Numerical Methods for Engineers

Fifth Edition

Steven C. Chapra
Berger Chair in Computing and Engineering
Tufts University

Raymond P. Canale
Professor Emeritus of Civil Engineering
University of Michigan

Boston Burr Ridge, IL Dubuque, IA Madison, WI New York San Francisco St. Louis
Bangkok Bogotá Caracas Kuala Lumpur Lisbon London Madrid Mexico City
Milan Montreal New Delhi Santiago Seoul Singapore Sydney Taipei Toronto

CONTENTS

PREFACE xiii

GUIDED TOUR xvi

ABOUT THE AUTHORS xviii

PART ONE

MODELING, COMPUTERS, AND ERROR ANALYSIS 3

PT 1.1 Motivation 3
PT 1.2 Mathematical Background 5
PT 1.3 Orientation 8

CHAPTER 1
Mathematical Modeling and Engineering Problem Solving 11

1.1 A Simple Mathematical Model 11
1.2 Conservation Laws and Engineering 18
Problems 21

CHAPTER 2
Programming and Software 25

2.1 Packages and Programming 25
2.2 Structured Programming 26
2.3 Modular Programming 35
2.4 Excel 37
2.5 MATLAB 41
2.6 Other Languages and Libraries 45
Problems 46

CHAPTER 3
Approximations and Round-Off Errors 50

3.1 Significant Figures 51
3.2 Accuracy and Precision 53
3.3 Error Definitions 54
3.4 Round-Off Errors 57
Problems 72

CHAPTER 4
Truncation Errors and the Taylor Series 73

4.1 The Taylor Series 73
4.2 Error Propagation 89
4.3 Total Numerical Error 93
4.4 Blunders, Formulation Errors, and Data Uncertainty 95
Problems 97

EPILOGUE: PART ONE 99
PT 1.4 Trade-Offs 99
PT 1.5 Important Relationships and Formulas 102
PT 1.6 Advanced Methods and Additional References 102

PART TWO

ROOTS OF EQUATIONS 105

PT 2.1 Motivation 105
PT 2.2 Mathematical Background 107
PT 2.3 Orientation 108

CHAPTER 5
Bracketing Methods 112

5.1 Graphical Methods 112
5.2 The Bisection Method 116
5.3 The False-Position Method 124
5.4 Incremental Searches and Determining Initial Guesses 130
Problems 131

CHAPTER 6
Open Methods 133

6.1 Simple Fixed-Point Iteration 134
6.2 The Newton-Raphson Method 139
6.3 The Secant Method 145
6.4 Multiple Roots 150
6.5 Systems of Nonlinear Equations 153
Problems 157

CHAPTER 7
Roots of Polynomials 160

7.1 Polynomials in Engineering and Science 160
7.2 Computing with Polynomials 163
7.3 Conventional Methods 166
7.4 Müller's Method 167
7.5 Bairstow's Method 171
7.6 Other Methods 176

7.7 Root Location with Libraries and Packages 176
Problems 185

CHAPTER 8
Case Studies: Roots of Equations 187

8.1 Ideal and Nonideal Gas Laws (Chemical/Bio Engineering) 187
8.2 Open-Channel Flow (Civil/Environmental Engineering) 190
8.3 Design of an Electric Circuit (Electrical Engineering) 194
8.4 Vibration Analysis (Mechanical/Aerospace Engineering) 196
Problems 203

EPILOGUE: PART TWO 212
PT 2.4 Trade-Offs 212
PT 2.5 Important Relationships and Formulas 213
PT 2.6 Advanced Methods and Additional References 213

PART THREE

LINEAR ALGEBRAIC EQUATIONS 217

PT 3.1 Motivation 217
PT 3.2 Mathematical Background 219
PT 3.3 Orientation 227

CHAPTER 9
Gauss Elimination 231

9.1 Solving Small Numbers of Equations 231
9.2 Naive Gauss Elimination 238
9.3 Pitfalls of Elimination Methods 244
9.4 Techniques for Improving Solutions 250
9.5 Complex Systems 257
9.6 Nonlinear Systems of Equations 257
9.7 Gauss-Jordan 259
9.8 Summary 261
Problems 261

CHAPTER 10
***LU* Decomposition and Matrix Inversion 264**

10.1 *LU* Decomposition 264
10.2 The Matrix Inverse 273
10.3 Error Analysis and System Condition 277
Problems 283

CHAPTER 11
Special Matrices and Gauss-Seidel 285

11.1 Special Matrices 285
11.2 Gauss-Seidel 289

11.3 Linear Algebraic Equations with Libraries and Packages 296
Problems 303

CHAPTER 12
Case Studies: Linear Algebraic Equations 305

12.1 Steady-State Analysis of a System of Reactors (Chemical/Bio Engineering) 305
12.2 Analysis of a Statically Determinate Truss (Civil/Environmental Engineering) 308
12.3 Currents and Voltages in Resistor Circuits (Electrical Engineering) 312
12.4 Spring-Mass Systems (Mechanical/Aerospace Engineering) 314
Problems 317

EPILOGUE: PART THREE 327
PT 3.4 Trade-Offs 327
PT 3.5 Important Relationships and Formulas 328
PT 3.6 Advanced Methods and Additional References 328

PART FOUR

OPTIMIZATION 331

PT 4.1 Motivation 331
PT 4.2 Mathematical Background 336
PT 4.3 Orientation 337

CHAPTER 13
One-Dimensional Unconstrained Optimization 341

13.1 Golden-Section Search 342
13.2 Quadratic Interpolation 349
13.3 Newton's Method 351
Problems 353

CHAPTER 14
Multidimensional Unconstrained Optimization 355

14.1 Direct Methods 356
14.2 Gradient Methods 360
Problems 373

CHAPTER 15
Constrained Optimization 375

15.1 Linear Programming 375
15.2 Nonlinear Constrained Optimization 386
15.3 Optimization with Packages 387
Problems 398

CHAPTER 16
Case Studies: Optimization 400

16.1 Least-Cost Design of a Tank (Chemical/Bio Engineering) 400
16.2 Least-Cost Treatment of Wastewater (Civil/Environmental Engineering) 405
16.3 Maximum Power Transfer for a Circuit (Electrical Engineering) 409
16.4 Mountain Bike Design (Mechanical/Aerospace Engineering) 413
Problems 415

EPILOGUE: PART FOUR 422
PT 4.4 Trade-Offs 422
PT 4.5 Additional References 423

PART FIVE

CURVE FITTING 425
PT 5.1 Motivation 425
PT 5.2 Mathematical Background 427
PT 5.3 Orientation 436

CHAPTER 17
Least-Squares Regression 440

17.1 Linear Regression 440
17.2 Polynomial Regression 456
17.3 Multiple Linear Regression 460
17.4 General Linear Least Squares 463
17.5 Nonlinear Regression 468
Problems 471

CHAPTER 18
Interpolation 474

18.1 Newton's Divided-Difference Interpolating Polynomials 475
18.2 Lagrange Interpolating Polynomials 486
18.3 Coefficients of an Interpolating Polynomial 491
18.4 Inverse Interpolation 491
18.5 Additional Comments 492
18.6 Spline Interpolation 495
Problems 505

CHAPTER 19
Fourier Approximation 507

19.1 Curve Fitting with Sinusoidal Functions 508
19.2 Continuous Fourier Series 514
19.3 Frequency and Time Domains 517

19.4 Fourier Integral and Transform 521
19.5 Discrete Fourier Transform (DFT) 523
19.6 Fast Fourier Transform (FFT) 525
19.7 The Power Spectrum 532
19.8 Curve Fitting with Libraries and Packages 533
Problems 542

CHAPTER 20
Case Studies: Curve Fitting 544

20.1 Linear Regression and Population Models (Chemical/Bio Engineering) 544
20.2 Use of Splines to Estimate Heat Transfer (Civil/Environmental Engineering) 548
20.3 Fourier Analysis (Electrical Engineering) 550
20.4 Analysis of Experimental Data (Mechanical/Aerospace Engineering) 551
Problems 553

EPILOGUE: PART FIVE 563
PT 5.4 Trade-Offs 563
PT 5.5 Important Relationships and Formulas 564
PT 5.6 Advanced Methods and Additional References 566

PART SIX

NUMERICAL DIFFERENTIATION AND INTEGRATION 569

PT 6.1 Motivation 569
PT 6.2 Mathematical Background 578
PT 6.3 Orientation 581

CHAPTER 21
Newton-Cotes Integration Formulas 584

21.1 The Trapezoidal Rule 586
21.2 Simpson's Rules 596
21.3 Integration with Unequal Segments 605
21.4 Open Integration Formulas 608
21.5 Multiple Integrals 608
Problems 610

CHAPTER 22
Integration of Equations 613

22.1 Newton-Cotes Algorithms for Equations 613
22.2 Romberg Integration 615
22.3 Gauss Quadrature 620
22.4 Improper Integrals 627
Problems 631

CHAPTER 23
Numerical Differentiation 632

23.1 High-Accuracy Differentiation Formulas 632
23.2 Richardson Extrapolation 635
23.3 Derivatives of Unequally Spaced Data 637
23.4 Derivatives and Integrals for Data with Errors 638
23.5 Numerical Integration/Differentiation with Libraries and Packages 639
Problems 643

CHAPTER 24
Case Studies: Numerical Integration and Differentiation 646

24.1 Integration to Determine the Total Quantity of Heat (Chemical/Bio Engineering) 646
24.2 Effective Force on the Mast of a Racing Sailboat (Civil/Environmental Engineering) 648
24.3 Root-Mean-Square Current by Numerical Integration (Electrical Engineering) 650
24.4 Numerical Integration to Compute Work (Mechanical/Aerospace Engineering) 653
Problems 657

EPILOGUE: PART SIX 667
PT 6.4 Trade-Offs 667
PT 6.5 Important Relationships and Formulas 668
PT 6.6 Advanced Methods and Additional References 668

PART SEVEN

ORDINARY DIFFERENTIAL EQUATIONS 671

PT 7.1 Motivation 671
PT 7.2 Mathematical Background 675
PT 7.3 Orientation 677

CHAPTER 25
Runge-Kutta Methods 681

25.1 Euler's Method 682
25.2 Improvements of Euler's Method 693
25.3 Runge-Kutta Methods 701
25.4 Systems of Equations 711
25.5 Adaptive Runge-Kutta Methods 716
Problems 724

CHAPTER 26
Stiffness and Multistep Methods 726

26.1 Stiffness 726
26.2 Multistep Methods 730
Problems 750

CHAPTER 27
Boundary-Value and Eigenvalue Problems 752

27.1 General Methods for Boundary-Value Problems 753
27.2 Eigenvalue Problems 759
27.3 ODEs and Eigenvalues with Libraries and Packages 771
Problems 779

CHAPTER 28
Case Studies: Ordinary Differential Equations 781

28.1 Using ODEs to Analyze the Transient Response of a Reactor (Chemical/Bio Engineering) 781
28.2 Predator-Prey Models and Chaos (Civil/Environmental Engineering) 788
28.3 Simulating Transient Current for an Electric Circuit (Electrical Engineering) 792
28.4 The Swinging Pendulum (Mechanical/Aerospace Engineering) 797
Problems 801

EPILOGUE: PART SEVEN 808

PT 7.4 Trade-Offs 808
PT 7.5 Important Relationships and Formulas 809
PT 7.6 Advanced Methods and Additional References 809

PART EIGHT

PARTIAL DIFFERENTIAL EQUATIONS 813

PT 8.1 Motivation 813
PT 8.2 Orientation 816

CHAPTER 29
Finite Difference: Elliptic Equations 820

29.1 The Laplace Equation 820
29.2 Solution Techniques 822
29.3 Boundary Conditions 828
29.4 The Control-Volume Approach 834
29.5 Software to Solve Elliptic Equations 837
Problems 838

CHAPTER 30
Finite Difference: Parabolic Equations 840

30.1 The Heat Conduction Equation 840
30.2 Explicit Methods 841
30.3 A Simple Implicit Method 845
30.4 The Crank-Nicolson Method 849
30.5 Parabolic Equations in Two Spatial Dimensions 852
Problems 855

CHAPTER 31
Finite-Element Method 857

31.1 The General Approach 858
31.2 Finite-Element Application in One Dimension 862
31.3 Two-Dimensional Problems 871
31.4 Solving PDEs with Libraries and Packages 875
Problems 881

CHAPTER 32
Case Studies: Partial Differential Equations 884

32.1 One-Dimensional Mass Balance of a Reactor (Chemical/Bio Engineering) 884
32.2 Deflections of a Plate (Civil/Environmental Engineering) 888
32.3 Two-Dimensional Electrostatic Field Problems (Electrical Engineering) 890
32.4 Finite-Element Solution of a Series of Springs (Mechanical/Aerospace Engineering) 893
Problems 897

EPILOGUE: PART EIGHT 899
PT 8.3 Trade-Offs 899
PT 8.4 Important Relationships and Formulas 899
PT 8.5 Advanced Methods and Additional References 900

APPENDIX A: THE FOURIER SERIES 901

APPENDIX B: GETTING STARTED WITH MATLAB 903

BIBLIOGRAPHY 911

INDEX 915

CHAPTER 4

Truncation Errors and the Taylor Series

Truncation errors are those that result from using an approximation in place of an exact mathematical procedure. For example, in Chap. 1 we approximated the derivative of velocity of a falling parachutist by a finite-divided-difference equation of the form [Eq. (1.11)]

$$\frac{dv}{dt} \cong \frac{\Delta v}{\Delta t} = \frac{v(t_{i+1}) - v(t_i)}{t_{i+1} - t_i} \tag{4.1}$$

A truncation error was introduced into the numerical solution because the difference equation only approximates the true value of the derivative (recall Fig. 1.4). In order to gain insight into the properties of such errors, we now turn to a mathematical formulation that is used widely in numerical methods to express functions in an approximate fashion—the Taylor series.

4.1 THE TAYLOR SERIES

Taylor's theorem (Box 4.1) and its associated formula, the Taylor series, is of great value in the study of numerical methods. In essence, the *Taylor series* provides a means to predict a function value at one point in terms of the function value and its derivatives at another point. In particular, the theorem states that any smooth function can be approximated as a polynomial.

A useful way to gain insight into the Taylor series is to build it term by term. For example, the first term in the series is

$$f(x_{i+1}) \cong f(x_i) \tag{4.2}$$

This relationship, called the *zero-order approximation*, indicates that the value of f at the new point is the same as its value at the old point. This result makes intuitive sense because if x_i and x_{i+1} are close to each other, it is likely that the new value is probably similar to the old value.

Equation (4.2) provides a perfect estimate if the function being approximated is, in fact, a constant. However, if the function changes at all over the interval, additional terms

Box 4.1 Taylor's Theorem

Taylor's Theorem

If the function f and its first $n + 1$ derivatives are continuous on an interval containing a and x, then the value of the function at x is given by

$$f(x) = f(a) + f'(a)(x - a) + \frac{f''(a)}{2!}(x - a)^2 + \frac{f^{(3)}(a)}{3!}(x - a)^3 + \cdots + \frac{f^{(n)}(a)}{n!}(x - a)^n + R_n \tag{B4.1.1}$$

where the remainder R_n is defined as

$$R_n = \int_a^x \frac{(x - t)^n}{n!} f^{(n+1)}(t)\, dt \tag{B4.1.2}$$

where t = a dummy variable. Equation (B4.1.1) is called the *Taylor series* or *Taylor's formula*. If the remainder is omitted, the right side of Eq. (B4.1.1) is the Taylor polynomial approximation to $f(x)$. In essence, the theorem states that any smooth function can be approximated as a polynomial.

Equation (B4.1.2) is but one way, called the *integral form*, by which the remainder can be expressed. An alternative formulation can be derived on the basis of the integral mean-value theorem.

First Theorem of Mean for Integrals

If the function g is continuous and integrable on an interval containing a and x, then there exists a point ξ between a and x such that

$$\int_a^x g(t)\, dt = g(\xi)(x - a) \tag{B4.1.3}$$

In other words, this theorem states that the integral can be r sented by an average value for the function $g(\xi)$ times the in length $x - a$. Because the average must occur between the mum and maximum values for the interval, there is a point x which the function takes on the average value.

The first theorem is in fact a special case of a second r value theorem for integrals.

Second Theorem of Mean for Integrals

If the functions g and h are continuous and integrable on an in containing a and x, and h does not change sign in the interval there exists a point ξ between a and x such that

$$\int_a^x g(t)h(t)\, dt = g(\xi) \int_a^x h(t)\, dt \tag{B}$$

Thus, Eq. (B4.1.3) is equivalent to Eq. (B4.1.4) with $h(t) = 1$

The second theorem can be applied to Eq. (B4.1.2) with

$$g(t) = f^{(n+1)}(t) \qquad h(t) = \frac{(x - t)^n}{n!}$$

As t varies from a to x, $h(t)$ is continuous and does not change Therefore, if $f^{(n+1)}(t)$ is continuous, then the integral mean- theorem holds and

$$R_n = \frac{f^{(n+1)}(\xi)}{(n + 1)!}(x - a)^{n+1}$$

This equation is referred to as the *derivative* or *Lagrange fo* the remainder.

of the Taylor series are required to provide a better estimate. For example, the *first-* *approximation* is developed by adding another term to yield

$$f(x_{i+1}) \cong f(x_i) + f'(x_i)(x_{i+1} - x_i)$$

The additional first-order term consists of a slope $f'(x_i)$ multiplied by the distance bet x_i and x_{i+1}. Thus, the expression is now in the form of a straight line and is capable o dicting an increase or decrease of the function between x_i and x_{i+1}.

Although Eq. (4.3) can predict a change, it is exact only for a straight-line, or *l* trend. Therefore, a *second-order* term is added to the series to capture some of the c ture that the function might exhibit:

$$f(x_{i+1}) \cong f(x_i) + f'(x_i)(x_{i+1} - x_i) + \frac{f''(x_i)}{2!}(x_{i+1} - x_i)^2$$

In a similar manner, additional terms can be included to develop the complete Taylor series expansion:

$$f(x_{i+1}) = f(x_i) + f'(x_i)(x_{i+1} - x_i) + \frac{f''(x_i)}{2!}(x_{i+1} - x_i)^2 + \frac{f^{(3)}(x_i)}{3!}(x_{i+1} - x_i)^3 + \cdots + \frac{f^{(n)}(x_i)}{n!}(x_{i+1} - x_i)^n + R_n \tag{4.5}$$

Note that because Eq. (4.5) is an infinite series, an equal sign replaces the approximate sign that was used in Eqs. (4.2) through (4.4). A remainder term is included to account for all terms from $n + 1$ to infinity:

$$R_n = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x_{i+1} - x_i)^{n+1} \tag{4.6}$$

where the subscript n connotes that this is the remainder for the nth-order approximation and ξ is a value of x that lies somewhere between x_i and x_{i+1}. The introduction of the ξ is so important that we will devote an entire section (Sec. 4.1.1) to its derivation. For the time being, it is sufficient to recognize that there is such a value that provides an exact determination of the error.

It is often convenient to simplify the Taylor series by defining a step size $h = x_{i+1} - x_i$ and expressing Eq. (4.5) as

$$f(x_{i+1}) = f(x_i) + f'(x_i)h + \frac{f''(x_i)}{2!}h^2 + \frac{f^{(3)}(x_i)}{3!}h^3 + \cdots + \frac{f^{(n)}(x_i)}{n!}h^n + R_n \tag{4.7}$$

where the remainder term is now

$$R_n = \frac{f^{(n+1)}(\xi)}{(n+1)!}h^{n+1} \tag{4.8}$$

EXAMPLE 4.1 Taylor Series Approximation of a Polynomial

Problem Statement. Use zero- through fourth-order Taylor series expansions to approximate the function

$$f(x) = -0.1x^4 - 0.15x^3 - 0.5x^2 - 0.25x + 1.2$$

from $x_i = 0$ with $h = 1$. That is, predict the function's value at $x_{i+1} = 1$.

Solution. Because we are dealing with a known function, we can compute values for $f(x)$ between 0 and 1. The results (Fig. 4.1) indicate that the function starts at $f(0) = 1.2$ and then curves downward to $f(1) = 0.2$. Thus, the true value that we are trying to predict is 0.2.

The Taylor series approximation with $n = 0$ is [Eq. (4.2)]

$$f(x_{i+1}) \simeq 1.2$$

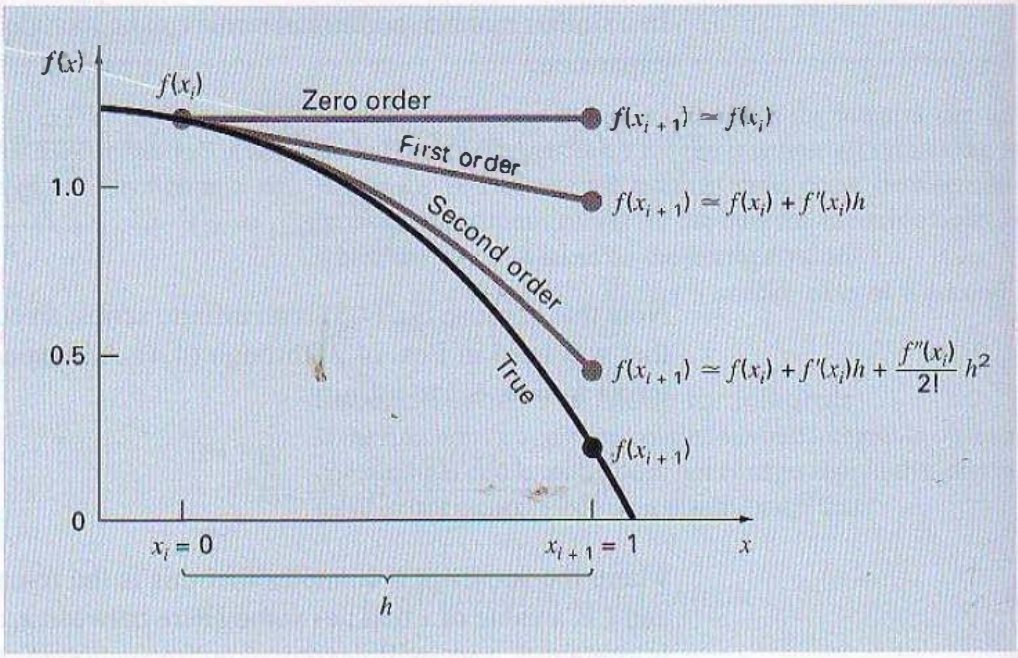


FIGURE 4.1
The approximation of $f(x) = -0.1x^4 - 0.15x^3 - 0.5x^2 - 0.25x + 1.2$ at $x = 1$ by zero- first-order, and second-order Taylor series expansions.

Thus, as in Fig. 4.1, the zero-order approximation is a constant. Using this formulati sults in a truncation error [recall Eq. (3.2)] of

$$E_t = 0.2 - 1.2 = -1.0$$

at $x = 1$.

For $n = 1$, the first derivative must be determined and evaluated at $x = 0$:

$$f'(0) = -0.4(0.0)^3 - 0.45(0.0)^2 - 1.0(0.0) - 0.25 = -0.25$$

Therefore, the first-order approximation is [Eq. (4.3)]

$$f(x_{i+1}) \simeq 1.2 - 0.25h$$

which can be used to compute $f(1) = 0.95$. Consequently, the approximation beg capture the downward trajectory of the function in the form of a sloping straigh (Fig. 4.1). This results in a reduction of the truncation error to

$$E_t = 0.2 - 0.95 = -0.75$$

For $n = 2$, the second derivative is evaluated at $x = 0$:

$$f'(0) = -1.2(0.0)^2 - 0.9(0.0) - 1.0 = -1.0$$

Therefore, according to Eq. (4.4),

$$f(x_{i+1}) \simeq 1.2 - 0.25h - 0.5h^2$$

and substituting $h = 1$, $f(1) = 0.45$. The inclusion of the second derivative now adds downward curvature resulting in an improved estimate, as seen in Fig. 4.1. The trunc error is reduced further to $0.2 - 0.45 = -0.25$.

Additional terms would improve the approximation even more. In fact, the inclusion of the third and the fourth derivatives results in exactly the same equation we started with:

$$f(x) = 1.2 - 0.25h - 0.5h^2 - 0.15h^3 - 0.1h^4$$

where the remainder term is

$$R_4 = \frac{f^{(5)}(\xi)}{5!}h^5 = 0$$

because the fifth derivative of a fourth-order polynomial is zero. Consequently, the Taylor series expansion to the fourth derivative yields an exact estimate at $x_{i+1} = 1$:

$$f(1) = 1.2 - 0.25(1) - 0.5(1)^2 - 0.15(1)^3 - 0.1(1)^4 = 0.2$$

In general, the nth-order Taylor series expansion will be exact for an nth-order polynomial. For other differentiable and continuous functions, such as exponentials and sinusoids, a finite number of terms will not yield an exact estimate. Each additional term will contribute some improvement, however slight, to the approximation. This behavior will be demonstrated in Example 4.2. Only if an infinite number of terms are added will the series yield an exact result.

Although the above is true, the practical value of Taylor series expansions is that, in most cases, the inclusion of only a few terms will result in an approximation that is close enough to the true value for practical purposes. The assessment of how many terms are required to get "close enough" is based on the remainder term of the expansion. Recall that the remainder term is of the general form of Eq. (4.8). This relationship has two major drawbacks. First, ξ is not known exactly but merely lies somewhere between x_i and x_{i+1}. Second, to evaluate Eq. (4.8), we need to determine the $(n + 1)$th derivative of $f(x)$. To do this, we need to know $f(x)$. However, if we knew $f(x)$, there would be no need to perform the Taylor series expansion in the present context!

Despite this dilemma, Eq. (4.8) is still useful for gaining insight into truncation errors. This is because we *do* have control over the term h in the equation. In other words, we can choose how far away from x we want to evaluate $f(x)$, and we can control the number of terms we include in the expansion. Consequently, Eq. (4.8) is usually expressed as

$$R_n = O(h^{n+1})$$

where the nomenclature $O(h^{n+1})$ means that the truncation error is of the order of h^{n+1}. That is, the error is proportional to the step size h raised to the $(n + 1)$th power. Although this approximation implies nothing regarding the magnitude of the derivatives that multiply h^{n+1}, it is extremely useful in judging the comparative error of numerical methods based on Taylor series expansions. For example, if the error is $O(h)$, halving the step size will halve the error. On the other hand, if the error is $O(h^2)$, halving the step size will quarter the error.

In general, we can usually assume that the truncation error is decreased by the addition of terms to the Taylor series. In many cases, if h is sufficiently small, the first- and other lower-order terms usually account for a disproportionately high percent of the error. Thus, only a few terms are required to obtain an adequate estimate. This property is illustrated by the following example.

EXAMPLE 4.2 Use of Taylor Series Expansion to Approximate a Function with an Infinite N of Derivatives

Problem Statement. Use Taylor series expansions with $n = 0$ to 6 to appro $f(x) = \cos x$ at $x_{i+1} = \pi/3$ on the basis of the value of $f(x)$ and its derivatives a $\pi/4$. Note that this means that $h = \pi/3 - \pi/4 = \pi/12$.

Solution. As with Example 4.1, our knowledge of the true function means that determine the correct value $f(\pi/3) = 0.5$.

The zero-order approximation is [Eq. (4.3)]

$$f\left(\frac{\pi}{3}\right) \cong \cos\left(\frac{\pi}{4}\right) = 0.707106781$$

which represents a percent relative error of

$$\varepsilon_t = \frac{0.5 - 0.707106781}{0.5} 100\% = -41.4\%$$

For the first-order approximation, we add the first derivative term where $f'(x) =$

$$f\left(\frac{\pi}{3}\right) \cong \cos\left(\frac{\pi}{4}\right) - \sin\left(\frac{\pi}{4}\right)\left(\frac{\pi}{12}\right) = 0.521986659$$

which has $\varepsilon_t = -4.40$ percent.

For the second-order approximation, we add the second derivative term $f''(x) = -\cos x$:

$$f\left(\frac{\pi}{3}\right) \cong \cos\left(\frac{\pi}{4}\right) - \sin\left(\frac{\pi}{4}\right)\left(\frac{\pi}{12}\right) - \frac{\cos(\pi/4)}{2}\left(\frac{\pi}{12}\right)^2 = 0.497754491$$

with $\varepsilon_t = 0.449$ percent. Thus, the inclusion of additional terms results in an im estimate.

The process can be continued and the results listed, as in Table 4.1. Notice that rivatives never go to zero as was the case with the polynomial in Example 4.1. The each additional term results in some improvement in the estimate. However, also how most of the improvement comes with the initial terms. For this case, by the ti

TABLE 4.1 Taylor series approximation of $f(x) = \cos x$ at $x_{i+1} = \pi/3$ using a base p of $\pi/4$. Values are shown for various orders (n) of approximation.

Order n	$f^{(n)}(x)$	$f(\pi/3)$	ε
0	$\cos x$	0.707106781	−41.4
1	$-\sin x$	0.521986659	−4.4
2	$-\cos x$	0.497754491	0.44
3	$\sin x$	0.499869147	2.62
4	$\cos x$	0.500007551	−1.51
5	$-\sin x$	0.500000304	−6.08
6	$-\cos x$	0.499999988	2.44

have added the third-order term, the error is reduced to 2.62×10^{-2} percent, which means that we have attained 99.9738 percent of the true value. Consequently, although the addition of more terms will reduce the error further, the improvement becomes negligible.

4.1.1 The Remainder for the Taylor Series Expansion

Before demonstrating how the Taylor series is actually used to estimate numerical errors, we must explain why we included the argument ξ in Eq. (4.8). A mathematical derivation is presented in Box 4.1. We will now develop an alternative exposition based on a somewhat more visual interpretation. Then we can extend this specific case to the more general formulation.

Suppose that we truncated the Taylor series expansion [Eq. (4.7)] after the zero-order term to yield

$$f(x_{i+1}) \cong f(x_i)$$

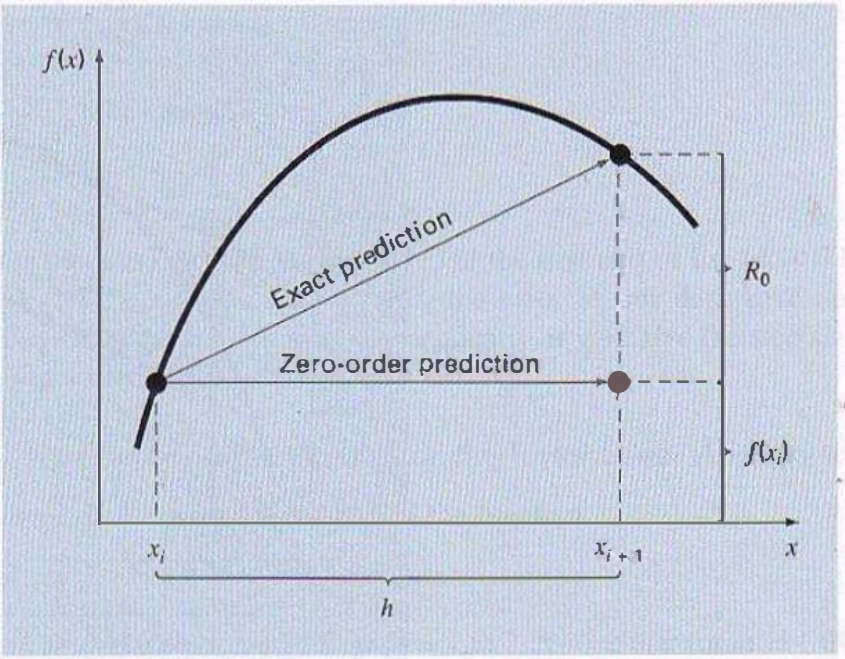
A visual depiction of this zero-order prediction is shown in Fig. 4.2. The remainder, or error, of this prediction, which is also shown in the illustration, consists of the infinite series of terms that were truncated:

$$R_0 = f'(x_i)h + \frac{f''(x_i)}{2!}h^2 + \frac{f^{(3)}(x_i)}{3!}h^3 + \cdots$$

It is obviously inconvenient to deal with the remainder in this infinite series format. One simplification might be to truncate the remainder itself, as in

$$R_0 \cong f'(x_i)h \tag{4.9}$$

FIGURE 4.2
Graphical depiction of a zero-order Taylor series prediction and remainder.



Although, as stated in the previous section, lower-order derivatives usually accoun greater share of the remainder than the higher-order terms, this result is still i because of the neglected second- and higher-order terms. This "inexactness" is impl the approximate equality symbol ($\cong$) employed in Eq. (4.9).

An alternative simplification that transforms the approximation into an equivale based on a graphical insight. As in Fig. 4.3, the *derivative mean-value theorem* stat if a function $f(x)$ and its first derivative are continuous over an interval from x_i to x_{i+} there exists at least one point on the function that has a slope, designated by $f'(\xi)$, parallel to the line joining $f(x_i)$ and $f(x_{i+1})$. The parameter ξ marks the x value whe slope occurs (Fig. 4.3). A physical illustration of this theorem is that, if you travel be two points with an average velocity, there will be at least one moment during the cou the trip when you will be moving at that average velocity.

By invoking this theorem it is simple to realize that, as illustrated in Fig. 4.3, the $f'(\xi)$ is equal to the rise R_0 divided by the run h, or

$$f'(\xi) = \frac{R_0}{h}$$

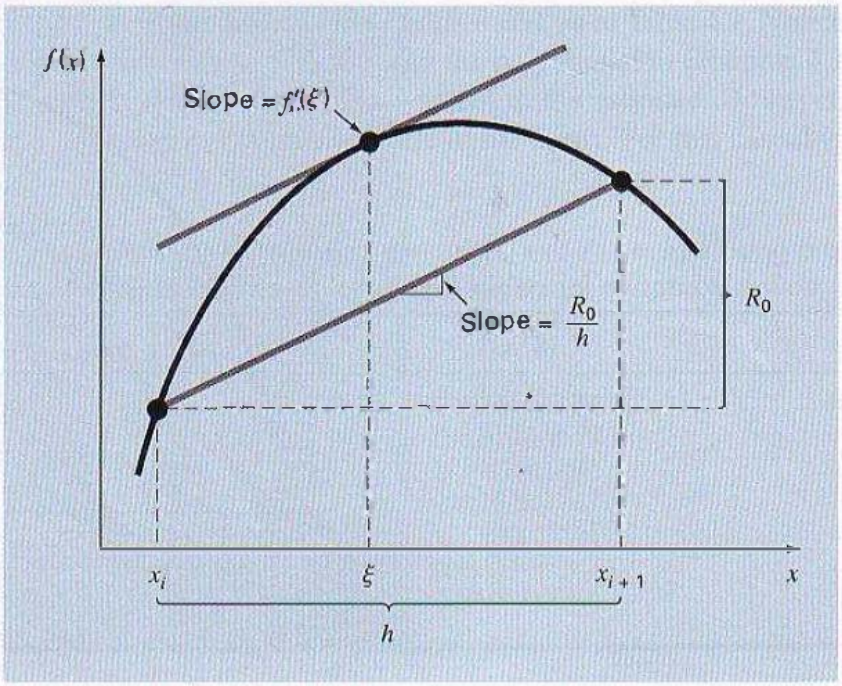
which can be rearranged to give

$$R_0 = f'(\xi)h$$

Thus, we have derived the zero-order version of Eq. (4.8). The higher-order versio merely a logical extension of the reasoning used to derive Eq. (4.10). The first-orde sion is

$$R_1 = \frac{f''(\xi)}{2!}h^2$$

FIGURE 4.3
Graphical depiction of the derivative mean-value theorem.



For this case, the value of ξ conforms to the x value corresponding to the second derivative that makes Eq. (4.11) exact. Similar higher-order versions can be developed from Eq. (4.8).

4.1.2 Using the Taylor Series to Estimate Truncation Errors

Although the Taylor series will be extremely useful in estimating truncation errors throughout this book, it may not be clear to you how the expansion can actually be applied to numerical methods. In fact, we have already done so in our example of the falling parachutist. Recall that the objective of both Examples 1.1 and 1.2 was to predict velocity as a function of time. That is, we were interested in determining $v(t)$. As specified by Eq. (4.5), $v(t)$ can be expanded in a Taylor series:

$$v(t_{i+1}) = v(t_i) + v'(t_i)(t_{i+1} - t_i) + \frac{v''(t_i)}{2!}(t_{i+1} - t_i)^2 + \cdots + R_n \tag{4.12}$$

Now let us truncate the series after the first derivative term:

$$v(t_{i+1}) = v(t_i) + v'(t_i)(t_{i+1} - t_i) + R_1 \tag{4.13}$$

Equation (4.13) can be solved for

$$v'(t_i) = \underbrace{\frac{v(t_{i+1}) - v(t_i)}{t_{i+1} - t_i}}_{\substack{\text{First-order} \\ \text{approximation}}} - \underbrace{\frac{R_1}{t_{i+1} - t_i}}_{\substack{\text{Truncation} \\ \text{error}}} \tag{4.14}$$

The first part of Eq. (4.14) is exactly the same relationship that was used to approximate the derivative in Example 1.2 [Eq. (1.11)]. However, because of the Taylor series approach, we have now obtained an estimate of the truncation error associated with this approximation of the derivative. Using Eqs. (4.6) and (4.14) yields

$$\frac{R_1}{t_{i+1} - t_i} = \frac{v''(\xi)}{2!}(t_{i+1} - t_i) \tag{4.15}$$

or

$$\frac{R_1}{t_{i+1} - t_i} = O(t_{i+1} - t_i) \tag{4.16}$$

Thus, the estimate of the derivative [Eq. (1.11) or the first part of Eq. (4.14)] has a truncation error of order $t_{i+1} - t_i$. In other words, the error of our derivative approximation should be proportional to the step size. Consequently, if we halve the step size, we would expect to halve the error of the derivative.

EXAMPLE 4.3 The Effect of Nonlinearity and Step Size on the Taylor Series Approximation

Problem Statement. Figure 4.4 is a plot of the function

$$f(x) = x^m \tag{E4.3.1}$$

for $m = 1, 2, 3$, and 4 over the range from $x = 1$ to 2. Notice that for $m = 1$ the function is linear, and as m increases, more curvature or nonlinearity is introduced into the function.

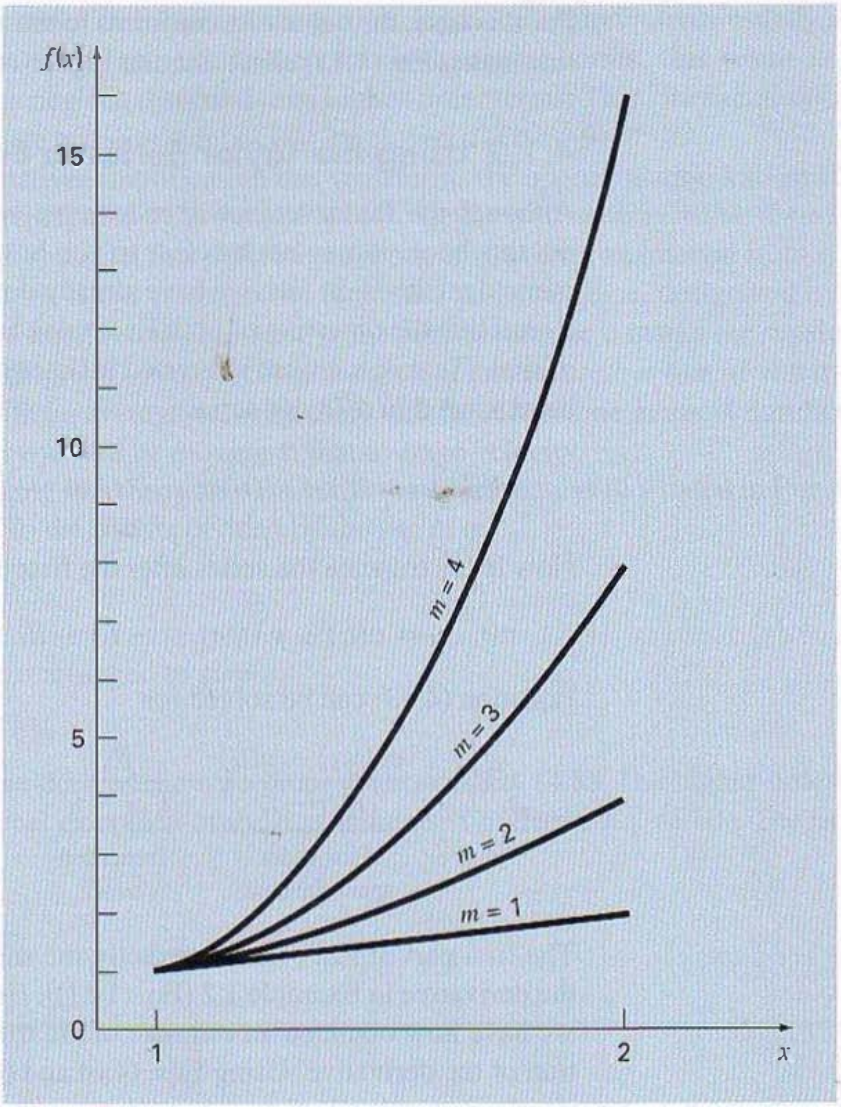


FIGURE 4.4
Plot of the function $f(x) = x^m$ for $m = 1$, 2, 3, and 4. Notice that the function becomes mo nonlinear as m increases.

Employ the first-order Taylor series to approximate this function for various values exponent m and the step size h.

Solution. Equation (E4.3.1) can be approximated by a first-order Taylor series expa as in

$$f(x_{i+1}) = f(x_i) + m x_i^{m-1} h$$

which has a remainder

$$R_1 = \frac{f''(x_i)}{2!}h^2 + \frac{f^{(3)}(x_i)}{3!}h^3 + \frac{f^{(4)}(x_i)}{4!}h^4 + \cdots$$

First, we can examine how the approximation performs as m increases—that is, function becomes more nonlinear. For $m = 1$, the actual value of the function at $x =$

The Taylor series yields

$$f(2) = 1 + 1(1) = 2$$

and

$$R_1 = 0$$

The remainder is zero because the second and higher derivatives of a linear function are zero. Thus, as expected, the first-order Taylor series expansion is perfect when the underlying function is linear.

For $m = 2$, the actual value is $f(2) = 2^2 = 4$. The first-order Taylor series approximation is

$$f(2) = 1 + 2(1) = 3$$

and

$$R_1 = \tfrac{2}{2}(1)^2 + 0 + 0 + \cdots = 1$$

Thus, because the function is a parabola, the straight-line approximation results in a discrepancy. Note that the remainder is determined exactly.

For $m = 3$, the actual value is $f(2) = 2^3 = 8$. The Taylor series approximation is

$$f(2) = 1 + 3(1)^2(1) = 4$$

and

$$R_1 = \tfrac{6}{2}(1)^2 + \tfrac{6}{6}(1)^3 + 0 + 0 + \cdots = 4$$

Again, there is a discrepancy that can be determined exactly from the Taylor series.

For $m = 4$, the actual value is $f(2) = 2^4 = 16$. The Taylor series approximation is

$$f(2) = 1 + 4(1)^3(1) = 5$$

and

$$R_1 = \tfrac{12}{2}(1)^2 + \tfrac{24}{6}(1)^3 + \tfrac{24}{24}(1)^4 + 0 + 0 + \cdots = 11$$

On the basis of these four cases, we observe that R_1 increases as the function becomes more nonlinear. Furthermore, R_1 accounts exactly for the discrepancy. This is because Eq. (E4.3.1) is a simple monomial with a finite number of derivatives. This permits a complete determination of the Taylor series remainder.

Next, we will examine Eq. (E4.3.2) for the case $m = 4$ and observe how R_1 changes as the step size h is varied. For $m = 4$, Eq. (E4.3.2) is

$$f(x + h) = f(x) + 4x_i^3 h$$

If $x = 1, f(1) = 1$ and this equation can be expressed as

$$f(1 + h) = 1 + 4h$$

with a remainder of

$$R_1 = 6h^2 + 4h^3 + h^4$$

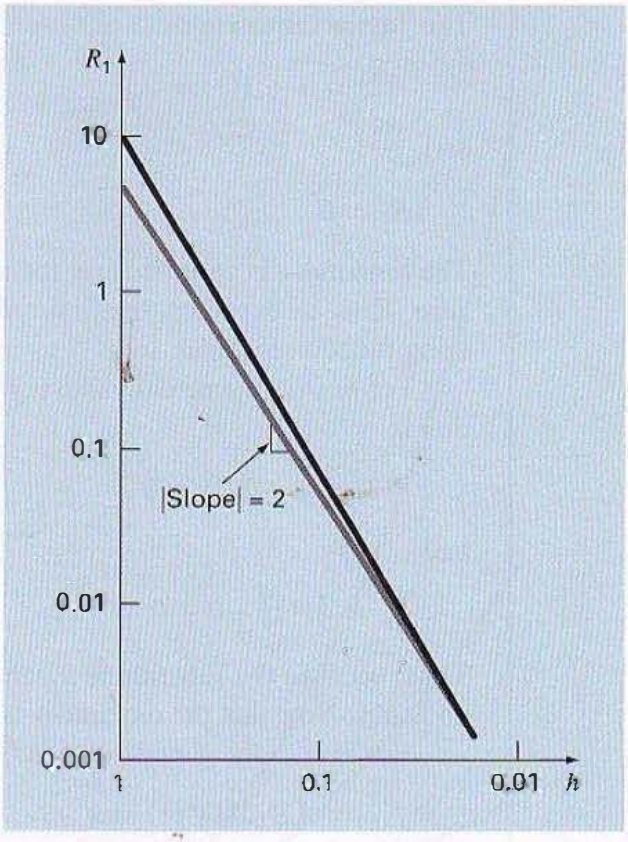


FIGURE 4.5
Log-log plot of the remainder R_1 of the first-order Taylor series approximation of the function $f(x) = x^4$ versus step size h. A line with a slope of 2 is also shown to indicate that as h decreases, the error becomes proportional to h^2.

TABLE 4.2 Comparison of the exact value of the function $f(x) = x^4$ with the first-order Taylor series approximation. Both the function and the approximation are evaluated at $x + h$, where $x = 1$.

h	True	First-Order Approximation	R_1
1	16	5	11
0.5	5.0625	3	2.0625
0.25	2.441406	2	0.44140
0.125	1.601807	1.5	0.10180
0.0625	1.274429	1.25	0.02442
0.03125	1.130982	1.125	0.00598
0.015625	1.063980	1.0625	0.00148

This leads to the conclusion that the discrepancy will decrease as h is reduced. Also, a ficiently small values of h, the error should become proportional to h^2. That is, as halved, the error will be quartered. This behavior is confirmed by Table 4.2 and Fig.

Thus, we conclude that the error of the first-order Taylor series approximatio creases as m approaches 1 and as h decreases. Intuitively, this means that the Taylor

becomes more accurate when the function we are approximating becomes more like a straight line over the interval of interest. This can be accomplished either by reducing the size of the interval or by "straightening" the function by reducing m. Obviously, the latter option is usually not available in the real world because the functions we analyze are typically dictated by the physical problem context. Consequently, we do not have control of their lack of linearity, and our only recourse is reducing the step size or including additional terms in the Taylor series expansion.

4.1.3 Numerical Differentiation

Equation (4.14) is given a formal label in numerical methods—it is called a *finite divided difference*. It can be represented generally as

$$f'(x_i) = \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} + O(x_{i+1} - x_i) \tag{4.17}$$

or

$$f'(x_i) = \frac{\Delta f_i}{h} + O(h) \tag{4.18}$$

where Δf_i is referred to as the *first forward difference* and h is called the step size, that is, the length of the interval over which the approximation is made. It is termed a "forward" difference because it utilizes data at i and $i + 1$ to estimate the derivative (Fig. 4.6*a*). The entire term $\Delta f/h$ is referred to as a *first finite divided difference*.

This forward divided difference is but one of many that can be developed from the Taylor series to approximate derivatives numerically. For example, backward and centered difference approximations of the first derivative can be developed in a fashion similar to the derivation of Eq. (4.14). The former utilizes values at x_{i-1} and x_i (Fig. 4.6*b*), whereas the latter uses values that are equally spaced around the point at which the derivative is estimated (Fig. 4.6*c*). More accurate approximations of the first derivative can be developed by including higher-order terms of the Taylor series. Finally, all the above versions can also be developed for second, third, and higher derivatives. The following sections provide brief summaries illustrating how some of these cases are derived.

Backward Difference Approximation of the First Derivative. The Taylor series can be expanded backward to calculate a previous value on the basis of a present value, as in

$$f(x_{i-1}) = f(x_i) - f'(x_i)h + \frac{f''(x_i)}{2!}h^2 - \cdots \tag{4.19}$$

Truncating this equation after the first derivative and rearranging yields

$$f'(x_i) \cong \frac{f(x_i) - f(x_{i-1})}{h} = \frac{\nabla f_i}{h} \tag{4.20}$$

where the error is $O(h)$, and ∇f_i is referred to as the *first backward difference*. See Fig. 4.6*b* for a graphical representation.

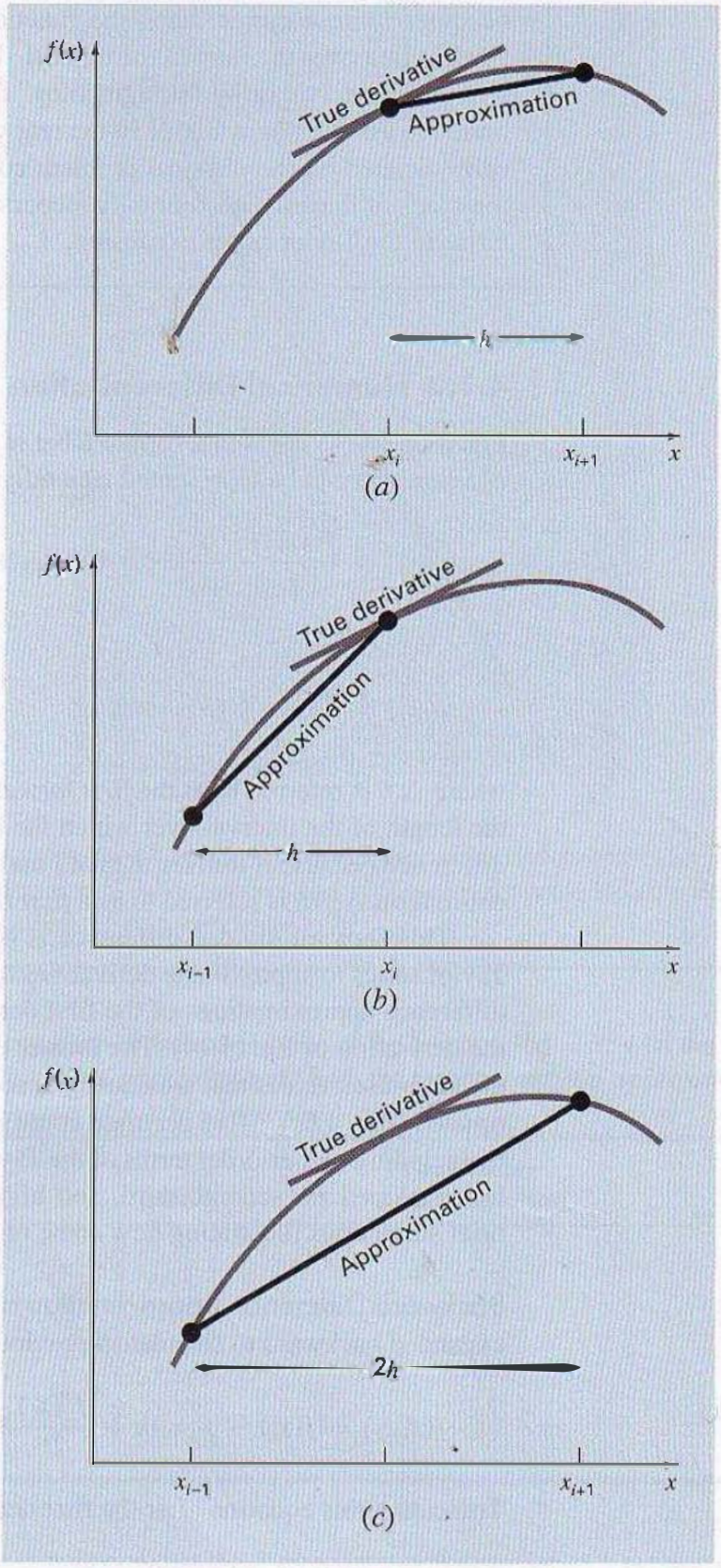


FIGURE 4.6
Graphical depiction of (a) forward, (b) backward, and (c) centered finite-divided-difference approximations of the first derivative.

Centered Difference Approximation of the First Derivative. A third way to approximate the first derivative is to subtract Eq. (4.19) from the forward Taylor series expansion:

$$f(x_{i+1}) = f(x_i) + f'(x_i)h + \frac{f''(x_i)}{2!}h^2 + \cdots \tag{4.21}$$

to yield

$$f(x_{i+1}) = f(x_{i-1}) + 2f'(x_i)h + \frac{2f^{(3)}(x_i)}{3!}h^3 + \cdots$$

which can be solved for

$$f'(x_i) = \frac{f(x_{i+1}) - f(x_{i-1})}{2h} - \frac{f^{(3)}(x_i)}{6}h^2 - \cdots$$

or

$$f'(x_i) = \frac{f(x_{i+1}) - f(x_{i-1})}{2h} - O(h^2) \tag{4.22}$$

Equation (4.22) is a *centered difference* representation of the first derivative. Notice that the truncation error is of the order of h^2 in contrast to the forward and backward approximations that were of the order of h. Consequently, the Taylor series analysis yields the practical information that the centered difference is a more accurate representation of the derivative (Fig. 4.6*c*). For example, if we halve the step size using a forward or backward difference, we would approximately halve the truncation error, whereas for the central difference, the error would be quartered.

EXAMPLE 4.4 Finite-Divided-Difference Approximations of Derivatives

Problem Statement. Use forward and backward difference approximations of $O(h)$ and a centered difference approximation of $O(h^2)$ to estimate the first derivative of

$$f(x) = -0.1x^4 - 0.15x^3 - 0.5x^2 - 0.25x + 1.2$$

at $x = 0.5$ using a step size $h = 0.5$. Repeat the computation using $h = 0.25$. Note that the derivative can be calculated directly as

$$f'(x) = -0.4x^3 - 0.45x^2 - 1.0x - 0.25$$

and can be used to compute the true value as $f'(0.5) = -0.9125$.

