathbf{C}) &= \alpha\mathbf{AB}+\beta\mathbf{AC}. && \text{(linearity)} && (30e)
\end{aligned}$$

Proof is left for the exercises.

Partitioning. Let us close this section with a discussion of the **partitioning** of matrices. The idea is that any matrix $\mathbf{A}$ (which is larger that 1×1) may be partitioned into a number of smaller matrices called **blocks** by vertical lines that extend from bottom to top, and horizontal lines that extend from left to right.

EXAMPLE 7. *Partitioning.*

$$\mathbf{A} = \begin{bmatrix} 2 & 0 & -3 \\ 5 & 2 & 7 \\ 1 & 3 & 0 \\ 0 & 4 & 6 \end{bmatrix} = \left[\begin{array}{cc|c} 2 & 0 & -3 \\ 5 & 2 & 7 \\ \hline 1 & 3 & 0 \\ \hline 0 & 4 & 6 \end{array}\right] = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \\ \mathbf{A}_{31} & \mathbf{A}_{32} \end{bmatrix}, \tag{31}$$

where the blocks are

$$\mathbf{A}_{11} = \begin{bmatrix} 2 & 0 \\ 5 & 2 \end{bmatrix}, \quad \mathbf{A}_{12} = \begin{bmatrix} -3 \\ 7 \end{bmatrix}, \quad \mathbf{A}_{21} = [1, 3] \quad \mathbf{A}_{22} = [0],$$

and so on. Clearly, the partition is not unique. In the present example we could also have set

$$\mathbf{A} = \begin{bmatrix} 2 & 0 & -3 \\ 5 & 2 & 7 \\ 1 & 3 & 0 \\ 0 & 4 & 6 \end{bmatrix} = \left[\begin{array}{c|c|c} 2 & 0 & -3 \\ 5 & 2 & 7 \\ 1 & 3 & 0 \\ 0 & 4 & 6 \end{array}\right] = [\mathbf{A}_{11}, \mathbf{A}_{12}, \mathbf{A}_{13}], \tag{32}$$

say. ∎

While the matrices used here as illustrations are kept small for convenience, those encountered in modern applications may be quite large, for example 600×800. Even with modern computers such large matrices create special computational problems, and it is often advantageous to work instead with a number of smaller matrices through the use of partitioning. Such advantages might well prove illusory, however, were it not for the fact that the usual matrix arithmetic *can* be carried out with partitioned matrices.

Specifically, if $\mathbf{A}$ and $\mathbf{B}$ are partitioned as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \cdots & \mathbf{A}_{1n} \\ \vdots & \vdots & & \vdots \\ \mathbf{A}_{m1} & \mathbf{A}_{m2} & \cdots & \mathbf{A}_{mn} \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} & \cdots & \mathbf{B}_{1q} \\ \vdots & \vdots & & \vdots \\ \mathbf{B}_{p1} & \mathbf{B}_{p2} & \cdots & \mathbf{B}_{pq} \end{bmatrix}, \tag{33}$$

then

$$\alpha\mathbf{A} = \begin{bmatrix} \alpha\mathbf{A}_{11} & \alpha\mathbf{A}_{12} & \cdots & \alpha\mathbf{A}_{1n} \\ \vdots & \vdots & & \vdots \\ \alpha\mathbf{A}_{m1} & \alpha\mathbf{A}_{m2} & \cdots & \alpha\mathbf{A}_{mn} \end{bmatrix}; \tag{34}$$

if $m = p$ and $n = q$ and each $\mathbf{A}_{ij}$ block is of the same form as the corresponding $\mathbf{B}_{ij}$ block, then

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} \mathbf{A}_{11} + \mathbf{B}_{11} & \mathbf{A}_{12} + \mathbf{B}_{12} & \cdots & \mathbf{A}_{1n} + \mathbf{B}_{1n} \\ \vdots & \vdots & & \vdots \\ \mathbf{A}_{m1} + \mathbf{B}_{m1} & \mathbf{A}_{m2} + \mathbf{B}_{m2} & \cdots & \mathbf{A}_{mn} + \mathbf{B}_{mn} \end{bmatrix}; \tag{35}$$

and if $n = p$ and we denote $\mathbf{AB} = \mathbf{C}$, then

$$\mathbf{C}_{ij} = \sum_{k=1}^{n} \mathbf{A}_{ik}\mathbf{B}_{kj}, \tag{36}$$

provided that the number of columns in each $\mathbf{A}_{ik}$ is the same as the number of rows in the corresponding $\mathbf{B}_{kj}$, so that the products in (36) are defined.

Verification of these three claims, (34) to (36), is left for the exercises.

EXAMPLE 8. If

$$\mathbf{A} = \left[\begin{array}{cc|c} 2 & 4 & 1 \\ -1 & 3 & 0 \\ \hline 5 & 4 & 6 \end{array}\right] = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}, \tag{37}$$

and

$$\mathbf{B} = \left[\begin{array}{cc|c} 0 & 1 & 3 \\ 2 & -4 & -1 \\ \hline 5 & 8 & 2 \end{array}\right] = \begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix}, \tag{38}$$

then

$$\mathbf{AB} = \begin{bmatrix} \mathbf{A}_{11}\mathbf{B}_{11} + \mathbf{A}_{12}\mathbf{B}_{21} & \mathbf{A}_{11}\mathbf{B}_{12} + \mathbf{A}_{12}\mathbf{B}_{22} \\ \mathbf{A}_{21}\mathbf{B}_{11} + \mathbf{A}_{22}\mathbf{B}_{21} & \mathbf{A}_{21}\mathbf{B}_{12} + \mathbf{A}_{22}\mathbf{B}_{22} \end{bmatrix}.$$

Working out the the elements,

$$\mathbf{A}_{11}\mathbf{B}_{11} + \mathbf{A}_{12}\mathbf{B}_{21} = \begin{bmatrix} 2 & 4 \\ -1 & 3 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 2 & -4 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} [5, 8]$$