Solution. For $h = 0.5$, the function can be employed to determine

$$\begin{aligned} x_{i-1} &= 0 & f(x_{i-1}) &= 1.2 \\ x_i &= 0.5 & f(x_i) &= 0.925 \\ x_{i+1} &= 1.0 & f(x_{i+1}) &= 0.2 \end{aligned}$$

These values can be used to compute the forward divided difference [Eq. (4.17)],

$$f'(0.5) \cong \frac{0.2 - 0.925}{0.5} = -1.45 \qquad |\varepsilon_t| = 58.9\%$$

the backward divided difference [Eq. (4.20)],

$$f'(0.5) \cong \frac{0.925 - 1.2}{0.5} = -0.55 \qquad |\varepsilon_t| = 39.7\%$$

and the centered divided difference [Eq. (4.22)],

$$f'(0.5) \cong \frac{0.2 - 1.2}{1.0} = -1.0 \qquad |\varepsilon_t| = 9.6\%$$

For $h = 0.25$,

$$x_{i-1} = 0.25 \qquad f(x_{i-1}) = 1.10351563$$
$$x_i = 0.5 \qquad f(x_i) = 0.925$$
$$x_{i+1} = 0.75 \qquad f(x_{i+1}) = 0.63632813$$

which can be used to compute the forward divided difference,

$$f'(0.5) \cong \frac{0.63632813 - 0.925}{0.25} = -1.155 \qquad |\varepsilon_t| = 26.5\%$$

the backward divided difference,

$$f'(0.5) \cong \frac{0.925 - 1.10351563}{0.25} = -0.714 \qquad |\varepsilon_t| = 21.7\%$$

and the centered divided difference,

$$f'(0.5) \cong \frac{0.63632813 - 1.10351563}{0.5} = -0.934 \qquad |\varepsilon_t| = 2.4\%$$

For both step sizes, the centered difference approximation is more accurate th ward or backward differences. Also, as predicted by the Taylor series analysis, halv step size approximately halves the error of the backward and forward differences an ters the error of the centered difference.

Finite Difference Approximations of Higher Derivatives. Besides first derivativ Taylor series expansion can be used to derive numerical estimates of higher derivati do this, we write a forward Taylor series expansion for $f(x_{i+2})$ in terms of $f(x_i)$:

$$f(x_{i+2}) = f(x_i) + f'(x_i)(2h) + \frac{f''(x_i)}{2!}(2h)^2 + \cdots$$

Equation (4.21) can be multiplied by 2 and subtracted from Eq. (4.23) to give

$$f(x_{i+2}) - 2f(x_{i+1}) = -f(x_i) + f''(x_i)h^2 + \cdots$$

which can be solved for

$$f''(x_i) = \frac{f(x_{i+2}) - 2f(x_{i+1}) + f(x_i)}{h^2} + O(h)$$

This relationship is called the *second forward finite divided difference.* Similar manipulations can be employed to derive a backward version

$$f''(x_i) = \frac{f(x_i) - 2f(x_{i-1}) + f(x_{i-2})}{h^2} + O(h)$$

and a centered version

$$f''(x_i) = \frac{f(x_{i+1}) - 2f(x_i) + f(x_{i-1})}{h^2} + O(h^2)$$

As was the case with the first-derivative approximations, the centered case is more accurate. Notice also that the centered version can be alternatively expressed as

$$f''(x_i) \cong \frac{\dfrac{f(x_{i+1}) - f(x_i)}{h} - \dfrac{f(x_i) - f(x_{i-1})}{h}}{h}$$

Thus, just as the second derivative is a derivative of a derivative, the second divided difference approximation is a difference of two first divided differences.

We will return to the topic of numerical differentiation in Chap. 23. We have introduced you to the topic at this point because it is a very good example of why the Taylor series is important in numerical methods. In addition, several of the formulas introduced in this section will be employed prior to Chap. 23.

4.2 ERROR PROPAGATION

The purpose of this section is to study how errors in numbers can propagate through mathematical functions. For example, if we multiply two numbers that have errors, we would like to estimate the error in the product.

4.2.1 Functions of a Single Variable

Suppose that we have a function $f(x)$ that is dependent on a single independent variable x. Assume that $\tilde{x}$ is an approximation of x. We, therefore, would like to assess the effect of the discrepancy between x and $\tilde{x}$ on the value of the function. That is, we would like to estimate

$$\Delta f(\tilde{x}) = |f(x) - f(\tilde{x})|$$

The problem with evaluating $\Delta f(\tilde{x})$ is that $f(x)$ is unknown because x is unknown. We can overcome this difficulty if $\tilde{x}$ is close to x and $f(\tilde{x})$ is continuous and differentiable. If these conditions hold, a Taylor series can be employed to compute $f(x)$ near $f(\tilde{x})$, as in

$$f(x) = f(\tilde{x}) + f'(\tilde{x})(x - \tilde{x}) + \frac{f''(\tilde{x})}{2}(x - \tilde{x})^2 + \cdots$$

Dropping the second- and higher-order terms and rearranging yields

$$f(x) - f(\tilde{x}) \cong f'(\tilde{x})(x - \tilde{x})$$

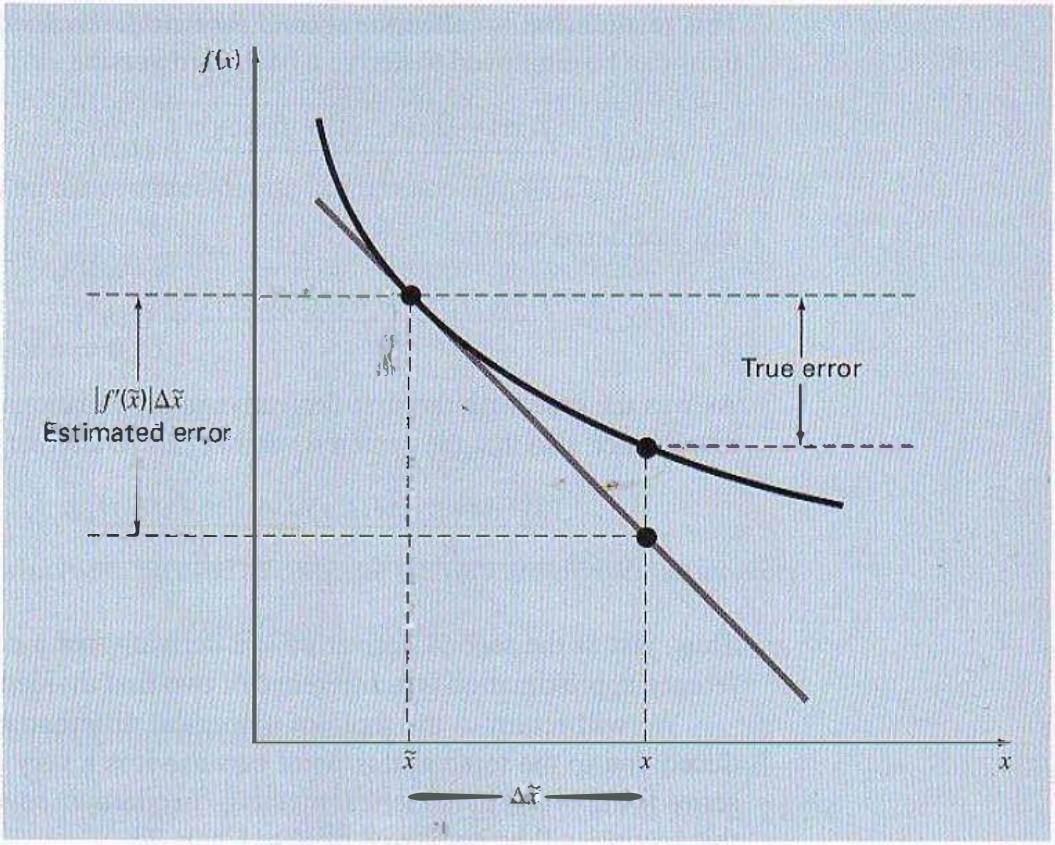


FIGURE 4.7
Graphical depiction of first-order error propagation.

or

$$\Delta f(\tilde{x}) = |f'(\tilde{x})|\Delta\tilde{x}$$

where $\Delta f(\tilde{x}) = |f(x) - f(\tilde{x})|$ represents an estimate of the error of the function and $|x - \tilde{x}|$ represents an estimate of the error of x. Equation (4.25) provides the capabi approximate the error in $f(x)$ given the derivative of a function and an estimate of the in the independent variable. Figure 4.7 is a graphical illustration of the operation.

EXAMPLE 4.5 Error Propagation in a Function of a Single Variable

Problem Statement. Given a value of $\tilde{x} = 2.5$ with an error of $\Delta\tilde{x} = 0.01$, estima resulting error in the function, $f(x) = x^3$.

Solution. Using Eq. (4.25),

$$\Delta f(\tilde{x}) \cong 3(2.5)^2(0.01) = 0.1875$$

Because $f(2.5) = 15.625$, we predict that

$$f(2.5) = 15.625 \pm 0.1875$$

or that the true value lies between 15.4375 and 15.8125. In fact, if x were actually 2.4 function could be evaluated as 15.4382, and if x were 2.51, it would be 15.8132. Fo case, the first-order error analysis provides a fairly close estimate of the true error.

4.2.2 Functions of More than One Variable

The foregoing approach can be generalized to functions that are dependent on more than one independent variable. This is accomplished with a multivariable version of the Taylor series. For example, if we have a function of two independent variables u and v, the Taylor series can be written as

$$f(u_{i+1}, v_{i+1}) = f(u_i, v_i) + \frac{\partial f}{\partial u}(u_{i+1} - u_i) + \frac{\partial f}{\partial v}(v_{i+1} - v_i) + \frac{1}{2!}\left[\frac{\partial^2 f}{\partial u^2}(u_{i+1} - u_i)^2 + 2\frac{\partial^2 f}{\partial u \partial v}(u_{i+1} - u_i)(v_{i+1} - v_i) + \frac{\partial^2 f}{\partial v^2}(v_{i+1} - v_i)^2\right] + \cdots \tag{4.26}$$

where all partial derivatives are evaluated at the base point i. If all second-order and higher terms are dropped, Eq. (4.26) can be solved for

$$\Delta f(\tilde{u}, \tilde{v}) = \left|\frac{\partial f}{\partial u}\right| \Delta\tilde{u} + \left|\frac{\partial f}{\partial v}\right| \Delta\tilde{v}$$

where $\Delta\tilde{u}$ and $\Delta\tilde{v}$ = estimates of the errors in u and v, respectively.

For n independent variables $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n$ having errors $\Delta\tilde{x}_1, \Delta\tilde{x}_2, \ldots, \Delta x_n$, the following general relationship holds:

$$\Delta f(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n) \cong \left|\frac{\partial f}{\partial x_1}\right| \Delta\tilde{x}_1 + \left|\frac{\partial f}{\partial x_2}\right| \Delta\tilde{x}_2 + \cdots + \left|\frac{\partial f}{\partial x_n}\right| \Delta\tilde{x}_n \tag{4.27}$$

EXAMPLE 4.6 Error Propagation in a Multivariable Function

Problem Statement. The deflection y of the top of a sailboat mast is

$$y = \frac{FL^4}{8EI}$$

where F = a uniform side loading (lb/ft), L = height (ft), E = the modulus of elasticity (lb/ft^2), and I = the moment of inertia (ft^4). Estimate the error in y given the following data:

$\tilde{F} = 50$ lb/ft	$\Delta\tilde{F} = 2$ lb/ft
$\tilde{L} = 30$ ft	$\Delta\tilde{L} = 0.1$ ft
$\tilde{E} = 1.5 \times 10^8$ lb/ft^2	$\Delta\tilde{E} = 0.01 \times 10^8$ lb/ft^2
$\tilde{I} = 0.06$ ft^4	$\Delta\tilde{I} = 0.0006$ ft^4

Solution. Employing Eq. (4.27) gives

$$\Delta y(\tilde{F}, \tilde{L}, \tilde{E}, \tilde{I}) = \left|\frac{\partial y}{\partial F}\right| \Delta\tilde{F} + \left|\frac{\partial y}{\partial L}\right| \Delta\tilde{L} + \left|\frac{\partial y}{\partial E}\right| \Delta\tilde{E} + \left|\frac{\partial y}{\partial I}\right| \Delta\tilde{I}$$

or

$$\Delta y(\tilde{F}, \tilde{L}, \tilde{E}, \tilde{I}) \cong \frac{\tilde{L}^4}{8\tilde{E}\tilde{I}}\Delta\tilde{F} + \frac{\tilde{F}\tilde{L}^3}{2\tilde{E}\tilde{I}}\Delta\tilde{L} + \frac{\tilde{F}\tilde{L}^4}{8\tilde{E}^2\tilde{I}}\Delta\tilde{E} + \frac{\tilde{F}\tilde{L}^4}{8\tilde{E}\tilde{I}^2}\Delta\tilde{I}$$

Substituting the appropriate values gives

$$\Delta y = 0.0225 + 0.0075 + 0.00375 + 0.005625 = 0.039375$$

Therefore, $y = 0.5625 \pm 0.039375$. In other words, y is between 0.523125 and 0.601
The validity of these estimates can be verified by substituting the extreme values f
variables into the equation to generate an exact minimum of

$$y_{min} = \frac{48(29.9)^4}{8(1.51 \times 10^8)0.0606} = 0.52407$$

and

$$y_{max} = \frac{52(30.1)^4}{8(1.49 \times 10^8)0.0594} = 0.60285$$

Thus, the first-order estimates are reasonably close to the exact values.

Equation (4.27) can be employed to define error propagation relationships for
mon mathematical operations. The results are summarized in Table 4.3. We will lea
derivation of these formulas as a homework exercise.

4.2.3 Stability and Condition

The *condition* of a mathematical problem relates to its sensitivity to changes in its
values. We say that a computation is *numerically unstable* if the uncertainty of the
values is grossly magnified by the numerical method.

These ideas can be studied using a first-order Taylor series

$$f(x) = f(\tilde{x}) + f'(\tilde{x})(x - \tilde{x})$$

This relationship can be employed to estimate the *relative error* of $f(x)$ as in

$$\frac{f(x) - f(\tilde{x})}{f(x)} \cong \frac{f'(\tilde{x})(x - \tilde{x})}{f(\tilde{x})}$$

The *relative error* of x is given by

$$\frac{x - \tilde{x}}{\tilde{x}}$$

TABLE 4.3 Estimated error bounds associated with common mathematical operations using inexact numbers $\tilde{u}$ and $\tilde{v}$.

Operation		Estimated Error
Addition	$\Delta(\tilde{u} + \tilde{v})$	$\Delta\tilde{u} + \Delta\tilde{v}$
Subtraction	$\Delta(\tilde{u} - \tilde{v})$	$\Delta\tilde{u} + \Delta\tilde{v}$
Multiplication	$\Delta(\tilde{u} \times \tilde{v})$	$\lvert\tilde{u}\rvert\Delta\tilde{v} + \lvert\tilde{v}\rvert\Delta\tilde{u}$
Division	$\Delta\left(\frac{\tilde{u}}{\tilde{v}}\right)$	$\frac{\lvert\tilde{u}\rvert\Delta\tilde{v} + \lvert\tilde{v}\rvert\Delta\tilde{u}}{\lvert\tilde{v}\rvert^2}$

A *condition number* can be defined as the ratio of these relative errors

$$\text{Condition number} = \frac{\tilde{x}f'(\tilde{x})}{f(\tilde{x})} \tag{4.28}$$

The condition number provides a measure of the extent to which an uncertainty in x is magnified by $f(x)$. A value of 1 tells us that the function's relative error is identical to the relative error in x. A value greater than 1 tells us that the relative error is amplified, whereas a value less than 1 tells us that it is attenuated. Functions with very large values are said to be *ill-conditioned*. Any combination of factors in Eq. (4.28) that increases the numerical value of the condition number will tend to magnify uncertainties in the computation of $f(x)$.

EXAMPLE 4.7 Condition Number

Problem Statement. Compute and interpret the condition number for

$$f(x) = \tan x \qquad \text{for } \tilde{x} = \frac{\pi}{2} + 0.1\left(\frac{\pi}{2}\right)$$

$$f(x) = \tan x \qquad \text{for } \tilde{x} = \frac{\pi}{2} + 0.01\left(\frac{\pi}{2}\right)$$

Solution. The condition number is computed as

$$\text{Condition number} = \frac{\tilde{x}(1/\cos^2 x)}{\tan \tilde{x}}$$

For $\tilde{x} = \pi/2 + 0.1(\pi/2)$,

$$\text{Condition number} = \frac{1.7279(40.86)}{-6.314} = -11.2$$

Thus, the function is ill-conditioned. For $\tilde{x} = \pi/2 + 0.01(\pi/2)$, the situation is even worse:

$$\text{Condition number} = \frac{1.5865(4053)}{-63.66} = -101$$

For this case, the major cause of ill conditioning appears to be the derivative. This makes sense because in the vicinity of $\pi/2$, the tangent approaches both positive and negative infinity.

4.3 TOTAL NUMERICAL ERROR

The *total numerical error* is the summation of the truncation and round-off errors. In general, the only way to minimize round-off errors is to increase the number of significant figures of the computer. Further, we have noted that round-off error will *increase* due to subtractive cancellation or due to an increase in the number of computations in an analysis. In contrast, Example 4.4 demonstrated that the truncation error can be reduced by decreasing the step size. Because a decrease in step size can lead to subtractive cancellation or to an increase in computations, the truncation errors are *decreased* as the round-off errors are *increased*. Therefore, we are faced by the following dilemma: The strategy for decreasing

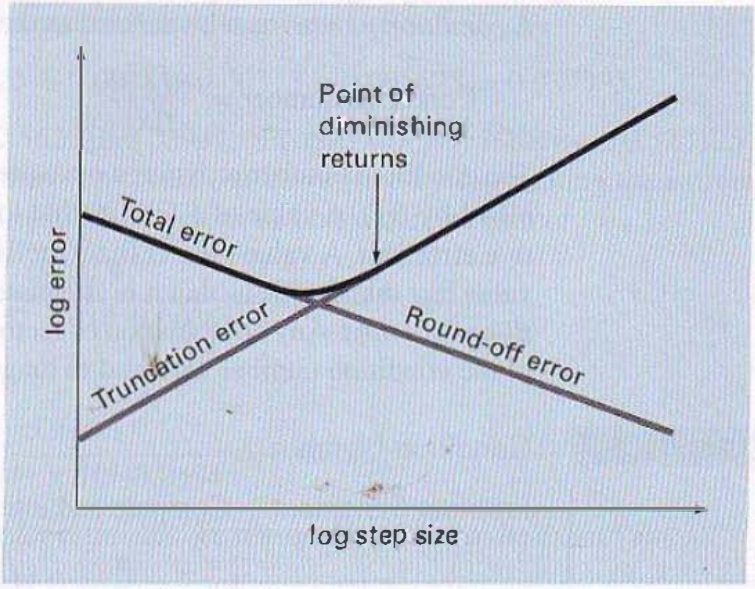


FIGURE 4.8
A graphical depiction of the trade-off between round-off and truncation error that sometin comes into play in the course of a numerical method. The point of diminishing returns is s where round-off error begins to negate the benefits of step-size reduction.

one component of the total error leads to an increase of the other component. In a tation, we could conceivably decrease the step size to minimize truncation errors discover that in doing so, the round-off error begins to dominate the solution and t error grows! Thus, our remedy becomes our problem (Fig. 4.8). One challenge that is to determine an appropriate step size for a particular computation. We would choose a large step size in order to decrease the amount of calculations and ro errors without incurring the penalty of a large truncation error. If the total error is as in Fig. 4.8, the challenge is to identify the point of diminishing returns where ro error begins to negate the benefits of step-size reduction.

In actual cases, however, such situations are relatively uncommon because mo puters carry enough significant figures that round-off errors do not predominate. theless, they sometimes do occur and suggest a sort of "numerical uncertainty pr that places an absolute limit on the accuracy that may be obtained using certain co ized numerical methods.

4.3.1 Control of Numerical Errors

For most practical cases, we do not know the exact error associated with numerica ods. The exception, of course, is when we have obtained the exact solution that ma numerical approximations unnecessary. Therefore, for most engineering applicati must settle for some estimate of the error in our calculations.

There are no systematic and general approaches to evaluating numerical error problems. In many cases error estimates are based on the experience and judgmen engineer.

Although error analysis is to a certain extent an art, there are several practical programming guidelines we can suggest. First and foremost, avoid subtracting two nearly equal numbers. Loss of significance almost always occurs when this is done. Sometimes you can rearrange or reformulate the problem to avoid subtractive cancellation. If this is not possible, you may want to use extended-precision arithmetic. Furthermore, when adding and subtracting numbers, it is best to sort the numbers and work with the smallest numbers first. This avoids loss of significance.

Beyond these computational hints, one can attempt to predict total numerical errors using theoretical formulations. The Taylor series is our primary tool for analysis of both truncation and round-off errors. Several examples have been presented in this chapter. Prediction of total numerical error is very complicated for even moderately sized problems and tends to be pessimistic. Therefore, it is usually attempted for only small-scale tasks.

The tendency is to push forward with the numerical computations and try to estimate the accuracy of your results. This can sometimes be done by seeing if the results satisfy some condition or equation as a check. Or it may be possible to substitute the results back into the original equation to check that it is actually satisfied.

Finally you should be prepared to perform numerical experiments to increase your awareness of computational errors and possible ill-conditioned problems. Such experiments may involve repeating the computations with a different step size or method and comparing the results. We may employ sensitivity analysis to see how our solution changes when we change model parameters or input values. We may want to try different numerical algorithms that have different theoretical foundations, are based on different computational strategies, or have different convergence properties and stability characteristics.

When the results of numerical computations are extremely critical and may involve loss of human life or have severe economic ramifications, it is appropriate to take special precautions. This may involve the use of two or more independent groups to solve the same problem so that their results can be compared.

The roles of errors will be a topic of concern and analysis in all sections of this book. We will leave these investigations to specific sections.

4.4 BLUNDERS, FORMULATION ERRORS, AND DATA UNCERTAINTY

Although the following sources of error are not directly connected with most of the numerical methods in this book, they can sometimes have great impact on the success of a modeling effort. Thus, they must always be kept in mind when applying numerical techniques in the context of real-world problems.

4.4.1 Blunders

We are all familiar with gross errors, or blunders. In the early years of computers, erroneous numerical results could sometimes be attributed to malfunctions of the computer itself. Today, this source of error is highly unlikely, and most blunders must be attributed to human imperfection.

Blunders can occur at any stage of the mathematical modeling process and can contribute to all the other components of error. They can be avoided only by sound knowledge

OBLEMS

ving infinite series can be used to approximate e^x:

$$+\frac{x^2}{2}+\frac{x^3}{3!}+\cdots+\frac{x^n}{n!}$$

this Maclaurin series expansion is a special case of series expansion [(Eq. (4.7)] with $x_i = 0$ and $h = x$. ylor series to estimate $f(x) = e^{-x}$ at $x_{i+1} = 1$ for mploy the zero-, first-, second-, and third-order ver- ompute the $|\varepsilon_t|$ for each case.

urin series expansion for cos x is

$$\frac{x^2}{2}+\frac{x^4}{4!}-\frac{x^6}{6!}+\frac{x^8}{8!}-\cdots$$

ne simplest version, cos $x = 1$, add terms one at a e cos$(\pi/3)$. After each new term is added, compute proximate percent relative errors. Use your pocket termine the true value. Add terms until the absolute proximate error estimate falls below an error crite- g to two significant figures.

e same computation as in Prob. 4.2, but use the s expansion for the sin x to estimate sin$(\pi/3)$.

$$\frac{x^3}{3!}+\frac{x^5}{5!}-\frac{x^7}{7!}+\cdots$$

through third-order Taylor series expansions to

$$^3-6x^2+7x-88$$

int at $x = 1$. Compute the true percent relative error roximation.

hrough fourth-order Taylor series expansions to pre- $f(x) = \ln x$ using a base point at $x = 1$. Compute t relative error ε_t for each approximation. Discuss the results.

d and backward difference approximations of $O(h)$ difference approximation of $O(h^2)$ to estimate the of the function examined in Prob. 4.4. Evaluate the $= 2$ using a step size of $h = 0.2$. Compare your true value of the derivative. Interpret your results the remainder term of the Taylor series expansion.

ered difference approximation of $O(h^2)$ to estimate vative of the function examined in Prob. 4.4. Per- uation at $x = 2$ using step sizes of $h = 0.25$ and e your estimates with the true value of the second rpret your results on the basis of the remainder term eries expansion.

4.8 Recall that the velocity of the falling parachutist can be computed by [Eq. (1.10)],

$$v(t) = \frac{gm}{c}\left(1 - e^{-(c/m)t}\right)$$

Use a first-order error analysis to estimate the error of v at $t = 6$, if $g = 9.8$ and $m = 50$ but $c = 12.5 \pm 1.5$.

4.9 Repeat Prob. 4.8 with $g = 9.8$, $t = 6$, $c = 12.5 \pm 1.5$, and $m = 50 \pm 2$.

4.10 The Stefan-Boltzmann law can be employed to estimate the rate of radiation of energy H from a surface, as in

$$H = A\varepsilon\sigma T^4$$

where H is in watts, A = the surface area (m^2), ε = the emissivity that characterizes the emitting properties of the surface (dimensionless), σ = a universal constant called the Stefan-Boltzmann constant ($= 5.67 \times 10^{-8}$ W m^{-2} K^{-4}), and T = absolute temperature (K). Determine the error of H for a steel plate with $A = 0.15$ m^2, $\varepsilon = 0.90$, and $T = 650 \pm 20$. Compare your results with the exact error. Repeat the computation but with $T = 650 \pm 40$. Interpret your results.

4.11 Repeat Prob. 4.10 but for a copper sphere with radius $= 0.15 \pm 0.01$ m, $\varepsilon = 0.90 \pm 0.05$, and $T = 550 \pm 20$.

4.12 Evaluate and interpret the condition numbers for

(a) $f(x) = \sqrt{|x-1|}+1$ for $x = 1.00001$

(b) $f(x) = e^{-x}$ for $x = 10$

(c) $f(x) = \sqrt{x^2+1}-x$ for $x = 300$

(d) $f(x) = \dfrac{e^{-x}-1}{x}$ for $x = 0.001$

(e) $f(x) = \dfrac{\sin x}{1+\cos x}$ for $x = 1.0001\pi$

4.13 Employing ideas from Sec. 4.2, derive the relationships from Table 4.3.

4.14 Prove that Eq. (4.4) is exact for all values of x if $f(x) = ax^2 + bx + c$.

4.15 Manning's formula for a rectangular channel can be written as

$$Q = \frac{1}{n}\frac{(BH)^{5/3}}{(B+2H)^{2/3}}\sqrt{S}$$

where Q = flow (m^3/s), n = a roughness coefficient, B = width (m), H = depth (m), and S = slope. You are applying this formula to a stream where you know that the width = 20 m and the depth = 0.3 m. Unfortunately, you know the roughness and the slope to only a $\pm$ 10% precision. That is, you know that the roughness is about 0.03 with a range from 0.027 to 0.033 and the slope is 0.0003 with a range from 0.00027 to 0.00033. Use a first-order

error analysis to determine the sensitivity of the flow prediction to each of these two factors. Which one should you attempt to measure with more precision?

4.16 If $|x| < 1$, it is known that

$$\frac{1}{1-x} = 1 + x + x^2 + x^3 + \cdots$$

Repeat Prob. 4.2 for this series for $x = 0.1$.

4.17 A missile leaves the ground with an initial velocity v_0 forming an angle ϕ_0 with the vertical as shown in Fig. P4.17. The maximum desired altitude is αR where R is the radius of the The laws of mechanics can be used to show that

$$\sin\phi_0 = (1+\alpha)\sqrt{1 - \frac{\alpha}{1+\alpha}\left(\frac{v_e}{v_0}\right)^2}$$

where v_e = the escape velocity of the missile. It is desired t the missile and reach the design maximum altitude within an racy of ±2%. Determine the range of values for ϕ_0 if $v_e/v_0 =$ $\alpha = 0.25$.

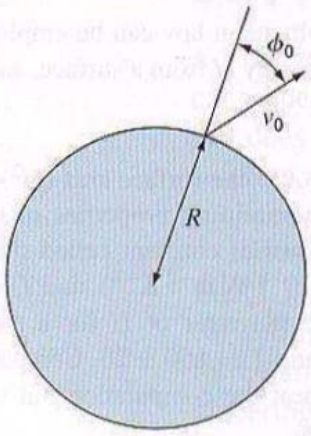


Figure P4.17

4.18 To calculate a planet's space coordinates, we have to solv function

$$f(x) = x - 1 - 0.5\sin x$$

Let the base point be $a = x_i = \pi/2$ on the interval $[0, \pi]$. D mine the highest-order Taylor series expansion resulting in a r mum error of 0.015 on the specified interval. The error is equ the absolute value of the difference between the given functio the specific Taylor series expansion. (Hint: Solve graphically.

4.19 Consider the function $f(x) = x^3 - 2x + 4$ on the int $[-2,2]$ with $h = 0.25$. Use the forward, backward, and cen finite difference approximations for the first and second deriva so as to graphically illustrate which approximation is most rate. Graph all three first derivative finite difference approx tions along with the theoretical, and do the same for the se derivative as well.

TABLE PT1.2 Summary of important information presented in Part One.

Error Definitions

True error: $E_t = \text{true value} - \text{approximation}$

True percent relative error: $\varepsilon_t = \dfrac{\text{true value} - \text{approximation}}{\text{true value}} 100\%$

Approximate percent relative error: $\varepsilon_a = \dfrac{\text{present approximation} - \text{previous approximation}}{\text{present approximation}} 100\%$

Stopping criterion: Terminate computation when $\varepsilon_a < \varepsilon_s$ where ε_s is the desired percent relative error

Taylor Series

Taylor series expansion:

$$f(x_{i+1}) = f(x_i) + f'(x_i)h + \frac{f''(x_i)}{2!}h^2 + \frac{f'''(x_i)}{3!}h^3 + \cdots + \frac{f^{(n)}(x_i)}{n!}h^n + R_n$$

where

Remainder:

$$R_n = \frac{f^{(n+1)}(\xi)}{(n+1)!}h^{n+1}$$

or

$$R_n = O(h^{n+1})$$

Numerical Differentiation

First forward finite divided difference:

$$f'(x_i) = \frac{f(x_{i+1}) - f(x_i)}{h} + O(h)$$

(Other divided differences are summarized in Chaps. 4 and 23.)

Error Propagation

For n independent variables $x_1, x_2, \ldots, x_n$ having errors $\Delta\tilde{x}_1, \Delta\tilde{x}_2, \ldots, \Delta\tilde{x}_n$, the error in the function f can be estimated via

$$\Delta f = \left|\frac{\partial f}{\partial x_1}\right| \Delta\tilde{x}_1 + \left|\frac{\partial f}{\partial x_2}\right| \Delta\tilde{x}_2 + \cdots + \left|\frac{\partial f}{\partial x_n}\right| \Delta\tilde{x}_n$$

Finally, although we hope that our book serves you well, it is always good to consult other sources when trying to master a new subject. Burden and Faires (1993); Ralston and Rabinowitz (1978); Hoffman (1992); and Carnahan, Luther, and Wilkes (1969) provide comprehensive discussions of most numerical methods, including some advanced methods that are beyond our scope. Other enjoyable books on the subject are Gerald and Wheatley (1989); Rice (1983); and Cheney and Kincaid (1985). In addition, Press et al. (1992) include computer codes to implement a variety of methods.

CHAPTER 5

Bracketing Methods

This chapter on roots of equations deals with methods that exploit the fact that a fun typically changes sign in the vicinity of a root. These techniques are called *brack methods* because two initial guesses for the root are required. As the name implies, guesses must "bracket," or be on either side of, the root. The particular methods desc herein employ different strategies to systematically reduce the width of the bracket hence, home in on the correct answer.

As a prelude to these techniques, we will briefly discuss graphical methods for de ing functions and their roots. Beyond their utility for providing rough guesses, grap techniques are also useful for visualizing the properties of the functions and the beh of the various numerical methods.

5.1 GRAPHICAL METHODS

A simple method for obtaining an estimate of the root of the equation $f(x) = 0$ is to m plot of the function and observe where it crosses the x axis. This point, which repre the x value for which $f(x) = 0$, provides a rough approximation of the root.

EXAMPLE 5.1 The Graphical Approach

Problem Statement. Use the graphical approach to determine the drag coeffici needed for a parachutist of mass $m = 68.1$ kg to have a velocity of 40 m/s after free-f for time $t = 10$ s. *Note:* The acceleration due to gravity is 9.8 m/s^2.

Solution. This problem can be solved by determining the root of Eq. (PT2.4) usir parameters $t = 10$, $g = 9.8$, $v = 40$, and $m = 68.1$:

$$f(c) = \frac{9.8(68.1)}{c}\left(1 - e^{-(c/68.1)10}\right) - 40$$

or

$$f(c) = \frac{667.38}{c}\left(1 - e^{-0.146843c}\right) - 40 \qquad (E$$

Various values of c can be substituted into the right-hand side of this equation to co

c	f(c)
4	34.115
8	17.653
12	6.067
16	−2.269
20	−8.401

These points are plotted in Fig. 5.1. The resulting curve crosses the c axis between 12 and 16. Visual inspection of the plot provides a rough estimate of the root of 14.75. The validity of the graphical estimate can be checked by substituting it into Eq. (E5.1.1) to yield

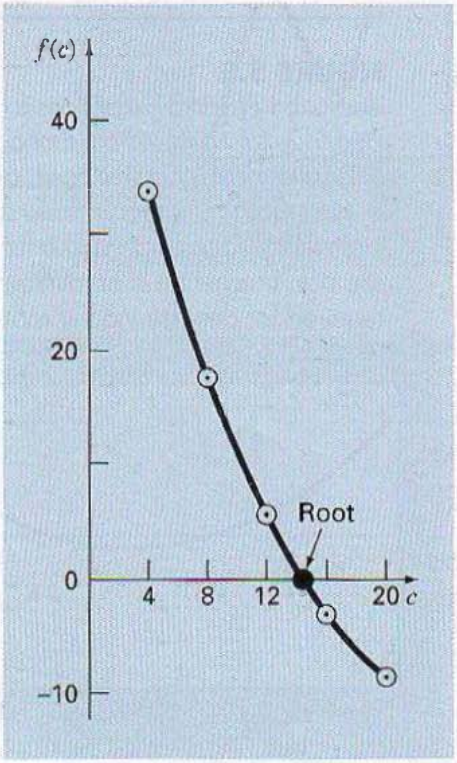
$$f(14.75) = \frac{667.38}{14.75}\left(1 - e^{-0.146843(14.75)}\right) - 40 = 0.059$$

which is close to zero. It can also be checked by substituting it into Eq. (PT2.4) along with the parameter values from this example to give

$$v = \frac{9.8(68.1)}{14.75}\left(1 - e^{-(14.75/68.1)10}\right) = 40.059$$

which is very close to the desired fall velocity of 40 m/s.

FIGURE 5.1
The graphical approach for determining the roots of an equation.



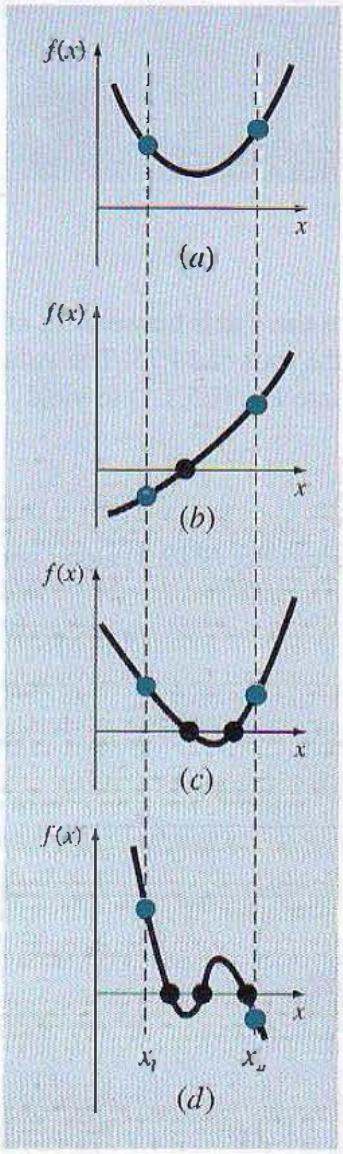


FIGURE 5.2
Illustration of a number of general ways that a root may occur in an interval prescribed by a lower bound x_l and an upper bound x_u. Parts (*a*) and (*c*) indicate that if both $f(x_l)$ and $f(x_u)$ have the same sign, either there will be no roots or there will be an even number of roots within the interval. Parts (*b*) and (*d*) indicate that if the function has different signs at the end points, there will be an odd number of roots in the interval.

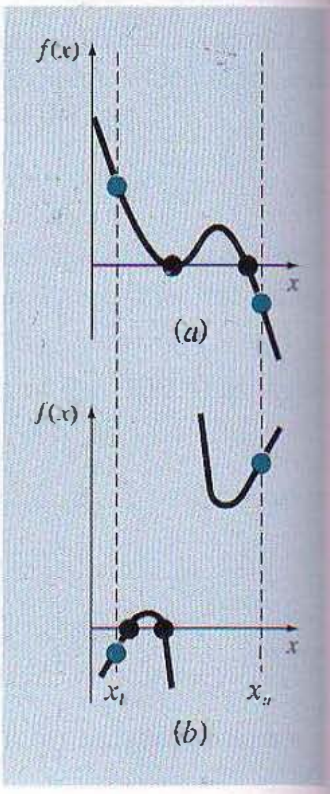
Graphical techniques are of limited practical value because they are not precise. However, graphical methods can be utilized to obtain rough estimates of roots. These estimates can be employed as starting guesses for numerical methods discussed in this and the next chapter.

Aside from providing rough estimates of the root, graphical interpretations are important tools for understanding the properties of the functions and anticipating the pitfalls of the numerical methods. For example, Fig. 5.2 shows a number of ways in which roots can occur (or be absent) in an interval prescribed by a lower bound x_l and an upper bound x_u. Figure 5.2*b* depicts the case where a single root is bracketed by negative and positive values of $f(x)$. However, Fig. 5.2*d*, where $f(x_l)$ and $f(x_u)$ are also on opposite sides of the x axis, shows three roots occurring within the interval. In general, if $f(x_l)$ and $f(x_u)$ have opposite signs, there are an odd number of roots in the interval. As indicated by Fig. 5.2*a* and *c*, if $f(x_l)$ and $f(x_u)$ have the same sign, there are either no roots or an even number of roots between the values.

Although these generalizations are usually true, there are cases where they do not hold. For example, functions that are tangential to the x axis (Fig. 5.3*a*) and discontinuous functions (Fig. 5.3*b*) can violate these principles. An example of a function that is tangential to the axis is the cubic equation $f(x) = (x - 2)(x - 2)(x - 4)$. Notice that $x = 2$ makes two terms in this polynomial equal to zero. Mathematically, $x = 2$ is called a *multiple root*. At the end of Chap. 6, we will present techniques that are expressly designed to locate multiple roots.

The existence of cases of the type depicted in Fig. 5.3 makes it difficult to develop general computer algorithms guaranteed to locate all the roots in an interval. However, when used in conjunction with graphical approaches, the methods described in the following

FIGURE 5.3
Illustration of some exceptions to the general cases depicted in Fig. 5.2. (*a*) Multiple root that occurs when the function is tangential to the x axis. For this case, although the end points are of opposite signs, there are an even number of axis intersections for the interval. (*b*) Discontinuous function where end points of opposite sign bracket an even number of roots. Special strategies are required for determining the roots for these cases.



sections are extremely useful for solving many roots of equations problems confronted routinely by engineers and applied mathematicians.

EXAMPLE 5.2 Use of Computer Graphics to Locate Roots

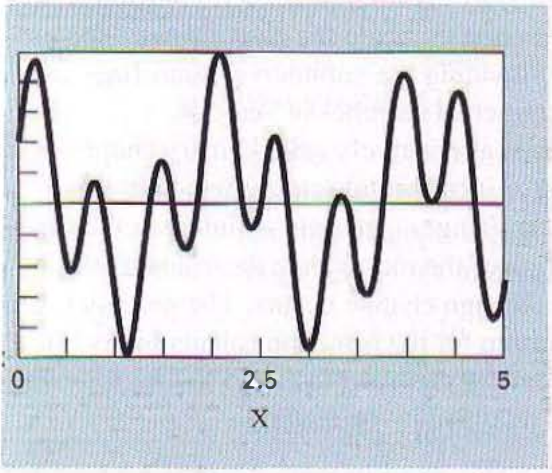
Problem Statement. Computer graphics can expedite and improve your efforts to locate roots of equations. The function

$$f(x) = \sin 10x + \cos 3x$$

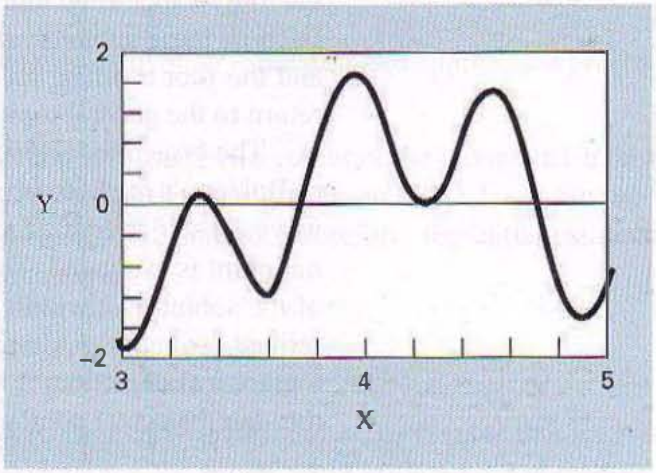
has several roots over the range $x = 0$ to $x = 5$. Use computer graphics to gain insight into the behavior of this function.

Solution. Packages such as Excel and MATLAB software can be used to generate plots. Figure 5.4*a* is a plot of $f(x)$ from $x = 0$ to $x = 5$. This plot suggests the presence of several roots, including a possible double root at about $x = 4.2$ where $f(x)$ appears to be tangent to

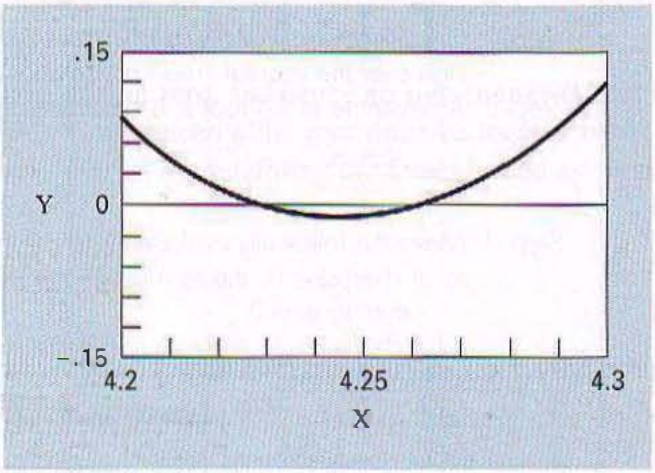
enlargement of $f(x) = \sin 10x + \cos 3x$ by the computer. Such interactive graphics
st to determine that two distinct roots exist between $x = 4.2$ and $x = 4.3$.



(*a*)



(*b*)



(*c*)

the x axis. A more detailed picture of the behavior of $f(x)$ is obtained by changing the p ting range from $x = 3$ to $x = 5$, as shown in Fig. 5.4*b*. Finally, in Fig. 5.4*c*, the vertical s is narrowed further to $f(x) = -0.15$ to $f(x) = 0.15$ and the horizontal scale is narrowe $x = 4.2$ to $x = 4.3$. This plot shows clearly that a double root does not exist in this reg and that in fact there are two distinct roots at about $x = 4.23$ and $x = 4.26$.

Computer graphics will have great utility in your studies of numerical methods. T capability will also find many other applications in your other classes and professio activities as well.

5.2 THE BISECTION METHOD

When applying the graphical technique in Example 5.1, you have observed (Fig. 5.1) $f(x)$ changed sign on opposite sides of the root. In general, if $f(x)$ is real and continuou the interval from x_l to x_u and $f(x_l)$ and $f(x_u)$ have opposite signs, that is,

$$f(x_l)f(x_u) < 0$$

then there is at least one real root between x_l and x_u.

Incremental search methods capitalize on this observation by locating an inte where the function changes sign. Then the location of the sign change (and consequen the root) is identified more precisely by dividing the interval into a number of subinterv Each of these subintervals is searched to locate the sign change. The process is repea and the root estimate refined by dividing the subintervals into finer increments. We return to the general topic of incremental searches in Sec. 5.4.

The *bisection method*, which is alternatively called binary chopping, interval halv or Bolzano's method, is one type of incremental search method in which the interval is ways divided in half. If a function changes sign over an interval, the function value at midpoint is evaluated. The location of the root is then determined as lying at the midp of the subinterval within which the sign change occurs. The process is repeated to ob refined estimates. A simple algorithm for the bisection calculation is listed in Fig. 5.5, a graphical depiction of the method is provided in Fig. 5.6. The following example g through the actual computations involved in the method.

FIGURE 5.5

Step 1: Choose lower x_l and upper x_u guesses for the root such that the function change sign over the interval. This can be checked by ensuring that $f(x_l)f(x_u) < 0$.

Step 2: An estimate of the root x_r is determined by

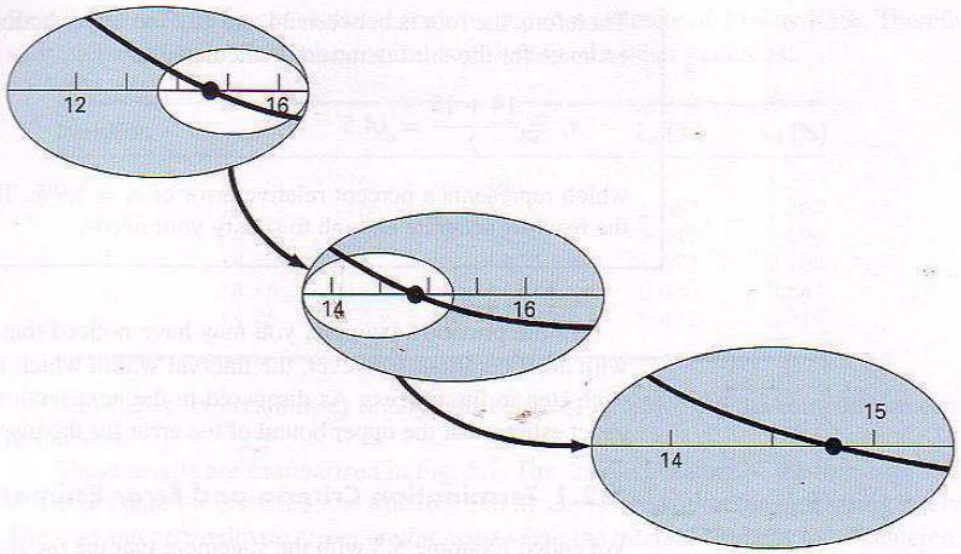
$$x_r = \frac{x_l + x_u}{2}$$

Step 3: Make the following evaluations to determine in which subinterval the root lies:

(a) If $f(x_l)f(x_r) < 0$, the root lies in the lower subinterval. Therefore, set $x_u = x_r$ an return to step 2.

(b) If $f(x_l)f(x_r) > 0$, the root lies in the upper subinterval. Therefore, set $x_l = x_r$ an return to step 2.

(c) If $f(x_l)f(x_r) = 0$, the root equals x_r; terminate the computation.



graphical interpretation of the bisection method. This plot shows the first three iterations of Example 5.3.

EXAMPLE 5.3 Bisection

Problem Statement. Use bisection to solve the same problem approached graphically in Example 5.1.

Solution. The first step in bisection is to guess two values of the unknown (in the present problem, c) that give values for $f(c)$ with different signs. From Fig. 5.1, we can see that the function changes sign between values of 12 and 16. Therefore, the initial estimate of the root x_r lies at the midpoint of the interval

$$x_r = \frac{12 + 16}{2} = 14$$

This estimate represents a true percent relative error of $\varepsilon_t = 5.3\%$ (note that the true value of the root is 14.7802). Next we compute the product of the function value at the lower bound and at the midpoint:

$$f(12)f(14) = 6.067(1.569) = 9.517$$

which is greater than zero, and hence no sign change occurs between the lower bound and the midpoint. Consequently, the root must be located between 14 and 16. Therefore, we create a new interval by redefining the lower bound as 14 and determining a revised root estimate as

$$x_r = \frac{14 + 16}{2} = 15$$

which represents a true percent error of $\varepsilon_t = 1.5\%$. The process can be repeated to obtain refined estimates. For example,

$$f(14)f(15) = 1.569(-0.425) = -0.666$$

Therefore, the root is between 14 and 15. The upper bound is redefined as 15, and the new estimate for the third iteration is calculated as

$$x_r = \frac{14 + 15}{2} = 14.5$$

which represents a percent relative error of $\varepsilon_t = 1.9\%$. The method can be repeated until the result is accurate enough to satisfy your needs.

In the previous example, you may have noticed that the true error does not decrease with each iteration. However, the interval within which the root is located is halved with each step in the process. As discussed in the next section, the interval width provides an exact estimate of the upper bound of the error for the bisection method.

5.2.1 Termination Criteria and Error Estimates

We ended Example 5.3 with the statement that the method could be continued to obtain a refined estimate of the root. We must now develop an objective criterion for deciding when to terminate the method.

An initial suggestion might be to end the calculation when the true error falls below some prespecified level. For instance, in Example 5.3, the relative error dropped from 18 to 1.9 percent during the course of the computation. We might decide that we should terminate when the error drops below, say, 0.1 percent. This strategy is flawed because the error estimates in the example were based on knowledge of the true root of the function. This would not be the case in an actual situation because there would be no point in using the method if we already knew the root.

Therefore, we require an error estimate that is not contingent on foreknowledge of the root. As developed previously in Sec. 3.3, an approximate percent relative error ε_a can be calculated, as in [recall Eq. (3.5)]

$$\varepsilon_a = \left| \frac{x_r^{\text{new}} - x_r^{\text{old}}}{x_r^{\text{new}}} \right| 100\%$$

where x_r^{new} is the root for the present iteration and x_r^{old} is the root from the previous iteration. The absolute value is used because we are usually concerned with the magnitude of ε_a rather than with its sign. When ε_a becomes less than a prespecified stopping criterion, the computation is terminated.

EXAMPLE 5.4 Error Estimates for Bisection

Problem Statement. Continue Example 5.3 until the approximate error falls below a stopping criterion of $\varepsilon_s = 0.5\%$. Use Eq. (5.2) to compute the errors.

Solution. The results of the first two iterations for Example 5.3 were 14 and 15. Substituting these values into Eq. (5.2) yields

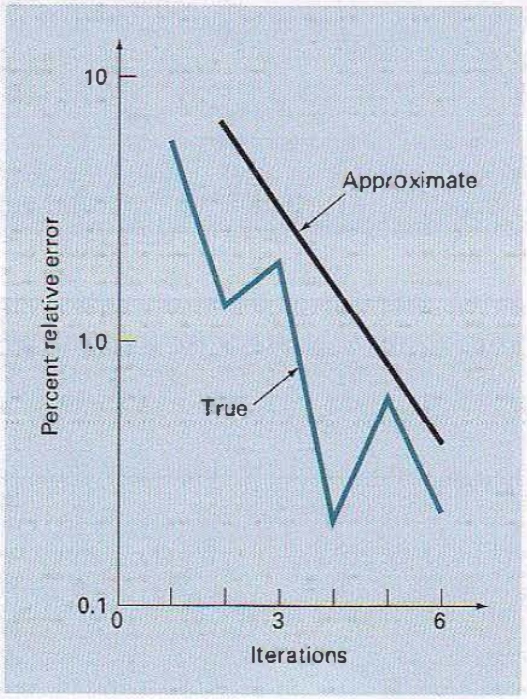
Recall that the true percent relative error for the root estimate of 15 was 1.5%. Therefore, ε_a is greater than ε_t. This behavior is manifested for the other iterations:

Iteration	x_l	x_u	x_r	ε_a (%)	ε_t (%)
1	12	16	14		5.279
2	14	16	15	6.667	1.487
3	14	15	14.5	3.448	1.896
4	14.5	15	14.75	1.695	0.204
5	14.75	15	14.875	0.840	0.641
6	14.75	14.875	14.8125	0.422	0.219

Thus, after six iterations ε_a finally falls below $\varepsilon_s = 0.5\%$, and the computation can be terminated.

These results are summarized in Fig. 5.7. The "ragged" nature of the true error is due to the fact that, for bisection, the true root can lie anywhere within the bracketing interval. The true and approximate errors are far apart when the interval happens to be centered on the true root. They are close when the true root falls at either end of the interval.

Although the approximate error does not provide an exact estimate of the true error, Fig. 5.7 suggests that ε_a captures the general downward trend of ε_t. In addition, the plot exhibits the extremely attractive characteristic that ε_a is always greater than ε_t. Thus, when



...bisection method. ...estimated errors are ...versus the number of ...

the root estimate is calculated. Previously calculated values are saved and merely reassigned as the bracket shrinks. Thus, $n + 1$ function evaluations are performed, rather than $2n$.

5.3 THE FALSE-POSITION METHOD

Although bisection is a perfectly valid technique for determining roots, its "brute-force" approach is relatively inefficient. False position is an alternative based on a graphical insight.

A shortcoming of the bisection method is that, in dividing the interval from x_l to x_u into equal halves, no account is taken of the magnitudes of $f(x_l)$ and $f(x_u)$. For example, if $f(x_l)$ is much closer to zero than $f(x_u)$, it is likely that the root is closer to x_l than to x_u (Fig. 5.11). An alternative method that exploits this graphical insight is to join $f(x_l)$ and $f(x_u)$ by a straight line. The intersection of this line with the x axis represents an improved estimate of the root. The fact that the replacement of the curve by a straight line gives a "false position" of the root is the origin of the name, *method of false position*, or in Latin, *regula falsi*. It is also called the *linear interpolation method*.

Using similar triangles (Fig. 5.12), the intersection of the straight line with the x axis can be estimated as

$$\frac{f(x_l)}{x_r - x_l} = \frac{f(x_u)}{x_r - x_u}$$

which can be solved for (see Box 5.1 for details).

$$x_r = x_u - \frac{f(x_u)(x_l - x_u)}{f(x_l) - f(x_u)}$$

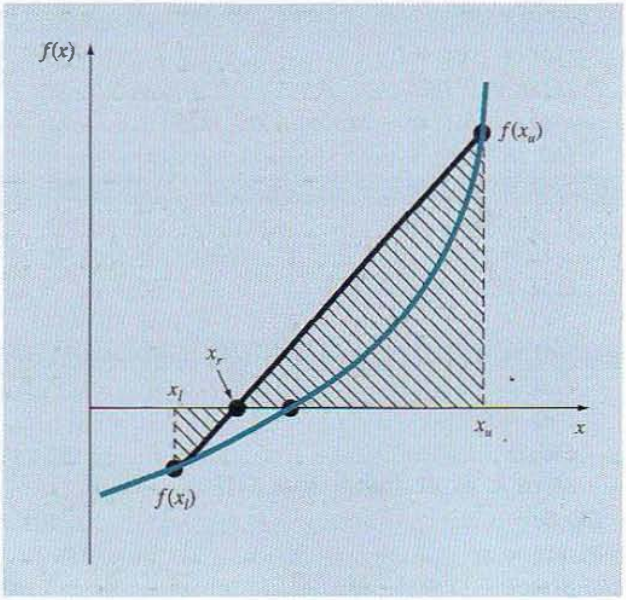


FIGURE 5.12
A graphical depiction of the method of false position. Similar triangles used to derive the formula for the method are shaded.

Box 5.1 Derivation of the Method of False Position

. (5.6) to yield

$$= f(x_u)(x_r - x_l)$$

rearrange:

$$_u)] = x_u f(x_l) - x_l f(x_u)$$

$f(x_u)$:

$$\frac{- x_l f(x_u)}{- f(x_l)} \qquad \text{(B5.1.1)}$$

f the method of false position. Note that it allows f the root x_r as a function of the lower and upper lt can be put in an alternative form by expanding

$$\frac{}{f(x_u)} - \frac{x_l f(x_u)}{f(x_l) - f(x_u)}$$

then adding and subtracting x_u on the right-hand side:

$$x_r = x_u + \frac{x_u f(x_l)}{f(x_l) - f(x_u)} - x_u - \frac{x_l f(x_u)}{f(x_l) - f(x_u)}$$

Collecting terms yields

$$x_r = x_u + \frac{x_u f(x_u)}{f(x_l) - f(x_u)} - \frac{x_l f(x_u)}{f(x_l) - f(x_u)}$$

or

$$x_r = x_u - \frac{f(x_u)(x_l - x_u)}{f(x_l) - f(x_u)}$$

which is the same as Eq. (5.7). We use this form because it involves one less function evaluation and one less multiplication than Eq. (B5.1.1). In addition, it is directly comparable with the secant method which will be discussed in Chap. 6.

This is the *false-position formula*. The value of x_r computed with Eq. (5.7) then replaces whichever of the two initial guesses, x_l or x_u, yields a function value with the same sign as $f(x_r)$. In this way, the values of x_l and x_u always bracket the true root. The process is repeated until the root is estimated adequately. The algorithm is identical to the one for bisection (Fig. 5.5) with the exception that Eq. (5.7) is used for step 2. In addition, the same stopping criterion [Eq. (5.2)] is used to terminate the computation.

XAMPLE 5.5 False Position

Problem Statement. Use the false-position method to determine the root of the same equation investigated in Example 5.1 [Eq. (E5.1.1)].

Solution. As in Example 5.3, initiate the computation with guesses of $x_l = 12$ and $x_u = 16$.

First iteration:

$$x_l = 12 \qquad f(x_l) = 6.0699$$
$$x_u = 16 \qquad f(x_u) = -2.2688$$
$$x_r = 16 - \frac{-2.2688(12 - 16)}{6.0669 - (-2.2688)} = 14.9113$$

which has a true relative error of 0.89 percent.

Second iteration:

$$f(x_l) f(x_r) = -1.5426$$

Therefore, the root lies in the first subinterval, and x_r becomes the upper limit for the next iteration, $x_u = 14.9113$:

$$x_l = 12 \qquad f(x_l) = 6.0699$$

$$x_u = 14.9113 \qquad f(x_u) = -0.2543$$

$$x_r = 14.9113 - \frac{-0.2543(12 - 14.9113)}{6.0669 - (-0.2543)} = 14.7942$$

which has true and approximate relative errors of 0.09 and 0.79 percent. Additional iterations can be performed to refine the estimate of the roots.

A feeling for the relative efficiency of the bisection and false-position methods can be appreciated by referring to Fig. 5.13, where we have plotted the true percent relative error for Examples 5.4 and 5.5. Note how the error for false position decreases much faster than for bisection because of the more efficient scheme for root location in the false-position method.

Recall in the bisection method that the interval between x_l and x_u grew smaller during the course of a computation. The interval, as defined by $\Delta x/2 = |x_u - x_l|/2$ for the first iteration, therefore provided a measure of the error for this approach. This is not the case

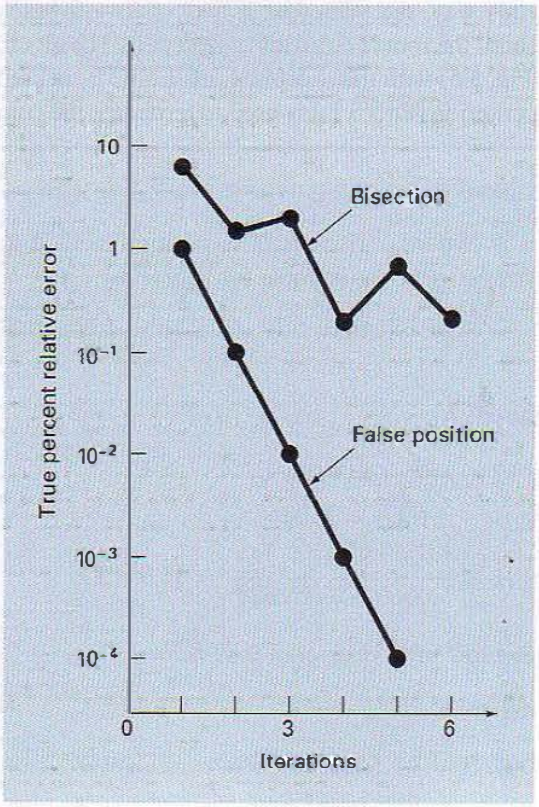


FIGURE 5.13
Comparison of the relative errors of the bisection and the false-position methods.

for the method of false position because one of the initial guesses may stay fixed throughout the computation as the other guess converges on the root. For instance, in Example 5.6 the lower guess x_l remained at 12 while x_u converged on the root. For such cases, the interval does not shrink but rather approaches a constant value.

Example 5.6 suggests that Eq. (5.2) represents a very conservative error criterion. In fact, Eq. (5.2) actually constitutes an approximation of the discrepancy of the previous iteration. This is because for a case such as Example 5.6, where the method is converging quickly (for example, the error is being reduced nearly an order of magnitude per iteration), the root for the present iteration x_r^{new} is a much better estimate of the true value than the result of the previous iteration x_r^{old}. Thus, the quantity in the numerator of Eq. (5.2) actually represents the discrepancy of the previous iteration. Consequently, we are assured that satisfaction of Eq. (5.2) ensures that the root will be known with greater accuracy than the prescribed tolerance. However, as described in the next section, there are cases where false position converges slowly. For these cases, Eq. (5.2) becomes unreliable, and an alternative stopping criterion must be developed.

5.3.1 Pitfalls of the False-Position Method

Although the false-position method would seem to always be the bracketing method of preference, there are cases where it performs poorly. In fact, as in the following example, there are certain cases where bisection yields superior results.

EXAMPLE 5.6 A Case Where Bisection Is Preferable to False Position

Problem Statement. Use bisection and false position to locate the root of

$$f(x) = x^{10} - 1$$

between $x = 0$ and 1.3.

Solution. Using bisection, the results can be summarized as

Iteration	x_l	x_u	x_r	ε_a (%)	ε_t (%)
1	0	1.3	0.65	100.0	35
2	0.65	1.3	0.975	33.3	2.5
3	0.975	1.3	1.1375	14.3	13.8
4	0.975	1.1375	1.05625	7.7	5.6
5	0.975	1.05625	1.015625	4.0	1.6

Thus, after five iterations, the true error is reduced to less than 2 percent. For false position, a very different outcome is obtained:

Iteration	x_l	x_u	x_r	ε_a (%)	ε_t (%)
1	0	1.3	0.09430		90.6
2	0.09430	1.3	0.18176	48.1	81.8
3	0.18176	1.3	0.26287	30.9	73.7
4	0.26287	1.3	0.33811	22.3	66.2
5	0.33811	1.3	0.40788	17.1	59.2

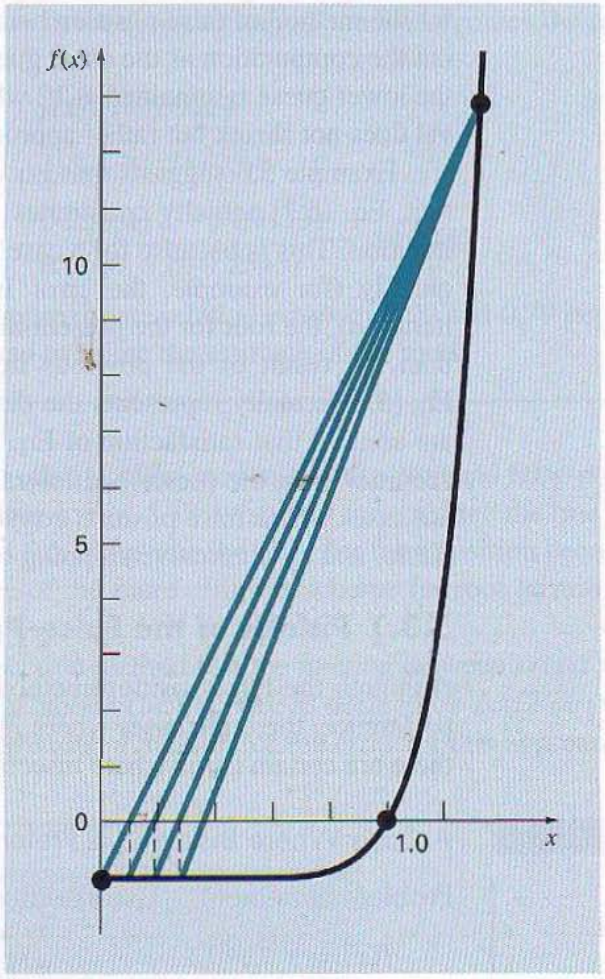


FIGURE 5.14
Plot of $f(x) = x^{10} - 1$, illustrating slow convergence of the false-position method.

After five iterations, the true error has only been reduced to about 59 percent. In ad
tion, note that $\varepsilon_a < \varepsilon_t$. Thus, the approximate error is misleading. Insight into these resu
can be gained by examining a plot of the function. As in Fig. 5.14, the curve violates t
premise upon which false position was based—that is, if $f(x_l)$ is much closer to zero th
$f(x_u)$, then the root is closer to x_l than to x_u (recall Fig. 5.12). Because of the shape of t
present function, the opposite is true.

The forgoing example illustrates that blanket generalizations regarding root-locati
methods are usually not possible. Although a method such as false position is often sup
rior to bisection, there are invariably cases that violate this general conclusion. Therefo
in addition to using Eq. (5.2), the results should always be checked by substituting the r
estimate into the original equation and determining whether the result is close to zero. Su
a check should be incorporated into all computer programs for root location.

The example also illustrates a major weakness of the false-position method: its o
sidedness. That is, as iterations are proceeding, one of the bracketing points will tend

stay fixed. This can lead to poor convergence, particularly for functions with significant curvature. The following section provides a remedy.

5.3.2 Modified False Position

One way to mitigate the "one-sided" nature of false position is to have the algorithm detect when one of the bounds is stuck. If this occurs, the function value at the stagnant bound can be divided in half. This is called the *modified false-position method*.

The algorithm in Fig. 5.15 implements this strategy. Notice how counters are used to determine when one of the bounds stays fixed for two iterations. If this occurs, the function value at this stagnant bound is halved.

The effectiveness of this algorithm can be demonstrated by applying it to Example 5.6. If a stopping criterion of 0.01% is used, the bisection and standard false-position methods

```
FUNCTION ModFalsePos(xl, xu, es, imax, xr, iter, ea)
  iter = 0
  fl = f(xl)
  fu = f(xu)
  DO
    xrold = xr
    xr = xu - fu * (xl - xu) / (fl - fu)
    fr = f(xr)
    iter = iter + 1
    IF xr <> 0 THEN
      ea = Abs((xr - xrold) / xr) * 100
    END IF
    test = fl * fr
    IF test < 0 THEN
      xu = xr
      fu = f(xu)
      iu = 0
      il = il + 1
      IF il >= 2 THEN fl = fl / 2
    ELSE IF test > 0 THEN
      xl = xr
      fl = f(xl)
      il = 0
      iu = iu + 1
      IF iu >= 2 THEN fu = fu / 2
    ELSE
      ea = 0
    END IF
    IF ea < es OR iter >= imax THEN EXIT
  END DO
  ModFalsePos = xr
END ModFalsePos
```

would converge in 14 and 39 iterations, respectively. In contrast, the modified position method would converge in 12 iterations. Thus, for this example, it is som more efficient than bisection and is vastly superior to the unmodified false-p method.

5.4 INCREMENTAL SEARCHES AND DETERMINING INITIAL GUESSES

Besides checking an individual answer, you must determine whether all possible root been located. As mentioned previously, a plot of the function is usually very useful i ing you in this task. Another option is to incorporate an incremental search at the beg of the computer program. This consists of starting at one end of the region of intere then making function evaluations at small increments across the region. When the fu changes sign, it is assumed that a root falls within the increment. The x values at t ginning and the end of the increment can then serve as the initial guesses for one bracketing techniques described in this chapter.

A potential problem with an incremental search is the choice of the increment l If the length is too small, the search can be very time consuming. On the other hand length is too great, there is a possibility that closely spaced roots might be (Fig. 5.16). The problem is compounded by the possible existence of multiple roots. tial remedy for such cases is to compute the first derivative of the function $f'(x)$ beginning and the end of each interval. If the derivative changes sign, it suggests minimum or maximum may have occurred and that the interval should be examine closely for the existence of a possible root.

Although such modifications or the employment of a very fine increment can all the problem, it should be clear that brute-force methods such as incremental search foolproof. You would be wise to supplement such automatic techniques with any information that provides insight into the location of the roots. Such information found in plotting and in understanding the physical problem from which the eq originated.

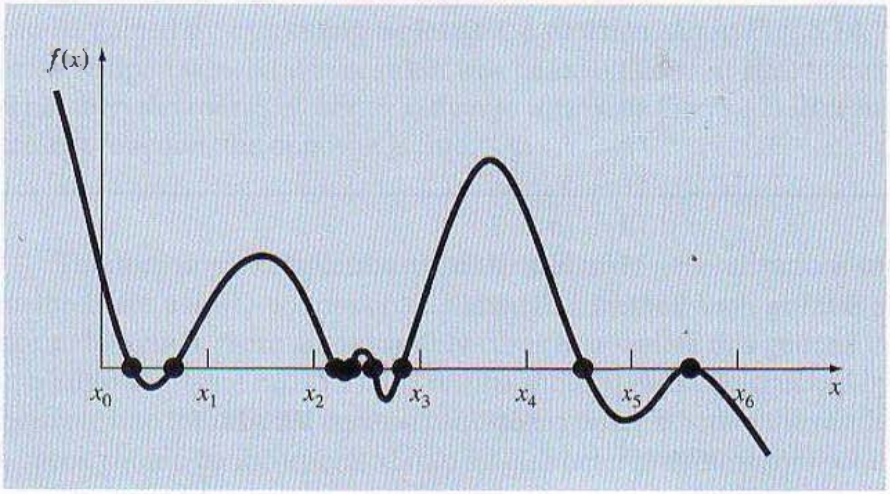


FIGURE 5.16
Cases where roots could be missed because the increment length of the search procedure is too large. Note that the last root on the right is multiple and would be missed regardless of increment length.

OBLEMS

he real roots of $f(x) = -0.5x^2 + 2.5x + 4.5$:

uadratic formula.

: iterations of the bisection method to determine root. Employ initial guesses of $x_l = 5$ and $x_u = 10$. e estimated error ε_a and the true error ε_t after each

he real root of $f(x) = 5x^3 - 5x^2 + 6x - 2$:

tion to locate the root. Employ initial guesses of $x_u = 1$ and iterate until the estimated error ε_a falls el of $\varepsilon_s = 10\%$.

the real root of $f(x) = -25 + 82x - 90x^2 + .7x^5$:

tion to determine the root to $\varepsilon_s = 10\%$. Employ ini- of $x_l = 0.5$ and $x_u = 1.0$.

: same computation as in (b) but use the false- thod and $\varepsilon_s = 0.2\ \%$.

line the roots of $f(x) = -12 - 21x + 18x^2 -$ ally. In addition, determine the first root of the) bisection, and (c) false position. For (b) and (c) es of $x_l = -1$ and $x_u = 0$, and a stopping criterion

first nontrivial root of $\sin x = x^3$, where x is in graphical technique and bisection with the initial 5 to 1. Perform the computation until ε_a is less than perform an error check by substituting your final original equation.

he positive real root of $\ln(x^4) = 0.7$ (a) graphi- hree iterations of the bisection method, with initial 0.5 and $x_u = 2$, and (c) using three iterations of the ethod, with the same initial guesses as in (b).

he real root of $f(x) = (0.8 - 0.3x)/x$:

iterations of the false-position method and initial and 3. Compute the approximate error ε_a and the after each iteration. Is there a problem with the

sitive square root of 18 using the false-position n $\varepsilon_s = 0.5\%$. Employ initial guesses of $x_l = 4$ and

allest positive root of the function (x is in radians) using the false-position method. To locate the the root lies, first plot this function for values of x 5. Perform the computation until ε_a falls below $\varepsilon_s = 1\%$. Check your final answer by substituting it into the original function.

5.10 Find the positive real root of $f(x) = x^4 - 8x^3 - 35x^2 + 450x - 1001$ using the false-position method. Use initial guesses of $x_l = 4.5$ and $x_u = 6$ and performs five iterations. Compute both the true and approximate errors based on the fact that the root is 5.60979. Use a plot to explain your results and perform the computation to within $\varepsilon_s = 1.0\%$.

5.11 Determine the real root of $x^{3.5} = 80$: (a) analytically, and (b) with the false-position method to within $\varepsilon_s = 2.5\%$. Use initial guesses of 2.0 and 5.0.

5.12 Given

$$f(x) = -2x^6 - 1.5x^4 + 10x + 2$$

Use bisection to determine the *maximum* of this function. Employ initial guesses of $x_l = 0$ and $x_u = 1$, and perform iterations until the approximate relative error falls below 5%.

5.13 The velocity v of a falling parachutist is given by

$$v = \frac{gm}{c}\left(1 - e^{-(c/m)t}\right)$$

where $g = 9.8\ \text{m/s}^2$. For a parachutist with a drag coefficient $c = 15$ kg/s, compute the mass m so that the velocity is $v = 35$ m/s at $t = 9$ s. Use the false-position method to determine m to a level of $\varepsilon_s = 0.1\%$.

5.14 A beam is loaded as shown in Fig. P5.14. Use the bisection method to solve for the position inside the beam where there is no moment.

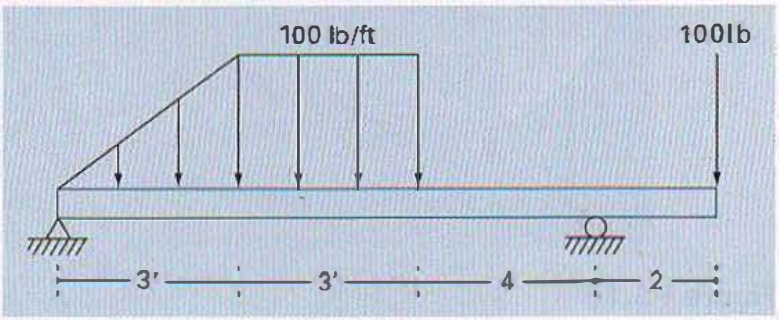


Figure P5.14

5.15 Water is flowing in a trapezoidal channel at a rate of $Q = 20\ \text{m}^3/\text{s}$. The critical depth y for such a channel must satisfy the equation

$$0 = 1 - \frac{Q^2}{gA_c^3}B$$

where $g = 9.81$ m/s^2, A_c = the cross-sectional area (m^2), and B = the width of the channel at the surface (m). For this case, the width and the cross-sectional area can be related to depth y by

$$B = 3 + y \qquad \text{and} \qquad A_c = 3y + \frac{y^2}{2}$$

Solve for the critical depth using (a) the graphical method, (b) bisection, and (c) false position. For (b) and (c) use initial guesses of $x_l = 0.5$ and $x_u = 2.5$, and iterate until the approximate error falls below 1% or the number of iterations exceeds 10. Discuss your results.

5.16 You are designing a spherical tank (Fig. P5.16) to hold water for a small village in a developing country. The volume of liquid it can hold can be computed as

$$V = \pi h^2 \frac{[3R - h]}{3}$$

where V = volume [m^3], h = depth of water in tank [m], and R = the tank radius [m].

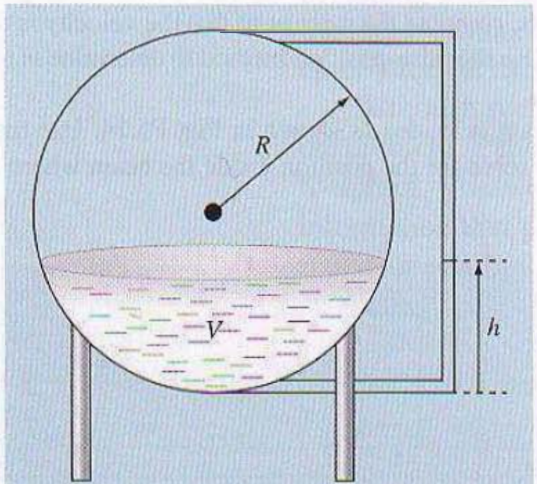


Figure P5.16

If $R = 3$ m, to what depth must the tank be filled so that it holds 30 m^3? Use three iterations of the false-position method to determine your answer. Determine the approximate relative error after each iteration.

5.17 The saturation concentration of dissolved oxygen in freshwater can be calculated with the equation (APHA, 1992)

$$\ln o_{sf} = -139.34411 + \frac{1.575701 \times 10^5}{T_a} - \frac{6.642308}{T_a^2} + \frac{1.243800 \times 10^{10}}{T_a^3} - \frac{8.621949 \times 10^{11}}{T_a^4}$$

where o_{sf} = the saturation concentration of dissolved o freshwater at 1 atm (mg/L) and T_a = absolute temperat Remember that $T_a = T + 273.15$, where T = temperatu According to this equation, saturation decreases with in temperature. For typical natural waters in temperate clim equation can be used to determine that oxygen conc ranges from 14.621 mg/L at 0°C to 6.413 mg/L at 40°C. value of oxygen concentration, this formula and the method can be used to solve for temperature in °C.

(a) If the initial guesses are set as 0 and 40°C, how many iterations would be required to determine temperatu absolute error of 0.05°C?

(b) Develop and test a bisection program to determine T a tion of a given oxygen concentration to a prespecified error as in (a). Given initial guesses of 0 and 40°C, program for an absolute error = 0.05°C and the f cases: o_{sf} = 8, 10 and 12 mg/L. Check your results.

5.18 Integrate the algorithm outlined in Fig. 5.10 into a user-friendly bisection subprogram. Among other things:

(a) Place documentation statements throughout the subpr identify what each section is intended to accomplish.

(b) Label the input and output.

(c) Add an answer check that substitutes the root estimate original function to verify whether the final result is clos

(d) Test the subprogram by duplicating the computatio Examples 5.3 and 5.4.

5.19 Develop a subprogram for the bisection method th mizes function evaluations based on the pseudocode from Determine the number of function evaluations (n) per to tions. Test the program by duplicating Example 5.6.

5.20 Develop a user-friendly program for the false method. The structure of your program should be simil bisection algorithm outlined in Fig. 5.10. Test the pro duplicating Example 5.5.

5.21 Develop a subprogram for the false-position met minimizes function evaluations in a fashion similar to F Determine the number of function evaluations (n) per to tions. Test the program by duplicating Example 5.6.

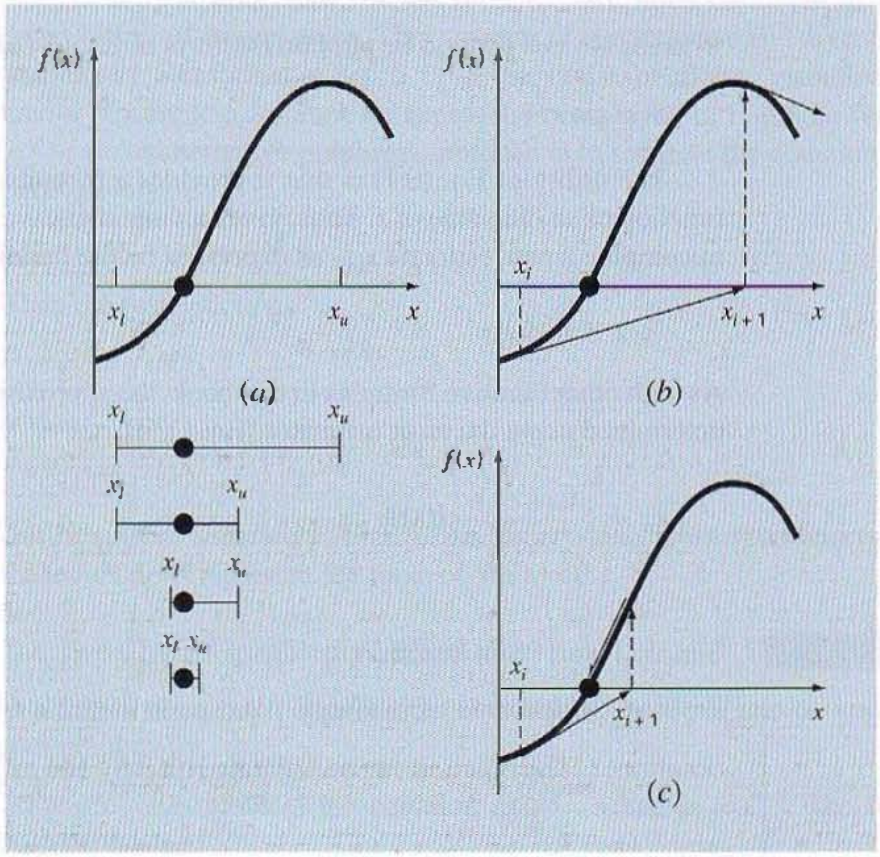
5.22 Develop a user-friendly subprogram for the modifi position method based on Fig. 5.15. Test the program by d ing the root of the function described in Example 5.6. R number of runs until the true percent relative error fal 0.01%. Plot the true and approximate percent relative erro number of iterations on semilog paper. Interpret your resul

TER 6

Open Methods

For the bracketing methods in the previous chapter, the root is located within an interval prescribed by a lower and an upper bound. Repeated application of these methods always results in closer estimates of the true value of the root. Such methods are said to be *convergent* because they move closer to the truth as the computation progresses (Fig. 6.1*a*).

In contrast, the *open methods* described in this chapter are based on formulas that require only a single starting value of x or two starting values that do not necessarily bracket



n of the
nce between
nd (b) and
or root
ch is the
ne root is
he interval
d x_u. In
an method
(c), a
roject from
itive fashion.
an either (b)
erge rapidly,
value of the

the root. As such, they sometimes *diverge* or move away from the tru computation progresses (Fig. 6.1*b*). However, when the open methods conve they usually do so much more quickly than the bracketing methods. We will cussion of open techniques with a simple version that is useful for illustratin form and also for demonstrating the concept of convergence.

6.1 SIMPLE FIXED-POINT ITERATION

As mentioned above, open methods employ a formula to predict the root. S can be developed for simple *fixed-point iteration* (or, as it is also called, one- or successive substitution) by rearranging the function $f(x) = 0$ so that x is o side of the equation:

$$x = g(x)$$

This transformation can be accomplished either by algebraic manipulation adding x to both sides of the original equation. For example,

$$x^2 - 2x + 3 = 0$$

can be simply manipulated to yield

$$x = \frac{x^2 + 3}{2}$$

whereas $\sin x = 0$ could be put into the form of Eq. (6.1) by adding x to both

$$x = \sin x + x$$

The utility of Eq. (6.1) is that it provides a formula to predict a new v function of an old value of x. Thus, given an initial guess at the root x_i, Eq. (6. to compute a new estimate x_{i+1} as expressed by the iterative formula

$$x_{i+1} = g(x_i)$$

As with other iterative formulas in this book, the approximate error for this e determined using the error estimator [Eq. (3.5)]:

$$\varepsilon_a = \left| \frac{x_{i+1} - x_i}{x_{i+1}} \right| 100\%$$

EXAMPLE 6.1 Simple Fixed-Point Iteration

Problem Statement. Use simple fixed-point iteration to locate the root of f

Solution. The function can be separated directly and expressed in the form

$$x_{i+1} = e^{-x_i}$$

Starting with an initial guess of $x_0 = 0$, this iterative equation can be applied to compute

i	x_i	ε_a (%)	ε_t (%)
0	0		100.0
1	1.000000	100.0	76.3
2	0.367879	171.8	35.1
3	0.692201	46.9	22.1
4	0.500473	38.3	11.8
5	0.606244	17.4	6.89
6	0.545396	11.2	3.83
7	0.579612	5.90	2.20
8	0.560115	3.48	1.24
9	0.571143	1.93	0.705
10	0.564879	1.11	0.399

Thus, each iteration brings the estimate closer to the true value of the root: 0.56714329.

6.1.1 Convergence

Notice that the true percent relative error for each iteration of Example 6.1 is roughly proportional (by a factor of about 0.5 to 0.6) to the error from the previous iteration. This property, called *linear convergence*, is characteristic of fixed-point iteration.

Aside from the "rate" of convergence, we must comment at this point about the "possibility" of convergence. The concepts of convergence and divergence can be depicted graphically. Recall that in Sec. 5.1, we graphed a function to visualize its structure and behavior (Example 5.1). Such an approach is employed in Fig. 6.2*a* for the function $f(x) = e^{-x} - x$. An alternative graphical approach is to separate the equation into two component parts, as in

$$f_1(x) = f_2(x)$$

Then the two equations

$$y_1 = f_1(x) \tag{6.3}$$

and

$$y_2 = f_2(x) \tag{6.4}$$

can be plotted separately (Fig. 6.2*b*). The x values corresponding to the intersections of these functions represent the roots of $f(x) = 0$.

AMPLE 6.2 The Two-Curve Graphical Method

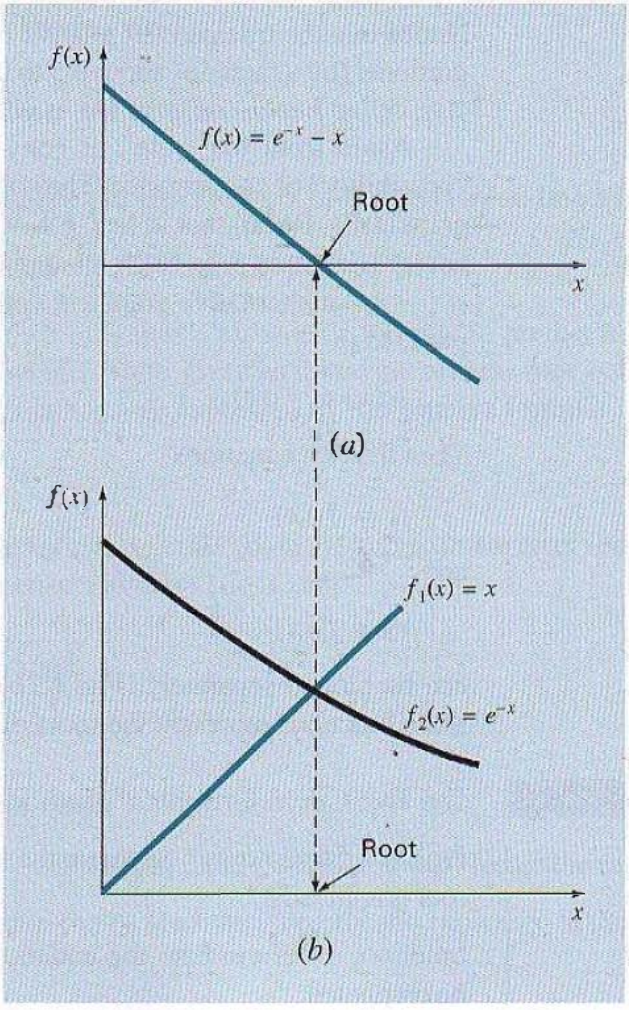
Problem Statement. Separate the equation $e^{-x} - x = 0$ into two parts and determine its root graphically.

Solution. Reformulate the equation as $y_1 = x$ and $y_2 = e^{-x}$. The following values can be computed:

x	y_1	y_2
0.0	0.0	1.000
0.2	0.2	0.819
0.4	0.4	0.670
0.6	0.6	0.549
0.8	0.8	0.449
1.0	1.0	0.368

These points are plotted in Fig. 6.2*b*. The intersection of the two curves indicates a r estimate of approximately $x = 0.57$, which corresponds to the point where the single cu in Fig. 6.2*a* crosses the x axis.

FIGURE 6.2
Two alternative graphical methods for determining the root of $f(x) = e^{-x} - x$. (*a*) Root at the point where it crosses the x axis; (*b*) root at the intersection of the component functions.



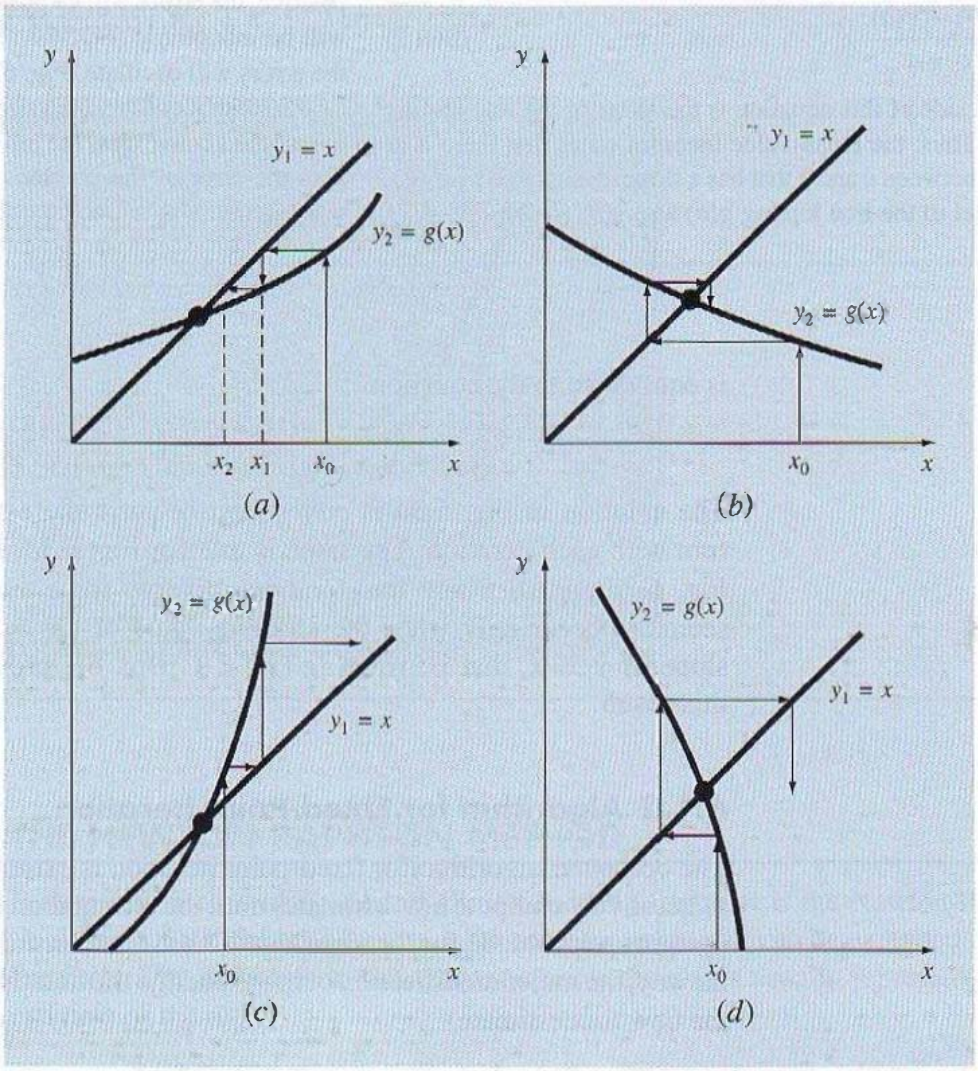
The two-curve method can now be used to illustrate the convergence and divergence of fixed-point iteration. First, Eq. (6.1) can be re-expressed as a pair of equations $y_1 = x$ and $y_2 = g(x)$. These two equations can then be plotted separately. As was the case with Eqs. (6.3) and (6.4), the roots of $f(x) = 0$ correspond to the abscissa value at the intersection of the two curves. The function $y_1 = x$ and four different shapes for $y_2 = g(x)$ are plotted in Fig. 6.3.

For the first case (Fig. 6.3*a*), the initial guess of x_0 is used to determine the corresponding point on the y_2 curve $[x_0, g(x_0)]$. The point (x_1, x_1) is located by moving left horizontally to the y_1 curve. These movements are equivalent to the first iteration in the fixed-point method:

$$x_1 = g(x_0)$$

Thus, in both the equation and in the plot, a starting value of x_0 is used to obtain an estimate of x_1. The next iteration consists of moving to $[x_1, g(x_1)]$ and then to (x_2, x_2). This iteration

on of (*a*) and
nd (*c*) and (*d*)
ple fixed-point
(*a*) and (*c*) are
patterns,
d) are called
il patterns.
lence occurs



Box 6.1 Convergence of Fixed-Point Iteration

From studying Fig. 6.3, it should be clear that fixed-point iteration converges if, in the region of interest, $|g'(x)| < 1$. In other words, convergence occurs if the magnitude of the slope of $g(x)$ is less than the slope of the line $f(x) = x$. This observation can be demonstrated theoretically. Recall that the iterative equation is

$$x_{i+1} = g(x_i)$$

Suppose that the true solution is

$$x_r = g(x_r)$$

Subtracting these equations yields

$$x_r - x_{i+1} = g(x_r) - g(x_i) \tag{B6.1.1}$$

The *derivative mean-value theorem* (recall Sec. 4.1.1) states that if a function $g(x)$ and its first derivative are continuous over an interval $a \le x \le b$, then there exists at least one value of $x = \xi$ within the interval such that

$$g'(\xi) = \frac{g(b) - g(a)}{b - a} \tag{B6.1.2}$$

The right-hand side of this equation is the slope of the line joining $g(a)$ and $g(b)$. Thus, the mean-value theorem states that there is at least one point between a and b that has a slope, designated by $g'(\xi)$, which is parallel to the line joining $g(a)$ and $g(b)$ (recall Fig. 4.3).

Now, if we let $a = x_i$ and $b = x_r$, the right-hand side (B6.1.1) can be expressed as

$$g(x_r) - g(x_i) = (x_r - x_i)g'(\xi)$$

where ξ is somewhere between x_i and x_r. This result can substituted into Eq. (B6.1.1) to yield

$$x_r - x_{i+1} = (x_r - x_i)g'(\xi)$$

If the true error for iteration i is defined as

$$E_{t,i} = x_r - x_i$$

then Eq. (B6.1.3) becomes

$$E_{t,i+1} = g'(\xi)E_{t,i}$$

Consequently, if $|g'(x)| < 1$, the errors decrease with each it For $|g'(x)| > 1$, the errors grow. Notice also that if the deriv positive, the errors will be positive, and hence, the iterative s will be monotonic (Fig. 6.3*a* and *c*). If the derivative is ne the errors will oscillate (Fig. 6.3*b* and *d*).

An offshoot of the analysis is that it also demonstrates tha the method converges, the error is roughly proportional to a than the error of the previous step. For this reason, simple point iteration is said to be *linearly convergent*.

is equivalent to the equation

$$x_2 = g(x_1)$$

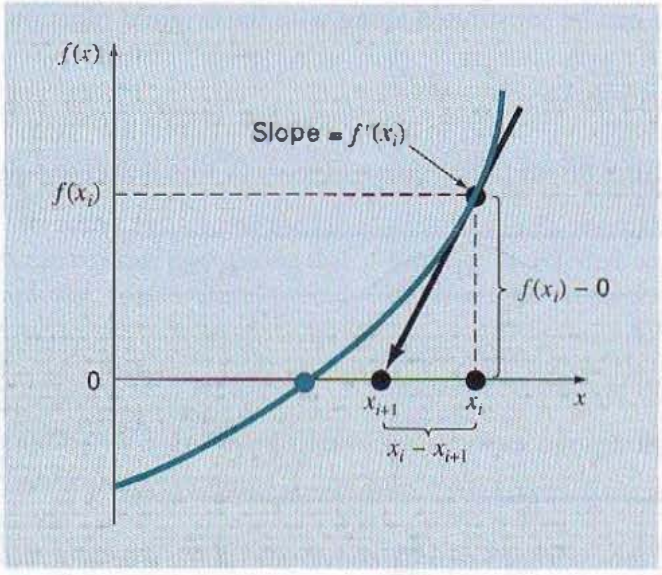
The solution in Fig. 6.3*a* is *convergent* because the estimates of x move closer root with each iteration. The same is true for Fig. 6.3*b*. However, this is not the ca Fig. 6.3*c* and *d*, where the iterations diverge from the root. Notice that conve seems to occur only when the absolute value of the slope of $y_2 = g(x)$ is less th slope of $y_1 = x$, that is, when $|g'(x)| < 1$. Box 6.1 provides a theoretical derivat this result.

6.1.2 Algorithm for Fixed-Point Iteration

The computer algorithm for fixed-point iteration is extremely simple. It consists of to iteratively compute new estimates until the termination criterion has been met. Figu presents pseudocode for the algorithm. Other open methods can be programmed in a lar way, the major modification being to change the iterative formula that is used to co the new root estimate.

```
FUNCTION Fixpt(x0, es, imax, iter, ea)
  xr = x0
  iter = 0
  DO
    xrold = xr
    xr = g(xrold)
    iter = iter + 1
    IF xr ≠ 0 THEN
      ea = |(xr - xrold)/xr| · 100
    END IF
    IF ea < es OR iter ≥ imax EXIT
  END DO
  Fixpt = xr
END Fixpt
```

ixed-point
ot other open
cast in this gen-



tion of the
method.
function of x_i
extrapolated
is to provide
e root at x_{i+1}.

6.2 THE NEWTON-RAPHSON METHOD

Perhaps the most widely used of all root-locating formulas is the Newton-Raphson equation (Fig. 6.5). If the initial guess at the root is x_i, a tangent can be extended from the point $[x_i, f(x_i)]$. The point where this tangent crosses the x axis usually represents an improved estimate of the root.

The Newton-Raphson method can be derived on the basis of this geometrical i pretation (an alternative method based on the Taylor series is described in Box 6.2). A Fig. 6.5, the first derivative at x is equivalent to the slope:

$$f'(x_i) = \frac{f(x_i) - 0}{x_i - x_{i+1}}$$

which can be rearranged to yield

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}$$

which is called the *Newton-Raphson formula.*

EXAMPLE 6.3 Newton-Rophson Method

Problem Statement. Use the Newton-Raphson method to estimate the root of $f($ $e^{-x} - x$, employing an initial guess of $x_0 = 0$.

Solution. The first derivative of the function can be evaluated as

$$f'(x) = -e^{-x} - 1$$

which can be substituted along with the original function into Eq. (6.6) to give

$$x_{i+1} = x_i - \frac{e^{-x_i} - x_i}{-e^{-x_i} - 1}$$

Starting with an initial guess of $x_0 = 0$, this iterative equation can be applied to com

i	x_i	ϵ_t (%)
0	0	100
1	0.500000000	11.8
2	0.566311003	0.147
3	0.567143165	0.0000220
4	0.567143290	$< 10^{-8}$

Thus, the approach rapidly converges on the true root. Notice that the true percent rel error at each iteration decreases much faster than it does in simple fixed-point iter (compare with Example 6.1).

6.2.1 Termination Criteria and Error Estimates

As with other root-location methods, Eq. (3.5) can be used as a termination criterion. I dition, however, the Taylor series derivation of the method (Box 6.2) provides theore insight regarding the rate of convergence as expressed by $E_{i+1} = O(E_i^2)$. Thus the should be roughly proportional to the square of the previous error. In other words

Box 6.2 Derivation and Error Analysis of the Newton-Raphson Method

geometric derivation [Eqs. (6.5) and (6.6)], the
n method may also be developed from the Taylor
. This alternative derivation is useful in that it also
into the rate of convergence of the method.
Chap. 4 that the Taylor series expansion can be rep-

$$x_i) + f'(x_i)(x_{i+1} - x_i)$$

$$\frac{f''(\xi)}{2!}(x_{i+1} - x_i)^2 \qquad \text{(B6.2.1)}$$

ewhere in the interval from x_i to x_{i+1}. An approxi-
obtainable by truncating the series after the first

$$x_i) + f'(x_i)(x_{i+1} - x_i)$$

ion with the x axis, $f(x_{i+1})$ would be equal to

$$f'(x_i)(x_{i+1} - x_i) \qquad \text{(B6.2.2)}$$

lved for

$$\frac{f(x_i)}{f'(x_i)}$$

d to Eq. (6.6). Thus, we have derived the Newton-
a using a Taylor series.
e derivation, the Taylor series can also be used to
r of the formula. This can be done by realizing that
Taylor series were employed, an exact result would be obtained. For this situation $x_{i+1} = x_r$, where x is the true value of the root. Substituting this value along with $f(x_r) = 0$ into Eq. (B6.2.1) yields

$$0 = f(x_i) + f'(x_i)(x_r - x_i) + \frac{f''(\xi)}{2!}(x_r - x_i)^2 \qquad \text{(B6.2.3)}$$

Equation (B6.2.2) can be subtracted from Eq. (B6.2.3) to give

$$0 = f'(x_i)(x_r - x_{i+1}) + \frac{f''(\xi)}{2!}(x_r - x_i)^2 \qquad \text{(B6.2.4)}$$

Now, realize that the error is equal to the discrepancy between x_{i+1} and the true value x_r, as in

$$E_{t,i+1} = x_r - x_{i+1}$$

and Eq. (B6.2.4) can be expressed as

$$0 = f'(x_i)E_{t,i+1} + \frac{f''(\xi)}{2!}E_{t,i}^2 \qquad \text{(B6.2.5)}$$

If we assume convergence, both x_i and ξ should eventually be approximated by the root x_r, and Eq. (B6.2.5) can be rearranged to yield

$$E_{t,i+1} = \frac{-f''(x_r)}{2f'(x_r)}E_{t,i}^2 \qquad \text{(B6.2.6)}$$

According to Eq. (B6.2.6), the error is roughly proportional to the square of the previous error. This means that the number of correct decimal places approximately doubles with each iteration. Such behavior is referred to as *quadratic convergence*. Example 6.4 manifests this property.

number of significant figures of accuracy approximately doubles with each iteration. This behavior is examined in the following example.

XAMPLE 6.4 Error Anolysis of Newton-Rophson Method

Problem Statement. As derived in Box 6.2, the Newton-Raphson method is quadratically convergent. That is, the error is roughly proportional to the square of the previous error, as in

$$E_{t,i+1} \cong \frac{-f''(x_r)}{2f'(x_r)}E_{t,i}^2 \qquad \text{(E6.4.1)}$$

Examine this formula and see if it applies to the results of Example 6.3.

Solution. The first derivative of $f(x) = e^{-x} - x$ is

$$f'(x) = -e^{-x} - 1$$

which can be evaluated at $x_r = 0.56714329$ as $f'(0.56714329) = -1.56714329$. second derivative is

$$f''(x) = e^{-x}$$

which can be evaluated as $f''(0.56714329) = 0.56714329$. These results can be substit into Eq. (E6.4.1) to yield

$$E_{t,i+1} \cong -\frac{0.56714329}{2(-1.56714329)} E_{t,i}^2 = 0.18095 E_{t,i}^2$$

From Example 6.3, the initial error was $E_{t,0} = 0.56714329$, which can be substituted the error equation to predict

$$E_{t,1} \cong 0.18095(0.56714329)^2 = 0.0582$$

which is close to the true error of 0.06714329. For the next iteration,

$$E_{t,2} \cong 0.18095(0.06714329)^2 = 0.0008158$$

which also compares favorably with the true error of 0.0008323. For the third itera

$$E_{t,3} \cong 0.18095(0.0008323)^2 = 0.000000125$$

which is the error obtained in Example 6.3. The error estimate improves in this ma because, as we come closer to the root, x and ξ are better approximated by x_r [recall assumption in going from Eq. (B6.2.5) to Eq. (B6.2.6) in Box 6.2]. Finally,

$$E_{t,4} \cong 0.18095(0.000000125)^2 = 2.83 \times 10^{-15}$$

Thus, this example illustrates that the error of the Newton-Raphson method for this ca in fact, roughly proportional (by a factor of 0.18095) to the square of the error of the vious iteration.

6.2.2 Pitfalls of the Newton-Raphson Method

Although the Newton-Raphson method is often very efficient, there are situations wh performs poorly. A special case—multiple roots—will be addressed later in this cha However, even when dealing with simple roots, difficulties can also arise, as in the fol ing example.

EXAMPLE 6.5 Example of a Slowly Converging Function with Newton-Raphson

Problem Statement. Determine the positive root of $f(x) = x^{10} - 1$ using the New Raphson method and an initial guess of $x = 0.5$.

Solution. The Newton-Raphson formula for this case is

$$x_{i+1} = x_i - \frac{x_i^{10} - 1}{10x_i^9}$$

which can be used to compute

Iteration	x
0	0.5
1	51.65
2	46.485
3	41.8365
4	37.65285
5	33.887565
.	
.	
.	
∞	1.0000000

Thus, after the first poor prediction, the technique is converging on the true root of 1, but at a very slow rate.

Aside from slow convergence due to the nature of the function, other difficulties can arise, as illustrated in Fig. 6.6. For example, Fig. 6.6*a* depicts the case where an inflection point [that is, $f''(x) = 0$] occurs in the vicinity of a root. Notice that iterations beginning at x_0 progressively diverge from the root. Figure 6.6*b* illustrates the tendency of the Newton-Raphson technique to oscillate around a local maximum or minimum. Such oscillations may persist, or as in Fig. 6.6*b*, a near-zero slope is reached, whereupon the solution is sent far from the area of interest. Figure 6.6*c* shows how an initial guess that is close to one root can jump to a location several roots away. This tendency to move away from the area of interest is because near-zero slopes are encountered. Obviously, a zero slope [$f'(x) = 0$] is truly a disaster because it causes division by zero in the Newton-Raphson formula [Eq. (6.6)]. Graphically (see Fig 6.6*d*), it means that the solution shoots off horizontally and never hits the x axis.

Thus, there is no general convergence criterion for Newton-Raphson. Its convergence depends on the nature of the function and on the accuracy of the initial guess. The only remedy is to have an initial guess that is "sufficiently" close to the root. And for some functions, no guess will work! Good guesses are usually predicated on knowledge of the physical problem setting or on devices such as graphs that provide insight into the behavior of the solution. The lack of a general convergence criterion also suggests that good computer software should be designed to recognize slow convergence or divergence. The next section addresses some of these issues.

6.2.3 Algorithm for Newton-Raphson

An algorithm for the Newton-Raphson method is readily obtained by substituting Eq. (6.6) for the predictive formula [Eq. (6.2)] in Fig. 6.4. Note, however, that the program must also be modified to compute the first derivative. This can be simply accomplished by the inclusion of a user-defined function.

Additionally, in light of the foregoing discussion of potential problems of the Newton-Raphson method, the program would be improved by incorporating several additional features:

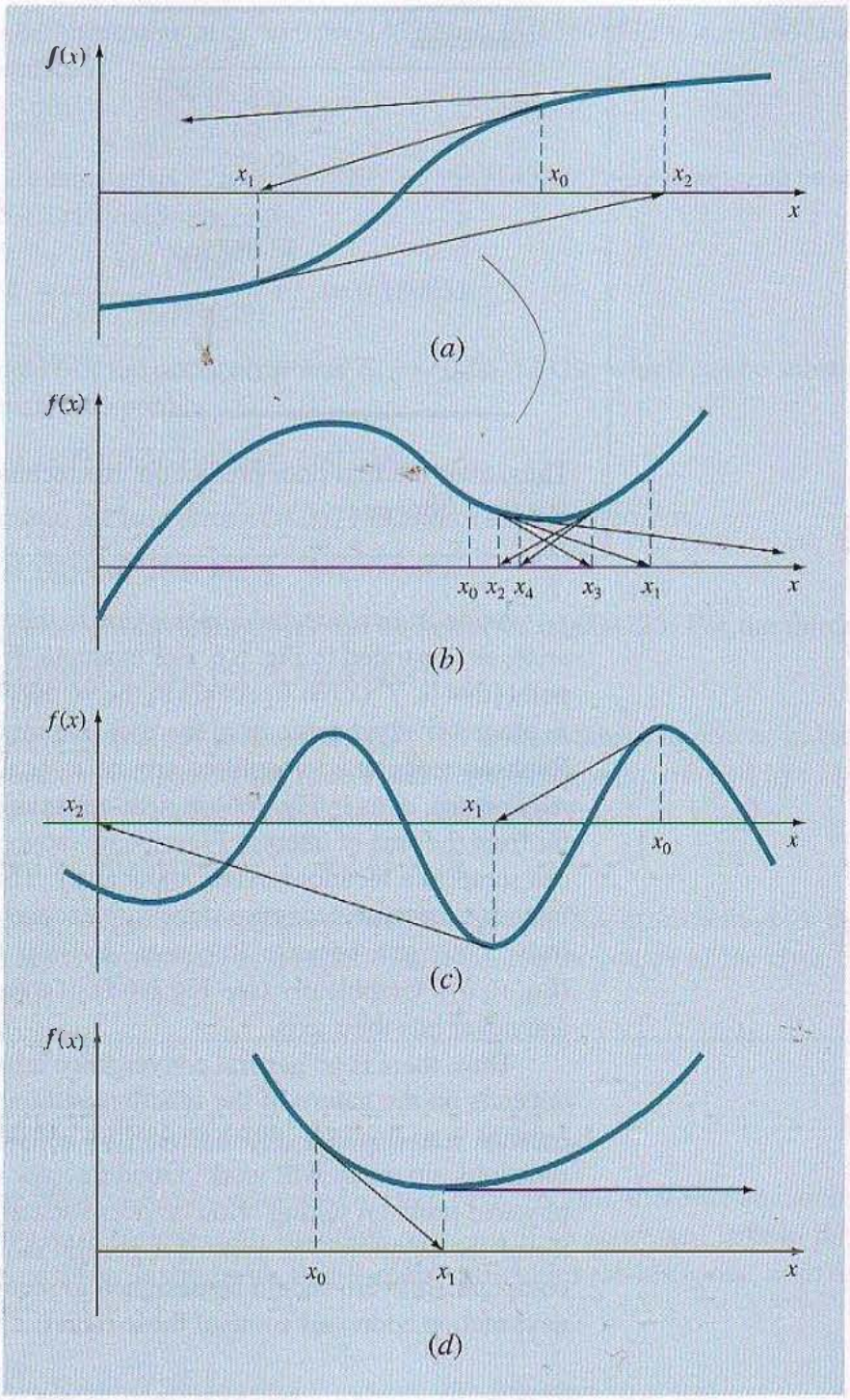


FIGURE 6.6
Four cases where the Newton-Raphson method exhibits poor convergence.

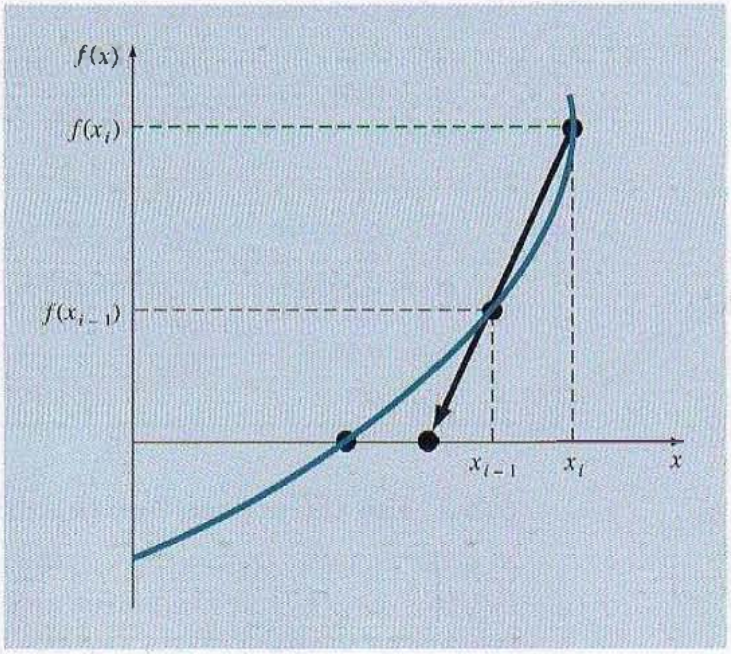
1. A plotting routine should be included in the program.
2. At the end of the computation, the final root estimate should always be substituted into the original function to compute whether the result is close to zero. This check partially guards against those cases where slow or oscillating convergence may lead to a small value of ε_a while the solution is still far from a root.
3. The program should always include an upper limit on the number of iterations to guard against oscillating, slowly convergent, or divergent solutions that could persist interminably.
4. The program should alert the user and take account of the possibility that $f'(x)$ might equal zero at any time during the computation.

6.3 THE SECANT METHOD

A potential problem in implementing the Newton-Raphson method is the evaluation of the derivative. Although this is not inconvenient for polynomials and many other functions, there are certain functions whose derivatives may be extremely difficult or inconvenient to evaluate. For these cases, the derivative can be approximated by a backward finite divided difference, as in (Fig. 6.7)

$$f'(x_i) \cong \frac{f(x_{i-1}) - f(x_i)}{x_{i-1} - x_i}$$

FIGURE 6.7
Graphical depiction of the secant method. This technique is similar to the Newton-Raphson technique (Fig. 6.5) in the sense that an estimate of the root is predicted by extrapolating a tangent of the function to the x axis. However, the secant method uses a difference rather than a derivative to estimate the slope.



This approximation can be substituted into Eq. (6.6) to yield the following iter equation:

$$x_{i+1} = x_i - \frac{f(x_i)(x_{i-1} - x_i)}{f(x_{i-1}) - f(x_i)}$$

Equation (6.7) is the formula for the *secant method*. Notice that the approach requires initial estimates of x. However, because $f(x)$ is not required to change signs betwee estimates, it is not classified as a bracketing method.

EXAMPLE 6.6 The Secant Method

Problem Statement. Use the secant method to estimate the root of $f(x) = e^{-x} - x$. with initial estimates of $x_{-1} = 0$ and $x_0 = 1.0$.

Solution. Recall that the true root is 0.56714329. . . .

First iteration:

$$x_{-1} = 0 \qquad f(x_{-1}) = 1.00000$$

$$x_0 = 1 \qquad f(x_0) = -0.63212$$

$$x_1 = 1 - \frac{-0.63212(0 - 1)}{1 - (-0.63212)} = 0.61270 \qquad \varepsilon_t = 8.0\%$$

Second iteration:

$$x_0 = 1 \qquad f(x_0) = -0.63212$$

$$x_1 = 0.61270 \qquad f(x_1) = -0.07081$$

(Note that both estimates are now on the same side of the root.)

$$x_2 = 0.61270 - \frac{-0.07081(1 - 0.61270)}{-0.63212 - (-0.07081)} = 0.56384 \qquad \varepsilon_t = 0.58\%$$

Third iteration:

$$x_1 = 0.61270 \qquad f(x_1) = -0.07081$$

$$x_2 = 0.56384 \qquad f(x_2) = 0.00518$$

$$x_3 = 0.56384 - \frac{0.00518(0.61270 - 0.56384)}{-0.07081 - (-0.00518)} = 0.56717 \qquad \varepsilon_t = 0.0048\%$$

6.3.1 The Difference Between the Secant and False-Position Meth

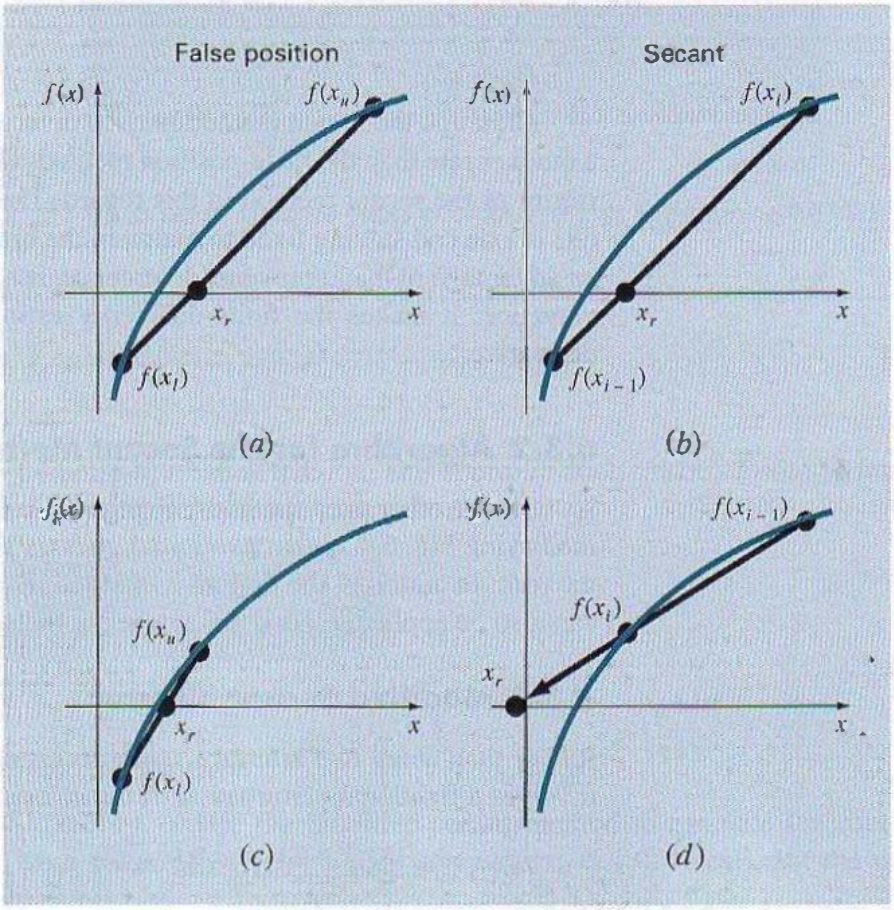
Note the similarity between the secant method and the false-position method. For exam Eqs. (6.7) and (5.7) are identical on a term-by-term basis. Both use two initial estimat compute an approximation of the slope of the function that is used to project to the x for a new estimate of the root. However, a critical difference between the methods is

one of the initial values is replaced by the new estimate. Recall that in the false-position method the latest estimate of the root replaces whichever of the original values yielded a function value with the same sign as $f(x_r)$. Consequently, the two estimates always bracket the root. Therefore, for all practical purposes, the method always converges because the root is kept within the bracket. In contrast, the secant method replaces the values in strict sequence, with the new value x_{i+1} replacing x_i and x_i replacing x_{i-1}. As a result, the two values can sometimes lie on the same side of the root. For certain cases, this can lead to divergence.

EXAMPLE 6.7 Comparison of Convergence of the Secant and False-Position Techniques

Problem Statement. Use the false-position and secant methods to estimate the root of $f(x) = \ln x$. Start the computation with values of $x_l = x_{i-1} = 0.5$ and $x_u = x_i = 5.0$.

FIGURE 6.8
Comparison of the false-position and the secant methods. The first iterations (*a*) and (*b*) for both techniques are identical. However, for the second iterations (*c*) and (*d*), the points used differ. As a consequence, the secant method can diverge, as indicated in (*d*).



Solution. For the false-position method, the use of Eq. (5.7) and the bracketing criterion for replacing estimates results in the following iterations:

Iteration	x_l	x_u	x_r
1	0.5	5.0	1.8546
2	0.5	1.8546	1.2163
3	0.5	1.2163	1.0585

As can be seen (Fig. 6.8*a* and *c*), the estimates are converging on the true root which equal to 1.

For the secant method, using Eq. (6.7) and the sequential criterion for replacing estimates results in

Iteration	x_{i-1}	x_i	x_{i+1}
1	0.5	5.0	1.8546
2	5.0	1.8546	−0.10438

As in Fig. 6.8*d*, the approach is divergent.

Although the secant method may be divergent, when it converges it usually does so at a quicker rate than the false-position method. For instance, Fig. 6.9 demonstrates the superiority of the secant method in this regard. The inferiority of the false-position method is due to one end staying fixed to maintain the bracketing of the root. This property, which is an advantage in that it prevents divergence, is a shortcoming with regard to the rate of convergence; it makes the finite-difference estimate a less-accurate approximation of the derivative.

6.3.2 Algorithm for the Secant Method

As with the other open methods, an algorithm for the secant method is obtained simply by modifying Fig. 6.4 so that two initial guesses are input and by using Eq. (6.7) to calculate the root. In addition, the options suggested in Sec. 6.2.3 for the Newton-Raphson method can also be applied to good advantage for the secant program.

6.3.3 Modified Secant Method

Rather than using two arbitrary values to estimate the derivative, an alternative approach involves a fractional perturbation of the independent variable to estimate $f'(x)$,

$$f'(x_i) \cong \frac{f(x_i + \delta x_i) - f(x_i)}{\delta x_i}$$

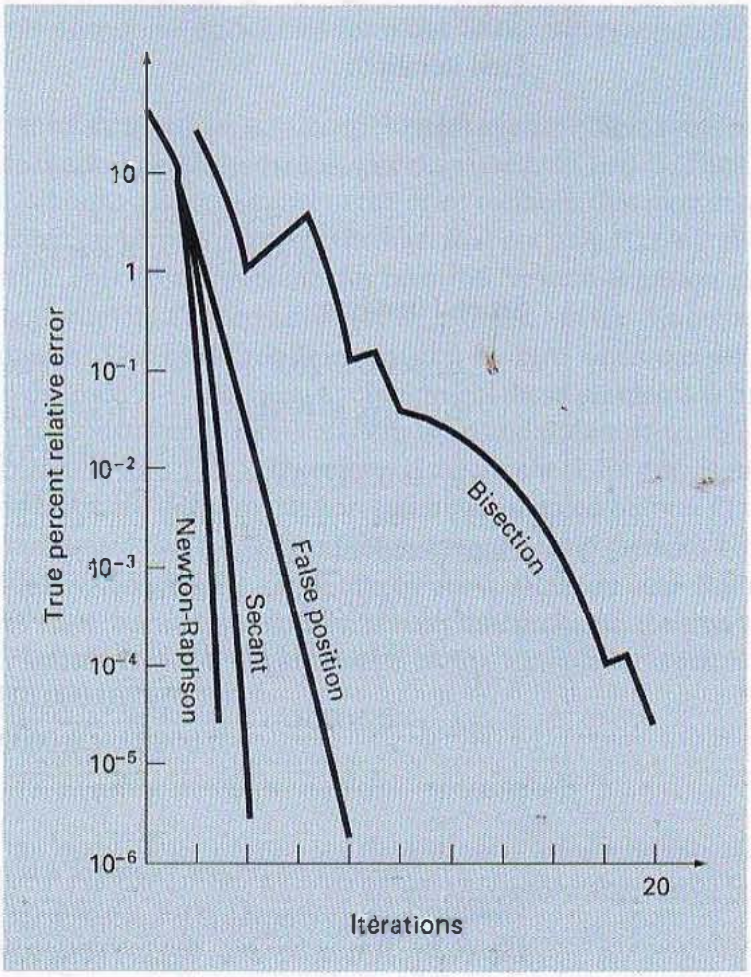


FIGURE 6.9
Comparison of the true percent relative errors ε_t for the methods to determine the roots of $f(x) = e^{-x} - x$.

where δ = a small perturbation fraction. This approximation can be substituted into Eq. (6.6) to yield the following iterative equation:

$$x_{i+1} = x_i - \frac{\delta x_i f(x_i)}{f(x_i + \delta x_i) - f(x_i)} \tag{6.8}$$

EXAMPLE 6.8 Modified Secant Method

Problem Statement. Use the modified secant method to estimate the root of $f(x) = e^{-x} - x$. Use a value of 0.01 for δ and start with $x_0 = 1.0$. Recall that the true root is 0.56714329....

Solution.
First iteration:

$$x_0 = 1 \qquad f(x_0) = -0.63212$$
$$x_0 + \delta x_0 = 1.01 \qquad f(x_0 + \delta x_0) = -0.64578$$
$$x_1 = 1 - \frac{0.01(-0.63212)}{-0.64578 - (-0.63212)} = 0.537263 \qquad |\varepsilon_t| = 5.3\%$$

Second iteration:

$$x_0 = 0.537263 \qquad f(x_0) = 0.047083$$
$$x_0 + \delta x_0 = 0.542635 \qquad f(x_0 + \delta x_0) = 0.038579$$
$$x_1 = 0.537263 - \frac{0.005373(0.047083)}{0.038579 - 0.047083} = 0.56701 \qquad |\varepsilon_t| = 0.0236\%$$

Third iteration:

$$x_0 = 0.56701 \qquad f(x_0) = 0.000209$$
$$x_0 + \delta x_0 = 0.572680 \qquad f(x_0 + \delta x_0) = -0.00867$$
$$x_1 = 0.56701 - \frac{0.00567(0.000209)}{-0.00867 - 0.000209} = 0.567143 \qquad |\varepsilon_t| = 2.365 \times 10^{-5}\%$$

The choice of a proper value for δ is not automatic. If δ is too small, the method can be swamped by round-off error caused by subtractive cancellation in the denominator of Eq. (6.8). If it is too big, the technique can become inefficient and even divergent. However, if chosen correctly, it provides a nice alternative for cases where evaluating the derivative is difficult and developing two initial guesses is inconvenient.

6.4 MULTIPLE ROOTS

A *multiple root* corresponds to a point where a function is tangent to the x axis. For example, a double root results from

$$f(x) = (x - 3)(x - 1)(x - 1)$$

or, multiplying terms, $f(x) = x^3 - 5x^2 + 7x - 3$. The equation has a *double root* because one value of x makes two terms in Eq. (6.9) equal to zero. Graphically, this corresponds to the curve touching the x axis tangentially at the double root. Examine Fig. 6.10*a* at $x =$ Notice that the function touches the axis but does not cross it at the root.

A *triple root* corresponds to the case where one x value makes three terms in an equation equal to zero, as in

$$f(x) = (x - 3)(x - 1)(x - 1)(x - 1)$$

or, multiplying terms, $f(x) = x^4 - 6x^3 + 12x^2 - 10x + 3$. Notice that the graphical depiction (Fig. 6.10*b*) again indicates that the function is tangent to the axis at the root, but that for this case the axis is crossed. In general, odd multiple roots cross the axis, whereas even ones do not. For example, the quadruple root in Fig. 6.10*c* does not cross the axis

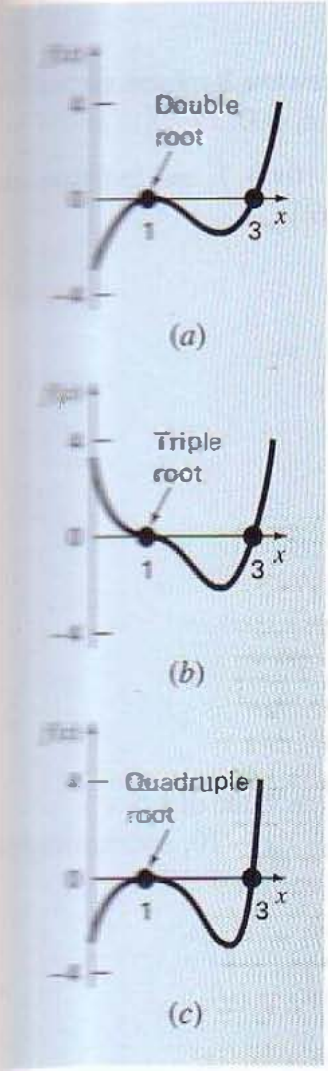


FIGURE 6.10
Examples of multiple roots that are tangential to the x axis. (a) and (c) The function does not cross the axis on either side of these multiple roots (a) and (c), whereas it crosses the axis at both roots (b).

Multiple roots pose some difficulties for many of the numerical methods described in Part Two:

1. The fact that the function does not change sign at even multiple roots precludes the use of the reliable bracketing methods that were discussed in Chap. 5. Thus, of the methods covered in this book, you are limited to the open methods that may diverge.
2. Another possible problem is related to the fact that not only $f(x)$ but also $f'(x)$ goes to zero at the root. This poses problems for both the Newton-Raphson and secant methods, which both contain the derivative (or its estimate) in the denominator of their respective formulas. This could result in division by zero when the solution converges very close to the root. A simple way to circumvent these problems is based on the fact that it can be demonstrated theoretically (Ralston and Rabinowitz, 1978) that $f(x)$ will always reach zero before $f'(x)$. Therefore, if a zero check for $f(x)$ is incorporated into the computer program, the computation can be terminated before $f'(x)$ reaches zero.
3. It can be demonstrated that the Newton-Raphson and secant methods are linearly, rather than quadratically, convergent for multiple roots (Ralston and Rabinowitz, 1978). Modifications have been proposed to alleviate this problem. Ralston and Rabinowitz (1978) have indicated that a slight change in the formulation returns it to quadratic convergence, as in

$$x_{i+1} = x_i - m\frac{f(x_i)}{f'(x_i)} \tag{6.9a}$$

where m is the multiplicity of the root (that is, $m = 2$ for a double root, $m = 3$ for a triple root, etc.). Of course, this may be an unsatisfactory alternative because it hinges on foreknowledge of the multiplicity of the root.

Another alternative, also suggested by Ralston and Rabinowitz (1978), is to define a new function $u(x)$, that is, the ratio of the function to its derivative, as in

$$u(x) = \frac{f(x)}{f'(x)} \tag{6.10}$$

It can be shown that this function has roots at all the same locations as the original function. Therefore, Eq. (6.10) can be substituted into Eq. (6.6) to develop an alternative form of the Newton-Raphson method:

$$x_{i+1} = x_i - \frac{u(x_i)}{u'(x_i)} \tag{6.11}$$

Equation (6.10) can be differentiated to give

$$u'(x) = \frac{f'(x)f'(x) - f(x)f''(x)}{[f'(x)]^2} \tag{6.12}$$

Equations (6.10) and (6.12) can be substituted into Eq. (6.11) and the result simplified to yield

$$x_{i+1} = x_i - \frac{f(x_i)f'(x_i)}{[f'(x_i)]^2 - f(x_i)f''(x_i)} \tag{6.13}$$

EXAMPLE 6.9 Modified Newton-Raphson Method for Multiple Roots

Problem Statement. Use both the standard and modified Newton-Raphson method evaluate the multiple root of Eq. (6.9), with an initial guess of $x_0 = 0$.

Solution. The first derivative of Eq. (6.9) is $f'(x) = 3x^2 - 10x + 7$, and therefore, standard Newton-Raphson method for this problem is [Eq. (6.6)]

$$x_{i+1} = x_i - \frac{x_i^3 - 5x_i^2 + 7x_i - 3}{3x_i^2 - 10x_i + 7}$$

which can be solved iteratively for

i	x_i	ε_t (%)
0	0	100
1	0.4285714	57
2	0.6857143	31
3	0.8328654	17
4	0.9133290	8.7
5	0.9557833	4.4
6	0.9776551	2.2

As anticipated, the method is linearly convergent toward the true value of 1.0.

For the modified method, the second derivative is $f''(x) = 6x - 10$, and the iterative relationship is [Eq. (6.13)]

$$x_{i+1} = x_i - \frac{(x_i^3 - 5x_i^2 + 7x_i - 3)(3x_i^2 - 10x_i + 7)}{(3x_i^2 - 10x_i + 7)^2 - (x_i^3 - 5x_i^2 + 7x_i - 3)(6x_i - 10)}$$

which can be solved for

i	x_i	ε_t (%)
0	0	100
1	1.105263	11
2	1.003082	0.31
3	1.000002	0.00024

Thus, the modified formula is quadratically convergent. We can also use both methods to search for the single root at $x = 3$. Using an initial guess of $x_0 = 4$ gives the following results:

i	Standard	ε_t (%)	Modified	ε_t (%)
0	4	33	4	33
1	3.4	13	2.636364	12
2	3.1	3.3	2.820225	6.0
3	3.008696	0.29	2.961728	1.3
4	3.000075	0.0025	2.998479	0.051
5	3.000000	2×10^{-7}	2.999998	7.7×10^{-5}

Thus, both methods converge quickly, with the standard method being somewhat more efficient.

The above example illustrates the trade-offs involved in opting for the modified Newton-Raphson method. Although it is preferable for multiple roots, it is somewhat less efficient and requires more computational effort than the standard method for simple roots.

It should be noted that a modified version of the secant method suited for multiple roots can also be developed by substituting Eq. (6.10) into Eq. (6.7). The resulting formula is (Ralston and Rabinowitz, 1978)

$$x_{i+1} = x_i - \frac{u(x_i)(x_{i-1} - x_i)}{u(x_{i-1}) - u(x_i)}$$

6.5 SYSTEMS OF NONLINEAR EQUATIONS

To this point, we have focused on the determination of the roots of a single equation. A related problem is to locate the roots of a set of simultaneous equations,

$$\begin{aligned} f_1(x_1, x_2, \ldots, x_n) &= 0 \\ f_2(x_1, x_2, \ldots, x_n) &= 0 \\ &\vdots \\ f_n(x_1, x_2, \ldots, x_n) &= 0 \end{aligned} \tag{6.14}$$

The solution of this system consists of a set of x values that simultaneously result in all the equations equaling zero.

In Part Three, we will present methods for the case where the simultaneous equations are linear—that is, they can be expressed in the general form

$$f(x) = a_1x_1 + a_2x_2 + \cdots + a_nx_n - b = 0 \tag{6.15}$$

where the b and the a's are constants. Algebraic and transcendental equations that do not fit this format are called *nonlinear equations*. For example,

$$x^2 + xy = 10$$

and

$$y + 3xy^2 = 57$$

are two simultaneous nonlinear equations with two unknowns, x and y. They can be expressed in the form of Eq. (6.14) as

$$u(x, y) = x^2 + xy - 10 = 0 \tag{6.16a}$$

$$v(x, y) = y + 3xy^2 - 57 = 0 \tag{6.16b}$$

Thus, the solution would be the values of x and y that make the functions $u(x, y)$ and $v(x, y)$ equal to zero. Most approaches for determining such solutions are extensions of the open

methods for solving single equations. In this section, we will investigate two of t fixed-point iteration and Newton-Raphson.

6.5.1 Fixed-Point Iteration

The fixed-point-iteration approach (Sec. 6.1) can be modified to solve two simultane nonlinear equations. This approach will be illustrated in the following example.

EXAMPLE 6.10 Fixed-Point Iteration for a Nonlinear System

Problem Statement. Use fixed-point iteration to determine the roots of Eq. (6.16). that a correct pair of roots is $x = 2$ and $y = 3$. Initiate the computation with guess $x = 1.5$ and $y = 3.5$.

Solution. Equation (6.16*a*) can be solved for

$$x_{i+1} = \frac{10 - x_i^2}{y_i} \tag{E6.}$$

and Eq. (6.16*b*) can be solved for

$$y_{i+1} = 57 - 3x_i y_i^2 \tag{E6.}$$

Note that we will drop the subscripts for the remainder of the example.

On the basis of the initial guesses, Eq. (E6.10.1) can be used to determine a new of x:

$$x = \frac{10 - (1.5)^2}{3.5} = 2.21429$$

This result and the initial value of $y = 3.5$ can be substituted into Eq. (E6.10.2) to d mine a new value of y:

$$y = 57 - 3(2.21429)(3.5)^2 = -24.37516$$

Thus, the approach seems to be diverging. This behavior is even more pronounced o second iteration:

$$x = \frac{10 - (2.21429)^2}{-24.37516} = -0.20910$$

$$y = 57 - 3(-0.20910)(-24.37516)^2 = 429.709$$

Obviously, the approach is deteriorating.

Now we will repeat the computation but with the original equations set up in a di ent format. For example, an alternative formulation of Eq. (6.16*a*) is

$$x = \sqrt{10 - xy}$$

and of Eq. (6.16*b*) is

$$y = \sqrt{\frac{57 - y}{3x}}$$

Now the results are more satisfactory:

$$x = \sqrt{10 - 1.5(3.5)} = 2.17945$$

$$y = \sqrt{\frac{57 - 3.5}{3(2.17945)}} = 2.86051$$

$$x = \sqrt{10 - 2.17945(2.86051)} = 1.94053$$

$$y = \sqrt{\frac{57 - 2.86051}{3(1.94053)}} = 3.04955$$

Thus, the approach is converging on the true values of $x = 2$ and $y = 3$.

The previous example illustrates the most serious shortcoming of simple fixed-point iteration—that is, convergence often depends on the manner in which the equations are formulated. Additionally, even in those instances where convergence is possible, divergence can occur if the initial guesses are insufficiently close to the true solution. Using reasoning similar to that in Box 6.1, it can be demonstrated that sufficient conditions for convergence for the two-equation case are

$$\left|\frac{\partial u}{\partial x}\right| + \left|\frac{\partial u}{\partial y}\right| < 1$$

and

$$\left|\frac{\partial v}{\partial x}\right| + \left|\frac{\partial v}{\partial y}\right| < 1$$

These criteria are so restrictive that fixed-point iteration has limited utility for solving nonlinear systems. However, as we will describe later in the book, it can be very useful for solving linear systems.

6.5.2 Newton-Raphson

Recall that the Newton-Raphson method was predicated on employing the derivative (that is, the slope) of a function to estimate its intercept with the axis of the independent variable—that is, the root (Fig. 6.5). This estimate was based on a first-order Taylor series expansion (recall Box 6.2),

$$f(x_{i+1}) = f(x_i) + (x_{i+1} - x_i) f'(x_i) \tag{6.17}$$

where x_i is the initial guess at the root and x_{i+1} is the point at which the slope intercepts the x axis. At this intercept, $f(x_{i+1})$ by definition equals zero and Eq. (6.17) can be rearranged to yield

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} \tag{6.18}$$

which is the single-equation form of the Newton-Raphson method.

CHAPTER 7

Roots of Polynomials

In this chapter, we will discuss methods to find the roots of polynomial equations o general form

$$f_n(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$$

where n = the order of the polynomial and the a's = constant coefficients. Althoug coefficients can be complex numbers, we will limit our discussion to cases where the real. For such cases, the roots can be real and/or complex.

The roots of such polynomials follow these rules:

1. For an nth-order equation, there are n real or complex roots. It should be noted these roots will not necessarily be distinct.
2. If n is odd, there is at least one real root.
3. If complex roots exist, they exist in conjugate pairs (that is, $\lambda + \mu i$ and $\lambda - \mu i$). w $i = \sqrt{-1}$.

Before describing the techniques for locating the roots of polynomials, we will pro some background. The first section offers some motivation for studying the techniques second deals with some fundamental computer manipulations involving polynomials.

7.1 POLYNOMIALS IN ENGINEERING AND SCIENCE

Polynomials have many applications in engineering and science. For example, they are extensively in curve-fitting. However, we believe that one of their most interesting and p erful applications is in characterizing dynamic systems and, in particular, linear syst Examples include mechanical devices, structures, and electrical circuits. We will be exp ing specific examples throughout the remainder of this text. In particular, they will be focus of several of the engineering applications throughout the remainder of this text.

For the time being, we will keep the discussion simple and general by focusing simple second-order system defined by the following linear *ordinary differential equa* (or ODE):

$$a_2 \frac{d^2y}{dt^2} + a_1 \frac{dy}{dt} + a_0 y = F(t)$$

When complex roots are possible, the bracketing methods cannot be used because of the obvious problem that the criterion for defining a bracket (that is, sign change) does not translate to complex guesses.

Of the open methods, the conventional Newton-Raphson method would provide a viable approach. In particular, concise code including deflation can be developed. If a language that accommodates complex variables (like Fortran) is used, such an algorithm will locate both real and complex roots. However, as might be expected, it would be susceptible to convergence problems. For this reason, special methods have been developed to find the real and complex roots of polynomials. We describe two—the Müller and Bairstow methods—in the following sections. As you will see, both are related to the more conventional open approaches described in Chap. 6.

7.4 MÜLLER'S METHOD

Recall that the secant method obtains a root estimate by projecting a straight line to the x axis through two function values (Fig. 7.3a). Müller's method takes a similar approach, but projects a parabola through three points (Fig. 7.3b).

The method consists of deriving the coefficients of the parabola that goes through the three points. These coefficients can then be substituted into the quadratic formula to obtain the point where the parabola intercepts the x axis—that is, the root estimate. The approach is facilitated by writing the parabolic equation in a convenient form,

$$f_2(x) = a(x - x_2)^2 + b(x - x_2) + c \tag{7.17}$$

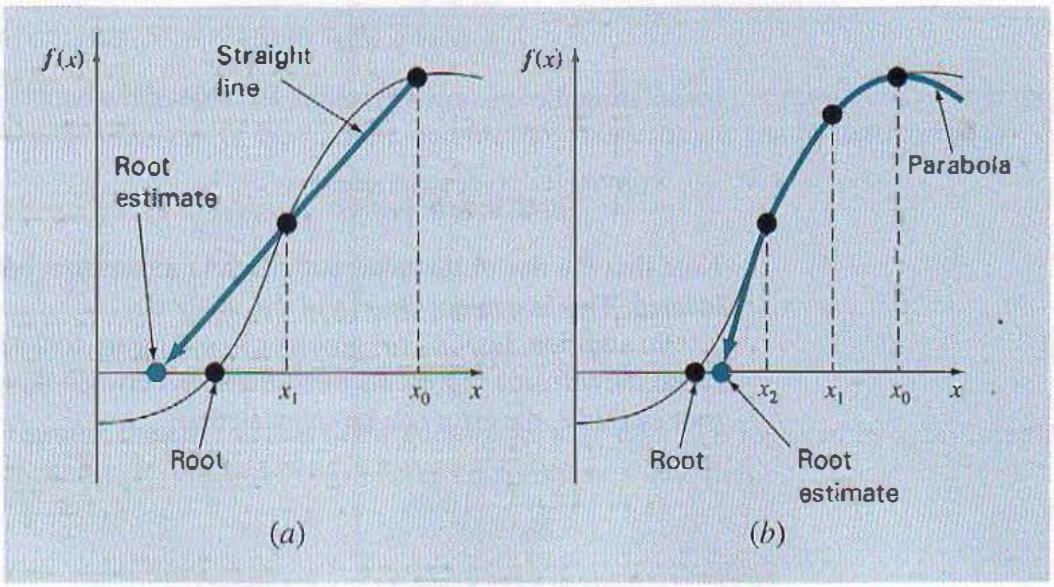
We want this parabola to intersect the three points $[x_0, f(x_0)]$, $[x_1, f(x_1)]$, and $[x_2, f(x_2)]$. The coefficients of Eq. (7.17) can be evaluated by substituting each of the three points to give

$$f(x_0) = a(x_0 - x_2)^2 + b(x_0 - x_2) + c \tag{7.18}$$

$$f(x_1) = a(x_1 - x_2)^2 + b(x_1 - x_2) + c \tag{7.19}$$

$$f(x_2) = a(x_2 - x_2)^2 + b(x_2 - x_2) + c \tag{7.20}$$

two related
cating roots:
hod and
d.



Note that we have dropped the subscript "2" from the function for conciseness. Bec have three equations, we can solve for the three unknown coefficients, a, b, and c. E two of the terms in Eq. (7.20) are zero, it can be immediately solved for $c = f(x_2)$. T coefficient c is merely equal to the function value evaluated at the third guess, result can then be substituted into Eqs. (7.18) and (7.19) to yield two equations unknowns:

$$f(x_0) - f(x_2) = a(x_0 - x_2)^2 + b(x_0 - x_2)$$
$$f(x_1) - f(x_2) = a(x_1 - x_2)^2 + b(x_1 - x_2)$$

Algebraic manipulation can then be used to solve for the remaining coefficien b. One way to do this involves defining a number of differences,

$$h_0 = x_1 - x_0 \qquad h_1 = x_2 - x_1$$
$$\delta_0 = \frac{f(x_1) - f(x_0)}{x_1 - x_0} \qquad \delta_1 = \frac{f(x_2) - f(x_1)}{x_2 - x_1}$$

These can be substituted into Eqs. (7.21) and (7.22) to give

$$(h_0 + h_1)b - (h_0 + h_1)^2 a = h_0\delta_0 + h_1\delta_1$$
$$h_1 \quad b - \quad h_1^2 \quad a = \quad h_1\delta_1$$

which can be solved for a and b. The results can be summarized as

$$a = \frac{\delta_1 - \delta_0}{h_1 + h_0}$$
$$b = ah_1 + \delta_1$$
$$c = f(x_2)$$

To find the root, we apply the quadratic formula to Eq. (7.17). However, bec potential round-off error, rather than using the conventional form, we use the alt formulation [Eq. (3.13)] to yield

$$x_3 - x_2 = \frac{-2c}{b \pm \sqrt{b^2 - 4ac}}$$

or isolating the unknown x_3 on the left side of the equal sign,

$$x_3 = x_2 + \frac{-2c}{b \pm \sqrt{b^2 - 4ac}}$$

Note that the use of the quadratic formula means that both real and complex roots located. This is a major benefit of the method.

In addition, Eq. (7.27a) provides a neat means to determine the approximate er cause the left side represents the difference between the present (x_3) and the previ root estimate, the error can be calculated as

$$\varepsilon_a = \left|\frac{x_3 - x_2}{x_3}\right| 100\%$$

Now, a problem with Eq. (7.27*a*) is that it yields two roots, corresponding to the ± term in the denominator. In Müller's method, the sign is chosen to agree with the sign of b. This choice will result in the largest denominator, and hence, will give the root estimate that is closest to x_2.

Once x_3 is determined, the process is repeated. This brings up the issue of which point is discarded. Two general strategies are typically used:

1. If only real roots are being located, we choose the two original points that are nearest the new root estimate, x_3.
2. If both real and complex roots are being evaluated, a sequential approach is employed. That is, just like the secant method, x_1, x_2, and x_3 take the place of x_0, x_1, and x_2.

EXAMPLE 7.2 Müller's Method

Problem Statement. Use Müller's method with guesses of x_0, x_1, and $x_2 = 4.5$, 5.5, and 5, respectively, to determine a root of the equation

$$f(x) = x^3 - 13x - 12$$

Note that the roots of this equation are -3, -1, and 4.

Solution. First, we evaluate the function at the guesses

$$f(4.5) = 20.625 \qquad f(5.5) = 82.875 \qquad f(5) = 48$$

which can be used to calculate

$$h_0 = 5.5 - 4.5 = 1 \qquad h_1 = 5 - 5.5 = -0.5$$

$$\delta_0 = \frac{82.875 - 20.625}{5.5 - 4.5} = 62.25 \qquad \delta_1 = \frac{48 - 82.875}{5 - 5.5} = 69.75$$

These values in turn can be substituted into Eqs. (7.24) through (7.26) to compute

$$a = \frac{69.75 - 62.25}{-0.5 + 1} = 15 \qquad b = 15(-0.5) + 69.75 = 62.25 \qquad c = 48$$

The square root of the discriminant can be evaluated as

$$\sqrt{62.25^2 - 4(15)48} = 31.54461$$

Then, because $|62.25 + 31.54451| > |62.25 - 31.54451|$, a positive sign is employed in the denominator of Eq. (7.27*b*), and the new root estimate can be determined as

$$x_3 = 5 + \frac{-2(48)}{62.25 + 31.54451} = 3.976487$$

and develop the error estimate

$$\varepsilon_a = \left|\frac{-1.023513}{3.976487}\right| 100\% = 25.74\%$$

Because the error is large, new guesses are assigned; x_0 is replaced by x_1, x_1 is replaced by x_2, and x_2 is replaced by x_3. Therefore, for the new iteration,

$$x_0 = 5.5 \qquad x_1 = 5 \qquad x_2 = 3.976487$$

and the calculation is repeated. The results, tabulated below, show that the metho verges rapidly on the root, $x_r = 4$:

i	x_r	ε_a (%)
0	5	
1	3.976487	25.74
2	4.001065	0.6139
3	4	0.0262
4	4	0.0000119

Pseudocode to implement Müller's method for real roots is presented in Fig. 7. tice that this routine is set up to take a single initial nonzero guess that is then pertur

FIGURE 7.4
Pseudocode for Müller's method.

```
SUB Muller(xr, h, eps, maxit)
x2 = xr
x1 = xr + h*xr
x0 = xr - h*xr
DO
  iter = iter + 1
  h0 = x1 - x0
  h1 = x2 - x1
  d0 = (f(x1) - f(x0)) / h0
  d1 = (f(x2) - f(x1)) / h1
  a = (d1 - d0) / (h1 + h0)
  b = a*h1 + d1
  c = f(x2)
  rad = SQRT(b*b - 4*a*c)
  If |b+rad| > |b-rad| THEN
    den = b + rad
  ELSE
    den = b - rad
  END IF
  dxr = -2*c / den
  xr = x2 + dxr
  PRINT iter, xr
  IF (|dxr| < eps*xr OR iter >= maxit) EXIT
  x0 = x1
  x1 = x2
  x2 = xr
END DO
END Muller
```

OBLEMS

olynomial $f(x) = x^4 - 7.5x^3 + 14.5x^2 + 3x - 20$
al factor $x - 2$. Is $x = 2$ a root?
olynomial $f(x) = x^5 - 5x^4 + x^3 - 6x^2 - 7x + 10$ by
actor $x - 2$.
's method to determine the positive real root of
$x^2 - 3x - 5$
$0.5x^2 + 4x - 3$
's method or MATLAB to determine the real and
of
$x^2 + 3x - 2$
$+ 6x^2 + 10$
$2x^3 + 6x^2 - 8x + 8$
ow's method to determine the roots of
$+ 6.2x - 4x^2 + 0.7x^3$
$- 21.97x + 16.3x^2 - 3.704x^3$
$3x^3 + 5x^2 - x - 10$
program to implement Müller's method. Test it by
mple 7.2.
gram developed in Prob. 7.6 to determine the real
7.4a. Construct a graph (by hand or with Excel or
phics package) to develop suitable starting guesses.
program to implement Bairstow's method. Test it by
mple 7.3.
gram developed in Prob. 7.8 to determine the roots
in Prob. 7.5.
the real root of $x^{3.5} = 80$ with the Goal Seek capa-
r a library or package of your choice.
ity of a falling parachutist is given by

$$- e^{-(c/m)t})$$

m/s^2. For a parachutist with a drag coefficient $c =$
te the mass m so that the velocity is $v = 35$ m/s at
Goal Seek capability of Excel or a library or pack-
ice to determine m.
the roots of the simultaneous nonlinear equations

$$+ 0.75$$
$$.2$$

uesses of $x = y = 1.2$ and use the Solver tool from
y or package of your choice.
the roots of the simultaneous nonlinear equations

$$y - 4)^2 = 5$$
$$5$$

approach to obtain your initial guesses. Determine
s with the Solver tool from Excel or a library or
choice.

7.14 Perform the identical MATLAB operations as those in Example 7.7 or use a library or package of your choice to find all the roots of the polynomial

$$f(x) = (x - 4)(x + 2)(x - 1)(x + 5)(x - 7)$$

Note that the `poly` function can be used to convert the roots to a polynomial.

7.15 Use MATLAB or a library or package of your choice to determine the roots for the equations in Prob. 7.5.

7.16 Develop a subprogram to solve for the roots of a polynomial using the IMSL routine, ZREAL or a library or package of your choice. Test it by determining the real roots of the equations from Probs. 7.4 and 7.5.

7.17 A two-dimensional circular cylinder is placed in a high-speed uniform flow. Vortices shed from the cylinder at a constant frequency, and pressure sensors on the rear surface of the cylinder detect this frequency by calculating how often the pressure oscillates. Given three data points, use Müller's method to find the time where the pressure was zero.

Time	0.60	0.62	0.64
Pressure	20	50	60

7.18 When trying to find the acidity of a solution of magnesium hydroxide in hydrochloric acid, we obtain the following equation

$$A(x) = x^3 + 3.5x^2 - 40$$

where x is the hydronium ion concentration. Find the hydronium ion concentration for a saturated solution (acidity equals zero) using two different methods in MATLAB (for example, graphically and the `roots` function).

7.19 Consider the following system with three unknowns a, u, and v:

$$u^2 - 2v^2 = a^2$$
$$u + v = 2$$
$$a^2 - 2a - u = 0$$

Solve for the real values of the unknowns using: (a) the Excel Solver and (b) a symbolic manipulator software package.

7.20 In control systems analysis, transfer functions are developed that mathematically relate the dynamics of a system's input to its output. A transfer function for a robotic positioning system is given by

$$G(s) = \frac{C(s)}{N(s)} = \frac{s^3 + 12.5s^2 + 50.5s + 66}{s^4 + 19s^3 + 122s^2 + 296s + 192}$$

where $G(s)$ = system gain, $C(s)$ = system output, $N(s)$ = system input, and s = Laplace transform complex frequency. Use a numerical technique to find the roots of the numerator and denominator and factor these into the form

$$G(s) = \frac{(s+a_1)(s+a_2)(s+a_3)}{(s+b_1)(s+b_2)(s+b_3)(s+b_4)}$$

where a_i and b_i = the roots of the numerator and denominator, respectively.

7.21 Develop an M-file function for bisection in a similar fashion to Fig. 5.10. Test the function by duplicating the computations from Examples 5.3 and 5.4.

7.22 Develop an M-file function for the false-position method. The structure of your function should be similar to the bisection algorithm outlined in Fig. 5.10. Test the program by dupl Example 5.5.

7.23 Develop an M-file function for the Newton-Raphson based on Fig. 6.4 and Sec. 6.2.3. Along with the initial gue the function and its derivative as arguments. Test it by dupl the computation from Example 6.3.

7.24 Develop an M-file function for the secant method b Fig. 6.4 and Sec. 6.3.2. Along with the two initial guesses, p function as an argument. Test it by duplicating the comp from Example 6.6.

7.25 Develop an M-file function for the modified secant based on Fig. 6.4 and Sec. 6.3.2. Along with the initial gu the perturbation fraction, pass the function as an argument. by duplicating the computation from Example 6.8.

CHAPTER 8

Case Studies: Roots of Equations

The purpose of this chapter is to use the numerical procedures discussed in Chaps. 5, 6, and 7 to solve actual engineering problems. Numerical techniques are important for practical applications because engineers frequently encounter problems that cannot be approached using analytical techniques. For example, simple mathematical models that can be solved analytically may not be applicable when real problems are involved. Thus, more complicated models must be employed. For these cases, it is appropriate to implement a numerical solution on a computer. In other situations, engineering design problems may require solutions for implicit variables in complicated equations.

The following case studies are typical of those that are routinely encountered during upper-class courses and graduate studies. Furthermore, they are representative of problems you will address professionally. The problems are drawn from the four major disciplines of engineering: chemical, civil, electrical, and mechanical. These applications also serve to illustrate the trade-offs among the various numerical techniques.

The first application, taken from chemical engineering, provides an excellent example of how root-location methods allow you to use realistic formulas in engineering practice. In addition, it also demonstrates how the efficiency of the Newton-Raphson technique is used to advantage when a large number of root-location computations is required.

The following engineering design problems are taken from civil, electrical, and mechanical engineering. Section 8.2 uses both bracketing and open methods to determine the depth and velocity of water flowing in an open channel. Section 8.3 shows how the roots of transcendental equations can be used in the design of an electrical circuit. Sections 8.2 and 8.3 also illustrate how graphical methods provide insight into the root-location process. Finally, Sec. 8.4 uses polynomial root location to analyze the vibrations of an automobile.

8.1 IDEAL AND NONIDEAL GAS LAWS (CHEMICAL/BIO ENGINEERING)

Background. The *ideal gas law* is given by

$$pV = nRT \tag{8.1}$$

where p is the absolute pressure, V is the volume, n is the number of moles, R is the universal gas constant, and T is the absolute temperature. Although this equation is widely

used by engineers and scientists, it is accurate over only a limited range of pres temperature. Furthermore, Eq. (8.1) is more appropriate for some gases than fo

An alternative equation of state for gases is given by

$$\left(p + \frac{a}{v^2}\right)(v - b) = RT$$

known as the *van der Waals equation*, where $v = V/n$ is the molal volume and a a empirical constants that depend on the particular gas.

A chemical engineering design project requires that you accurately estimate t volume (v) of both carbon dioxide and oxygen for a number of different tempera pressure combinations so that appropriate containment vessels can be selected. It i interest to examine how well each gas conforms to the ideal gas law by compa molal volume as calculated by Eqs. (8.1) and (8.2). The following data are provid

$$R = 0.082054 \text{ L atm/(mol K)}$$

$$\left.\begin{aligned} a &= 3.592 \\ b &= 0.04267 \end{aligned}\right\} \text{ carbon dioxide}$$

$$\left.\begin{aligned} a &= 1.360 \\ b &= 0.03183 \end{aligned}\right\} \text{ oxygen}$$

The design pressures of interest are 1, 10, and 100 atm for temperature combina 300, 500, and 700 K.

Solution. Molal volumes for both gases are calculated using the ideal gas law, wit For example, if $p = 1$ atm and $T = 300$ K,

$$v = \frac{V}{n} = \frac{RT}{p} = 0.082054 \frac{\text{L atm}}{\text{mol K}} \frac{300 \text{ K}}{1 \text{ atm}} = 24.6162 \text{ L/mol}$$

These calculations are repeated for all temperature and pressure combinations sented in Table 8.1.

TABLE 8.1 Computations of molal volume.

Temperature, K	Pressure, atm	Molal Volume (Ideal Gas Law), L/mol	Molal Volume (van der Waals) Carbon Dioxide, L/mol	Molal V (van der Oxy L/m
300	1	24.6162	24.5126	24.5
	10	2.4616	2.3545	2.4
	100	0.2462	0.0795	0.2
500	1	41.0270	40.9821	41.0
	10	4.1027	4.0578	4.1
	100	0.4103	0.3663	0.4
700	1	57.4378	57.4179	57.4
	10	5.7438	5.7242	5.7
	100	0.5744	0.5575	0.5

The computation of molal volume from the van der Waals equation can be accomplished using any of the numerical methods for finding roots of equations discussed in Chaps. 5, 6, and 7, with

$$f(v) = \left(p + \frac{a}{v^2}\right)(v - b) - RT \tag{8.3}$$

In this case, the derivative of $f(v)$ is easy to determine and the Newton-Raphson method is convenient and efficient to implement. The derivative of $f(v)$ with respect to v is given by

$$f'(v) = p - \frac{a}{v^2} + \frac{2ab}{v^3} \tag{8.4}$$

The Newton-Raphson method is described by Eq. (6.6):

$$v_{i+1} = v_i - \frac{f(v_i)}{f'(v_i)}$$

which can be used to estimate the root. For example, using the initial guess of 24.6162, the molal volume of carbon dioxide at 300 K and 1 atm is computed as 24.5126 L/mol. This result was obtained after just two iterations and has an ε_a of less than 0.001 percent.

Similar computations for all combinations of pressure and temperature for both gases are presented in Table 8.1. It is seen that the results for the ideal gas law differ from those for van der Waals equation for both gases, depending on specific values for p and T. Furthermore, because some of these results are significantly different, your design of the containment vessels would be quite different, depending on which equation of state was used.

In this case, a complicated equation of state was examined using the Newton-Raphson method. The results varied significantly from the ideal gas law for several cases. From a practical standpoint, the Newton-Raphson method was appropriate for this application because $f'(v)$ was easy to calculate. Thus, the rapid convergence properties of the Newton-Raphson method could be exploited.

In addition to demonstrating its power for a single computation, the present design problem also illustrates how the Newton-Raphson method is especially attractive when numerous computations are required. Because of the speed of digital computers, the efficiency of various numerical methods for most roots of equations is indistinguishable for a single computation. Even a 1-s difference between the crude bisection approach and the efficient Newton-Raphson does not amount to a significant time loss when only one computation is performed. However, suppose that millions of root evaluations are required to solve a problem. In this case, the efficiency of the method could be a deciding factor in the choice of a technique.

For example, suppose that you are called upon to design an automatic computerized control system for a chemical production process. This system requires accurate estimates of molal volumes on an essentially continuous basis to properly manufacture the final product. Gauges are installed that provide instantaneous readings of pressure and temperature. Evaluations of v must be obtained for a variety of gases that are used in the process.

For such an application, bracketing methods such as bisection or false position would probably be too time-consuming. In addition, the two initial guesses that are required for

these approaches may also interject a critical delay in the procedure. This shortcom relevant to the secant method, which also needs two initial estimates.

In contrast, the Newton-Raphson method requires only one guess for the roo ideal gas law could be employed to obtain this guess at the initiation of the process. assuming that the time frame is short enough so that pressure and temperature do no wildly between computations, the previous root solution would provide a good gue the next application. Thus, the close guess that is often a prerequisite for converge the Newton-Raphson method would automatically be available. All the above cons tions would greatly favor the Newton-Raphson technique for such problems.

8.2 OPEN-CHANNEL FLOW (CIVIL/ENVIRONMENTAL ENGINEERING)

Background. Civil engineering is a broad field that includes such diverse areas as tural, geotechnical, transportation, environmental, and water-resources engineering last two specialties deal with both water pollution and water supply, and hence, extensive use of the science of fluid mechanics.

One general problem relates to the flow of water in open channels such as rive canals. The flow rate, which is routinely measured in most major rivers and streams, fined as the volume of water passing a particular point in a channel per unit time, Q (

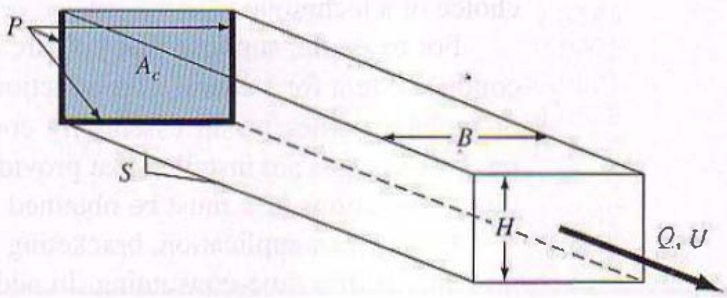
Although the flow rate is a useful quantity, a further question relates to what ha when you put a specific flow rate into a sloping channel (Fig. 8.1). In fact, two thing pen: the water will reach a specific depth H (m) and move at a specific velocity U Environmental engineers might be interested in knowing these quantities to pred transport and fate of pollutants in a river. So the general question is: If you are giv flow rate for a channel, how do you compute the depth and velocity?

Solution. The most fundamental relationship between flow and depth is the *con equation:*

$$Q = UA_c$$

where A_c = the cross-sectional area of the channel (m^2). Depending on the channel the area can be related to the depth by some functional relationship. For the recta

FIGURE 8.1



channel depicted in Fig. 8.1, $A_c = BH$. Substituting this relationship into Eq. (8.5) gives

$$Q = UBH \tag{8.6}$$

where B = the width (m). It should be noted that the continuity equation derives from the *conservation of mass* (recall Table 1.1).

Now, although Eq. (8.6) certainly relates the channel parameters, it is not sufficient to answer our question. Assuming that B is specified, we have one equation and two unknowns (U and H). We therefore require an additional equation. For uniform flow (meaning that the flow does not vary spatially and temporally), the Irish engineer Robert Manning proposed the following semiempirical formula (appropriately called the *Manning equation*)

$$U = \frac{1}{n} R^{2/3} S^{1/2} \tag{8.7}$$

where n = the Manning roughness coefficient (a dimensionless number used to parameterize the channel friction), S = the channel slope (dimensionless, meters drop per meter length), and R = the hydraulic radius (m), which is related to more fundamental parameters by

$$R = \frac{A_c}{P} \tag{8.8}$$

where P = the wetted perimeter (m). As the name implies, the wetted perimeter is the length of the channel sides and bottom that is under water. For example, for a rectangular channel, it is defined as

$$P = B + 2H \tag{8.9}$$

It should be noted that just as the continuity equation derives from the conservation of mass, the Manning equation is an expression of the *conservation of momentum*. In particular, it indicates how velocity is dependent on roughness, a manifestation of friction.

Although the system of nonlinear equations (8.6 and 8.7) can be solved simultaneously (for example, using the multidimensional Newton-Raphson approach described in Sec. 6.5.2), an easier approach would be to combine the equations. Equation (8.7) can be substituted into Eq. (8.6) to give

$$Q = \frac{BH}{n} R^{2/3} S^{1/2} \tag{8.10}$$

Then the hydraulic radius, Eq. (8.8), along with the various relationships for the rectangular channel can be substituted,

$$Q = \frac{S^{1/2}}{n} \frac{(BH)^{5/3}}{(B + 2H)^{2/3}} \tag{8.11}$$

Thus, the equation now contains a single unknown H along with the given value for Q and the channel parameters (n, S, and B).

Although we have one equation with an unknown, it is impossible to solve explicitly for H. However, the depth can be determined numerically by reformulating the equation as

a roots problem,

$$f(H) = \frac{S^{1/2}}{n}\frac{(BH)^{5/3}}{(B+2H)^{2/3}} - Q = 0$$

Equation (8.12) can be solved readily with any of the root-location methods des in Chaps. 5 and 6. For example, if $Q = 5$ m^3/s, $B = 20$ m, $n = 0.03$, and $S = 0.000$ equation is

$$f(H) = 0.471405\frac{(20H)^{5/3}}{(20+2H)^{2/3}} - 5 = 0$$

It can be solved for $H = 0.7023$m. The result can be checked by substitutio Eq. (8.13) to give

$$f(H) = 0.471405\frac{(20 \times 0.7023)^{5/3}}{(20+2 \times 0.7023)^{2/3}} - 5 = 7.8 \times 10^{-5}$$

which is quite close to zero.

Our other unknown, the velocity, can now be determined by substitution bac Eq. (8.6),

$$U = \frac{Q}{BH} = \frac{5}{20(0.7023)} = 0.356 \text{ m/s}$$

Thus, we have successfully solved for the depth and velocity.

Now let us delve a little deeper into the numerical aspects of this problem. One nent question might be: How do we come up with good initial guesses for our nun method? The answer depends on the type of method.

For bracketing methods such as bisection and false position, one approach woul determine whether we can estimate lower and upper guesses that always bracket a root. A conservative approach might be to choose zero as our lower bound. Then, knew the maximum possible depth that could occur, this value could serve as the guess. For example, all but the world's biggest rivers are less than about 10 m deep. fore, we might choose 0 and 10 as our bracket for H.

If $Q > 0$ and $H = 0$, Eq. (8.12) will always be negative for the lower guess. A increased, Eq. (8.12) will increase monotonically and eventually become positive. fore, the guesses should bracket a single root for most cases routinely confronted in rivers and streams.

Now, a technique like bisection should very reliably home in on the root. Bu price is paid? By using such a wide bracket and a technique like bisection, the nun iterations to attain a desirable precision could be computationally excessive. For ex if a tolerance of 0.001 m were chosen, Eq. (5.5) could be used to calculate

$$n = \frac{\log(10/0.001)}{\log 2} = 13.3$$

Thus, 14 iterations would be required. Although this would certainly not be costly fo gle calculation, it could be exorbitant if many such evaluations were made. The alter would be to narrow the initial bracket (based on system-specific knowledge), chan more efficient bracketing technique (like false position), or accept a coarser precisi

Another way to get better efficiency would be to use an open method like the Newton-Raphson or secant methods. Of course, for these cases, the problem of initial guesses is complicated by the issue of convergence.

Insight into these issues can be attained by examining the least efficient of the open approaches—fixed-point iteration. Examining Eq. (8.11), there are two straightforward ways to solve for H, that is, we can solve for either the H in the numerator,

$$H = \frac{(Q_n)^{3/5}(B + 2H)^{2/5}}{BS^{3/10}} \tag{8.16}$$

or the H in the denominator

$$H = \frac{1}{2}\left[\frac{S^3(BH)^{5/2}}{(Q_n)^{3/2}} - B\right] \tag{8.17}$$

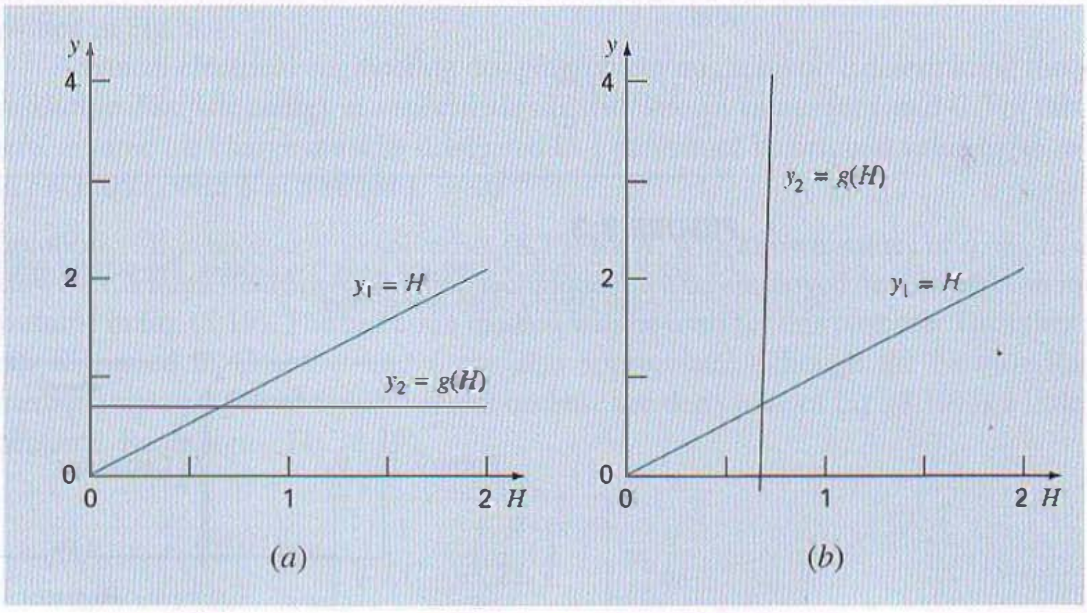
Now, here is where physical reasoning can be helpful. For most rivers and streams, the width is much greater than the depth. Thus, the quantity $B + 2H$ should not vary much. In fact, it should be roughly equal to B. In comparison, BH is directly proportional to H. Consequently, Eq. (8.16) should home in more rapidly on the root. This can be verified by substituting the brackets of $H = 0$ and 10 into both equations. For Eq. (8.16), the results are 0.6834 and 0.9012, which are both close to the true value of 0.7023. In contrast, the results for Eq. (8.17) are -10 and 8,178, which clearly are distant from the root.

The superiority of Eq. (8.16) is further supported by component plots (recall Fig. 6.3). As in Fig. 8.2, the $g(H)$ component for Eq. (8.16) is almost flat. Thus, it will not only converge, but should do so rapidly. In contrast, the $g(H)$ component for Eq. (8.17) is almost vertical, connoting strong and rapid divergence.

There are two practical benefits to such an analysis:

1. In the event that a more refined open method were used, Eq. (8.16) provides a means to develop an excellent starting guess. For example, if H is chosen as zero, Eq. (8.12)

ls for two cases
eration, one that
a), Eq. (8.16)]
ill diverge
].



becomes

$$H_0 = \frac{(Qn/B)^{3/5}}{S^{3/10}}$$

where H_0 would be the initial value used in the Newton-Raphson or secant m

2. We have demonstrated that fixed-point iteration provides a viable option f particular problem. For example, using an initial guess of $H = 0$, Eq. (8.16) six digits of precision in four iterations for the case we are examining. One where a fixed-point formula might come in handy would be a spreadsheet cation. That is, spreadsheets are ideal for a convergent, iterative formula that d on a single cell.

8.3 DESIGN OF AN ELECTRIC CIRCUIT (ELECTRICAL ENGINEERING)

Background. Electrical engineers often use Kirchhoff's laws to study the steac (not time-varying) behavior of electric circuits. Such steady-state behavior will be ined in Sec. 12.3. Another important problem involves circuits that are transient ir where sudden temporal changes take place. Such a situation occurs following the of the switch in Fig. 8.3. In this case, there will be a period of adjustment follow closing of the switch as a new steady state is reached. The length of this adjustment is closely related to the storage properties of the capacitor and the inductor. Energy may oscillate between these two elements during a transient period. However, resist the circuit will dissipate the magnitude of the oscillations.

The flow of current through the resistor causes a voltage drop (V_R) given by

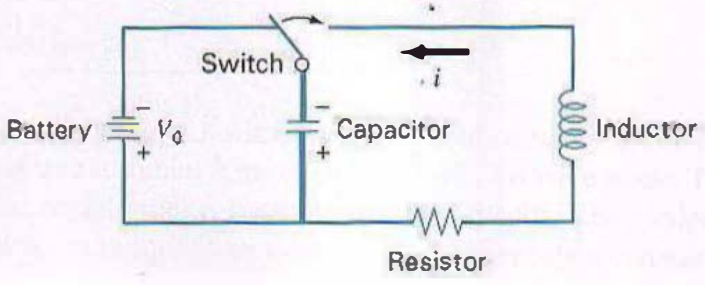
$$V_R = iR$$

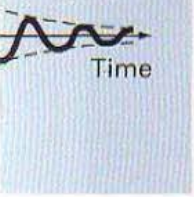
where i = the current and R = the resistance of the resistor. When R and i have ohms and amperes, respectively, V_R has units of volts.

Similarly, an inductor resists changes in current, such that the voltage drop V it is

$$V_L = L\frac{di}{dt}$$

FIGURE 8.3
An electric circuit. When the switch is closed, the current will undergo a series of oscillat until a new steady state is reached.





where L = the inductance. When L and i have units of henrys and amperes, respectively, V_L has units of volts and t has units of seconds.

The voltage drop across the capacitor (V_C) depends on the charge (q) on it:

$$V_C = \frac{q}{C}$$

where C = the capacitance. When the charge is expressed in units of coulombs, the unit of C is the farad.

Kirchhoff's second law states that the algebraic sum of voltage drops around a closed circuit is zero. After the switch is closed we have

$$L\frac{di}{dt} + Ri + \frac{q}{C} = 0$$

However, the current is related to the charge according to

$$i = \frac{dq}{dt}$$

Therefore,

$$L\frac{d^2q}{dt^2} + R\frac{dq}{dt} + \frac{1}{C}q = 0 \tag{8.18}$$

This is a second-order linear ordinary differential equation that can be solved using the methods of calculus (see Sec. 8.4). This solution is given by

$$q(t) = q_0 e^{-Rt/(2L)} \cos\left[\sqrt{\frac{1}{LC} - \left(\frac{R}{2L}\right)^2}\, t\right] \tag{8.19}$$

where at $t = 0$, $q = q_0 = V_0 C$, and V_0 = the voltage from the charging battery. Equation (8.19) describes the time variation of the charge on the capacitor. The solution $q(t)$ is plotted in Fig. 8.4.

A typical electrical engineering design problem might involve determining the proper resistor to dissipate energy at a specified rate, with known values for L and C. For this problem, assume the charge must be dissipated to 1 percent of its original value ($q/q_0 = 0.01$) in $t = 0.05$ s, with $L = 5$ H and $C = 10^{-4}$ F.

Solution. It is necessary to solve Eq. (8.19) for R, with known values of q, q_0, L, and C. However, a numerical approximation technique must be employed because R is an implicit variable in Eq. (8.19). The bisection method will be used for this purpose. The other methods discussed in Chaps. 5 and 6 are also appropriate, although the Newton-Raphson method might be deemed inconvenient because the derivative of Eq. (8.19) is a little cumbersome. Rearranging Eq. (8.19),

$$f(R) = e^{-Rt/(2L)} \cos\left[\sqrt{\frac{1}{LC} - \left(\frac{R}{2L}\right)^2}\, t\right] - \frac{q}{q_0}$$

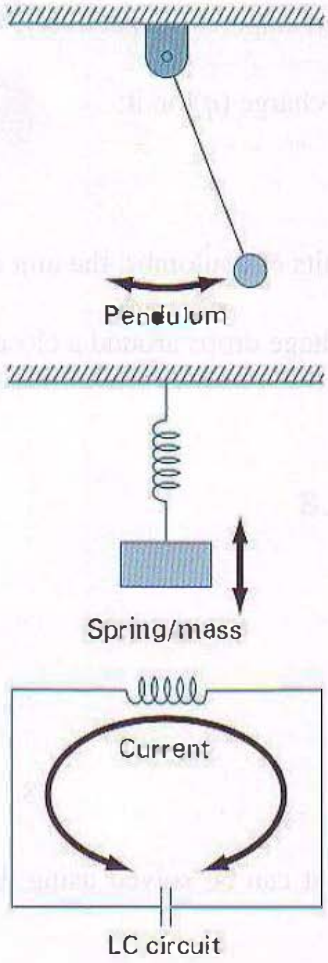


FIGURE 8.6
Three examples of simple harmonic oscillators. The two-way arrows illustrate the oscillations for each system.

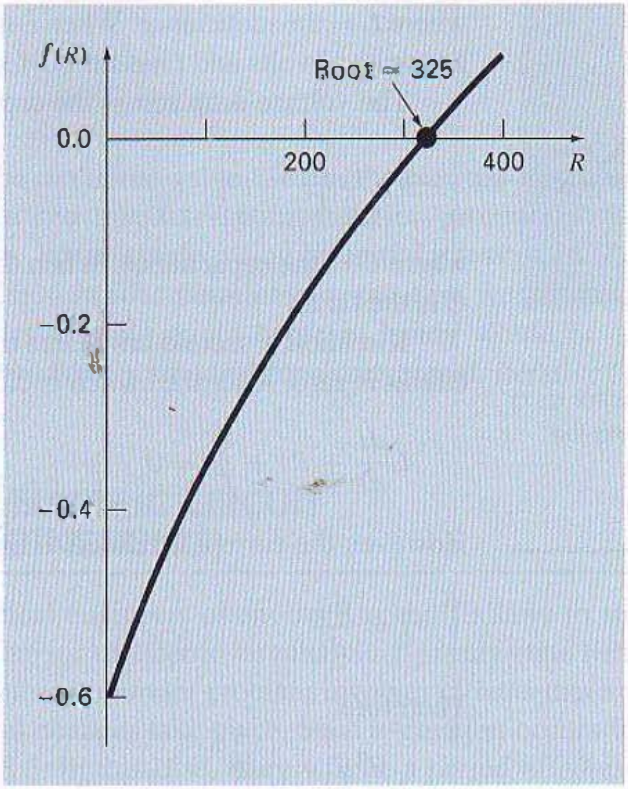


FIGURE 8.5
Plot of Eq. (8.20) used to obtain initial guesses for R that bracket the root.

or using the numerical values given,

$$f(R) = e^{-0.005R}\cos[\sqrt{2000 - 0.01R^2}\,(0.05)] - 0.01$$

Examination of this equation suggests that a reasonable initial range for R is 0 to (because $2000 - 0.01R^2$ must be greater than zero). Figure 8.5, a plot of Eq. (8.2 firms this. Twenty-one iterations of the bisection method give $R = 328.1515\ \Omega$, error of less than 0.0001 percent.

Thus, you can specify a resistor with this rating for the circuit shown in Fig. expect to achieve a dissipation performance that is consistent with the requirement problem. This design problem could not be solved efficiently without using the nu methods in Chaps. 5 and 6.

8.4 VIBRATION ANALYSIS (MECHANICAL/AEROSPACE ENGINEERING)

Background. Differential equations are often used to model the vibration of engi systems. Some examples (Fig. 8.6) are a simple pendulum, a mass on a spring, inductance-capacitance electric circuit (recall Sec. 8.3). The vibration of these may be damped by some energy-absorbing mechanism. In addition, the vibrati be free or subject to some external periodic disturbance. In the latter case the m said to be *forced*. In this section, we will examine the free and forced vibration

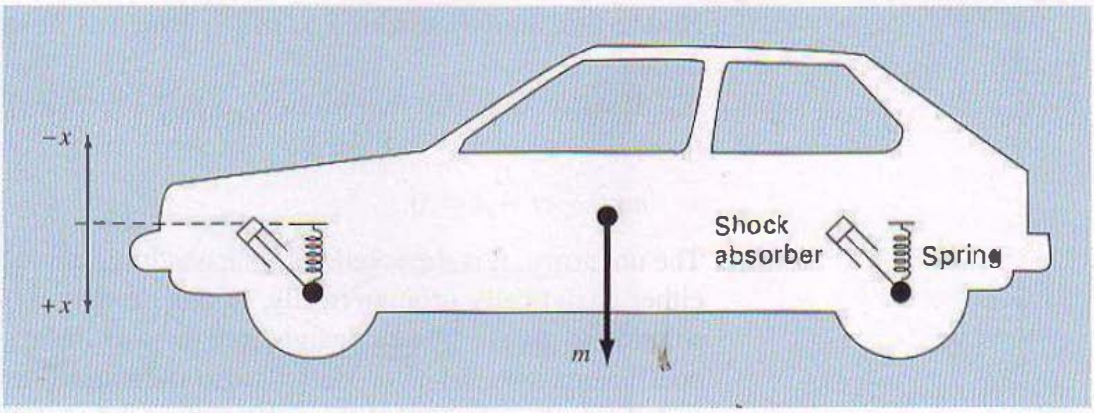


FIGURE 8.7
A car of mass m.

automobile shown in Fig. 8.7. The general approach is applicable to various other engineering problems.

As shown in Fig. 8.7, a car of mass m is supported by springs and shock absorbers. Shock absorbers offer resistance to the motion that is p oportional to the vertical speed (up-and-down motion). Free vibrations result when the car is disturbed from equilib ium, such as after encountering a pothole. At any instant after hitting the pothole the net forces acting on m are the resistance of the springs and the damping force of the shock absorbers. These forces tend to return the car to the original equilibrium state. According to *Hooke's law*, the resistance of the spring is proportional to the spring constant k and the distance from the equilibrium position, x. Therefore,

$$\text{Spring force} = -kx$$

where the negative sign indicates that the restoring force acts to return the car toward the position of equilibrium (that is, the negative x direction). The damping force of the shock absorbers is given by

$$\text{Damping force} = -c\frac{dx}{dt}$$

where c is a damping coefficient and dx/dt is the vertical velocity. The negative sign indicates that the damping force acts in the opposite direction against the velocity.

The equations of motion for the system are given by Newton's second law ($F = ma$), which for the present problem is expressed as

$$\underbrace{m}_{\text{Mass}} \times \underbrace{\frac{d^2x}{dt^2}}_{\text{acceleration}} = \underbrace{-c\frac{dx}{dt}}_{\text{damping force}} + \underbrace{(-kx)}_{\text{spring force}}$$

or

$$m\frac{d^2x}{dt^2} + c\frac{dx}{dt} + kx = 0$$

Thus, using the above results and Fig. 8.10, it is found that the proposed car design will behave acceptably for common driving speeds. At this point, the designer must be aware that the design would not meet suitability requirements at extremely high speeds (for example, racing).

This design problem has presented an extremely simple example that has allowed us to obtain some analytical results that were used to evaluate the accuracy of our numerical methods for finding roots. Real cases can quickly become so complicated that solutions can be obtained only by using numerical methods.

OBLEMS

Engineering

same computation as in Sec. 8.1, but for ethyl al- 2 and $b = 0.08407$) at a temperature of 400 K and mpare your results with the ideal gas law. Use any l methods discussed in Chaps. 5 and 6 to perform . Justify your choice of technique.

engineering, plug flow reactors (that is, those in vs from one end to the other with minimal mixing udinal axis) are often used to convert reactants into been determined that the efficiency of the conver- nes be improved by recycling a portion of the prod- at it returns to the entrance for an additional pass tor (Fig. P8.2). The recycle rate is defined as

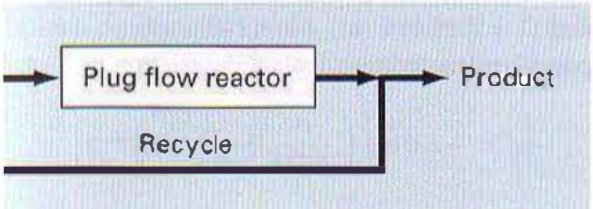
$$\frac{\text{of fluid returned to entrance}}{\text{ume leaving the system}}$$

are processing a chemical A to generate a product B. re A forms B according to an autocatalytic reaction n one of the products acts as a catalyst or stimulus). it can be shown that an optimal recycle rate

$$\frac{X_{Af})}{X_{Af})} = \frac{R+1}{R[1+R(1-X_{Af})]}$$

e fraction of reactant A that is converted to product recycle rate corresponds to the minimum-sized to attain the desired level of conversion. Use a

esentation of a plug flow reactor with recycle.



numerical method to determine the recycle ratios needed to minimize reactor size for a fractional conversion of $X_{Af} = 0.95$.

8.3 In a chemical engineering process, water vapor (H_2O) is heated to sufficiently high temperatures that a significant portion of the water dissociates, or splits apart, to form oxygen (O_2) and hydrogen (H_2):

$$H_2O \rightleftharpoons H_2 + \tfrac{1}{2}O_2$$

If it is assumed that this is the only reaction involved, the mole fraction x of H_2O that dissociates can be represented by

$$K = \frac{x}{1-x}\sqrt{\frac{2p_t}{2+x}} \tag{P8.3}$$

where K = the reaction equilibrium constant and p_t = the total pressure of the mixture. If $p_t = 3.5$ atm and $K = 0.04$, determine the value of x that satisfies Eq. (P8.3).

8.4 The following equation pertains to the concentration of a chemical in a completely mixed reactor:

$$c = c_{in}(1 - e^{-0.04t}) + c_0 e^{-0.04t}$$

If the initial concentration $c_0 = 5$ and the inflow concentration $c_{in} = 12$, compute the time required for c to be 85 percent of c_{in}.

8.5 A reversible chemical reaction

$$2A + B \rightleftharpoons C$$

can be characterized by the equilibrium relationship

$$K = \frac{c_c}{c_a^2 c_b}$$

where the nomenclature c_i represents the concentration of constituent i. Suppose that we define a variable x as representing the number of moles of C that are produced. Conservation of mass can be used to reformulate the equilibrium relationship as

$$K = \frac{(c_{c,0} + x)}{(c_{a,0} - 2x)^2 (c_{b,0} - x)}$$

where the subscript 0 designates the initial concentration of each constituent. If $K = 0.016$, $c_{a,0} = 42$, $c_{b,0} = 28$, and $c_{c,0} = 4$,

determine the value of x. (a) Obtain the solution graphically. (b) On the basis of (a), solve for the root with initial guesses of $x_l = 0$ and $x_u = 20$ to $\varepsilon_s = 0.5\%$. Choose either bisection or false position to obtain your solution. Justify your choice.

8.6 The following chemical reactions take place in a closed system

$$2A + B \rightleftharpoons C$$
$$A + D \rightleftharpoons C$$

At equilibrium, they can be characterized by

$$K_1 = \frac{c_c}{c_a^2 c_b}$$

$$K_2 = \frac{c_c}{c_a c_d}$$

where the nomenclature c_i represents the concentration of constituent i. If x_1 and x_2 are the number of moles of C that are produced due to the first and second reactions, respectively, use an approach similar to that of Prob. 8.5 to reformulate the equilibrium relationships in terms of the initial concentrations of the constituents. Then, use the Newton-Raphson method to solve the pair of simultaneous nonlinear equations for x_1 and x_2 if $K_1 = 4 \times 10^{-4}$, $K_2 = 3.7 \times 10^{-2}$, $c_{a,0} = 50$, $c_{b,0} = 20$, $c_{c,0} = 5$, and $c_{d,0} = 10$. Use a graphical approach to develop your initial guesses.

8.7 The Redlich-Kwong equation of state is given by

$$p = \frac{RT}{v - b} - \frac{a}{v(v + b)\sqrt{T}}$$

where R = the universal gas constant [= 0.518 kJ/(kg K)], T = absolute temperature (K), p = absolute pressure (kPa), and v = the volume of a kg of gas (m^3/kg). The parameters a and b are calculated by

$$a = 0.427\frac{R^2 T_c^{2.5}}{p_c} \qquad b = 0.0866\frac{RT_c}{p_c}$$

where p_c = critical pressure (kPa) and T_c = critical temperature (K). As a chemical engineer, you are asked to determine the amount of methane fuel (p_c = 4580 kPa and T_c = 191 K) that can be held in a 3-m^3 tank at a temperature of −50°C with a pressure of 65,000 kPa. Use a root-locating method of your choice to calculate v and then determine the mass of methane contained in the tank.

8.8 The volume V of liquid in a hollow horizontal cylinder of radius r and length L is related to the depth of the liquid h by

$$V = \left[r^2 \cos^{-1}\left(\frac{r - h}{r}\right) - (r - h)\sqrt{2rh - h^2}\right] L$$

Determine h given r = 2 m, L = 5 m, and V = 8.5 m^3. Note that if you are using a programming language or software tool that is not rich in trigonometric functions, the arc cosine can be co with

$$\cos^{-1} x = \frac{\pi}{2} - \tan^{-1}\left(\frac{x}{\sqrt{1 - x^2}}\right)$$

8.9 The volume V of liquid in a spherical tank of radius r is to the depth h of the liquid by

$$V = \frac{\pi h^2 (3r - h)}{3}$$

Determine h given r = 1 m and V = 0.75 m^3.

8.10 For the spherical tank in Prob. 8.9, it is possible to the following two fixed-point formulas:

$$h = \sqrt{\frac{h^3 + (3V/\pi)}{3r}}$$

and

$$h = \sqrt[3]{3\left(rh^2 - \frac{V}{\pi}\right)}$$

If r = 1 m and V = 0.75 m^3, determine whether either of stable, and the range of initial guesses for which they are st

8.11 The Ergun equation, shown below, is used to descr flow of a fluid through a packed bed. ΔP is the pressure dr the density of the fluid, G_o is the mass velocity (mass fl divided by cross-sectional area), D_p is the diameter of the p within the bed, μ is the fluid viscosity, L is the length of t and ε is the void fraction of the bed.

$$\frac{\Delta P \rho}{G_o^2}\frac{D_p}{L}\frac{\varepsilon^3}{1 - \varepsilon} = 150\frac{1 - \varepsilon}{(D_p G_o/\mu)} + 1.75$$

Given the parameter values listed below, find the void fracti the bed.

$$\frac{D_p G_o}{\mu} = 1000$$

$$\frac{\Delta P \rho D_p}{G_o^2 L} = 20$$

8.12 The pressure drop in a section of pipe can be calcul

$$\Delta p = f\frac{L\rho V^2}{2D}$$

where Δp = the pressure drop (Pa), f = the friction fact the length of pipe [m], ρ = density (kg/m^3), V = velocity and D = diameter (m). For turbulent flow, the *Colebrook e* provides a means to calculate the friction factor,

$$\frac{1}{\sqrt{f}} = -2.0\log\left(\frac{\varepsilon}{3.7D} + \frac{2.51}{\mathrm{Re}\sqrt{f}}\right)$$

ughness (m), and Re = the *Reynolds number*:

nic viscosity (N · s/m²).

Δp for a 0.2-m-long horizontal stretch of smooth given $\rho = 1.23$ kg/m³, $\mu = 1.79 \times 10^{-5}$ N · s/m², n, $V = 40$ m/s, and $\varepsilon = 0.0015$ mm. Use a numer- to determine the friction factor. Note that smooth $e < 10^5$, a good initial guess can be obtained using ormula, $f = 0.316/\mathrm{Re}^{0.25}$.

computation but for a rougher commercial steel 045 mm).

water has great significance to environmental and ers. It can be related to processes ranging from acid rain. The pH is related to the hydrogen ion

$[H^+]$

ive equations govern the concentrations of a dioxide and water for a closed system:

$\frac{[CO_3^-]}{O_2]}$

$\frac{CO_3^{2-}]}{O_3^-]}$

$OH^-]$

$[HCO_3^-] + [CO_3^{2-}]$

$^-] + 2[CO_3^{2-}] + [OH^-] - [H^+]$

alkalinity, c_T = total inorganic carbon, and the m coefficients. The five unknowns are $[CO_2] =$ $[HCO_3^-]$ = bicarbonate, $[CO_3^{2-}]$ = carbonate, n ion, and $[OH^-]$ = hydroxyl ion. Solve for the given that Alk $= 2 \times 10^{-3}$, $c_T = 3 \times 10^{-3}$, $= 10^{-10.3}$, and $K_w = 10^{-14}$. Also, calculate the

on of a constant density plug flow reactor for the ubstance via an enzymatic reaction is described by w, where V is the volume of the reactor, F is the nt C, C_{in} and C_{out} are the concentrations of reac- leaving the reactor, respectively, and K and k_{max} r a 500-L reactor, with an inlet concentration of let flow rate of 40 L/s, $k_{max} = 5 \times 10^{-3}$ s^{-1}, and the concentration of C at the outlet of the reactor.

$$\frac{K}{k_{max}C} + \frac{1}{k_{max}} dC$$

Civil and Environmental Engineering

8.15 The displacement of a structure is defined by the following equation for a damped oscillation:

$$y = 9e^{-kt}\cos\omega t$$

where $k = 0.7$ and $\omega = 4$.

(a) Use the graphical method to make an initial estimate of the time required for the displacement to decrease to 3.5.

(b) Use the Newton-Raphson method to determine the root to $\varepsilon_s = 0.01\%$.

(c) Use the secant method to determine the root to $\varepsilon_s = 0.01\%$.

8.16 In structural engineering, the *secant formula* defines the force per unit area, P/A, that causes a maximum stress σ_m in a column of given slenderness ratio L/k:

$$\frac{P}{A} = \frac{\sigma_m}{1 + (ec/k^2)\sec[0.5\sqrt{P/(EA)}(L/k)]}$$

where ec/k^2 = the eccentricity ratio and E = the modulus of elasticity. If for a steel beam, $E = 200{,}000$ MPa, $ec/k^2 = 0.4$, and $\sigma_m = 250$ MPa, compute P/A for $L/k = 50$. Recall that $\sec x = 1/\cos x$.

8.17 A catenary cable is one that is hung between two points not in the same vertical line. As depicted in Fig. P8.17*a*, it is subject to no loads other than its own weight. Thus, its weight (N/m) acts as a uniform load per unit length along the cable. A free-body diagram of a section AB is depicted in Fig. P8.17*b*, where T_A and T_B are the tension forces at the end. Based on horizontal and vertical force balances, the following differential equation model of the cable can be derived:

$$\frac{d^2y}{dx^2} = \frac{w}{T_A}\sqrt{1 + \left(\frac{dy}{dx}\right)^2}$$

Calculus can be employed to solve this equation for the height y of the cable as a function of distance x,

$$y = \frac{T_A}{w}\cosh\left(\frac{w}{T_A}x\right) + y_0 - \frac{T_A}{w}$$

where the hyperbolic cosine can be computed by

$$\cosh x = \frac{1}{2}(e^x + e^{-x})$$

Use a numerical method to calculate a value for the parameter T_A given values for the parameters $w = 12$ and $y_0 = 6$, such that the cable has a height of $y = 15$ at $x = 50$.

8.18 Figure P8.18*a* shows a uniform beam subject to a linearly increasing distributed load. The equation for the resulting elastic curve is (see Fig. P8.18*b*)

$$y = \frac{w_0}{120EIL}(-x^5 + 2L^2x^3 - L^4x) \qquad \text{(P8.18)}$$

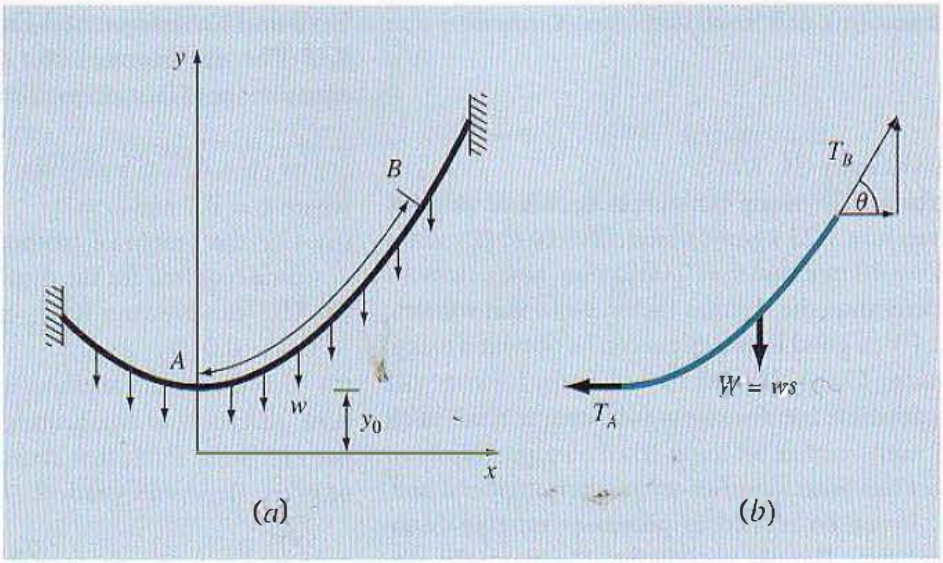


Figure P8.17
(a) Forces acting on a section AB of a flexible hanging cable. The load is uniform along the cable (but not uniform per the horizontal distance x). (b) A free-body diagram of section AB.

Use bisection to determine the point of maximum deflection (that is, the value of x where $dy/dx = 0$). Then substitute this value into Eq. (P8.18) to determine the value of the maximum deflection. Use the following parameter values in your computation: $L = 600$ cm, $E = 50{,}000$ kN/cm^2, $I = 30{,}000$ cm^4, and $w_0 = 2.5$ kN/cm.

8.19 In environmental engineering (a specialty area in civil engineering), the following equation can be used to compute the oxygen level c (mg/L) in a river downstream from a sewage discharge:

$$c = 10 - 20(e^{-0.15x} - e^{-0.5x})$$

where x is the distance downstream in kilometers.

(a) Determine the distance downstream where the oxyge first falls to a reading of 5 mg/L. (Hint: It is within 2 kr discharge.) Determine your answer to a 1% error. Note t els of oxygen below 5 mg/L are generally harmful to g such as trout and salmon.

(b) Determine the distance downstream at which the oxyge minimum. What is the concentration at that location?

8.20 The concentration of pollutant bacteria c in a lake de according to

$$c = 75e^{-1.5t} + 20e^{-0.075t}$$

Determine the time required for the bacteria concentratio reduced to 15 using (a) the graphical method and (b) us Newton-Raphson method with an initial guess of $t = 6$ stopping criterion of 0.5%. Check your result.

8.21 In ocean engineering, the equation for a reflected s wave in a harbor is given by $\lambda = 16$, $t = 12$, $v = 48$:

$$h = h_0\left[\sin\left(\frac{2\pi x}{\lambda}\right)\cos\left(\frac{2\pi t v}{\lambda}\right) + e^{-x}\right]$$

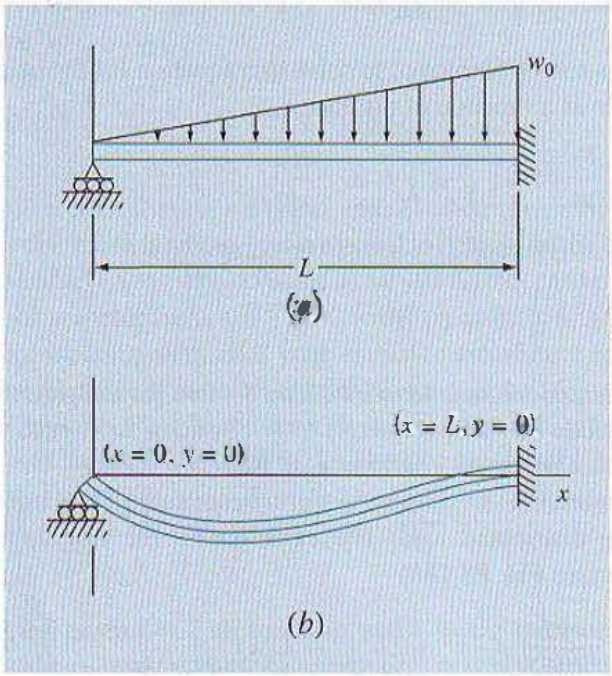
Solve for the lowest positive value of x if $h = 0.5h_0$.

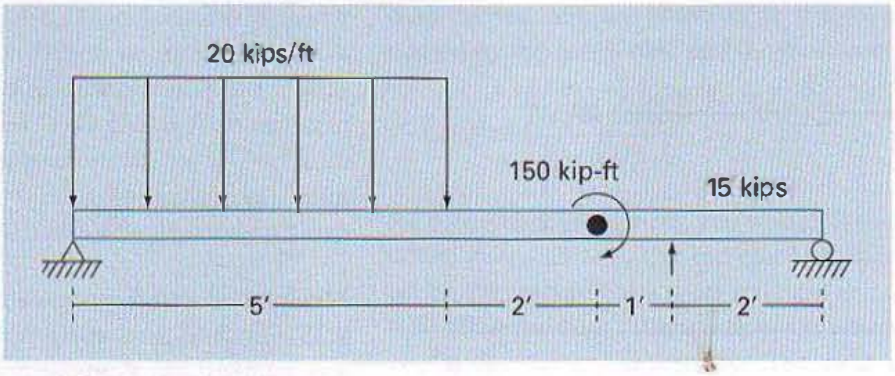
8.22 You buy a $25,000 piece of equipment for nothing do $5,500 per year for 6 years. What interest rate are you payi formula relating present worth P, annual payments A, nu years n, and interest rate i is

$$A = P\frac{i(1+i)^n}{(1+i)^n - 1}$$

8.23 Many fields of engineering require accurate pop estimates. For example, transportation engineers might

Figure P8.18





mine separately the population growth trends of a t suburb. The population of the urban area is ne according to

$$e^{-k_u t} + P_{u,\min}$$

n population is growing, as in

$$\frac{P_{s,\max}}{P_{s,\max}/P_0 - 1]e^{-k_s t}}$$

$P_{s,\max}$, P_0, and k_s = empirically derived parame- ne time and corresponding values of $P_u(t)$ and burbs are 20% larger than the city. The parameter $_s = 75{,}000$, $k_u = 0.045/\text{yr}$, $P_{u,\min} = 100{,}000$ 300,000 people, $P_0 = 10{,}000$ people, $k_s =$ n your solutions, use (a) graphical, (b) false- odified secant methods.

pported beam is loaded as shown in Fig. P8.24. functions, the shear along the beam can be quation:

$$- 0\rangle^1 - \langle x - 5\rangle^1] - 15\langle x - 8\rangle^0 - 57$$

singularity function can be expressed as follows:

$$\left.\begin{array}{ll} x - a)^n & \text{when } x > a \\ 0 & \text{when } x \le a \end{array}\right\}$$

ethod to find the point(s) where the shear equals

imply supported beam from Prob. 8.24, the beam, $M(x)$, is given by:

$$x - 0\rangle^2 - \langle x - 5\rangle^2] + 15\langle x - 8\rangle^1$$

$$\langle x - 7\rangle^0 + 57x$$

nethod to find the point(s) where the moment

8.26 Using the simply supported beam from Prob. 8.24, the slope along the beam is given by:

$$\frac{du_y}{dx}(x) = \frac{-10}{3}[\langle x - 0\rangle^3 - \langle x - 5\rangle^3] + \frac{15}{2}\langle x - 8\rangle^2 + 150\langle x - 7\rangle^1 + \frac{57}{2}x^2 - 238.25$$

Use a numerical method to find the point(s) where the slope equals zero.

8.27 Using the simply supported beam from Prob. 8.24, the displacement along the beam is given by:

$$u_y(x) = \frac{-5}{6}[\langle x - 0\rangle^4 - \langle x - 5\rangle^4] + \frac{15}{6}\langle x - 8\rangle^3 + 75\langle x - 7\rangle^2 + \frac{57}{6}x^3 - 238.25x$$

(a) Find the point(s) where the displacement equals zero.
(b) How would you use a root location technique to determine the location of the minimum displacement?

Electrical Engineering

8.28 Perform the same computation as in Sec. 8.3, but determine the value of C required for the circuit to dissipate to 1% of its original value in $t = 0.05$ s, given $R = 280\Omega$, and $L = 7.5$ H. Use (a) a graphical approach, (b) bisection, and (c) root location software such as the Excel Solver or the MATLAB function `fzero`.

8.29 An oscillating current in an electric circuit is described by $i = 9e^{-t}\cos(2\pi t)$, where t is in seconds. Determine all values of t such that $i = 3$.

8.30 The resistivity ρ of doped silicon is based on the charge q on an electron, the electron density n, and the electron mobility μ. The electron density is given in terms of the doping density N and the intrinsic carrier density n_i. The electron mobility is described by the temperature T, the reference temperature T_0, and the

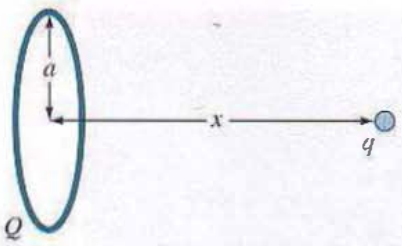


Figure P8.31

reference mobility μ_0. The equations required to compute the resistivity are

$$\rho = \frac{1}{qn\mu}$$

where

$$n = \frac{1}{2}\left(N + \sqrt{N^2 + 4n_i^2}\right) \qquad \text{and} \qquad \mu = \mu_0 \left(\frac{T}{T_0}\right)^{-2.42}$$

Determine N, given $T_0 = 300$ K, $T = 1000$ K, $\mu_0 = 1350$ cm^2 (V s)$^{-1}$, $q = 1.7 \times 10^{-19}$ C, $n_i = 6.21 \times 10^9$ cm^{-3}, and a desired $\rho = 6.5 \times 10^6$ V s cm/C. Use (a) bisection and (b) the modified secant method.

8.31 A total charge Q is uniformly distributed around a ring-shaped conductor with radius a. A charge q is located at a distance x from the center of the ring (Fig. P8.31). The force exerted on the charge by the ring is given by

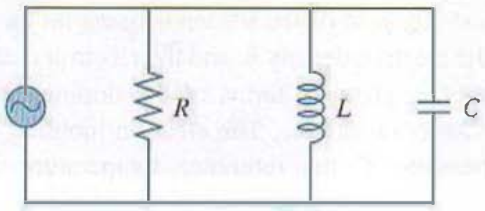
$$F = \frac{1}{4\pi\varepsilon_0} \frac{qQx}{(x^2 + a^2)^{3/2}}$$

where $\varepsilon_0 = 8.85 \times 10^{-12}$ C^2/(N m^2). Find the distance x where the force is 1.25 N if q and Q are 2×10^{-5} C for a ring with a radius of 0.9 m.

8.32 Figure P8.32 shows a circuit with a resistor, an inductor, and a capacitor in parallel. Kirchhoff's rules can be used to express the impedance of the system as

$$\frac{1}{Z} = \sqrt{\frac{1}{R^2} + \left(\omega C - \frac{1}{\omega L}\right)^2}$$

Figure P8.32



where Z = impedance (Ω) and ω = the angular frequency ω that results in an impedance of 75 Ω using both bise false position with initial guesses of 1 and 1000 for the parameters: $R = 225\ \Omega$, $C = 0.6 \times 10^{-6}$ F, and $L = 0.5$ mine how many iterations of each technique are necessar mine the answer to $\varepsilon_s = 0.1\%$. Use the graphical ap explain any difficulties that arise.

Mechanical and Aerospace Engineering

8.33 For fluid flow in pipes, friction is described by a d less number, the *Fanning friction factor f*. The Fannin factor is dependent on a number of parameters related to the pipe and the fluid, which can all be represented b dimensionless quantity, the *Reynolds number* Re. A for predicts f given Re is the *von Karman equation*,

$$\frac{1}{\sqrt{f}} = 4\log_{10}(\mathrm{Re}\sqrt{f}) - 0.4$$

Typical values for the Reynolds number for turbulent 10,000 to 500,000 and for the Fanning friction factor ar 0.01. Develop a function that uses bisection to solve for user-supplied value of Re between 2,500 and 1,000,000. D function so that it ensures that the absolute error in the $E_{a,d} < 0.000005$.

8.34 Real mechanical systems may involve the deflectio linear springs. In Fig. P8.34, a mass m is released a d above a nonlinear spring. The resistance force F of the given by

$$F = -\left(k_1 d + k_2 d^{3/2}\right)$$

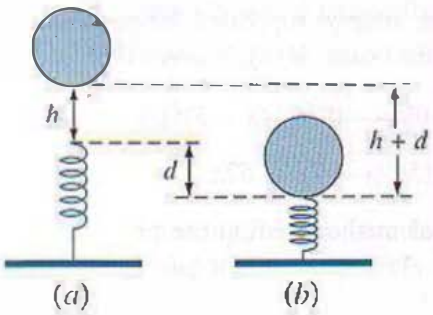
Conservation of energy can be used to show that

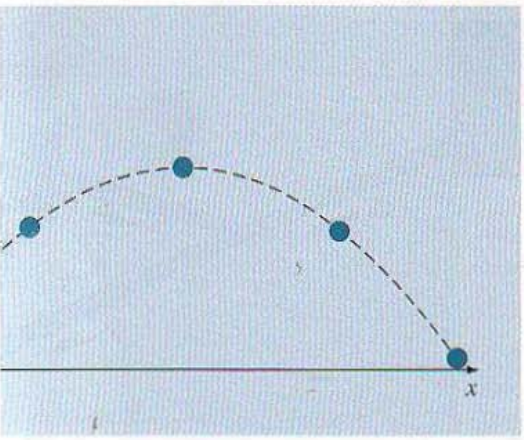
$$0 = \frac{2k_2 d^{5/2}}{5} + \frac{1}{2}k_1 d^2 - mgd - mgh$$

Solve for d, given the following parameter values: $k_1 = 5$ $k_2 = 40$ g/(s^2 m$^{0.5}$), $m = 90$ g, $g = 9.81$ m/s^2, and $h = 0$

8.35 Mechanical engineers, as well as most other engi thermodynamics extensively in their work. The

Figure P8.34





c used to relate the zero-pressure specific heat of
K), to temperature (K):

$$+ 1.671 \times 10^{-4}T + 9.7215 \times 10^{-8}T^2$$
$$\times 10^{-11}T^3 + 1.9520 \times 10^{-14}T^4$$

mperature that corresponds to a specific heat of

engineers sometimes compute the trajectories of
ckets. A related problem deals with the trajectory
The trajectory of a ball is defined by the (x, y)
isplayed in Fig. P8.36. The trajectory can be

$$- \frac{g}{2v_0^2 \cos^2 \theta_0} x^2 + y_0$$

riate initial angle θ_0, if the initial velocity
he distance to the catcher x is 35 m. Note that the
ower's hand at an elevation of $y_0 = 2$ m and the
at 1 m. Express the final result in degrees. Use a
's^2 for g and employ the graphical method to
al guesses.

velocity of a rocket can be computed by the
t:

$$\frac{}{qt} - gt$$

rd velocity, u = the velocity at which fuel is
o the rocket, m_0 = the initial mass of the rocket
= the fuel consumption rate, and g = the down-
of gravity (assumed constant = 9.81 m/s^2). If
= 150,000 kg, and q = 2700 kg/s, compute the
= 750 m/s. (Hint: t is somewhere between 10 and
our result so that it is within 1% of the true value.
er.

8.38 In Sec. 8.4, the phase angle ϕ between the forced vibration caused by the rough road and the motion of the car is given by

$$\tan\phi = \frac{2(c/c_c)(\omega/p)}{1 - (\omega/p)^2}$$

As a mechanical engineer, you would like to know if there are cases where $\phi = \omega/3 - 1$. Use the other parameters from the section to set up the equation as a roots problem and solve for ω.

8.39 Two fluids at different temperatures enter a mixer and come out at the same temperature. The heat capacity of fluid A is given by:

$$c_p = 3.381 + 1.804 \times 10^{-2}T - 4.300 \times 10^{-6}T^2$$

and the heat capacity of fluid B is given by:

$$c_p = 8.592 + 1.290 \times 10^{-1}T - 4.078 \times 10^{-5}T^2$$

where c_p is in units of cal/mol K, and T is in units of K. Note that

$$\Delta H = \int_{T_1}^{T_2} c_p dT$$

A enters the mixer at 400°C. B enters the mixer at 700°C. There is twice as much A as there is B entering into the mixer. At what temperature do the two fluids exit the mixer?

8.40 A compressor is operating at compression ratio R_c of 3.0 (the pressure of gas at the outlet is three times greater than the pressure of the gas at the inlet). The power requirements of the compressor H_p can be determined from the equation below. Assuming that the power requirements of the compressor are exactly equal to zRT_1/MW, find the polytropic efficiency n of the compressor. The parameter z is compressibility of the gas under operating conditions of the compressor, R is the gas constant, T_1 is the temperature of the gas at the compressor inlet, and MW is the molecular weight of the gas.

$$\text{HP} = \frac{zRT_1}{\text{MW}} \frac{n}{n-1} \left(R_c^{(n-1)/n} - 1\right)$$

8.41 In the thermos shown in Fig. P8.41, the innermost compartment is separated from the middle container by a vacuum. There is a final shell around the thermos. This final shell is separated from the middle layer by a thin layer of air. The outside of the final shell comes in contact with room air. Heat transfer from the inner compartment to the next layer q_1 is by radiation only (since the space is evacuated). Heat transfer between the middle layer and outside shell q_2 is by convection in a small space. Heat transfer from the outside shell to the air q_3 is by natural convection. The heat flux from each region of the thermos must be equal—that is, $q_1 = q_2 = q_3$. Find the temperatures T_1 and T_2 at steady state. T_0 is 450°C and T_3 = 25°C.

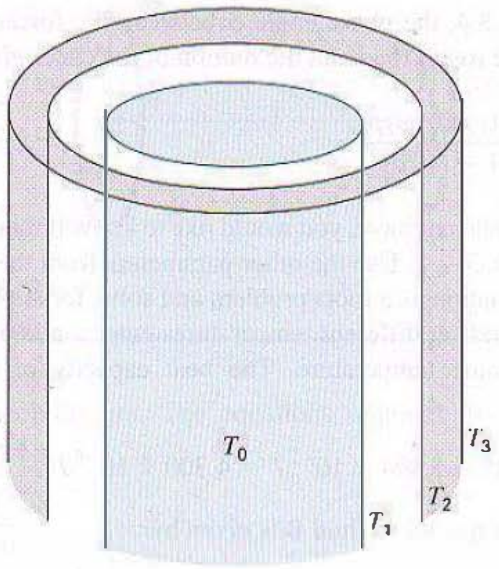


Figure P8.41

$$q_1 = 10^{-9}[(T_0 + 273)^4 - (T_1 + 273)^4]$$
$$q_2 = 4(T_1 - T_2)$$
$$q_3 = 1.3(T_2 - T_3)^{4/3}$$

8.42 The general form for a three-dimensional stress field is given by

$$\begin{bmatrix} \sigma_{xx} & \sigma_{xy} & \sigma_{xz} \\ \sigma_{xy} & \sigma_{yy} & \sigma_{yz} \\ \sigma_{xz} & \sigma_{yz} & \sigma_{zz} \end{bmatrix}$$

where the diagonal terms represent tensile or compressive stresses and the off-diagonal terms represent shear stresses. A stress field (in MPa) is given by

$$\begin{bmatrix} 10 & 14 & 25 \\ 14 & 7 & 15 \\ 25 & 15 & 16 \end{bmatrix}$$

To solve for the principal stresses, it is necessary to construct the following matrix (again in MPa):

$$\begin{bmatrix} 10-\sigma & 14 & 25 \\ 14 & 7-\sigma & 15 \\ 25 & 15 & 16-\sigma \end{bmatrix}$$

σ_1, σ_2, and σ_3 can be solved from the equation

$$\sigma^3 - I\sigma^2 + II\sigma - III = 0$$

where

$$I = \sigma_{xx} + \sigma_{yy} + \sigma_{zz}$$
$$II = \sigma_{xx}\sigma_{yy} + \sigma_{xx}\sigma_{zz} + \sigma_{yy}\sigma_{zz} - \sigma_{xy}^2 - \sigma_{xz}^2 - \sigma_{yz}^2$$
$$III = \sigma_{xx}\sigma_{yy}\sigma_{zz} - \sigma_{xx}\sigma_{yz}^2 - \sigma_{yy}\sigma_{xz}^2 - \sigma_{zz}\sigma_{xy}^2 + 2\sigma_{xy}\sigma_{xz}\sigma_{yz}$$

I, II, and III are known as the stress invariants. Find σ_1, σ_2, and σ_3 using a root-finding technique.

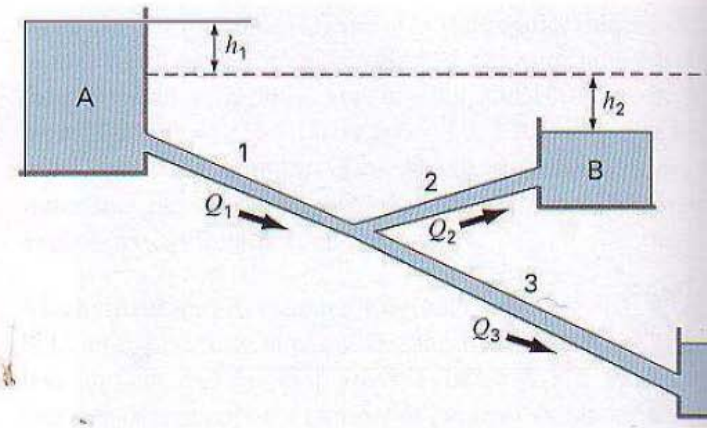


Figure P8.43

8.43 Figure P8.43 shows three reservoirs connected by pipes. The pipes, which are made of asphalt-dipped ca ($\varepsilon = 0.0012$ m), have the following characteristics:

Pipe	1	2
Length, m	1800	500
Diameter, m	0.4	0.25
Flow, m^3/s	?	0.1

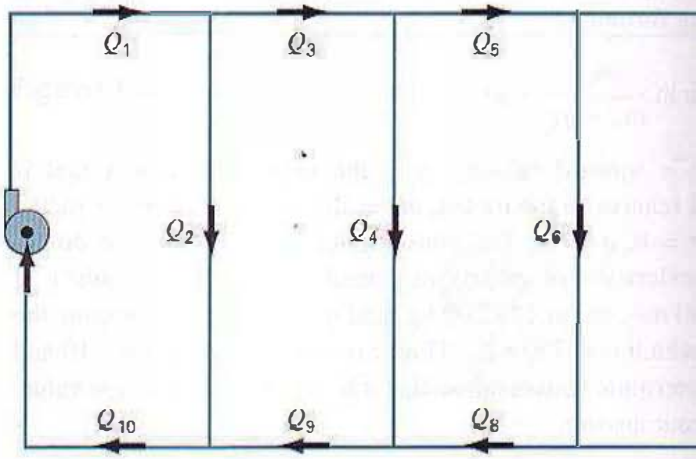
If the water surface elevations in Reservoirs A and C are 2 172.5 m, respectively, determine the elevation in Reservoi the flows in pipes 1 and 3. Note that the kinematic visc water is 1×10^{-6} m^2/s and use the Colebrook equation t mine the friction factor (recall Prob. 8.12).

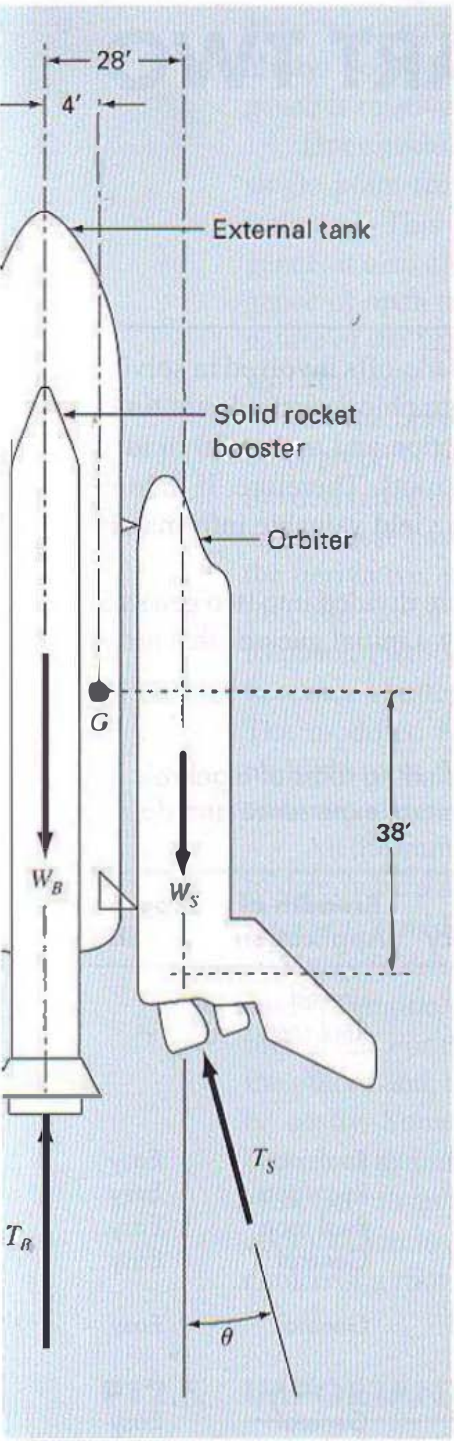
8.44 A fluid is pumped into the network of pipes sh Fig. P8.44. At steady state, the following flow balances mu

$$Q_1 = Q_2 + Q_3$$
$$Q_3 = Q_4 + Q_5$$
$$Q_5 = Q_6 + Q_7$$

where Q_i = flow in pipe i [m^3/s]. In addition, the pressur around the three right-hand loops must equal zero. The p

Figure P8.44





ılar pipe length can be computed with

$$\frac{\rho}{\sqrt{5}}Q^2$$

ıe pressure drop [Pa], f = the friction factor L = the pipe length [m], ρ = the fluid density = pipe diameter [m]. Write a program (or develop a mathematics software package) that will allow you to compute the flow in every pipe length given that $Q_1 = 1\ \text{m}^3/\text{s}$ and $\rho = 1.23\ \text{kg/m}^3$. All the pipes have $D = 500\text{mm}$ and $f = 0.005$. The pipe lengths are: $L_3 = L_5 = L_8 = L_9 = 2$ m; $L_2 = L_4 = L_6 = 4$ m; and $L_7 = 8$ m.

8.45 Repeat Prob. 8.44, but incorporate the fact that the friction factor can be computed with the *von Karman equation*,

$$\frac{1}{\sqrt{f}} = 4\log_{10}(\text{Re}\sqrt{f}) - 0.4$$

where Re = the Reynolds number

$$\text{Re} = \frac{\rho V D}{\mu}$$

where V = the velocity of the fluid in the pipe [m/s] and μ = dynamic viscosity ($\text{N} \cdot \text{s/m}^2$). Note that for a circular pipe $V = 4Q/\pi D^2$. Also, assume that the fluid has a viscosity of $1.79 \times 10^{-5}\ \text{N} \cdot \text{s/m}^2$.

8.46 The space shuttle, at lift-off from the launch pad, has four forces acting on it, which are shown on the free-body diagram (Fig. P8.46). The combined weight of the two solid rocket boosters and external fuel tank is $W_B = 1.663 \times 10^6$ lb. The weight of the orbiter with a full payload is $W_S = 0.23 \times 10^6$ lb. The combined thrust of the two solid rocket boosters is $T_B = 5.30 \times 10^6$ lb. The combined thrust of the three liquid fuel orbiter engines is $T_S = 1.125 \times 10^6$ lb.

At liftoff, the orbiter engine thrust is directed at angle θ to make the resultant moment acting on the entire craft assembly (external tank, solid rocket boosters, and orbiter) equal to zero. With the resultant moment equal to zero, the craft will not rotate about its mass center G at liftoff. With these forces, the craft will have a resultant force with components in both the vertical and horizontal direction. The vertical resultant force component is what allows the craft to lift off from the launch pad and fly vertically. The horizontal resultant force component causes the craft to fly horizontally. The resultant moment acting on the craft will be zero when θ is adjusted to the proper value. If this angle is not adjusted properly, and there is some resultant moment acting on the craft, the craft will tend to rotate about it mass center.

(a) Resolve the orbiter thrust T_S into horizontal and vertical components, and then sum moments about point G, the craft mass center. Set the resulting moment equation equal to zero. This equation can now be solved for the value of θ required for liftoff.

(b) Derive an equation for the resultant moment acting on the craft in terms of the angle θ. Plot the resultant moment as a function of the angle θ over a range of -5 radians to $+5$ radians.

(c) Write a computer program to solve for the angle θ using Newton's method to find the root of the resultant moment equation. Make an initial first guess at the root of interest using the plot. Terminate your iterations when the value of θ has better than five significant figures.

(d) Repeat the program for the minimum payload weight of the orbiter of $W_S = 195{,}000$ lb.

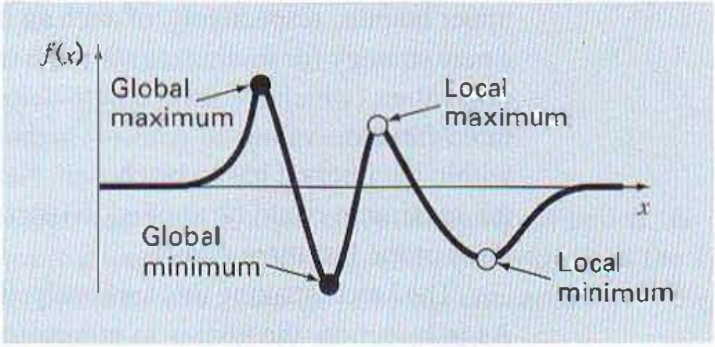
CHAPTER 13

One-Dimensional Unconstrained Optimization

This section will describe techniques to find the minimum or maximum of a function of a single variable, $f(x)$. A useful image in this regard is the one-dimensional, "roller coaster"–like function depicted in Fig. 13.1. Recall from Part Two that root location was complicated by the fact that several roots can occur for a single function. Similarly, both local and global optima can occur in optimization. Such cases are called *multimodal*. In almost all instances, we will be interested in finding the absolute highest or lowest value of a function. Thus, we must take care that we do not mistake a local result for the global optimum.

Distinguishing a global from a local extremum can be a very difficult problem for the general case. There are three usual ways to approach this problem. First, insight into the behavior of low-dimensional functions can sometimes be obtained graphically. Second, finding optima based on widely varying and perhaps randomly generated starting guesses, and then selecting the largest of these as global. Finally, perturbing the starting point associated with a local optimum and seeing if the routine returns a better point or always returns to the same point. Although all these approaches can have utility, the fact is that in some problems (usually the large ones), there may be no practical way to ensure that you have located a global optimum. However, although you should always be sensitive to the

FIGURE 13.1
A function that asymptotically approaches zero at plus and minus ∞ and has two maximum and two minimum points in the vicinity of the origin. The two points to the right are local optima, whereas the two to the left are global.



issue, it is fortunate that there are numerous engineering problems where you can locate global optimum in an unambiguous fashion.

Just as in root location, optimization in one dimension can be divided into bracke and open methods. As described in the next section, the golden-section search is an ex ple of a bracketing method that depends on initial guesses that bracket a single optim This is followed by a somewhat more sophisticated bracketing approach---quadratic i polation.

The final method described in this chapter is an open method based on the idea f calculus that the minimum or maximum can be found by solving $f'(x) = 0$. This red the optimization problem to finding the root of $f'(x)$ using techniques of the sort descr in Part Two. We will demonstrate one version of this approach—Newton's method.

13.1 GOLDEN-SECTION SEARCH

In solving for the root of a single nonlinear equation, the goal was to find the value o variable x that yields a *zero* of the function $f(x)$. *Single-variable optimization* has the of finding the value of x that yields an *extremum*, either a maximum or minimum of

The golden-section search is a simple, general-purpose, single-variable search t nique. It is similar in spirit to the bisection approach for locating roots in Chap. 5. Re that bisection hinged on defining an interval, specified by a lower guess (x_l) and an up guess (x_u), that bracketed a single root. The presence of a root between these bounds verified by determining that $f(x_l)$ and $f(x_u)$ had different signs. The root was then estim as the midpoint of this interval,

$$x_r = \frac{x_l + x_u}{2}$$

The final step in a bisection iteration involved determining a new smaller bracket. This done by replacing whichever of the bounds x_l or x_u had a function value with the same as $f(x_r)$. One advantage of this approach was that the new value x_r replaced one of the bounds.

Now we can develop a similar approach for locating the optimum of a one-dimensi function. For simplicity, we will focus on the problem of finding a maximum. When we cuss the computer algorithm, we will describe the minor modifications needed to simul minimum.

As with bisection, we can start by defining an interval that contains a single ans That is, the interval should contain a single maximum, and hence is called *unimodal*. can adopt the same nomenclature as for bisection, where x_l and x_u defined the lower upper bounds, respectively, of such an interval. However, in contrast to bisection, we a new strategy for finding a maximum within the interval. Rather than using only two f tion values (which are sufficient to detect a sign change, and hence a zero), we would three function values to detect whether a maximum occurred. Thus, an additional p within the interval has to be chosen. Next, we have to pick a fourth point. Then the tes the maximum could be applied to discern whether the maximum occurred within the three or the last three points.

The key to making this approach efficient is the wise choice of the intermediate po As in bisection, the goal is to minimize function evaluations by replacing old values

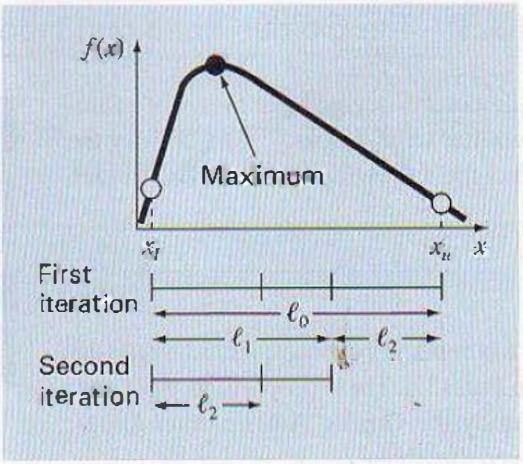


FIGURE 13.2
The initial step of the golden-section search algorithm involves choosing two interior points according to the golden ratio.

new values. This goal can be achieved by specifying that the following two conditions hold (Fig. 13.2):

$$\ell_0 = \ell_1 + \ell_2 \tag{13.1}$$

$$\frac{\ell_1}{\ell_0} = \frac{\ell_2}{\ell_1} \tag{13.2}$$

The first condition specifies that the sum of the two sublengths ℓ_1 and ℓ_2 must equal the original interval length. The second says that the ratio of the lengths must be equal. Equation (13.1) can be substituted into Eq. (13.2),

$$\frac{\ell_1}{\ell_1 + \ell_2} = \frac{\ell_2}{\ell_1} \tag{13.3}$$

If the reciprocal is taken and $R = \ell_2 / \ell_1$, we arrive at

$$1 + R = \frac{1}{R} \tag{13.4}$$

or

$$R^2 + R - 1 = 0 \tag{13.5}$$

which can be solved for the positive root

$$R = \frac{-1 + \sqrt{1 - 4(-1)}}{2} = \frac{\sqrt{5} - 1}{2} = 0.61803\ldots \tag{13.6}$$

This value, which has been known since antiquity, is called the *golden ratio* (see Box 13.1). Because it allows optima to be found efficiently, it is the key element of the golden-section method we have been developing conceptually. Now let us derive an algorithm to implement this approach on the computer.

Box 13.1 The Golden Ratio and Fibonacci Numbers

In many cultures, certain numbers are ascribed qualities. For example, we in the West are all familiar with "Lucky 7" and "Friday the 13th." Ancient Greeks called the following number the "golden ratio:"

$$\frac{\sqrt{5}-1}{2} = 0.61803\ldots$$

This ratio was employed for a number of purposes, including the development of the rectangle in Fig. 13.3. These proportions were considered aesthetically pleasing by the Greeks. Among other things, many of their temples followed this shape.

The golden ratio is related to an important mathematical series known as the *Fibonacci numbers*, which are

0, 1, 1, 2, 3, 5, 8, 13, 21, 34, ...

Thus, each number after the first two represents the sum of the preceding two. This sequence pops up in many diverse areas of science and engineering. In the context of the present discussion, an interesting property of the Fibonacci sequence relates to the ratio of consecutive numbers in the sequence; that is, $0/1 = 0$, $1/1 = 1$, $1/2 = 0.5$, $2/3 = 0.667$, $3/5 = 0.6$, $5/8 = 0.625$, $8/13 = 0.615$, and so on. As one proceeds, the ratio of consecutive numbers approaches the golden ratio!

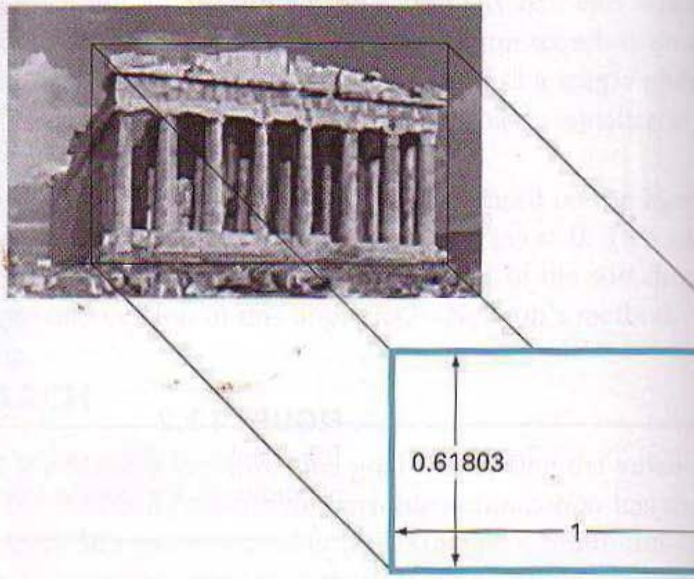


FIGURE 13.3
The Parthenon in Athens, Greece, was constructed in the 5th century B.C. Its front dimensions can be fit almost exactly within a golden rectangle.

As mentioned above and as depicted in Fig. 13.4, the method starts with two guesses, x_l and x_u, that bracket one local extremum of $f(x)$. Next, two interior points x_1 and x_2 are chosen according to the golden ratio,

$$d = \frac{\sqrt{5}-1}{2}(x_u - x_l)$$

$$x_1 = x_l + d$$

$$x_2 = x_u - d$$

The function is evaluated at these two interior points. Two results can occur:

1. If, as is the case in Fig. 13.4, $f(x_1) > f(x_2)$, then the domain of x to the left of x_2, from x_l to x_2, can be eliminated because it does not contain the maximum. For this case, x_1 becomes the new x_l for the next round.
2. If $f(x_2) > f(x_1)$, then the domain of x to the right of x_1, from x_1 to x_u would have been eliminated. In this case, x_1 becomes the new x_u for the next round.

Now, here is the real benefit from the use of the golden ratio. Because the original x_1 and x_2 were chosen using the golden ratio, we do not have to recalculate all the f

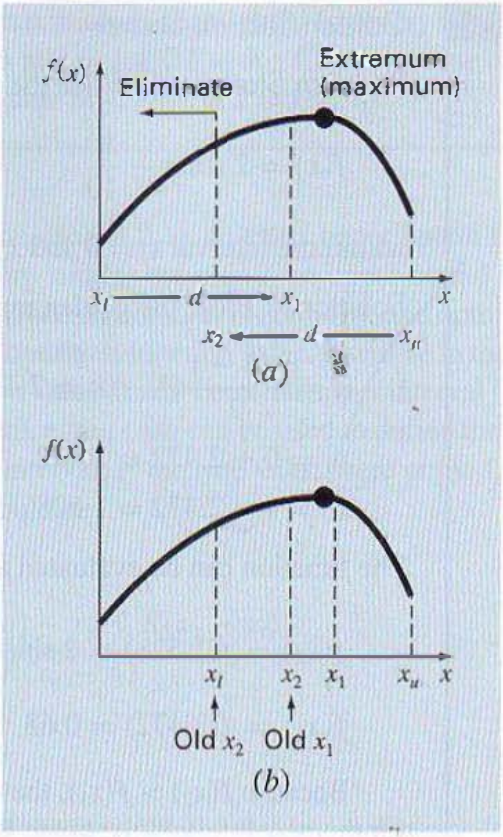


FIGURE 13.4
(a) The initial step of the golden-section search algorithm involves choosing two interior points according to the golden ratio. (b) The second step involves defining a new interval that includes the optimum.

values for the next iteration. For example, for the case illustrated in Fig. 13.4, the old x_l becomes the new x_2. This means that we already have the value for the new $f(x_2)$, since it is the same as the function value at the old x_l.

To complete the algorithm, we now only need to determine the new x_1. This is done with the same proportionality as before,

$$x_1 = x_l + \frac{\sqrt{5}-1}{2}(x_u - x_l)$$

A similar approach would be used for the alternate case where the optimum fell in the left subinterval.

As the iterations are repeated, the interval containing the extremum is reduced rapidly. In fact, each round the interval is reduced by a factor of the golden ratio (about 61.8%). That means that after 10 rounds, the interval is shrunk to about 0.618^{10} or 0.008 or 0.8% of its initial length. After 20 rounds, it is about 0.0066%. This is not quite as good as the reduction achieved with bisection, but this is a harder problem.

EXAMPLE 13.1 Golden-Section Search

Problem Statement. Use the golden-section search to find the maximum of

$$f(x) = 2\sin x - \frac{x^2}{10}$$

within the interval $x_l = 0$ and $x_u = 4$.

Solution. First, the golden ratio is used to create the two interior points

$$d = \frac{\sqrt{5}-1}{2}(4-0) = 2.472$$

$$x_1 = 0 + 2.472 = 2.472$$

$$x_2 = 4 - 2.472 = 1.528$$

The function can be evaluated at the interior points

$$f(x_2) = f(1.528) = 2\sin(1.528) - \frac{1.528^2}{10} = 1.765$$

$$f(x_1) = f(2.472) = 0.63$$

Because $f(x_2) > f(x_1)$, the maximum is in the interval defined by x_l, x_2, and x for the new interval, the lower bound remains $x_l = 0$, and x_1 becomes the upper bou is, $x_u = 2.472$. In addition, the former x_2 value becomes the new x_1, that is, x_1 Further, we do not have to recalculate $f(x_1)$ because it was determined on the iteration as $f(1.528) = 1.765$.

All that remains is to compute the new values of d and x_2,

$$d = \frac{\sqrt{5}-1}{2}(2.472 - 0) = 1.528$$

$$x_2 = 2.4721 - 1.528 = 0.944$$

The function evaluation at x_2 is $f(0.994) = 1.531$. Since this value is less than tion value at x_1, the maximum is in the interval prescribed by x_2, x_1, and x_u.

The process can be repeated, with the results tabulated below:

i	x_l	$f(x_l)$	x_2	$f(x_2)$	x_1	$f(x_1)$	x_u	$f(x_u)$
1	0	0	1.5279	1.7647	2.4721	0.6300	4.0000	−3.1136
2	0	0	0.9443	1.5310	1.5279	1.7647	2.4721	0.6300
3	0.9443	1.5310	1.5279	1.7647	1.8885	1.5432	2.4721	0.6300
4	0.9443	1.5310	1.3050	1.7595	1.5279	1.7647	1.8885	1.5432
5	1.3050	1.7595	1.5279	1.7647	1.6656	1.7136	1.8885	1.5432
6	1.3050	1.7595	1.4427	1.7755	1.5279	1.7647	1.6656	1.7136
7	1.3050	1.7595	1.3901	1.7742	1.4427	1.7755	1.5279	1.7647
8	1.3901	1.7742	1.4427	1.7755	1.4752	1.7732	1.5279	1.7647

Note that the current maximum is highlighted for every iteration. After the eighth iteration, the maximum occurs at $x = 1.4427$ with a function value of 1.7755. Thus, the result is converging on the true value of 1.7757 at $x = 1.4276$.

Recall that for bisection (Sec. 5.2.1), an exact upper bound for the error can be calculated at each iteration. Using similar reasoning, an upper bound for golden-section search can be derived as follows: Once an iteration is complete, the optimum will either fall in one of two intervals. If x_2 is the optimum function value, it will be in the lower interval (x_l, x_2, x_1). If x_1 is the optimum function value, it will be in the upper interval (x_2, x_1, x_u). Because the interior points are symmetrical, either case can be used to define the error.

Looking at the upper interval, if the true value were at the far left, the maximum distance from the estimate would be

$$\begin{aligned}\Delta x_a &= x_1 - x_2 \\ &= x_l + R(x_u - x_l) - x_u + R(x_u - x_l) \\ &= (x_l - x_u) + 2R(x_u - x_l) \\ &= (2R - 1)(x_u - x_l)\end{aligned}$$

or $0.236(x_u - x_l)$.

If the true value were at the far right, the maximum distance from the estimate would be

$$\begin{aligned}\Delta x_b &= x_u - x_1 \\ &= x_u - x_l - R(x_u - x_l) \\ &= (1 - R)(x_u - x_l)\end{aligned}$$

or $0.382(x_u - x_l)$. Therefore, this case would represent the maximum error. This result can then be normalized to the optimal value for that iteration, x_{opt}, to yield

$$\varepsilon_a = (1 - R)\left|\frac{x_u - x_l}{x_{opt}}\right| 100\%$$

This estimate provides a basis for terminating the iterations.

Pseudocode for the golden-section-search algorithm for maximization is presented in Fig. 13.5*a*. The minor modifications to convert the algorithm to minimization are listed in Fig. 13.5*b*. In both versions the x value for the optimum is returned as the function value (*gold*). In addition, the value of $f(x)$ at the optimum is returned as the variable (*fx*).

You may be wondering why we have stressed the reduced function evaluations of the golden-section search. Of course, for solving a single optimization, the speed savings would be negligible. However, there are two important contexts where minimizing the number of function evaluations can be important. These are

1. *Many evaluations.* There are cases where the golden-section-search algorithm may be a part of a much larger calculation. In such cases, it may be called many times. Therefore, keeping function evaluations to a minimum could pay great dividends for such cases.

FIGURE 13.5
Algorithm for the golden-section search.

```
FUNCTION Gold (xlow, xhigh, maxit, es, fx)
R = (5^0.5 - 1)/2
xℓ = xlow; xu = xhigh
iter = 1
d  = R * (xu - xℓ)
x1 = xℓ + d; x2 = xu - d
f1 = f(x1)
f2 = f(x2)
IF f1 > f2 THEN                                   IF f1 < f2 THEN
  xopt = x1
  fx = f1
ELSE
  xopt = x2
  fx = f2
END IF
DO
  d = R*d
  IF f1 > f2 THEN                                 IF f1 < f2 THEN
    xℓ = x2
    x2 = x1
    x1 = xℓ+d
    f2 = f1
    f1 = f(x1)
  ELSE
    xu = x1
    x1 = x2
    x2 = xu-d
    f1 = f2
    f2 = f(x2)
  END IF
  iter = iter+1
  IF f1 > f2 THEN                                 IF f1 < f2 THEN
    xopt = x1
    fx = f1
  ELSE
    xopt = x2
    fx = f2
  END IF
  IF xopt ≠ 0. THEN
    ea = (1.-R) *ABS((xu - xℓ)/xopt) * 100.
  END IF
  IF ea ≤ es OR iter ≥ maxit EXIT
END DO
Gold = xopt
END Gold
```

(*a*) **Maximization** (*b*) **Minimizati**

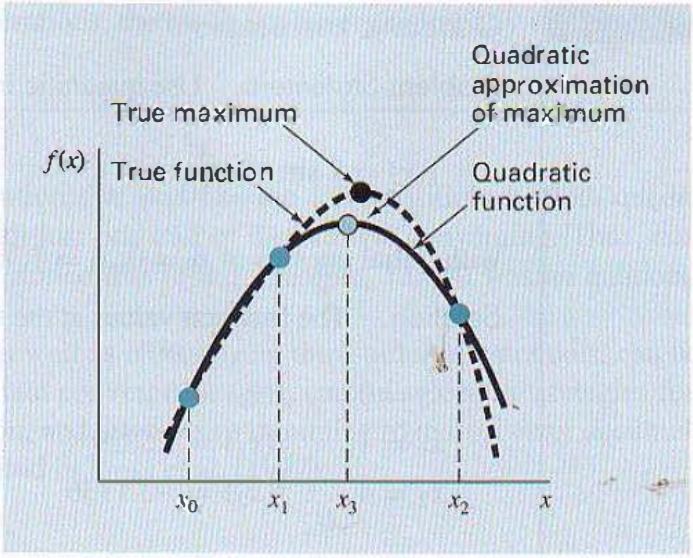


FIGURE 13.6
Graphical description of quadratic interpolation.

2. *Time-consuming evaluation.* For pedagogical reasons, we use simple functions in most of our examples. You should understand that a function can be very complex and time-consuming to evaluate. For example, in a later part of this book, we will describe how optimization can be used to estimate the parameters of a model consisting of a system of differential equations. For such cases, the "function" involves time-consuming model integration. Any method that minimizes such evaluations would be advantageous.

13.2 QUADRATIC INTERPOLATION

Quadratic interpolation takes advantage of the fact that a second-order polynomial often provides a good approximation to the shape of $f(x)$ near an optimum (Fig. 13.6).

Just as there is only one straight line connecting two points, there is only one quadratic or parabola connecting three points. Thus, if we have three points that jointly bracket an optimum, we can fit a parabola to the points. Then we can differentiate it, set the result equal to zero, and solve for an estimate of the optimal x. It can be shown through some algebraic manipulations that the result is

$$x_3 = \frac{f(x_0)(x_1^2 - x_2^2) + f(x_1)(x_2^2 - x_0^2) + f(x_2)(x_0^2 - x_1^2)}{2f(x_0)(x_1 - x_2) + 2f(x_1)(x_2 - x_0) + 2f(x_2)(x_0 - x_1)} \tag{13.7}$$

where x_0, x_1, and x_2 are the initial guesses, and x_3 is the value of x that corresponds to the maximum value of the quadratic fit to the guesses.

EXAMPLE 13.2 Quadratic Interpolation

Problem Statement. Use quadratic interpolation to approximate the maximum of

$$f(x) = 2\sin x - \frac{x^2}{10}$$

with initial guesses of $x_0 = 0$, $x_1 = 1$, and $x_2 = 4$.

Solution. The function values at the three guesses can be evaluated,

$$x_0 = 0 \qquad f(x_0) = 0$$
$$x_1 = 1 \qquad f(x_1) = 1.5829$$
$$x_2 = 4 \qquad f(x_2) = -3.1136$$

and substituted into Eq. (13.7) to give,

$$x_3 = \frac{0(1^2 - 4^2) + 1.5829(4^2 - 0^2) + (-3.1136)(0^2 - 1^2)}{2(0)(1 - 4) + 2(1.5829)(4 - 0) + 2(-3.1136)(0 - 1)} = 1.5055$$

which has a function value of $f(1.5055) = 1.7691$.

Next, a strategy similar to the golden-section search can be employed to determ which point should be discarded. Because the function value for the new point is hig than for the intermediate point (x_1) and the new x value is to the right of the intermedi point, the lower guess (x_0) is discarded. Therefore, for the next iteration,

$$x_0 = 1 \qquad f(x_0) = 1.5829$$
$$x_1 = 1.5055 \qquad f(x_1) = 1.7691$$
$$x_2 = 4 \qquad f(x_2) = -3.1136$$

which can be substituted into Eq. (13.7) to give

$$x_3 = \frac{1.5829(1.5055^2 - 4^2) + 1.7691(4^2 - 1^2) + (-3.1136)(1^2 - 1.5055^2)}{2(1.5829)(1.5055 - 4) + 2(1.7691)(4 - 1) + 2(-3.1136)(1 - 1.5055)}$$
$$= 1.4903$$

which has a function value of $f(1.4903) = 1.7714$.

The process can be repeated, with the results tabulated below:

i	x_0	$f(x_0)$	x_1	$f(x_1)$	x_2	$f(x_2)$	x_3	$f($
1	0.0000	0.0000	1.0000	1.5829	4.0000	−3.1136	1.5055	1.7
2	1.0000	1.5829	1.5055	1.7691	4.0000	−3.1136	1.4903	1.7
3	1.0000	1.5829	1.4903	1.7714	1.5055	1.7691	1.4256	1.7
4	1.0000	1.5829	1.4256	1.7757	1.4903	1.7714	1.4266	1.7
5	1.4256	1.7757	1.4266	1.7757	1.4903	1.7714	1.4275	1.7

Thus, within five iterations, the result is converging rapidly on the true value of 1.7757 at $x = 1.4276$.

We should mention that just like the false-position method, quadratic interpolation can get hung up with just one end of the interval converging. Thus, convergence can be slow. For example, notice that in our example, 1.0000 was an endpoint for most of the iterations.

This method, as well as others using third-order polynomials, can be formulated into algorithms that contain convergence tests, careful selection strategies for the points to retain on each iteration, and attempts to minimize round-off error accumulation. In particular, see Brent's method in Press et al. (1992).

13.3 NEWTON'S METHOD

Recall that the Newton-Raphson method of Chap. 6 is an open method that finds the root x of a function such that $f(x) = 0$. The method is summarized as

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}$$

A similar open approach can be used to find an optimum of $f(x)$ by defining a new function, $g(x) = f'(x)$. Thus, because the same optimal value x^* satisfies both

$$f'(x^*) = g(x^*) = 0$$

we can use the following,

$$x_{i+1} = x_i - \frac{f'(x_i)}{f''(x_i)} \tag{13.8}$$

as a technique to find the minimum or maximum of $f(x)$. It should be noted that this equation can also be derived by writing a second-order Taylor series for $f(x)$ and setting the derivative of the series equal to zero. Newton's method is an open method similar to Newton-Raphson because it does not require initial guesses that bracket the optimum. In addition, it also shares the disadvantage that it may be divergent. Finally, it is usually a good idea to check that the second derivative has the correct sign to confirm that the technique is converging on the result you desire.

EXAMPLE 13.3 Newton's Method

Problem Statement. Use Newton's method to find the maximum of

$$f(x) = 2 \sin x - \frac{x^2}{10}$$

with an initial guess of $x_0 = 2.5$.

Solution. The first and second derivatives of the function can be evaluated as

$$f'(x) = 2\cos x - \frac{x}{5}$$

$$f''(x) = -2\sin x - \frac{1}{5}$$

which can be substituted into Eq. (13.8) to give

$$x_{i+1} = x_i - \frac{2\cos x_i - x_i/5}{-2\sin x_i - 1/5}$$

Substituting the initial guess yields

$$x_1 = 2.5 - \frac{2\cos 2.5 - 2.5/5}{-2\sin 2.5 - 1/5} = 0.99508$$

which has a function value of 1.57859. The second iteration gives

$$x_1 = 0.995 - \frac{2\cos 0.995 - 0.995/5}{-2\sin 0.995 - 1/5} = 1.46901$$

which has a function value of 1.77385.

The process can be repeated, with the results tabulated below:

i	x	$f(x)$	$f'(x)$	f''
0	2.5	0.57194	−2.10229	−1
1	0.99508	1.57859	0.88985	−1
2	1.46901	1.77385	−0.09058	−2
3	1.42764	1.77573	−0.00020	−2
4	1.42755	1.77573	0.00000	−2

Thus, within four iterations, the result converges rapidly on the true value.

Although Newton's method works well in some cases, it is impractical for cases the derivatives cannot be conveniently evaluated. For these cases, other approach do not involve derivative evaluation are available. For example, a secant-like vers Newton's method can be developed by using finite-difference approximations for t rivative evaluations.

A bigger reservation regarding the approach is that it may diverge based on the of the function and the quality of the initial guess. Thus, it is usually employed only we are close to the optimum. Hybrid techniques that use *bracketing approaches* fa the optimum and *open methods* near the optimum attempt to exploit the strong po both approaches.

This concludes our treatment of methods to solve the optima of functions of a single variable. Some engineering examples are presented in Chap. 16. In addition, the techniques described here are an important element of some procedures to optimize multivariable functions, as discussed in the next chapter.

OBLEMS

formula

$+ 8x - 12$

the maximum and the corresponding value of x for
on analytically (i.e., using differentiation).
Eq. (13.7) yields the same results based on initial
$x_0 = 0, x_1 = 2$, and $x_2 = 6$.

$5x^6 - 2x^4 + 12x$

nction.
ical methods to prove that the function is concave
es of x.
te the function and then use a root-location method
or the maximum $f(x)$ and the corresponding value

the value of x that maximizes $f(x)$ in Prob. 13.2
en-section search. Employ initial guesses of $x_l = 0$
perform three iterations.
rob. 13.3, except use quadratic interpolation. Em-
sses of $x_0 = 0$, $x_1 = 1$, and $x_2 = 2$ and perform three

rob. 13.3 but use Newton's method. Employ an ini-
$= 2$ and perform three iterations.
he advantages and disadvantages of golden-section
ic interpolation, and Newton's method for locating
lue in one dimension.
he following methods to find the maximum of

$- 1.8x^2 + 1.2x^3 - 0.3x^4$

ction search ($x_l = -2$, $x_u = 4$, $\varepsilon_s = 1\%$).
interpolation ($x_0 = 1.75$, $x_1 = 2$, $x_2 = 2.5$, itera-

method ($x_0 = 3$, $\varepsilon_s = 1\%$).
the following function:

$- 2x^3 - 8x^2 - 5x$

and graphical methods to show the function has a
ome value of x in the range $-2 \le x \le 1$.

13.9 Employ the following methods to find the maximum of the function from Prob. 13.8:
(a) Golden-section search ($x_l = -2$, $x_u = 1$, $\varepsilon_s = 1\%$).
(b) Quadratic interpolation ($x_0 = -2$, $x_1 = -1$, $x_2 = 1$, iterations $= 4$).
(c) Newton's method ($x_0 = -1$, $\varepsilon_s = 1\%$).

13.10 Consider the following function:

$$f(x) = 2x + \frac{3}{x}$$

Perform 10 iterations of quadratic interpolation to locate the minimum. Comment on the convergence of your results. ($x_0 = 0.1$, $x_1 = 0.5$, $x_2 = 5$)

13.11 Consider the following function:

$$f(x) = 3 + 6x + 5x^2 + 3x^3 + 4x^4$$

Locate the minimum by finding the root of the derivative of this function. Use bisection with initial guesses of $x_l = -2$ and $x_u = 1$.

13.12 Determine the minimum of the function from Prob. 13.11 with the following methods:
(a) Newton's method ($x_0 = -1$, $\varepsilon_s = 1\%$).
(b) Newton's method, but using a finite difference approximation for the derivative estimates.

$$f'(x) = \frac{f(x_i + \delta x_i) - f(x_i - \delta x_i)}{2\delta x_i}$$

$$f''(x) = \frac{f(x_i + \delta x_i) - 2f(x_i) - f(x_i - \delta x_i)}{(\delta x_i)^2}$$

where δ = a perturbation fraction ($= 0.01$). Use an initial guess of $x_0 = -1$ and iterate to $\varepsilon_s = 1\%$.

13.13 Develop a program using a programming or macro language to implement the golden-section search algorithm. Design the program so that it is expressly designed to locate a maximum. The subroutine should have the following features:

- Iterate until the relative error falls below a stopping criterion or exceeds a maximum number of iterations.
- Return both the optimal x and $f(x)$.
- Minimize the number of function evaluations.

Test your program with the same problem as Example 13.1.

13.14 Develop a program as described in Prob. 13.13, but make it perform minimization or maximization depending on the user's preference.

13.15 Develop a program using a programming or macro language to implement the quadratic interpolation algorithm. Design the program so that it is expressly designed to locate a maximum. The subroutine should have the following features:

- Base it on two initial guesses, and have the program generate the third initial value at the midpoint of the interval.
- Check whether the guesses bracket a maximum. If not, the subroutine should not implement the algorithm, but should return an error message.
- Iterate until the relative error falls below a stopping criterion or exceeds a maximum number of iterations.
- Return both the optimal x and $f(x)$.
- Minimize the number of function evaluations.

Test your program with the same problem as Example 13.2.

13.16 Develop a program using a programming or macro language to implement Newton's method. The subroutine should have the following features:

- Iterate until the relative error falls below a stopping criterion or exceeds a maximum number of iterations.
- Returns both the optimal x and $f(x)$.

Test your program with the same problem as Example 13.3.

13.17 Pressure measurements are taken at certain points behind an airfoil over time. The data best fits the curve $y = 6\cos x - 1.5\sin x$ from $x = 0$ to 6 s. Use four iterations of the golden-search method to find the minimum pressure. Set $x_l = 2$ and $x_u = 4$.

13.18 The trajectory of a ball can be computed with

$$y = (\tan\theta_0)x - \frac{g}{2v_0^2\cos^2\theta_0}x^2 + y_0$$

where y = the height (m), θ_0 = the initial angle (radia initial velocity (m/s), g = the gravitational constant and y_0 = the initial height (m). Use the golden-secti determine the maximum height given $y_0 = 1$ m, $v_0 =$ $\theta_0 = 50°$. Iterate until the approximate error falls bel using initial guesses of $x_l = 0$ and $x_u = 60$ m.

13.19 The deflection of a uniform beam subject to creasing distributed load can be computed as

$$y = \frac{w_0}{120EIL}(-x^5 + 2L^2x^3 - L^4x)$$

Given that $L = 600$ cm, $E = 50{,}000$ kN/cm^2, $I = 30{,}$ $w_0 = 2.5$ kN/cm, determine the point of maximu (a) graphically, (b) using the golden-section search proximate error falls below $\varepsilon_s = 1\%$ with initial gues and $x_u = L$.

13.20 An object with a mass of 100 kg is projected upw surface of the earth at a velocity of 50 m/s. If the objec linear drag ($c = 15$ kg/s), use the golden-section search the maximum height the object attains. Hint: recall Sec

13.21 The normal distribution is a bell-shaped curv

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

Use the golden-section search to determine the locati flection point of this curve for positive x.

TER 14

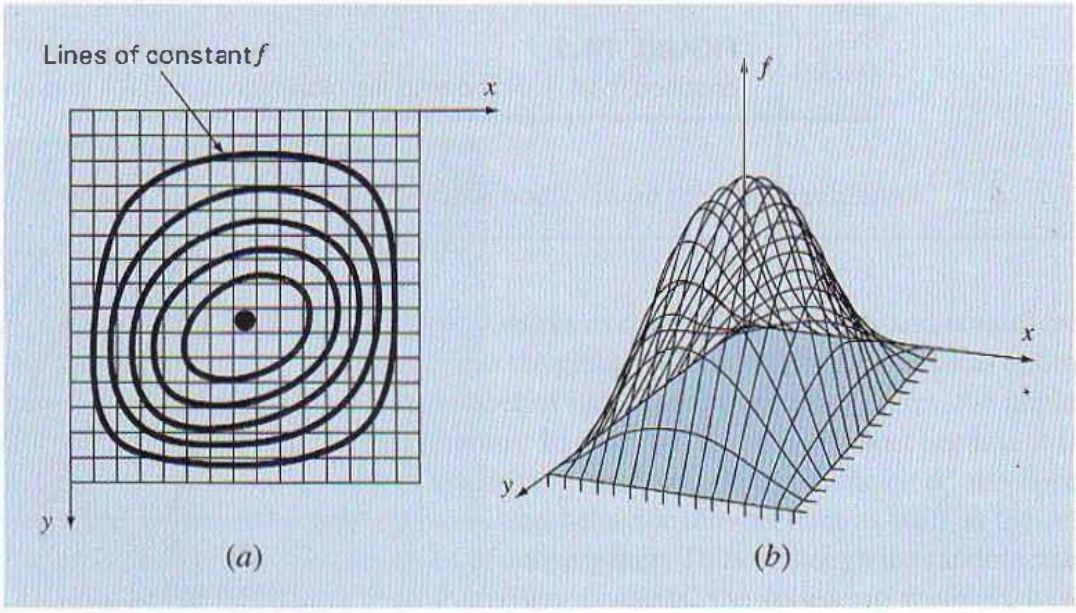
Multidimensional Unconstrained Optimization

This chapter describes techniques to find the minimum or maximum of a function of several variables. Recall from Chap. 13 that our visual image of a one-dimensional search was like a roller coaster. For two-dimensional cases, the image becomes that of mountains and valleys (Fig. 14.1). For higher-dimensional problems, convenient images are not possible.

We have chosen to limit this chapter to the two-dimensional case. We have adopted this approach because the essential features of multidimensional searches are often best communicated visually.

Techniques for multidimensional unconstrained optimization can be classified in a number of ways. For purposes of the present discussion, we will divide them depending on whether they require derivative evaluation. The approaches that do not require derivative evaluation are called *nongradient*, or *direct, methods*. Those that require derivatives are called *gradient*, or *descent* (or *ascent*), *methods*.

way to visual-
al searches is
scending a
zation) or de-
lley (mini-
D topo-
corresponds
in in (*b*).



14.1 DIRECT METHODS

These methods vary from simple brute force approaches to more elegant techniques that tempt to exploit the nature of the function. We will start our discussion with a brute approach.

14.1.1 Random Search

A simple example of a brute force approach is the *random search method*. As the name implies, this method repeatedly evaluates the function at randomly selected values of the independent variables. If a sufficient number of samples are conducted, the optimum will eventually be located.

EXAMPLE 14.1 Random Search Method

Problem Statement. Use a random number generator to locate the maximum of

$$f(x, y) = y - x - 2x^2 - 2xy - y^2 \qquad \text{(E1}$$

in the domain bounded by $x = -2$ to 2 and $y = 1$ to 3. The domain is depicted in Fig. Notice that a single maximum of 1.5 occurs at $x = -1$ and $y = 1.5$.

Solution. Random number generators typically generate values between 0 and 1. designate such a number as r, the following formula can be used to generate x values domly within a range between x_l to x_u:

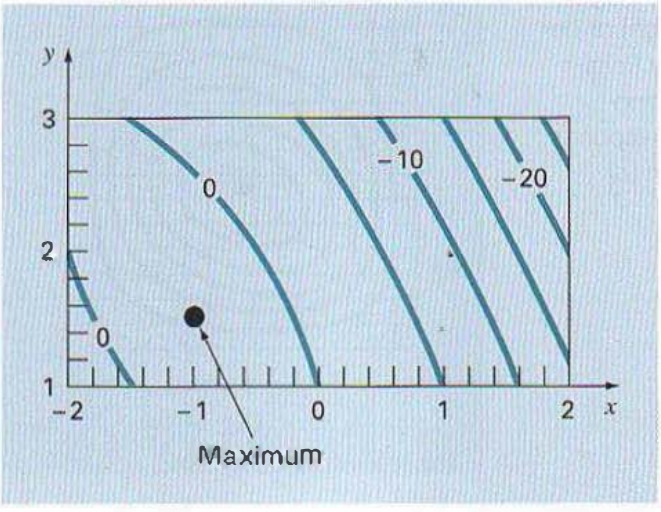
$$x = x_l + (x_u - x_l)r$$

For the present application, $x_l = -2$ and $x_u = 2$, and the formula is

$$x = -2 + (2 - (-2))r = -2 + 4r$$

This can be tested by substituting 0 and 1 to yield -2 and 2, respectively.

FIGURE 14.2
Equation (E14.1.1) showing the maximum at $x = -1$ and $y = 1.5$.



Similarly for y, a formula for the present example could be developed as

$$y = y_l + (y_u - y_l)r = 1 + (3 - 1)r = 1 + 2r$$

The following Excel VBA macrocode uses the VBA random number function Rnd, to generate (x, y) pairs. These are then substituted into Eq. (E14.1.1). The maximum value from among these random trials is stored in the variable *maxf*, and the corresponding x and y values in *maxx* and *maxy*, respectively.

```
maxf = -1E9
For j = 1 To n
  x = -2 + 4 * Rnd
  y = 1 + 2 * Rnd
  fn = y - x - 2 * x ^ 2 - 2 * x * y - y ^ 2
  If fn > maxf Then
    maxf = fn
    maxx = x
    maxy = y
  End If
Next j
```

A number of iterations yields

Iterations	x	y	f(x, y)
1000	−0.9886	1.4282	1.2462
2000	−1.0040	1.4724	1.2490
3000	−1.0040	1.4724	1.2490
4000	−1.0040	1.4724	1.2490
5000	−1.0040	1.4724	1.2490
6000	−0.9837	1.4936	1.2496
7000	−0.9960	1.5079	1.2498
8000	−0.9960	1.5079	1.2498
9000	−0.9960	1.5079	1.2498
10000	−0.9978	1.5039	1.2500

The results indicate that the technique homes in on the true maximum.

This simple brute force approach works even for discontinuous and nondifferentiable functions. Furthermore, it always finds the global optimum rather than a local optimum. Its major shortcoming is that as the number of independent variables grows, the implementation effort required can become onerous. In addition, it is not efficient because it takes no account of the behavior of the underlying function. The remainder of the approaches described in this chapter do take function behavior into account as well as the results of previous trials to improve the speed of convergence. Thus, although the random search can certainly prove useful in specific problem contexts, the following methods have more general utility and almost always lead to more efficient convergence.

It should be noted that more sophisticated search techniques are available. T heuristic approaches that were developed to handle either nonlinear and/or disc problems that classical optimization cannot usually handle well, if at all. Simu nealing, tabu search, artificial neural networks, and genetic algorithms are a few. widely applied is the *genetic algorithm*, with a number of commercial packages Holland (1975) pioneered the genetic algorithm approach and Davis (1991) and (1989) provide good overviews of the theory and application of the method.

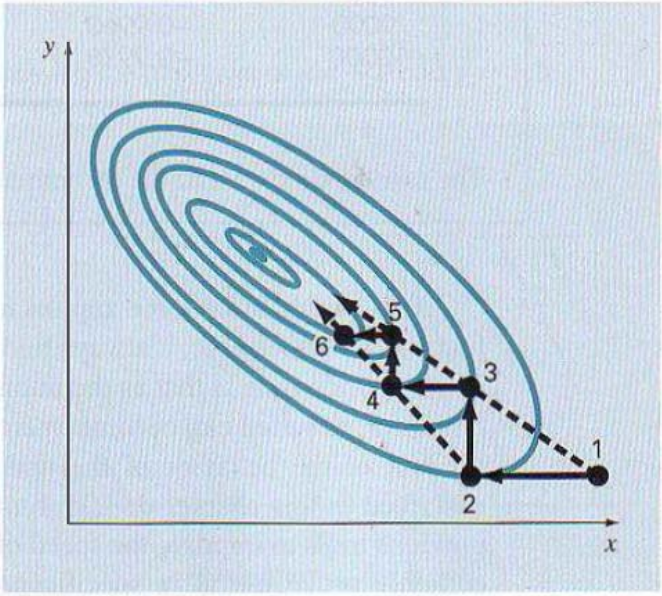
14.1.2 Univariate and Pattern Searches

It is very appealing to have an efficient optimization approach that does not req uation of derivatives. The random search method described above does no derivative evaluation, but it is not very efficient. This section describes an appr univariate search method, that is more efficient and still does not require deriva uation.

The basic strategy underlying the *univariate search method* is to change one at a time to improve the approximation while the other variables are held consta only one variable is changed, the problem reduces to a sequence of one-dim searches that can be solved using a variety of methods (including those des Chap. 13).

Let us perform a univariate search graphically, as shown in Fig. 14.3. Start a and move along the x axis with y constant to the maximum at point 2. You can point 2 is a maximum by noticing that the trajectory along the x axis just touches line at the point. Next, move along the y axis with x constant to point 3. Con process generating points 4, 5, 6, etc.

FIGURE 14.3
A graphical depiction of how a univariate search is conducted.



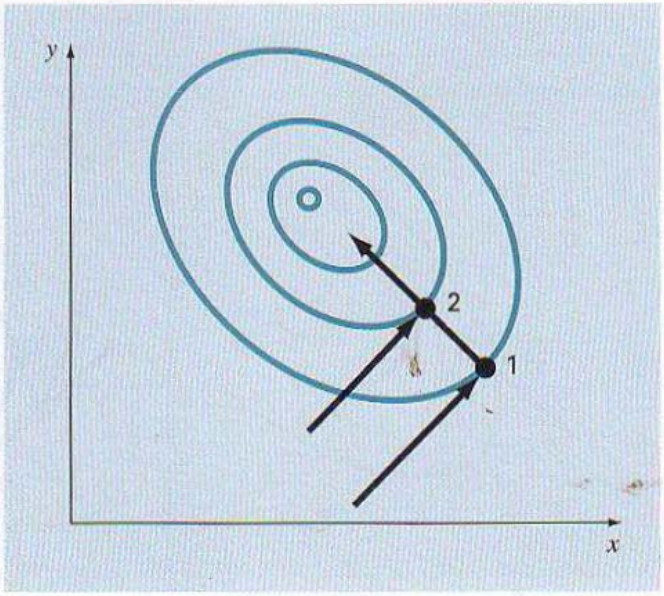


FIGURE 14.4
Conjugate directions.

Although we are gradually moving toward the maximum, the search becomes less efficient as we move along the narrow ridge toward the maximum. However, also note that lines joining alternate points such as 1-3, 3-5 or 2-4, 4-6 point in the general direction of the maximum. These trajectories present an opportunity to shoot directly along the ridge toward the maximum. Such trajectories are called *pattern directions*.

Formal algorithms are available that capitalize on the idea of pattern directions to find optimum values efficiently. The best known of these algorithms is called *Powell's method*. It is based on the observation (see Fig. 14.4) that if points 1 and 2 are obtained by one-dimensional searches in the same direction but from different starting points, then the line formed by 1 and 2 will be directed toward the maximum. Such lines are called *conjugate directions*.

In fact, it can be proved that if $f(x, y)$ is a quadratic function, sequential searches along conjugate directions will converge exactly in a finite number of steps regardless of the starting point. Since a general nonlinear function can often be reasonably approximated by a quadratic function, methods based on conjugate directions are usually quite efficient and are in fact quadratically convergent as they approach the optimum.

Let us graphically implement a simplified version of Powell's method to find the maximum of

$$f(x, y) = c - (x - a)^2 - (y - b)^2$$

where a, b, and c are positive constants. This equation results in circular contours in the x, y plane, as shown in Fig. 14.5.

Initiate the search at point 0 with starting directions h_1 and h_2. Note that h_1 and h_2 are not necessarily conjugate directions. From zero, move along h_1 until a maximum is located

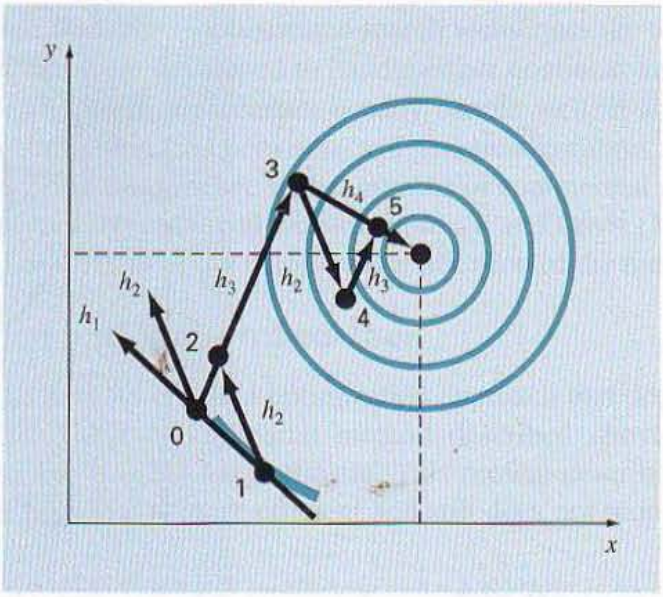


FIGURE 14.5
Powell's method.

at point 1. Then search from point 1 along direction h_2 to find point 2. Next, form a search direction h_3 through points 0 and 2. Search along this direction until the maxim at point 3 is located. Then search from point 3 in the h_2 direction until the maximu point 4 is located. From point 4 arrive at point 5 by again searching along h_3. Now, obs that both points 5 and 3 have been located by searching in the h_3 direction from two di ent points. Powell has shown that h_4 (formed by points 3 and 5) and h_3 are conju directions. Thus, searching from point 5 along h_4 brings us directly to the maximum.

Powell's method can be refined to make it more efficient, but the formal algorithm beyond the scope of this text. However, it is an efficient method that is quadratically vergent without requiring derivative evaluation.

14.2 GRADIENT METHODS

As the name implies, *gradient methods* explicitly use derivative information to gen efficient algorithms to locate optima. Before describing specific approaches, we must review some key mathematical concepts and operations.

14.2.1 Gradients and Hessians

Recall from calculus that the first derivative of a one-dimensional function provides a s or tangent to the function being differentiated. From the standpoint of optimization, th useful information. For example, if the slope is positive, it tells us that increasing the i pendent variable will lead to a higher value of the function we are exploring.

From calculus, also recall that the first derivative may tell us when we have reache optimal value since this is the point that the derivative goes to zero. Further, the sign o second derivative can tell us whether we have reached a minimum (positive second d ative) or a maximum (negative second derivative).

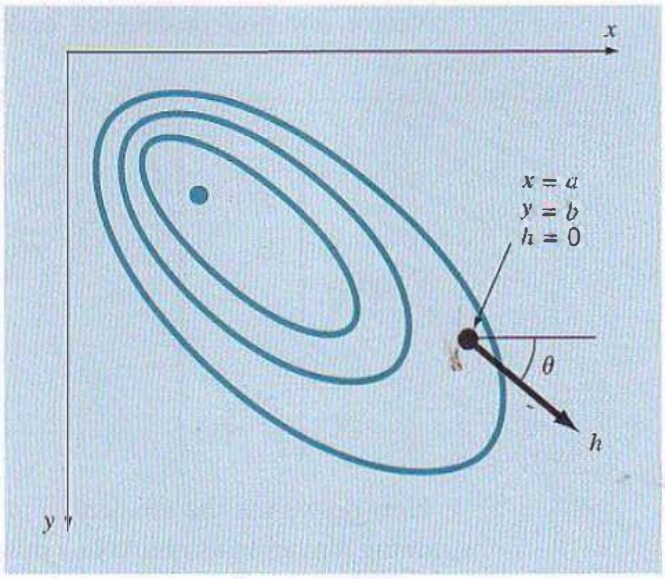


FIGURE 14.6
The directional gradient is defined along an axis h that forms an angle θ with the x axis.

These ideas were useful to us in the one-dimensional search algorithms we explored in the previous chapter. However, to fully understand multidimensional searches, we must first understand how the first and second derivatives are expressed in a multidimensional context.

The Gradient. Suppose we have a two-dimensional function $f(x, y)$. An example might be your elevation on a mountain as a function of your position. Suppose that you are at a specific location on the mountain (a, b) and you want to know the slope in an arbitrary direction. One way to define the direction is along a new axis h that forms an angle θ with the x axis (Fig. 14.6). The elevation along this new axis can be thought of as a new function $g(h)$. If you define your position as being the origin of this axis (that is, $h = 0$), the slope in this direction would be designated as $g'(0)$. This slope, which is called the *directional derivative*, can be calculated from the partial derivatives along the x and y axis by

$$g'(0) = \frac{\partial f}{\partial x}\cos\theta + \frac{\partial f}{\partial y}\sin\theta \tag{14.1}$$

where the partial derivatives are evaluated at $x = a$ and $y = b$.

Assuming that your goal is to gain the most elevation with the next step, the next logical question would be: what direction is the steepest ascent? The answer to this question is provided very neatly by what is referred to mathematically as the *gradient*, which is defined as

$$\nabla f = \frac{\partial f}{\partial x}\mathbf{i} + \frac{\partial f}{\partial y}\mathbf{j} \tag{14.2}$$

This vector is also referred to as "del f." It represents the directional derivative of $f(x, y)$ at point $x = a$ and $y = b$.

Vector notation provides a concise means to generalize the gradient to sions, as

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(\mathbf{x}) \\ \frac{\partial f}{\partial x_2}(\mathbf{x}) \\ \cdot \\ \cdot \\ \cdot \\ \frac{\partial f}{\partial x_n}(\mathbf{x}) \end{bmatrix}$$

How do we use the gradient? For the mountain-climbing problem, if we are in gaining elevation as quickly as possible, the gradient tells us what direction t cally and how much we will gain by taking it. Note, however, that this strateg necessarily take us on a direct path to the summit! We will discuss these ide depth later in this chapter.

EXAMPLE 14.2 Using the Gradient to Evaluate the Path of Steepest Ascent

Problem Statement. Employ the gradient to evaluate the steepest ascent direct function

$$f(x, y) = xy^2$$

at the point (2, 2). Assume that positive x is pointed east and positive y is poi

Solution. First, our elevation can be determined as

$$f(4, 2) = 2(2)^2 = 8$$

Next, the partial derivatives can be evaluated,

$$\frac{\partial f}{\partial x} = y^2 = 2^2 = 4$$

$$\frac{\partial f}{\partial y} = 2xy = 2(2)(2) = 8$$

which can be used to determine the gradient as

$$\nabla f = 4\mathbf{i} + 8\mathbf{j}$$

This vector can be sketched on a topographical map of the function, as in Fi immediately tells us that the direction we must take is

$$\theta = \tan^{-1}\left(\frac{8}{4}\right) = 1.107 \text{ radians } (= 63.4°)$$

relative to the x axis. The slope in this direction, which is the magnitude of ∇f culated as

$$\sqrt{4^2 + 8^2} = 8.944$$

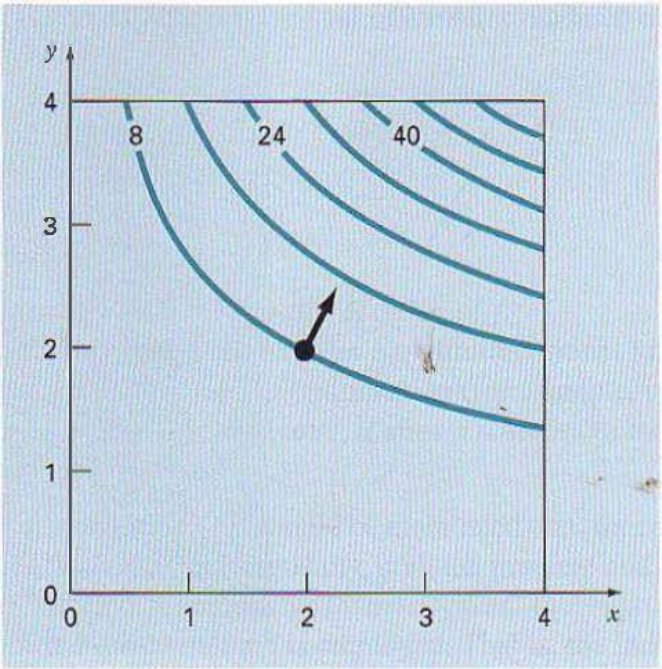


FIGURE 14.7
The arrow follows the direction of steepest ascent calculated with the gradient.

Thus, during our first step, we will initially gain 8.944 units of elevation rise for a unit distance advanced along this steepest path. Observe that Eq. (14.1) yields the same result,

$$g'(0) = 4\cos(1.107) + 8\sin(1.107) = 8.944$$

Note that for any other direction, say $\theta = 1.107/2 = 0.5235$, $g'(0) = 4\cos(0.5235) + 8\sin(0.5235) = 7.608$, which is smaller.

As we move forward, both the direction and magnitude of the steepest path will change. These changes can be quantified at each step using the gradient, and your climbing direction modified accordingly.

A final insight can be gained by inspecting Fig. 14.7. As indicated, the direction of steepest ascent is perpendicular, or *orthogonal,* to the elevation contour at the coordinate (2, 2). This is a general characteristic of the gradient.

Aside from defining a steepest path, the first derivative can also be used to discern whether an optimum has been reached. As is the case for a one-dimensional function, if the partial derivatives with respect to both x and y are zero, a two-dimensional optimum has been reached.

The Hessian. For one-dimensional problems, both the first and second derivatives provide valuable information for searching out optima. The first derivative (*a*) provides a steepest trajectory of the function and (*b*) tells us that we have reached an optimum. Once at an optimum, the second derivative tells us whether we are a maximum [negative $f''(x)$]

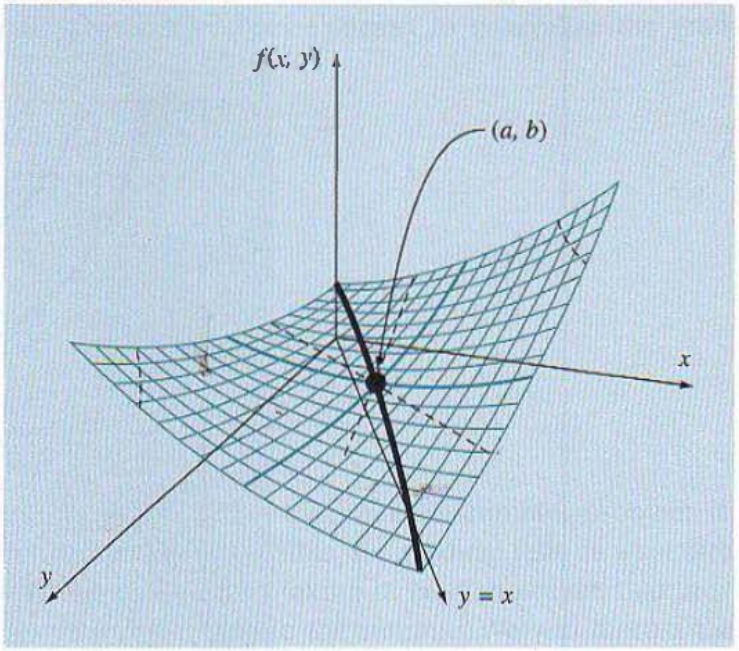


FIGURE 14.8
A saddle point ($x = a$ and $y = b$). Notice that when the curve is viewed along the x and y directions, the function appears to go through a minimum (positive second derivative), whereas when viewed along an axis $x = y$, it is concave downward (negative second derivative).

or a minimum [positive $f''(x)$]. In the previous paragraphs, we illustrated how the gradient provides best local trajectories for multidimensional problems. Now, we will examine how the second derivative is used in such contexts.

You might expect that if the partial second derivatives with respect to both x and y are both negative, then you have reached a maximum. Figure 14.8 shows a function where this is not true. The point (a, b) of this graph appears to be a minimum when observed along either the x dimension or the y dimension. In both instances, the second partial derivatives are positive. However, if the function is observed along the line $y = x$, it can be seen that a maximum occurs at the same point. This shape is called a *saddle*, and clearly, neither a maximum or a minimum occurs at the point.

Whether a maximum or a minimum occurs involves not only the partials with respect to x and y but also the second partial with respect to x and y. Assuming that the partial derivatives are continuous at and near the point being evaluated, the following quantity can be computed:

$$|H| = \frac{\partial^2 f}{\partial x^2}\frac{\partial^2 f}{\partial y^2} - \left(\frac{\partial^2 f}{\partial x \partial y}\right)^2$$

Three cases can occur

- If $|H| > 0$ and $\partial^2 f/\partial x^2 > 0$, then $f(x, y)$ has a local minimum.
- If $|H| > 0$ and $\partial^2 f/\partial x^2 < 0$, then $f(x, y)$ has a local maximum.
- If $|H| < 0$, then $f(x, y)$ has a saddle point.

The quantity $|H|$ is equal to the determinant of a matrix made up of the second derivatives,[1]

$$H = \begin{bmatrix} \dfrac{\partial^2 f}{\partial x^2} & \dfrac{\partial^2 f}{\partial x \partial y} \\ \dfrac{\partial^2 f}{\partial y \partial x} & \dfrac{\partial^2 f}{\partial y^2} \end{bmatrix} \tag{14.4}$$

where this matrix is formally referred to as the *Hessian* of f.

Besides providing a way to discern whether a multidimensional function has reached an optimum, the Hessian has other uses in optimization (for example, for the multidimensional form of Newton's method). In particular, it allows searches to include second-order curvature to attain superior results.

Finite-Difference Approximations. It should be mentioned that, for cases where they are difficult or inconvenient to compute analytically, both the gradient and the determinant of the Hessian can be evaluated numerically. In most cases, the approach introduced in Sec. 6.3.3 for the modified secant method is employed. That is, the independent variables can be perturbed slightly to generate the required partial derivatives. For example, if a centered-difference approach is adopted, they can be computed as

$$\frac{\partial f}{\partial x} = \frac{f(x + \delta x, y) - f(x - \delta x, y)}{2\delta x} \tag{14.5}$$

$$\frac{\partial f}{\partial y} = \frac{f(x, y + \delta y) - f(x, y - \delta y)}{2\delta y} \tag{14.6}$$

$$\frac{\partial^2 f}{\partial x^2} = \frac{f(x + \delta x, y) - 2f(x, y) + f(x - \delta x, y)}{\delta x^2} \tag{14.7}$$

$$\frac{\partial^2 f}{\partial y^2} = \frac{f(x, y + \delta y) - 2f(x, y) + f(x, y - \delta y)}{\delta y^2} \tag{14.8}$$

$$\frac{\partial^2 f}{\partial x \partial y} = \frac{f(x + \delta x, y + \delta y) - f(x + \delta x, y - \delta y) - f(x - \delta x, y + \delta y) + f(x - \delta x, y - \delta y)}{4\delta x \delta y} \tag{14.9}$$

where δ is some small fractional value.

Note that the methods employed in commercial software packages also use forward differences. In addition, they are usually more complicated than the approximations listed in Eqs. (14.5) through (14.9). For example, the IMSL library bases the perturbation on machine epsilon. Dennis and Schnabel (1996) provide more detail on the approach.

Regardless of how the approximation is implemented, the important point is that you may have the option of evaluating the gradient and/or the Hessian analytically. This can sometimes be an arduous task, but the performance of the algorithm may benefit enough

[1]Note that $\partial^2 f/(\partial x \partial y) = \partial^2 f/(\partial y \partial x)$.

to make your effort worthwhile. The closed-form derivatives will be exact, but mo portantly, you will reduce the number of function evaluations. This latter point can critical impact on the execution time.

On the other hand, you will often exercise the option of having the quantitie puted internally using numerical approaches. In many cases, the performance will b adequate and you will be saved the difficulty of numerous partial differentiations would be the case on the optimizers used in certain spreadsheets and mathematic ware packages (for example, Excel). In such cases, you may not even be given the of entering an analytically derived gradient and Hessian. However, for small to mod sized problems, this is usually not a major shortcoming.

14.2.2 Steepest Ascent Method

An obvious strategy for climbing a hill would be to determine the maximum slope starting position and then start walking in that direction. But clearly, another problem almost immediately. Unless you were really lucky and started on a ridge that poin rectly to the summit, as soon as you moved, your path would diverge from the steep cent direction.

Recognizing this fact, you might adopt the following strategy. You could walk distance along the gradient direction. Then you could stop, reevaluate the gradie walk another short distance. By repeating the process you would eventually get to of the hill.

Although this strategy sounds superficially sound, it is not very practical. In par the continuous reevaluation of the gradient can be computationally demanding. Apr approach involves moving in a fixed path along the initial gradient until $f(x, y)$ st creasing, that is, becomes level along your direction of travel. This stopping point be the starting point where ∇f is reevaluated and a new direction followed. The proces peated until the summit is reached. This approach is called the *steepest ascent met* is the most straightforward of the gradient search techniques. The basic idea behind proach is depicted in Fig. 14.9.

We start at an initial point (x_0, y_0) labeled "0" in the figure. At this point, we det the direction of steepest ascent, that is, the gradient. We then search along the direc the gradient, h_0, until we find a maximum, which is labeled "1" in the figure. The is then repeated.

Thus, the problem boils down to two parts: (1) determining the "best" direc search and (2) determining the "best value" along that search direction. As we will effectiveness of the various algorithms described in the coming pages depends clever we are at both parts.

For the time being, the steepest ascent method uses the gradient approach as its for the "best" direction. We have already shown how the gradient is evaluated in ple 14.1. Now, before examining how the algorithm goes about locating the ma along the steepest direction, we must pause to explore how to transform a function y into a function of h along the gradient direction.

[2]Because of our emphasis on maximization here, we use the terminology *steepest ascent*. The same app also be used for minimization, in which case the terminology *steepest descent* is used.

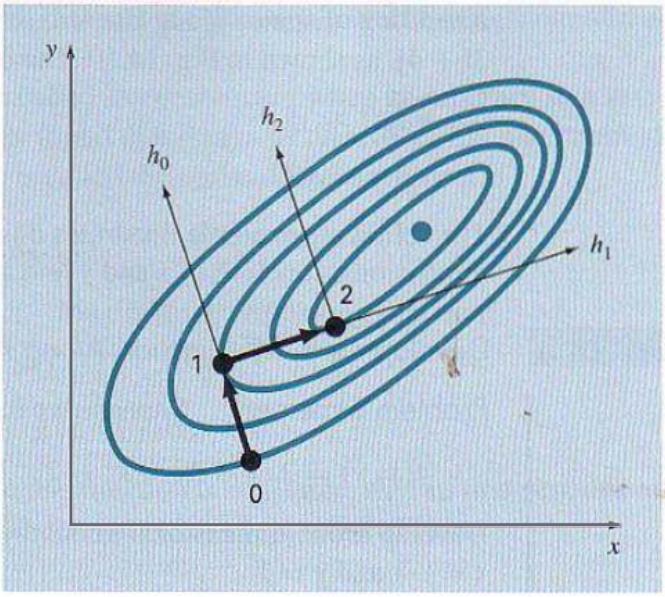


FIGURE 14.9
A graphical depiction of the method of steepest ascent.

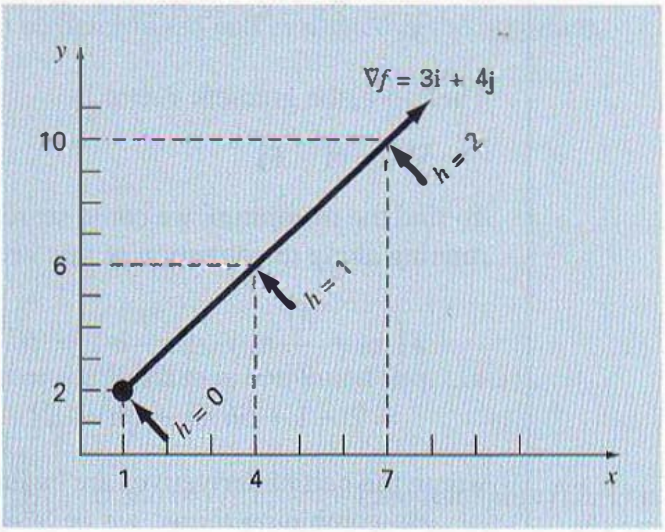


FIGURE 14.10
The relationship between an orbitrary direction h and x and y coordinates.

Starting at x_0, y_0 the coordinates of any point in the gradient direction can be expressed as

$$x = x_0 + \frac{\partial f}{\partial x}h \tag{14.10}$$

$$y = y_0 + \frac{\partial f}{\partial y}h \tag{14.11}$$

where h is distance along the h axis. For example, suppose $x_0 = 1$ and $y_0 = 2$ and $3\mathbf{i} + 4\mathbf{j}$, as shown in Fig. 14.10. The coordinates of any point along the h axis are given

$$x = 1 + 3h$$
$$y = 2 + 4h$$

The following example illustrates how we can use these transformations to convert a dimensional function of x and y into a one-dimensional function in h.

EXAMPLE 14.3 Developing a 1-D Function Along the Gradient Direction

Problem Statement. Suppose we have the following two-dimensional function:

$$f(x, y) = 2xy + 2x - x^2 - 2y^2$$

Develop a one-dimensional version of this equation along the gradient direction at $x = -1$ and $y = 1$.

Solution. The partial derivatives can be evaluated at $(-1, 1)$,

$$\frac{\partial f}{\partial x} = 2y + 2 - 2x = 2(1) + 2 - 2(-1) = 6$$
$$\frac{\partial f}{\partial y} = 2x - 4y = 2(-1) - 4(1) = -6$$

Therefore, the gradient vector is

$$\nabla f = 6\mathbf{i} - 6\mathbf{j}$$

To find the maximum, we could search along the gradient direction, that is, along an axis running along the direction of this vector. The function can be expressed along this axis

$$f\left(x_0 + \frac{\partial f}{\partial x}h, y_0 + \frac{\partial f}{\partial y}h\right) = f(-1 + 6h, 1 - 6h)$$
$$= 2(-1 + 6h)(1 - 6h) + 2(-1 + 6h) - (-1 + 6h)^2 - 2(1 - 6h)^2$$

where the partial derivatives are evaluated at $x = -1$ and $y = 1$.

By combining terms, we develop a one-dimensional function $g(h)$ that maps f along the h axis,

$$g(h) = -180h^2 + 72h - 7$$

Now that we have developed a function along the path of steepest ascent, we can explore how to answer the second question. That is, how far along this path do we travel? One approach might be to move along this path until we find the maximum of this function. We will call the location of this maximum h^*. This is the value of the step that maximizes (and hence, f) in the gradient direction. This problem is equivalent to finding the maximum of a function of a single variable h. This can be done using different one-dimensional search techniques like the ones we discussed in Chap. 13. Thus, we convert from a

the optimum of a two-dimensional function to performing a one-dimensional search along the gradient direction.

This method is called *steepest ascent* when an arbitrary step size h is used. If a value of a single step h^* is found that brings us directly to the maximum along the gradient direction, the method is called the *optimal steepest ascent.*

EXAMPLE 14.4 Optimal Steepest Ascent

Problem Statement. Maximize the following function:

$$f(x, y) = 2xy + 2x - x^2 - 2y^2$$

using initial guesses, $x = -1$ and $y = 1$.

Solution. Because this function is so simple, we can first generate an analytical solution. To do this, the partial derivatives can be evaluated as

$$\frac{\partial f}{\partial x} = 2y + 2 - 2x = 0$$

$$\frac{\partial f}{\partial y} = 2x - 4y = 0$$

This pair of equations can be solved for the optimum, $x = 2$ and $y = 1$. The second partial derivatives can also be determined and evaluated at the optimum,

$$\frac{\partial^2 f}{\partial x^2} = -2$$

$$\frac{\partial^2 f}{\partial y^2} = -4$$

$$\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x} = 2$$

and the determinant of the Hessian is computed [Eq. (14.3)],

$$|H| = -2(-4) - 2^2 = 4$$

Therefore, because $|H| > 0$ and $\partial^2 f/\partial x^2 < 0$, function value $f(2, 1)$ is a maximum.

Now let us implement steepest ascent. Recall that, at the end of Example 14.3, we had already implemented the initial steps of the problem by generating

$$g(h) = -180h^2 + 72h - 7$$

Now, because this is a simple parabola, we can directly locate the maximum (that is, $h = h^*$) by solving the problem,

$$g'(h^*) = 0$$

$$-360h^* + 72 = 0$$

$$h^* = 0.2$$

This means that if we travel along the h axis, $g(h)$ reaches a minimum value when $h = h^* = 0.2$. This result can be placed back into Eqs. (14.10) and (14.11) to solve for the

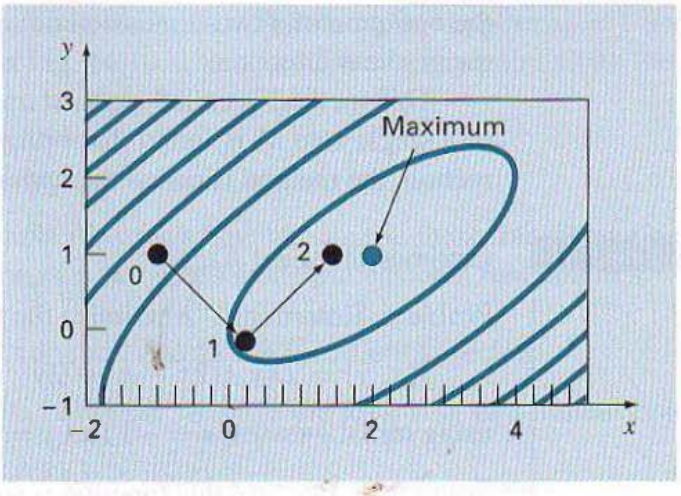


FIGURE 14.11
The method of optimal steepest ascent.

(x, y) coordinates corresponding to this point,

$$x = -1 + 6(0.2) = 0.2$$
$$y = 1 - 6(0.2) = -0.2$$

This step is depicted in Fig. 14.11 as the move from point 0 to 1.

The second step is merely implemented by repeating the procedure. First, the p derivatives can be evaluated at the new starting point (0.2, −0.2) to give

$$\frac{\partial f}{\partial x} = 2(-0.2) + 2 - 2(0.2) = 1.2$$
$$\frac{\partial f}{\partial y} = 2(0.2) - 4(-0.2) = 1.2$$

Therefore, the gradient vector is

$$\nabla f = 1.2\mathbf{i} + 1.2\mathbf{j}$$

This means that the steepest direction is now pointed up and to the right at a 45° angl the x axis (see Fig. 14.11). The coordinates along this new h axis can now be expres

$$x = 0.2 + 1.2h$$
$$y = -0.2 + 1.2h$$

Substituting these values into the function yields

$$f(0.2 + 1.2h, -0.2 + 1.2h) = g(h) = -1.44h^2 + 2.88h + 0.2$$

The step h^* to take us to the maximum along the search direction can then be directl puted as

$$g'(h^*) = -2.88h^* + 2.88 = 0$$
$$h^* = 1$$

This result can be placed back into Eqs. (14.10) and (14.11) to solve for the (x, y) coordinates corresponding to this new point.

$$x = 0.2 + 1.2(1) = 1.4$$
$$y = -0.2 + 1.2(1) = 1$$

As depicted in Fig. 14.11, we move to the new coordinates, labeled point 2 in the plot, and in so doing move closer to the maximum. The approach can be repeated with the final result converging on the analytical solution, $x = 2$ and $y = 1$.

It can be shown that the method of steepest descent is linearly convergent. Further, it tends to move very slowly along long, narrow ridges. This is because the new gradient at each maximum point will be perpendicular to the original direction. Thus, the technique takes many small steps criss-crossing the direct route to the summit. Hence, although it is reliable, there are other approaches that converge much more rapidly, particularly in the vicinity of an optimum. The remainder of the section is devoted to such methods.

14.2.3 Advanced Gradient Approaches

Conjugate Gradient Method (Fletcher-Reeves). In Sec. 14.1.2, we have seen how conjugate directions in Powell's method greatly improved the efficiency of a univariate search. In a similar manner, we can also improve the linearly convergent steepest ascent using conjugate gradients. In fact, an optimization method that makes use of conjugate gradients to define search directions can be shown to be quadratically convergent. This also ensures that the method will optimize a quadratic function exactly in a finite number of steps regardless of the starting point. Since most well-behaved functions can be approximated reasonably well by a quadratic in the vicinity of an optimum, quadratically convergent approaches are often very efficient near an optimum.

We have seen how starting with two arbitrary search directions, Powell's method produced new conjugate search directions. This method is quadratically convergent and does not require gradient information. On the other hand, if evaluation of derivatives is practical, we can devise algorithms that combine the ideas of steepest descent and conjugate directions to achieve robust initial performance and rapid convergence as the technique gravitates toward the optimum. The *Fletcher-Reeves conjugate gradient algorithm* modifies the steepest-ascent method by imposing the condition that successive gradient search directions be mutually conjugate. The proof and algorithm are beyond the scope of the text but are described by Rao (1996).

Newton's Method. Newton's method for a single variable (recall Sec. 13.3) can be extended to multivariate cases. Write a second-order Taylor series for $f(\mathbf{x})$ near $\mathbf{x} = \mathbf{x}_i$,

$$f(\mathbf{x}) = f(\mathbf{x}_i) + \nabla f^T(\mathbf{x}_i)(\mathbf{x} - \mathbf{x}_i) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_i)^T H_i(\mathbf{x} - \mathbf{x}_i)$$

where H_i is the Hessian matrix. At the minimum,

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}_j} = 0 \qquad \text{for } j = 1, 2, \ldots, n$$

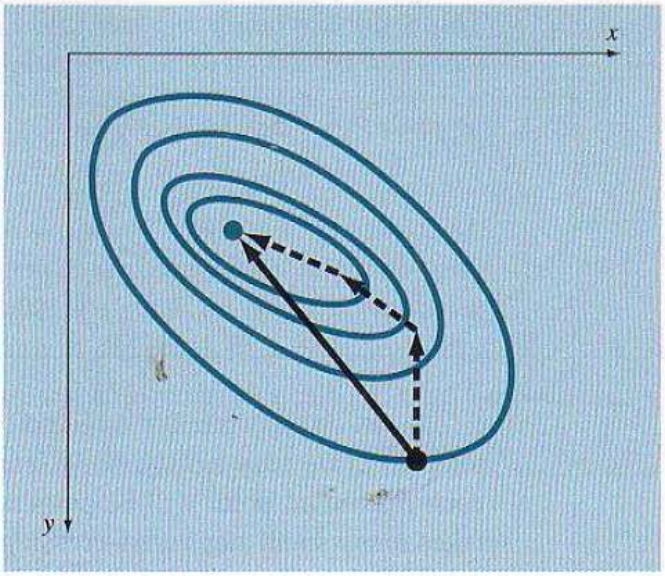


FIGURE 14.12
When the starting point is close to the optimal point, following the gradient can be inefficient. Newton methods attempt to search along a direct path to the optimum (solid line).

Thus,

$$\nabla f = \nabla f(\mathbf{x}_i) + H_i(\mathbf{x} - \mathbf{x}_i) = 0$$

If H is nonsingular,

$$\mathbf{x}_{i+1} = \mathbf{x}_i - H_i^{-1}\nabla f$$

which can be shown to converge quadratically near the optimum. This method ag[illegible] forms better than the steepest ascent method (see Fig. 14.12). However, note [illegible] method requires both the computation of second derivatives and matrix inversion a[illegible] iteration. Thus, the method is not very useful in practice for functions with large nu[illegible] of variables. Furthermore, Newton's method may not converge if the starting point is close to the optimum.

Marquardt Method. We know that the method of steepest ascent increases the func[illegible] value even if the starting point is far from an optimum. On the other hand, we have described Newton's method, which converges rapidly near the maximum. *Marqua[illegible] method* uses the steepest descent method when $\mathbf{x}$ is far from $\mathbf{x}^*$, and Newton's me[illegible] when $\mathbf{x}$ closes in on an optimum. This is accomplished by modifying the diagonal of Hessian in Eq. (14.14),

$$\tilde{H}_i = H_i + \alpha_i I$$

where α_i is a positive constant and I is the identity matrix. At the start of the procedur[illegible] is assumed to be large and

$$\tilde{H}_i^{-1} \approx \frac{1}{\alpha_i} I$$

which reduces Eq. (14.14) to the steepest ascent method. As the iterations proceed, α_i approaches zero and the method becomes Newton's method.

Thus, Marquardt's method offers the best of both worlds: it plods along reliably from poor initial starting values yet accelerates rapidly when it approaches the optimum. Unfortunately, the method still requires Hessian evaluation and matrix inversion at each step.

It should be noted that the Marquardt method is primarily used for nonlinear least-squares problems. For example, the IMSL library contains a subroutine for this purpose.

Quasi-Newton Methods. *Quasi-Newton*, or *variable metric, methods* seek to estimate the direct path to the optimum in a manner similar to Newton's method. However, notice that the Hessian matrix in Eq. (14.14) is composed of the second derivatives of f that vary from step to step. Quasi-Newton methods attempt to avoid these difficulties by approximating H with another matrix A using only first partial derivatives of f. The approach involves starting with an initial approximation of H^{-1} and updating and improving it with each iteration. The methods are called quasi-Newton because we do not use the true Hessian, rather an approximation. Thus, we have two approximations at work simultaneously: (1) the original Taylor-series approximation and (2) the Hessian approximation.

There are two primary methods of this type: the *Davidon-Fletcher-Powell* (DFP) and the *Broyden-Fletcher-Goldfarb-Shanno* (BFGS) algorithms. They are similar except for details concerning how they handle round-off error and convergence issues. BFGS is generally recognized as being superior in most cases. Rao (1996) provides details and formal statements of both the DFP and the BFGS algorithms.

OBLEMS

xample 14.2 for the following function at the point

$$xy + 1.5y - 1.25x^2 - 2y^2 + 5$$

directional derivative of

$$^2 + 2y^2$$

$= 2$ in the direction of $h = 2i + 3j$.

gradient vector and Hessian matrix for each of the ions:

$xy^2 + 2e^{xy}$

$2x^2 + y^2 + z^2$

$(x^2 + 3xy + 2y^2)$

$$.25xy + 1.75y - 1.5x^2 - 2y^2$$

solve a system of linear algebraic equations that . Note that this is done by setting the partial derivaespect to both x and y to zero.

n initial guess of $x = 1$ and $y = 1$ and apply two applihe steepest ascent method to $f(x, y)$ from Prob. 14.4.

(h) Construct a plot from the results of (a) showing the path of the search.

14.6 Find the minimum value of

$$f(x, y) = (x - 3)^2 + (y - 2)^2$$

starting at $x = 1$ and $y = 1$, using the steepest descent method with a stopping criterion of $\varepsilon_s = 1\%$. Explain your results.

14.7 Perform one iteration of the steepest ascent method to locate the maximum of

$$f(x, y) = 4x + 2y + x^2 - 2x^4 + 2xy - 3y^2$$

using initial guesses $x = 0$ and $y = 0$. Employ bisection to find the optimal step size in the gradient search direction.

14.8 Perform one iteration of the optimal gradient steepest descent method to locate the minimum of

$$f(x, y) = -8x + x^2 + 12y + 4y^2 - 2xy$$

using initial guesses $x = 0$ and $y = 0$.

14.9 Develop a program using a programming or macro language to implement the random search method. Design the subprogram so that it is expressly designed to locate a maximum. Test the program with $f(x, y)$ from Prob. 14.7. Use a range of -2 to 2 for both x and y.

14.10 The grid search is another brute force approach to optimization. The two-dimensional version is depicted in Fig. P14.10. The x and y dimensions are divided into increments to create a grid. The function is then evaluated at each node of the grid. The denser the grid, the more likely it would be to locate the optimum.

Develop a program using a programming or macro language to implement the grid search method. Design the program so that it is expressly designed to locate a maximum. Test it with the same problem as Example 14.1.

14.11 Develop a one-dimensional equation in the pressure gradient direction at the point (4, 2). The pressure function is

$$f(x, y) = 6x^2y - 9y^2 - 8x^2$$

14.12 A temperature function is

$$f(x, y) = 2x^3y^2 - 7xy + x^2 + 3y$$

Develop a one-dimensional function in the temperature gradient direction at the point (1, 1).

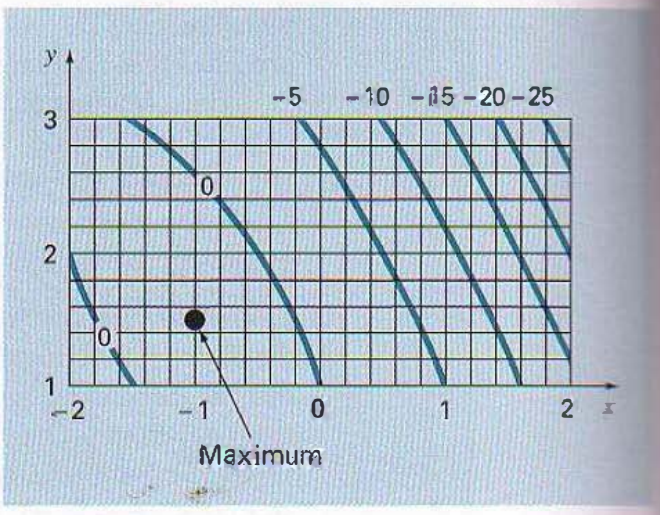


Figure P14.10
The grid search.

CURVE FITTING

PT5.1 MOTIVATION

Data is often given for discrete values along a continuum. However, you may require estimates at points between the discrete values. The present part of this book describes techniques to fit curves to such data to obtain intermediate estimates. In addition, you may require a simplified version of a complicated function. One way to do this is to compute values of the function at a number of discrete values along the range of interest. Then, a simpler function may be derived to fit these values. Both of these applications are known as *curve fitting*.

There are two general approaches for curve fitting that are distinguished from each other on the basis of the amount of error associated with the data. First, where the data exhibits a significant degree of error or "noise," the strategy is to derive a single curve that represents the general trend of the data. Because any individual data point may be incorrect, we make no effort to intersect every point. Rather, the curve is designed to follow the pattern of the points taken as a group. One approach of this nature is called *least-squares regression* (Fig. PT5.1*a*).

Second, where the data is known to be very precise, the basic approach is to fit a curve or a series of curves that pass directly through each of the points. Such data usually originates from tables. Examples are values for the density of water or for the heat capacity of gases as a function of temperature. The estimation of values between well-known discrete points is called *interpolation* (Fig. PT5.1*b* and *c*).

PT5.1.1 Noncomputer Methods for Curve Fitting

The simplest method for fitting a curve to data is to plot the points and then sketch a line that visually conforms to the data. Although this is a valid option when quick estimates are required, the results are dependent on the subjective viewpoint of the person sketching the curve.

For example, Fig. PT5.1 shows sketches developed from the same set of data by three engineers. The first did not attempt to connect the points, but rather, characterized the general upward trend of the data with a straight line (Fig. PT5.1*a*). The second engineer used straight-line segments or linear interpolation to connect the points (Fig. PT5.1*b*). This is a very common practice in engineering. If the values are truly close to being linear or are spaced closely, such an approximation provides estimates that are adequate for many engineering calculations. However, where the underlying relationship is highly curvilinear or the data is widely spaced, significant errors can be introduced by such linear interpolation. The third engineer used curves to try to capture the meanderings suggested by the data (Fig. PT5.1*c*). A fourth or fifth engineer would likely develop alternative fits. Obviously,

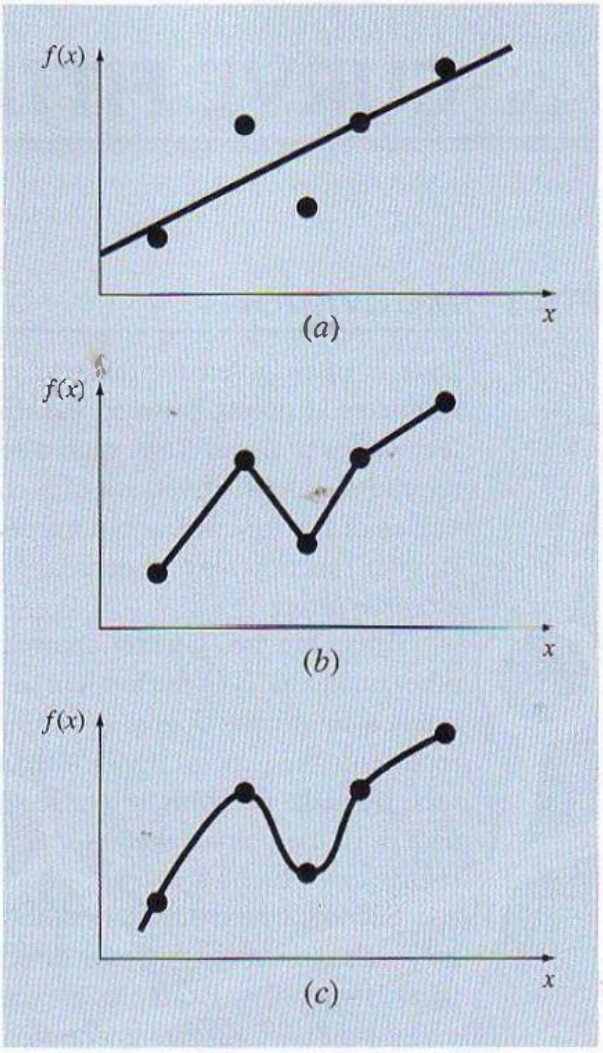


FIGURE PT5.1
Three attempts to fit a "best" curve through five data points. (a) Least-squares regression, (b) linear interpolation, and (c) curvilinear interpolation.

our goal here is to develop systematic and objective methods for the purpose of d such curves.

PT5.1.2 Curve Fitting and Engineering Practice

Your first exposure to curve fitting may have been to determine intermediate valu tabulated data—for instance, from interest tables for engineering economics or fron tables for thermodynamics. Throughout the remainder of your career, you will ha quent occasion to estimate intermediate values from such tables.

Although many of the widely used engineering properties have been tabulate are a great many more that are not available in this convenient form. Special cases a problem contexts often require that you measure your own data and develop your o dictive relationships. Two types of applications are generally encountered when experimental data: trend analysis and hypothesis testing.

Trend analysis represents the process of using the pattern of the data to make predictions. For cases where the data is measured with high precision, you might utilize interpolating polynomials. Imprecise data is often analyzed with least-squares regression.

Trend analysis may be used to predict or forecast values of the dependent variable. This can involve extrapolation beyond the limits of the observed data or interpolation within the range of the data. All fields of engineering commonly involve problems of this type.

A second engineering application of experimental curve fitting is *hypothesis testing*. Here, an existing mathematical model is compared with measured data. If the model coefficients are unknown, it may be necessary to determine values that best fit the observed data. On the other hand, if estimates of the model coefficients are already available, it may be appropriate to compare predicted values of the model with observed values to test the adequacy of the model. Often, alternative models are compared and the "best" one is selected on the basis of empirical observations.

In addition to the above engineering applications, curve fitting is important in other numerical methods such as integration and the approximate solution of differential equations. Finally, curve-fitting techniques can be used to derive simple functions to approximate complicated functions.

PT5.2 MATHEMATICAL BACKGROUND

The prerequisite mathematical background for interpolation is found in the material on Taylor series expansions and finite divided differences introduced in Chap. 4. Least-squares regression requires additional information from the field of statistics. If you are familiar with the concepts of the mean, standard deviation, residual sum of the squares, normal distribution, and confidence intervals, feel free to skip the following pages and proceed directly to PT5.3. If you are unfamiliar with these concepts or are in need of a review, the following material is designed as a brief introduction to these topics.

PT5.2.1 Simple Statistics

Suppose that in the course of an engineering study, several measurements were made of a particular quantity. For example, Table PT5.1 contains 24 readings of the coefficient of thermal expansion of a structural steel. Taken at face value, the data provides a limited amount of information—that is, that the values range from a minimum of 6.395 to a maximum of 6.775. Additional insight can be gained by summarizing the data in one or more well-chosen statistics that convey as much information as possible about specific characteristics of the data set. These descriptive statistics are most often selected to represent

TABLE PT5.1 Measurements of the coefficient of thermal expansion of structural steel [$\times 10^{-6}$ in/(in · °F)].

6.495	6.595	6.615	6.635	6.485	6.555
6.665	6.505	6.435	6.625	6.715	6.655
6.755	6.625	6.715	6.575	6.655	6.605
6.565	6.515	6.555	6.395	6.775	6.685

(1) the location of the center of the distribution of the data and (2) the degree of spread of the data set.

The most common location statistic is the arithmetic mean. The *arithmetic mean* of a sample is defined as the sum of the individual data points (y_i) divided by the number of points (n), or

$$\bar{y} = \frac{\Sigma y_i}{n}$$

where the summation (and all the succeeding summations in this introduction) is from $i = 1$ through n.

The most common measure of spread for a sample is the *standard deviation* about the mean,

$$s_y = \sqrt{\frac{S_t}{n-1}}$$

where S_t is the total sum of the squares of the residuals between the data points and the mean, or

$$S_t = \Sigma (y_i - \bar{y})^2$$

Thus, if the individual measurements are spread out widely around the mean, S_t (and consequently, s_y) will be large. If they are grouped tightly, the standard deviation will be small. The spread can also be represented by the square of the standard deviation, which is called the *variance*:

$$s_y^2 = \frac{\Sigma (y_i - \bar{y})^2}{n-1}$$

Note that the denominator in both Eqs. (PT5.2) and (PT5.4) is $n - 1$. The quantity $n - 1$ is referred to as the degrees of freedom. Hence S_t and s_y are said to be based on $n - 1$ *degrees of freedom*. This nomenclature derives from the fact that the sum of the quantities on which S_t is based (that is, $\bar{y} - y_1, \bar{y} - y_2, \ldots, \bar{y} - y_n$) is zero. Consequently, if $\bar{y}$ is known and $n - 1$ of the values are specified, the remaining value is fixed. Thus, only $n - 1$ values are said to be freely determined. Another justification for dividing by $n - 1$ is the fact that there is no such thing as the spread of a single data point. For the case where $n = 1$, Eqs. (PT5.2) and (PT5.4) yield a meaningless result of infinity.

It should be noted that an alternative, more convenient formula is available to compute the standard deviation,

$$s_y^2 = \frac{\Sigma y_i^2 - (\Sigma y_i)^2 / n}{n-1}$$

This version does not require precomputation of $\bar{y}$ and yields an identical result to Eq. (PT5.4).

A final statistic that has utility in quantifying the spread of data is the *coefficient of variation* (c.v.). This statistic is the ratio of the standard deviation to the mean. As such, it provides a normalized measure of the spread. It is often multiplied by 100 so that it can be expressed in the form of a percent:

$$\text{c.v.} = \frac{s_y}{\bar{y}} 100\% \tag{PT5.5}$$

Notice that the coefficient of variation is similar in spirit to the percent relative error (ε_t) discussed in Sec. 3.3. That is, it is the ratio of a measure of error (s_y) to an estimate of the true value ($\bar{y}$).

EXAMPLE PT5.1 Simple Statistics of a Sample

Problem Statement. Compute the mean, variance, standard deviation, and coefficient of variation for the data in Table PT5.1.

TABLE PT5.2 Computations for statistics for the readings of the coefficient of thermal expansion. The frequencies and bounds are developed to construct the histogram in Fig. PT5.2.

				Interval	
i	y_i	$(y_i - \bar{y})^2$	Frequency	Lower Bound	Upper Bound
1	6.395	0.042025	1	6.36	6.40
2	6.435	0.027225	1	6.40	6.44
3	6.485	0.013225			
4	6.495	0.011025	4	6.48	6.52
5	6.505	0.009025			
6	6.515	0.007225			
7	6.555	0.002025	2	6.52	6.56
8	6.555	0.002025			
9	6.565	0.001225			
10	6.575	0.000625	3	6.56	6.60
11	6.595	0.000025			
12	6.605	0.000025			
13	6.615	0.000225			
14	6.625	0.000625	5	6.60	6.64
15	6.625	0.000625			
16	6.635	0.001225			
17	6.655	0.003025			
18	6.655	0.003025	3	6.64	6.68
19	6.665	0.004225			
20	6.685	0.007225			
21	6.715	0.013225	3	6.68	6.72
22	6.715	0.013225			
23	6.755	0.024025	1	6.72	6.76
24	6.775	0.030625	1	6.76	6.80
Σ	158.4	0.217000			

Solution. The data is added (Table PT5.2), and the results are used to [Eq. (PT5.1)]

$$\bar{y} = \frac{158.4}{24} = 6.6$$

As in Table PT5.2, the sum of the squares of the residuals is 0.217000, which to compute the standard deviation [Eq. (PT5.2)]:

$$s_y = \sqrt{\frac{0.217000}{24 - 1}} = 0.097133$$

the variance [Eq. (PT5.4)]:

$$s_y^2 = 0.009435$$

and the coefficient of variation [Eq. (PT5.5)]:

$$\text{c.v.} = \frac{0.097133}{6.6} 100\% = 1.47\%$$

PT5.2.2 The Normal Distribution

Another characteristic that bears on the present discussion is the *data distribution*— the shape with which the data is spread around the mean. A histogram provides visual representation of the distribution. As seen in Table PT5.2, the histogram structed by sorting the measurements into intervals. The units of measurement are on the abscissa and the frequency of occurrence of each interval is plotted on the Thus, five of the measurements fall between 6.60 and 6.64. As in Fig. PT5.2, the his suggests that most of the data is grouped close to the mean value of 6.6.

If we have a very large set of data, the histogram often can be approxima smooth curve. The symmetric, bell-shaped curve superimposed on Fig. PT5.2 is characteristic shape—the *normal distribution*. Given enough additional measure histogram for this particular case could eventually approach the normal distribution

The concepts of the mean, standard deviation, residual sum of the squares, and distribution all have great relevance to engineering practice. A very simple exampl use to quantify the confidence that can be ascribed to a particular measurement. If tity is normally distributed, the range defined by $\bar{y} - s_y$ to $\bar{y} + s_y$ will encompass imately 68 percent of the total measurements. Similarly, the range defined by $\bar{y} - 2s_y$ to $\bar{y} + 2s_y$ will encompass approximately 95 percent.

For example, for the data in Table PT5.1 ($\bar{y} = 6.6$ and $s_y = 0.097133$), we can statement that approximately 95 percent of the readings should fall between 6.405 6.794266. If someone told us that they had measured a value of 7.35, we would susp the measurement might be erroneous. The following section elaborates on such evalu

PT5.2.3 Estimation of Confidence Intervals

As should be clear from the previous sections, one of the primary aims of statistics timate the properties of a *population* based on a limited *sample* drawn from that popu

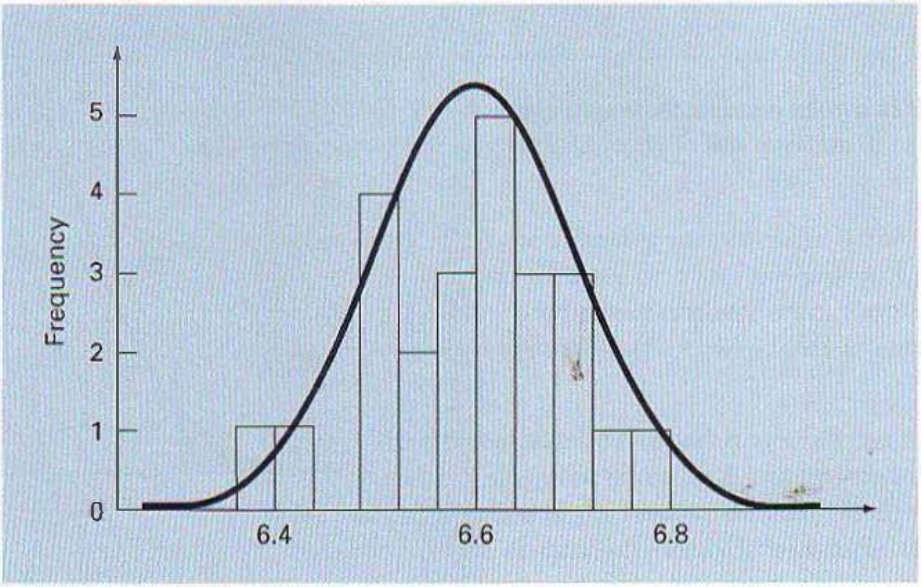


FIGURE PT5.2
A histogram used to depict the distribution of data. As the number of data points increases, the histogram could approach the smooth, bell-shaped curve called the normal distribution.

Clearly, it is impossible to measure the coefficient of thermal expansion for every piece of structural steel that has ever been produced. Consequently, as seen in Tables PT5.1 and PT5.2, we can randomly make a number of measurements and, on the basis of the sample, attempt to characterize the properties of the entire population.

Because we "infer" properties of the unknown population from a limited sample, the endeavor is called *statistical inference*. Because the results are often reported as estimates of the population parameters, the process is also referred to as *estimation*.

We have already shown how we estimate the central tendency (sample mean, $\bar{y}$) and spread (sample standard deviation and variance) of a limited sample. Now, we will briefly describe how we can attach probabilistic statements to the quality of these estimates. In particular, we will discuss how we can define a confidence interval around our estimate of the mean. We have chosen this particular topic because of its direct relevance to the regression models we will be describing in Chap. 17.

Note that in the following discussion, the nomenclature $\bar{y}$ and s_y refer to the sample mean and standard deviation, respectively. The nomenclature μ and σ refer to the population mean and standard deviation, respectively. The former are sometimes referred to as the "estimated" mean and standard deviation, whereas the latter are sometimes called the "true" mean and standard deviation.

An *interval estimator* gives the range of values within which the parameter is expected to lie with a given probability. Such intervals are described as being one-sided or two-sided. As the name implies, a *one-sided interval* expresses our confidence that the parameter estimate is less than or greater than the true value. In contrast, the *two-sided interval* deals with the more general proposition thatthe estimate agrees with the truth with no consideration to the sign of the discrepancy. Because it is more general, we will focus on the two-sided interval.

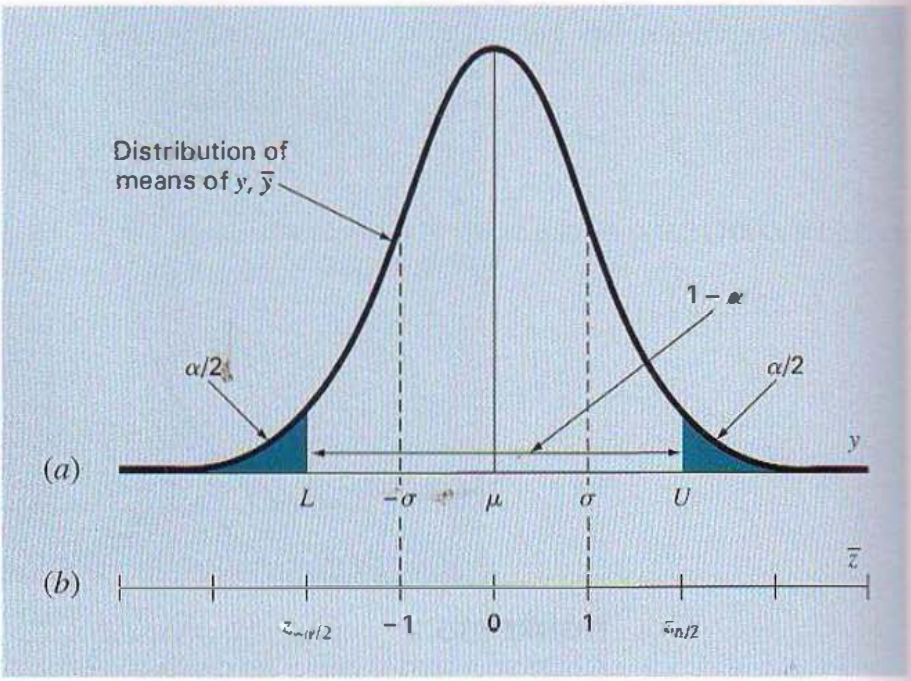


FIGURE PT5.3
A two-sided confidence interval. The abscissa scale in (*a*) is written in the natural units of random variable y. The normalized version of the abscissa in (*b*) has the mean at the origin and scales the axis so that the standard deviation corresponds to a unit value.

A two-sided interval can be described by the statement

$$P\{L \leq \mu \leq U\} = 1 - \alpha$$

which reads, "the probability that the true mean of y, μ, falls within the bound from L to U is $1 - \alpha$." The quantity α is called the *significance level*. So the problem of defining a confidence interval reduces to estimating L and U. Although it is not absolutely necessary, it is customary to view the two-sided interval with the α probability distributed evenly in each tail of the distribution, as in Fig. PT5.3.

If the true variance of the distribution of y, σ^2, is known (which is not usually the case), statistical theory states that the sample mean $\bar{y}$ comes from a normal distribution with mean μ and variance σ^2/n (Box PT5.1). In the case illustrated in Fig. PT5.3, we really want to know μ. Therefore, we do not know where the normal curve is exactly located with respect to $\bar{y}$. To circumvent this dilemma, we compute a new quantity, the *standard normal estimate*

$$\bar{z} = \frac{\bar{y} - \mu}{\sigma/\sqrt{n}}$$

which represents the normalized distance between $\bar{y}$ and μ. According to statistical theory this quantity should be normally distributed with a mean of 0 and a variance of 1. Furthermore, the probability that $\bar{z}$ would fall within the unshaded region of Fig.

Box PT5.1 A Little Statistics

...se several courses to become proficient at statis- ...may not have taken such a course yet, we would ...few ideas that might make this present section

...ated, the "game" of inferential statistics assumes ...riable you are sampling, y, has a true mean (μ). Further, in the present discussion, we also as- ...particular distribution: the normal distribution. ...normal distribution has a finite value that spec- ...the normal distribution. If the variance is large, ...broad. Conversely, if the variance is small, the ...row. Thus, the true variance quantifies the intrin- ...the random variable.

...statistics, we take a limited number of measure- ...ary called a sample. From this sample, we can ...ated mean ($\bar{y}$) and variance (s_y^2). The more mea- ..., the better the estimates approximate the true ... $n \to \infty$, $\bar{y} \to \mu$ and $s_y^2 \to \sigma^2$.

...e take n samples and compute an estimated mean ...another n samples and compute another, $\bar{y}_2$. We ...this process until we have generated a sample of ..., $\bar{y}_m$, where m is large. We can then develop ...ese means and determine a "distribution of the ...a "mean of the means" and a "standard deviation ...w the question arises: does this new distribution ...tatistics behave in a predictable fashion?

There is an extremely important theorem known as the *Central Limit Theorem* that speaks directly to this question. It can be stated as

Let $y_1, y_2, \ldots, y_n$ be a random sample of size n from a distribution with mean μ and variance σ^2. Then, for large n, $\bar{y}$ is approximately normal with mean μ and variance σ^2/n. Furthermore, for large n, the random variable $(\bar{y} - \mu)/(\sigma/\sqrt{n})$ is approximately standard normal.

Thus, the theorem states the remarkable result that the distribution of means will always be normally distributed regardless of the underlying distribution of the random variables! It also yields the expected result that given a sufficiently large sample, the mean of the means should converge on the true population mean μ.

Further, the theorem says that as the sample size gets larger, the variance of the means should approach zero. This makes sense, because if n is small, our individual estimates of the mean should be poor and the variance of the means should be large. As n increases, our estimates of the mean will improve and hence their spread should shrink. The Central Limit Theorem neatly defines exactly how this shrinkage relates to both the true variance and the sample size, that is, as σ^2/n.

Finally, the theorem states the important result that we have given as Eq. (PT5.6). As is shown in this section, this result is the basis for constructing confidence intervals for the mean.

should be $1 - \alpha$. Therefore, the statement can be made that

$$\frac{\bar{y} - \mu}{\sigma/\sqrt{n}} < -z_{\alpha/2} \quad \text{or} \quad \frac{\bar{y} - \mu}{\sigma/\sqrt{n}} > z_{\alpha/2}$$

with a probability of α.

The quantity $z_{\alpha/2}$ is a standard normal random variable. This is the distance measured along the normalized axis above and below the mean that encompasses $1 - \alpha$ probability (Fig. PT5.3*b*). Values of $z_{\alpha/2}$ are tabulated in statistics books (for example, Milton and Arnold, 1995). They can also be calculated using functions on software packages and libraries like Excel and IMSL. As an example, for $\alpha = 0.05$ (in other words, defining an interval encompassing 95%), $z_{\alpha/2}$ is equal to about 1.96. This means that an interval around the mean of width ± 1.96 times the standard deviation will encompass approximately 95% of the distribution.

These results can be rearranged to yield

$$L \le \mu \le U$$

with a probability of $1 - \alpha$, where

$$L = \bar{y} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \qquad U = \bar{y} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}$$

Now, although the foregoing provides an estimate of L and U, it is based on of the true variance σ. For our case, we know only the estimated variance s_y. A ward alternative would be to develop a version of Eq. (PT5.6) based on s_y,

$$t = \frac{\bar{y} - \mu}{s_y / \sqrt{n}}$$

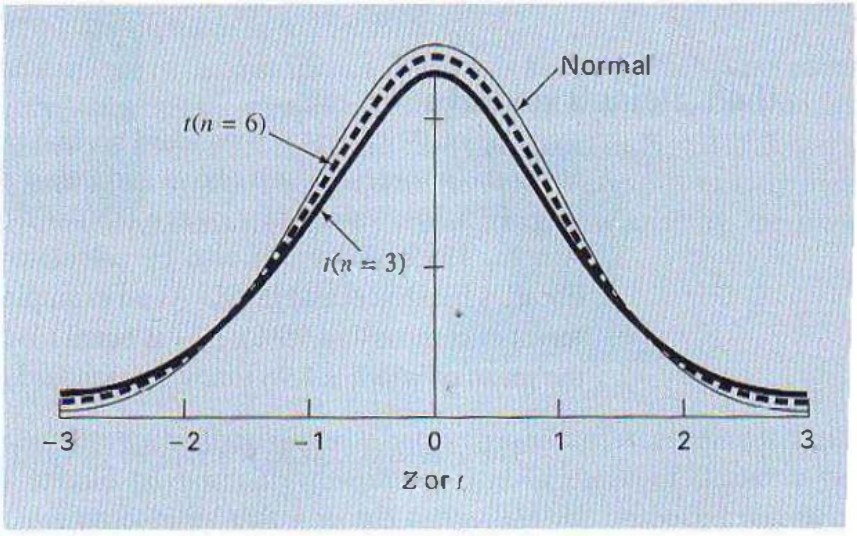
Even when we sample from a normal distribution, this fraction will not b distributed, particularly when n is small. It was found by W. S. Gossett that t variable defined by Eq. (PT5.8) follows the so-called Student-t, or simply, t d For this case,

$$L = \bar{y} - \frac{s_y}{\sqrt{n}} t_{\alpha/2,n-1} \qquad U = \bar{y} + \frac{s_y}{\sqrt{n}} t_{\alpha/2,n-1}$$

where $t_{\alpha/2,n-1}$ is the standard random variable for the t distribution for a pro $\alpha/2$. As was the case for $z_{\alpha/2}$, values are tabulated in statistics books and can a culated using software packages and libraries. For example, if $\alpha = 0.05$ a $t_{\alpha/2,n-1} = 2.086$.

The t distribution can be thought of as a modification of the normal distri accounts for the fact that we have an imperfect estimate of the standard devia n is small, it tends to be flatter than the normal (see Fig. PT5.4). Therefore

FIGURE PT5.4
Comparison of the normal distribution with the t distribution for $n = 3$ and $n = 6$. No the t distribution is generally flatter.



numbers of measurements, it yields wider and hence more conservative confidence intervals. As n grows larger, the t distribution converges on the normal.

EXAMPLE PT5.2 Confidence Interval on the Mean

Problem Statement. Determine the mean and the corresponding 95% confidence interval for the data from Table PT5.1. Perform three estimates based on (*a*) the first 8, (*b*) the first 16, and (*c*) all 24 measurements.

Solution. (*a*) The mean and standard deviation for the first 8 points is

$$\bar{y} = \frac{52.72}{8} = 6.59 \qquad s_y = \sqrt{\frac{347.4814 - (52.72)^2/8}{8-1}} = 0.089921$$

The appropriate t statistic can be calculated as

$$t_{0.05/2,8-1} = t_{0.025,7} = 2.364623$$

which can be used to compute the interval

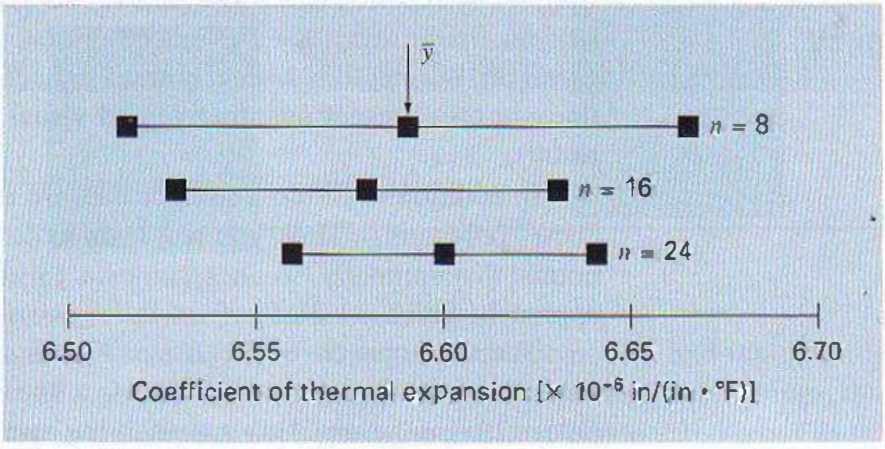
$$L = 6.59 - \frac{0.089921}{\sqrt{8}}2.364623 = 6.5148$$

$$U = 6.59 + \frac{0.089921}{\sqrt{8}}2.364623 = 6.6652$$

or

$$6.5148 \le \mu \le 6.6652$$

FIGURE PT5.5
Estimates of the mean and 95% confidence intervals for different numbers of sample size.



CHAPTER 17

Least-Squares Regression

Where substantial error is associated with data, polynomial interpolation is in and may yield unsatisfactory results when used to predict intermediate values. E tal data is often of this type. For example, Fig. 17.1*a* shows seven experimenta data points exhibiting significant variability. Visual inspection of the data sugg tive relationship between y and x. That is, the overall trend indicates that higher are associated with higher values of x. Now, if a sixth-order interpolating polyn ted to this data (Fig. 17.1*b*), it will pass exactly through all of the points. Howev of the variability in the data, the curve oscillates widely in the interval between In particular, the interpolated values at $x = 1.5$ and $x = 6.5$ appear to be well range suggested by the data.

A more appropriate strategy for such cases is to derive an approximating fu fits the shape or general trend of the data without necessarily matching the points. Figure 17.1*c* illustrates how a straight line can be used to generally char trend of the data without passing through any particular point.

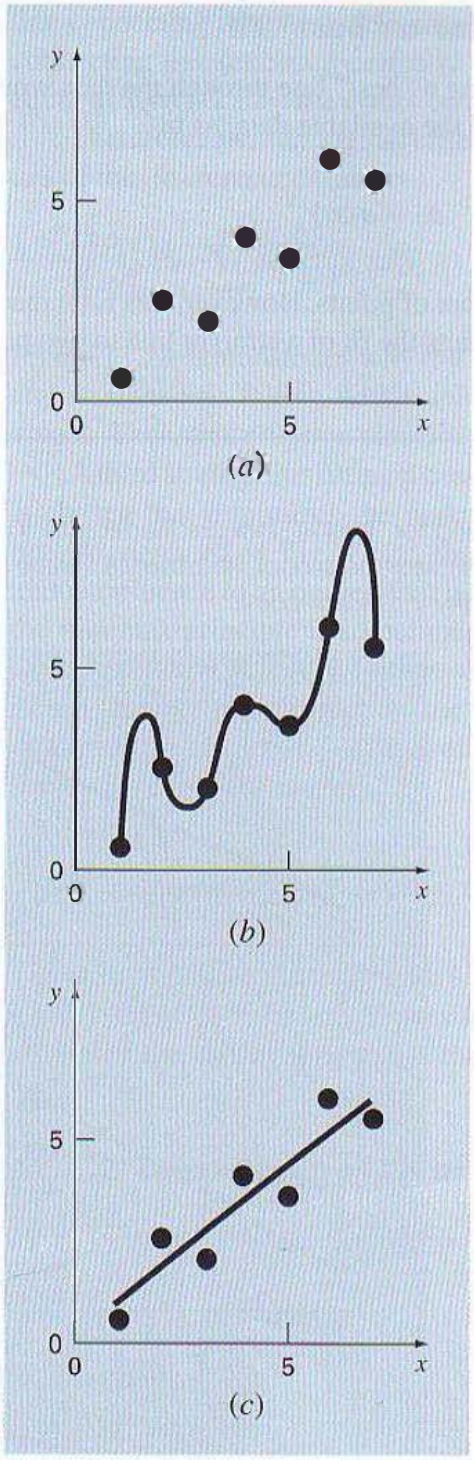
One way to determine the line in Fig. 17.1*c* is to visually inspect the plott then sketch a "best" line through the points. Although such "eyeball" appro commonsense appeal and are valid for "back-of-the-envelope" calculations, th ficient because they are arbitrary. That is, unless the points define a perfect s (in which case, interpolation would be appropriate), different analysts would d ent lines.

To remove this subjectivity, some criterion must be devised to establish a b fit. One way to do this is to derive a curve that minimizes the discrepancy betwe points and the curve. A technique for accomplishing this objective, called *least- gression*, will be discussed in the present chapter.

17.1 LINEAR REGRESSION

The simplest example of a least-squares approximation is fitting a straight line paired observations: (x_1, y_1), (x_2, y_2), . . . , (x_n, y_n). The mathematical expressi straight line is

$$y = a_0 + a_1 x + e$$



significant
al fit
the range of
e satisfactory
east-squares fit.

where a_0 and a_1 are coefficients representing the intercept and the slope, respectively, and e is the error, or residual, between the model and the observations, which can be represented by rearranging Eq. (17.1) as

$$e = y - a_0 - a_1 x$$

Thus, the error, or *residual,* is the discrepancy between the true value of y and the approximate value, $a_0 + a_1 x$, predicted by the linear equation.

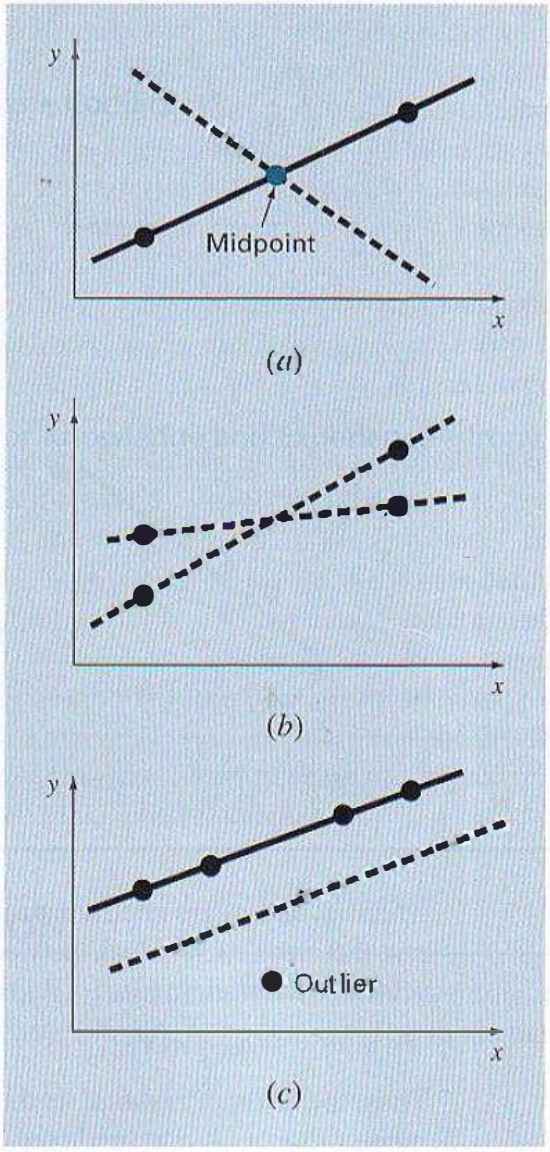
17.1.1 Criteria for a "Best" Fit

One strategy for fitting a "best" line through the data would be to minimize th residual errors for all the available data, as in

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)$$

where n = total number of points. However, this is an inadequate criterion, as Fig. 17.2a which depicts the fit of a straight line to two points. Obviously, the

FIGURE 17.2
Examples of some criteria for "best fit" that are inadequate for regression: (*a*) minim of the residuals, (*b*) minimizes the sum of the absolute values of the residuals, and (c the maximum error of any individual point.



line connecting the points. However, any straight line passing through the midpoint of the connecting line (except a perfectly vertical line) results in a minimum value of Eq. (17.2) equal to zero because the errors cancel.

Therefore, another logical criterion might be to minimize the sum of the absolute values of the discrepancies, as in

$$\sum_{i=1}^{n} |e_i| = \sum_{i=1}^{n} |y_i - a_0 - a_1 x_i|$$

Figure 17.2*b* demonstrates why this criterion is also inadequate. For the four points shown, any straight line falling within the dashed lines will minimize the sum of the absolute values. Thus, this criterion also does not yield a unique best fit.

A third strategy for fitting a best line is the *minimax* criterion. In this technique, the line is chosen that minimizes the maximum distance that an individual point falls from the line. As depicted in Fig. 17.2*c*, this strategy is ill-suited for regression because it gives undue influence to an outlier, that is, a single point with a large error. It should be noted that the minimax principle is sometimes well-suited for fitting a simple function to a complicated function (Carnahan, Luther, and Wilkes, 1969).

A strategy that overcomes the shortcomings of the aforementioned approaches is to minimize the sum of the squares of the residuals between the measured y and the y calculated with the linear model

$$S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_{i,\text{measured}} - y_{i,\text{model}})^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2 \tag{17.3}$$

This criterion has a number of advantages, including the fact that it yields a unique line for a given set of data. Before discussing these properties, we will present a technique for determining the values of a_0 and a_1 that minimize Eq. (17.3).

17.1.2 Least-Squares Fit of a Straight Line

To determine values for a_0 and a_1, Eq. (17.3) is differentiated with respect to each coefficient:

$$\frac{\partial S_r}{\partial a_0} = -2 \sum (y_i - a_0 - a_1 x_i)$$

$$\frac{\partial S_r}{\partial a_1} = -2 \sum [(y_i - a_0 - a_1 x_i) x_i]$$

Note that we have simplified the summation symbols; unless otherwise indicated, all summations are from $i = 1$ to n. Setting these derivatives equal to zero will result in a minimum S_r. If this is done, the equations can be expressed as

$$0 = \sum y_i - \sum a_0 - \sum a_1 x_i$$

$$0 = \sum y_i x_i - \sum a_0 x_i - \sum a_1 x_i^2$$

Now, realizing that $\Sigma a_0 = na_0$, we can express the equations as a set of two simu linear equations with two unknowns (a_0 and a_1):

$$na_0 + \left(\sum x_i\right) a_1 = \sum y_i$$

$$\left(\sum x_i\right) a_0 + \left(\sum x_i^2\right) a_1 = \sum x_i y_i$$

These are called the *normal equations*. They can be solved simultaneously

$$a_1 = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{n\Sigma x_i^2 - (\Sigma x_i)^2}$$

This result can then be used in conjunction with Eq. (17.4) to solve for

$$a_0 = \bar{y} - a_1 \bar{x}$$

where $\bar{y}$ and $\bar{x}$ are the means of y and x, respectively.

EXAMPLE 17.1 Linear Regression

Problem Statement. Fit a straight line to the x and y values in the first two col Table 17.1.

Solution. The following quantities can be computed:

$$n = 7 \qquad \sum x_i y_i = 119.5 \qquad \sum x_i^2 = 140$$

$$\sum x_i = 28 \qquad \bar{x} = \frac{28}{7} = 4$$

$$\sum y_i = 24 \qquad \bar{y} = \frac{24}{7} = 3.428571$$

Using Eqs. (17.6) and (17.7),

$$a_1 = \frac{7(119.5) - 28(24)}{7(140) - (28)^2} = 0.8392857$$

$$a_0 = 3.428571 - 0.8392857(4) = 0.07142857$$

TABLE 17.1 Computations for an error analysis of the linear fit.

x_i	y_i	$(y_i - \bar{y})^2$	$(y_i - a_0 - a_1 x_i)^2$
1	0.5	8.5765	0.1687
2	2.5	0.8622	0.5625
3	2.0	2.0408	0.3473
4	4.0	0.3265	0.3265
5	3.5	0.0051	0.5896
6	6.0	6.6122	0.7972
7	5.5	4.2908	0.1993
Σ	24.0	22.7143	2.9911

Therefore, the least-squares fit is

$$y = 0.07142857 + 0.8392857x$$

The line, along with the data, is shown in Fig. 17.1c.

17.1.3 Quantification of Error of Linear Regression

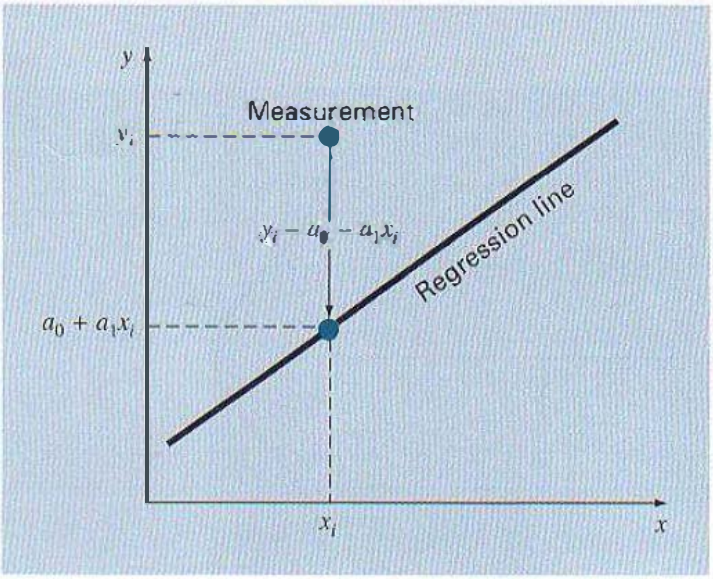
Any line other than the one computed in Example 17.1 results in a larger sum of the squares of the residuals. Thus, the line is unique and in terms of our chosen criterion is a "best" line through the points. A number of additional properties of this fit can be elucidated by examining more closely the way in which residuals were computed. Recall that the sum of the squares is defined as [Eq. (17.3)]

$$S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2 \tag{17.8}$$

Notice the similarity between Eqs. (PT5.3) and (17.8). In the former case, the square of the residual represented the square of the discrepancy between the data and a single estimate of the measure of central tendency—the mean. In Eq. (17.8), the square of the residual represents the square of the vertical distance between the data and another measure of central tendency—the straight line (Fig. 17.3).

The analogy can be extended further for cases where (1) the spread of the points around the line is of similar magnitude along the entire range of the data and (2) the distribution of these points about the line is normal. It can be demonstrated that if these criteria are met, least-squares regression will provide the best (that is, the most likely) estimates of a_0 and a_1 (Draper and Smith, 1981). This is called the *maximum likelihood principle* in

FIGURE 17.3
The residual in linear regression represents the vertical distance between a data point and the straight line.



statistics. In addition, if these criteria are met, a "standard deviation" for the re can be determined as [compare with Eq. (PT5.2)]

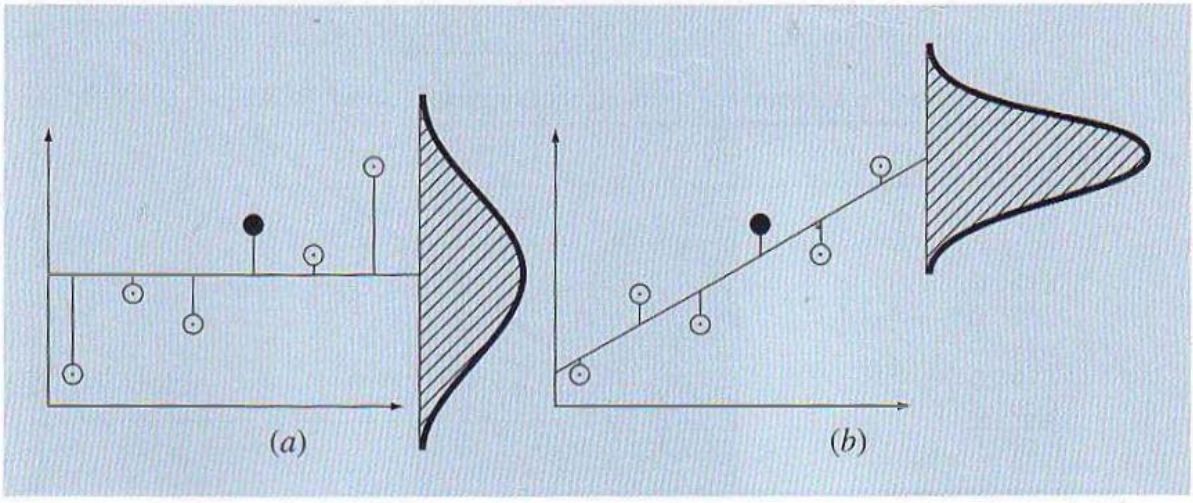
$$s_{y/x} = \sqrt{\frac{S_r}{n-2}}$$

where $s_{y/x}$ is called the *standard error of the estimate*. The subscript notation nates that the error is for a predicted value of y corresponding to a particula Also, notice that we now divide by $n - 2$ because two data-derived estimates— were used to compute S_r; thus, we have lost two degrees of freedom. As with sion of the standard deviation in PT5.2.1, another justification for dividing by there is no such thing as the "spread of data" around a straight line connecting Thus, for the case where $n = 2$, Eq. (17.9) yields a meaningless result of infin

Just as was the case with the standard deviation, the standard error of quantifies the spread of the data. However, $s_{y/x}$ quantifies the spread *around th line* as shown in Fig. 17.4*b* in contrast to the original standard deviation s_y tha the spread *around the mean* (Fig. 17.4*a*).

The above concepts can be used to quantify the "goodness" of our fit. Th larly useful for comparison of several regressions (Fig. 17.5). To do this, we original data and determine the *total sum of the squares* around the mean for th variable (in our case, y). As was the case for Eq. (PT5.3), this quantity is design is the magnitude of the residual error associated with the dependent variable gression. After performing the regression, we can compute S_r, the sum of the s residuals around the regression line. This characterizes the residual error that r the regression. It is, therefore, sometimes called the unexplained sum of the

FIGURE 17.4
Regression data showing (*a*) the spread of the data around the mean of the dependent variable and (*b*) the spread of the data around the best-fit line. The reduction in the spread in going from (*a*) to (*b*), as indicated by the bell-shaped curves at the right, represents the improvement due to linear regression.



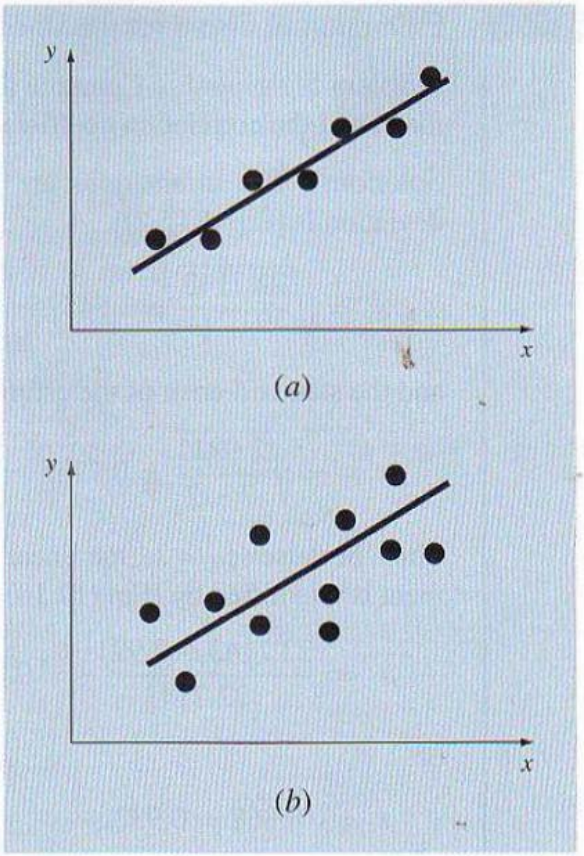


FIGURE 17.5
Examples of linear regression with (a) small and (b) large residual errors.

difference between the two quantities, $S_t - S_r$, quantifies the improvement or error reduction due to describing the data in terms of a straight line rather than as an average value. Because the magnitude of this quantity is scale-dependent, the difference is normalized to S_t to yield

$$r^2 = \frac{S_t - S_r}{S_t} \tag{17.10}$$

where r^2 is called the *coefficient of determination* and r is the *correlation coefficient* $(= \sqrt{r^2})$. For a perfect fit, $S_r = 0$ and $r = r^2 = 1$, signifying that the line explains 100 percent of the variability of the data. For $r = r^2 = 0$, $S_r = S_t$ and the fit represents no improvement. An alternative formulation for r that is more convenient for computer implementation is

$$r = \frac{n\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)}{\sqrt{n\Sigma x_i^2 - (\Sigma x_i)^2}\sqrt{n\Sigma y_i^2 - (\Sigma y_i)^2}} \tag{17.11}$$

EXAMPLE 17.2 Estimation of Errors for the Linear Least-Squares Fit

Problem Statement. Compute the total standard deviation, the standard er mate, and the correlation coefficient for the data in Example 17.1.

Solution. The summations are performed and presented in Table 17.1. deviation is [Eq. (PT5.2)]

$$s_y = \sqrt{\frac{22.7143}{7-1}} = 1.9457$$

and the standard error of the estimate is [Eq. (17.9)]

$$s_{y/x} = \sqrt{\frac{2.9911}{7-2}} = 0.7735$$

Thus, because $s_{y/x} < s_y$, the linear regression model has merit. The extent ment is quantified by [Eq. (17.10)]

$$r^2 = \frac{22.7143 - 2.9911}{22.7143} = 0.868$$

or

$$r = \sqrt{0.868} = 0.932.$$

These results indicate that 86.8 percent of the original uncertainty has beer the linear model.

Before proceeding to the computer program for linear regression, a wor in order. Although the correlation coefficient provides a handy measure of g you should be careful not to ascribe more meaning to it than is warranted. Ju "close" to 1 does not mean that the fit is necessarily "good." For example, i obtain a relatively high value of r when the underlying relationship betwee even linear. Draper and Smith (1981) provide guidance and additional mat assessment of results for linear regression. In addition, at the minimum, you inspect a plot of the data along with your regression curve. As described in th software packages include such a capability.

17.1.4 Computer Program for Linear Regression

It is a relatively trivial matter to develop a pseudocode for linear regression mentioned above, a plotting option is critical to the effective use and interp gression. Such capabilities are included in popular packages like MATLAB Excel. If your computer language has plotting capabilities, we recommend th your program to include a plot of y versus x, showing both the data and the r The inclusion of the capability will greatly enhance the utility of the progra solving contexts.

```
SUB Regress(x, y, n, a1, a0, syx, r2)

  sumx = 0: sumxy = 0: st = 0
  sumy = 0: sumx2 = 0: sr = 0
  DOFOR i = 1, n
    sumx = sumx + xi
    sumy = sumy + yi
    sumxy = sumxy + xi*yi
    sumx2 = sumx2 + xi*xi
  END DO
  xm = sumx/n
  ym = sumy/n
  a1 = (n*sumxy - sumx*sumy)/(n*sumx2 - sumx*sumx)
  a0 = ym - a1*xm
  DOFOR i = 1, n
    st = st + (yi - ym)^2
    sr = sr + (yi - a1*xi - a0)^2
  END DO
  syx = (sr/(n - 2))^0.5
  r2 = (st - sr)/st

END Regress
```

FIGURE 17.6
Algorithm for linear regression.

EXAMPLE 17.3 Linear Regression Using the Computer

Problem Statement. We can use software based on Fig. 17.6 to solve a hypothesis-testing problem associated with the falling parachutist discussed in Chap. 1. A theoretical mathematical model for the velocity of the parachutist was given as the following [Eq. (1.10)]:

$$v(t) = \frac{gm}{c}\left(1 - e^{(-c/m)t}\right)$$

where v = velocity (m/s), g = gravitational constant (9.8 m/s^2), m = mass of the parachutist equal to 68.1 kg, and c = drag coefficient of 12.5 kg/s. The model predicts the velocity of the parachutist as a function of time, as described in Example 1.1.

An alternative empirical model for the velocity of the parachutist is given by

$$v(t) = \frac{gm}{c}\left(\frac{t}{3.75 + t}\right) \tag{E17.3.1}$$

Suppose that you would like to test and compare the adequacy of these two mathematical models. This might be accomplished by measuring the actual velocity of the parachutist

TABLE 17.2 Measured and calculated velocities for the falling parachutist.

Time, s	Measured v, m/s (a)	Model-calculated v, m/s [Eq. (1.10)] (b)	Model-calcula m/s [Eq. (E17 (c)
1	10.00	8.953	11.240
2	16.30	16.405	18.570
3	23.00	22.607	23.729
4	27.50	27.769	27.556
5	31.00	32.065	30.509
6	35.60	35.641	32.855
7	39.00	38.617	34.766
8	41.50	41.095	36.351
9	42.90	43.156	37.687
10	45.00	44.872	38.829
11	46.00	46.301	39.816
12	45.50	47.490	40.678
13	46.00	48.479	41.437
14	49.00	49.303	42.110
15	50.00	49.988	42.712

at known values of time and comparing these results with the predicted velocities acc to each model.

Such an experimental-data-collection program was implemented, and the resu listed in column (*a*) of Table 17.2. Computed velocities for each model are lis columns (*b*) and (*c*).

Solution. The adequacy of the models can be tested by plotting the model-calcula locity versus the measured velocity. Linear regression can be used to calculate the and the intercept of the plot. This line will have a slope of 1, an intercept of 0, and an if the model matches the data perfectly. A significant deviation from these values used as an indication of the inadequacy of the model.

Figure 17.7*a* and *b* are plots of the line and data for the regressions of columns ((*c*), respectively, versus column (*a*). For the first model [Eq. (1.10) as depic Fig. 17.7*a*],

$$v_{\text{model}} = -0.859 + 1.032 v_{\text{measure}}$$

and for the second model [Eq. (E17.3.1) as depicted in Fig. 17.7*b*],

$$v_{\text{model}} = 5.776 + 0.752 v_{\text{measure}}$$

These plots indicate that the linear regression between the data and each of the is highly significant. Both models match the data with a correlation coefficient of than 0.99.

However, the model described by Eq. (1.10) conforms to our hypothesis test much better than that described by Eq. (E17.3.1) because the slope and intercept ar nearly equal to 1 and 0. Thus, although each plot is well described by a straigh Eq. (1.10) appears to be a better model than Eq. (E17.3.1).

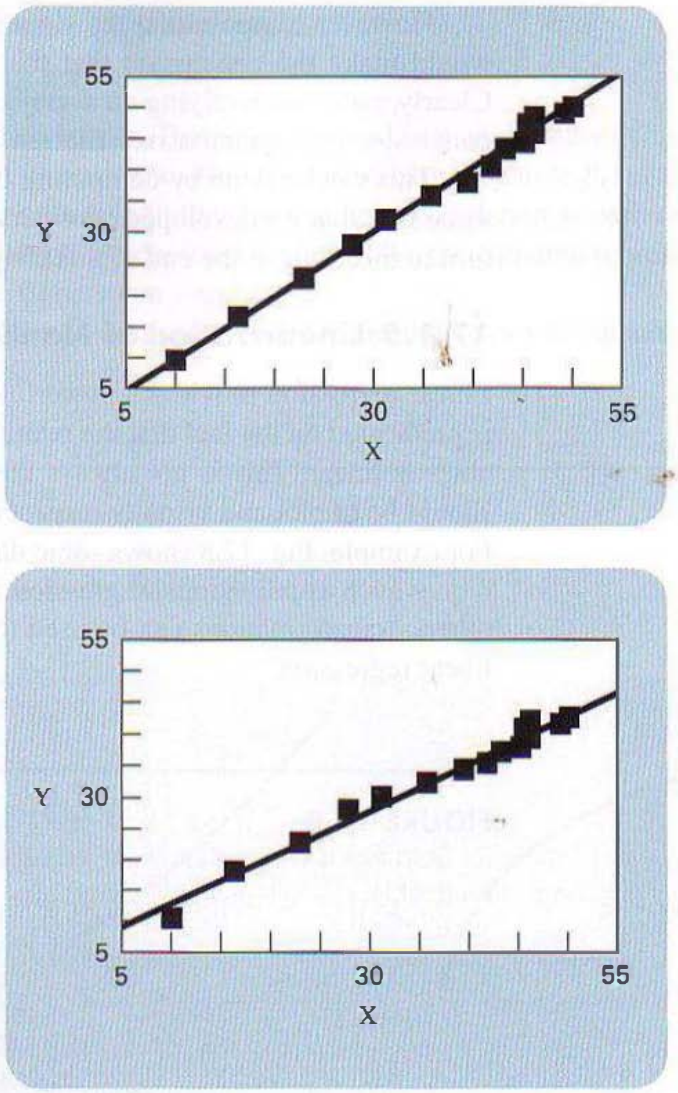


FIGURE 17.7
(a) Results using linear regression to compare predictions computed with the theoretical model [Eq. (1.10)] versus measured values. (b) Results using linear regression to compare predictions computed with the empirical model [Eq. (E17.3.1)] versus measured values.

Model testing and selection are common and extremely important activities performed in all fields of engineering. The background material provided in this chapter, together with your software, should allow you to address many practical problems of this type.

There is one shortcoming with the analysis in Example 17.3. The example was unambiguous because the empirical model [Eq. (E17.3.1)] was clearly inferior to Eq. (1.10). Thus, the slope and intercept for the former were so much closer to the desired result of 1 and 0, that it was obvious which model was superior.

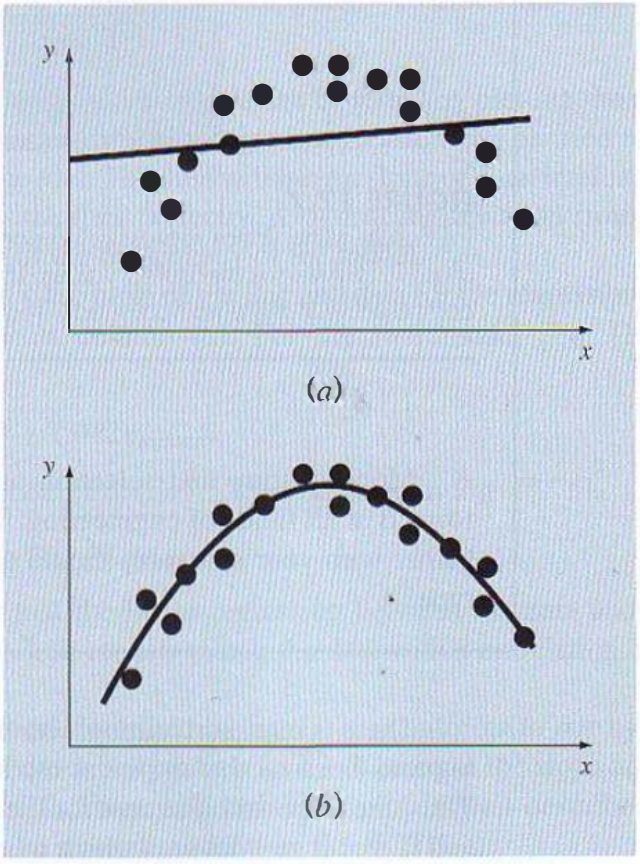
However, suppose that the slope were 0.85 and the intercept were 2. Obvio would make the conclusion that the slope and intercept were 1 and 0 open to Clearly, rather than relying on a subjective judgment, it would be preferable to ba conclusion on a quantitative criterion.

This can be done by computing confidence intervals for the model paramete same way that we developed confidence intervals for the mean in Sec. PT5.2.3. W turn to this topic at the end of this chapter.

17.1.5 Linearization of Nonlinear Relationships

Linear regression provides a powerful technique for fitting a best line to data. Ho is predicated on the fact that the relationship between the dependent and indepen ables is linear. This is not always the case, and the first step in any regression should be to plot and visually inspect the data to ascertain whether a linear mode For example, Fig. 17.8 shows some data that is obviously curvilinear. In some ca niques such as polynomial regression, which is described in Sec. 17.2, are approp others, transformations can be used to express the data in a form that is compat linear regression.

FIGURE 17.8
(a) Data that is ill-suited for linear least-squares regression. (b) Indication that a parabola preferable.



One example is the *exponential model*

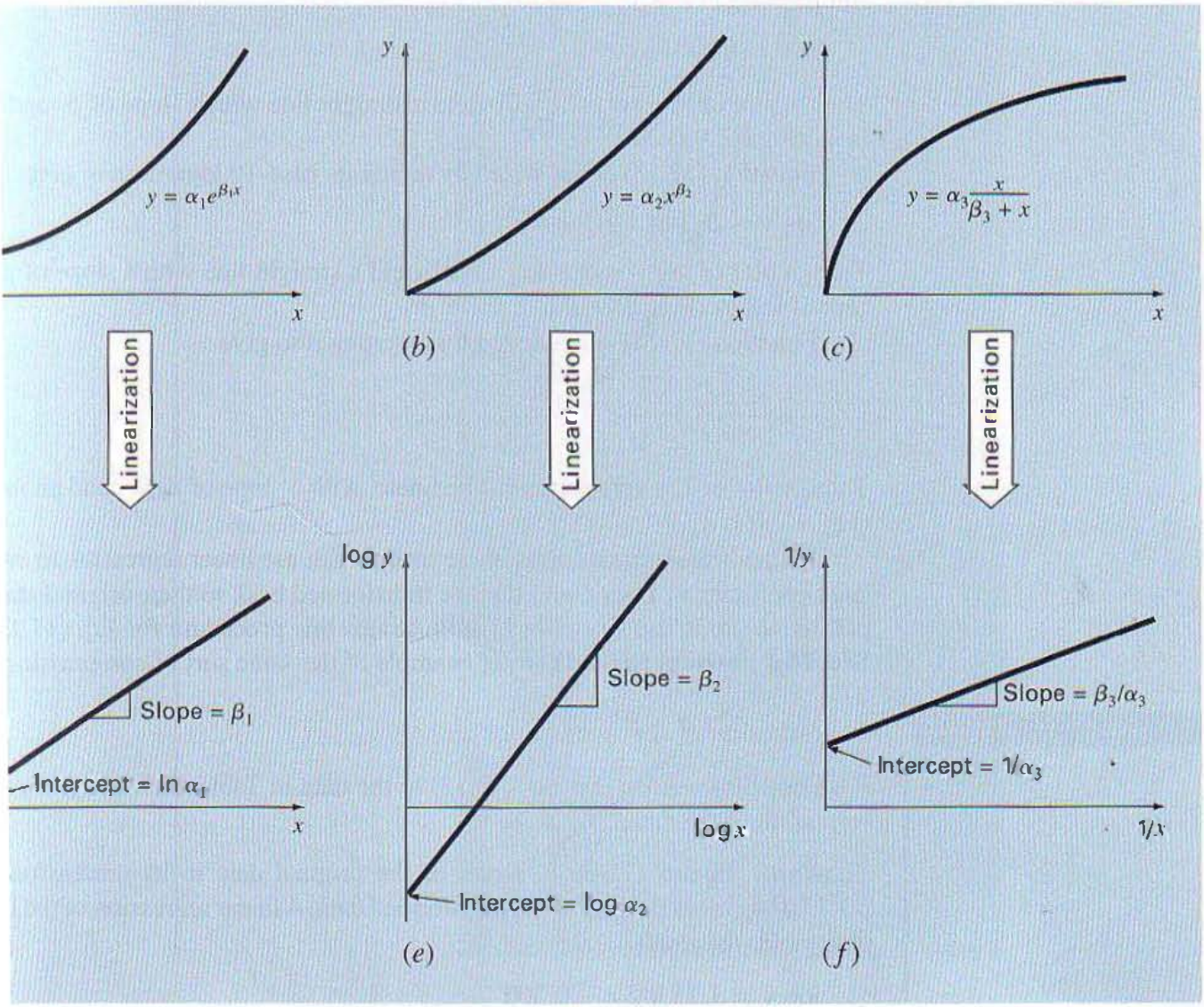
$$y = \alpha_1 e^{\beta_1 x} \tag{17.12}$$

where α_1 and β_1 are constants. This model is used in many fields of engineering to characterize quantities that increase (positive β_1) or decrease (negative β_1) at a rate that is directly proportional to their own magnitude. For example, population growth or radioactive decay can exhibit such behavior. As depicted in Fig. 17.9*a*, the equation represents a nonlinear relationship (for $\beta_1 \neq 0$) between y and x.

Another example of a nonlinear model is the simple power equation

$$y = \alpha_2 x^{\beta_2} \tag{17.13}$$

...l equation, (*b*) the power equation, and (*c*) the saturation-growth-rate ...c), (*e*), and (*f*) are linearized versions of these equations that result ...formations.



where α_2 and β_2 are constant coefficients. This model has wide applicability in a engineering. As depicted in Fig. 17.9*b*, the equation (for $\beta_2 \neq 0$ or 1) is nonline

A third example of a nonlinear model is the saturation-growth-rate equat Eq. (E17.3.1)]

$$y = \alpha_3 \frac{x}{\beta_3 + x}$$

where α_3 and β_3 are constant coefficients. This model, which is particularly well characterizing population growth rate under limiting conditions, also represents ear relationship between y and x (Fig. 17.9*c*) that levels off, or "saturates," as x

Nonlinear regression techniques are available to fit these equations to exp data directly. (Note that we will discuss nonlinear regression in Sec. 17.5.) Howe pler alternative is to use mathematical manipulations to transform the equations ear form. Then, simple linear regression can be employed to fit the equations to

For example, Eq. (17.12) can be linearized by taking its natural logarithm t

$$\ln y = \ln \alpha_1 + \beta_1 x \ln e$$

But because $\ln e = 1$,

$$\ln y = \ln \alpha_1 + \beta_1 x$$

Thus, a plot of $\ln y$ versus x will yield a straight line with a slope of β_1 and an i $\ln \alpha_1$ (Fig. 17.9*d*).

Equation (17.13) is linearized by taking its base-10 logarithm to give

$$\log y = \beta_2 \log x + \log \alpha_2$$

Thus, a plot of $\log y$ versus $\log x$ will yield a straight line with a slope of β_2 an cept of $\log \alpha_2$ (Fig. 17.9*e*).

Equation (17.14) is linearized by inverting it to give

$$\frac{1}{y} = \frac{\beta_3}{\alpha_3} \frac{1}{x} + \frac{1}{\alpha_3}$$

Thus, a plot of $1/y$ versus $1/x$ will be linear, with a slope of β_3/α_3 and an interc (Fig. 17.9*f*).

In their transformed forms, these models can use linear regression to evalu stant coefficients. They could then be transformed back to their original state a predictive purposes. Example 17.4 illustrates this procedure for Eq. (17.13). Sec. 20.1 provides an engineering example of the same sort of computation.

EXAMPLE 17.4 Linearization of a Power Equation

Problem Statement. Fit Eq. (17.13) to the data in Table 17.3 using a logarit formation of the data.

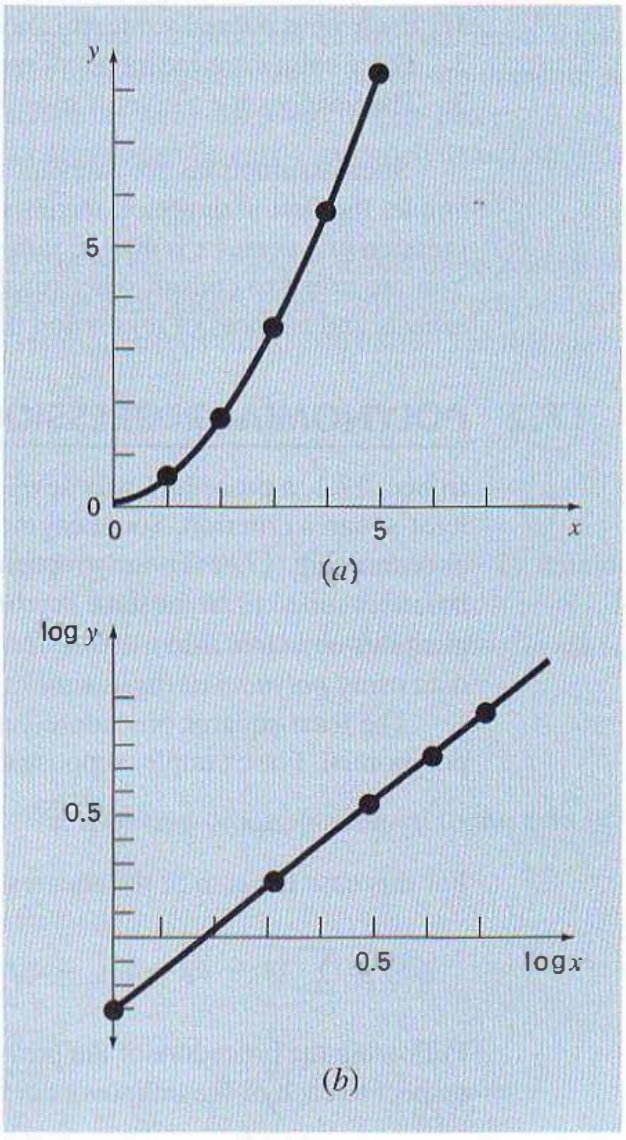
Solution. Figure 17.10*a* is a plot of the original data in its untransformed ure 17.10*b* shows the plot of the transformed data. A linear regression of the log- data yields the result

$$\log y = 1.75 \log x - 0.300$$

TABLE 17.3 Data to be fit to the power equation.

x	y	log x	log y
1	0.5	0	−0.301
2	1.7	0.301	0.226
3	3.4	0.477	0.534
4	5.7	0.602	0.753
5	8.4	0.699	0.922

FIGURE 17.10
(a) Plot of untransformed data with the power equation that fits the data. (b) Plot of transformed data used to determine the coefficients of the power equation.



Thus, the intercept, $\log \alpha_2$, equals -0.300, and therefore, by taking the antilogarithm $10^{-0.3} = 0.5$. The slope is $\beta_2 = 1.75$. Consequently, the power equation is

$$y = 0.5x^{1.75}$$

This curve, as plotted in Fig. 17.10*a*, indicates a good fit.

17.1.6 General Comments on Linear Regression

Before proceeding to curvilinear and multiple linear regression, we must emphasi introductory nature of the foregoing material on linear regression. We have focused simple derivation and practical use of equations to fit data. You should be cogniz the fact that there are theoretical aspects of regression that are of practical importan are beyond the scope of this book. For example, some statistical assumptions that herent in the linear least-squares procedures are

1. Each x has a fixed value; it is not random and is known without error.
2. The y values are independent random variables and all have the same variance.
3. The y values for a given x must be normally distributed.

Such assumptions are relevant to the proper derivation and use of regression. F ample, the first assumption means that (1) the x values must be error-free and (2) gression of y versus x is not the same as x versus y (try Prob. 17.4 at the end of the ch You are urged to consult other references such as Draper and Smith (1981) to app aspects and nuances of regression that are beyond the scope of this book.

17.2 POLYNOMIAL REGRESSION

In Sec. 17.1, a procedure was developed to derive the equation of a straight line usi least-squares criterion. Some engineering data, although exhibiting a marked patter as seen in Fig. 17.8, is poorly represented by a straight line. For these cases, a curve be better suited to fit the data. As discussed in the previous section, one method to a plish this objective is to use transformations. Another alternative is to fit polynomials data using *polynomial regression*.

The least-squares procedure can be readily extended to fit the data to a higher polynomial. For example, suppose that we fit a second-order polynomial or quadrat

$$y = a_0 + a_1 x + a_2 x^2 + e$$

For this case the sum of the squares of the residuals is [compare with Eq. (17.3)]

$$S_r = \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i - a_2 x_i^2\right)^2$$

Following the procedure of the previous section, we take the derivative of Eq. (17.1 respect to each of the unknown coefficients of the polynomial, as in

$$\frac{\partial S_r}{\partial a_0} = -2 \sum \left(y_i - a_0 - a_1 x_i - a_2 x_i^2\right)$$

$$\frac{\partial S_r}{\partial a_1} = -2\sum x_i\left(y_i - a_0 - a_1 x_i - a_2 x_i^2\right)$$

$$\frac{\partial S_r}{\partial a_2} = -2\sum x_i^2\left(y_i - a_0 - a_1 x_i - a_2 x_i^2\right)$$

These equations can be set equal to zero and rearranged to develop the following set of normal equations:

$$\begin{aligned}(n)a_0 + \left(\sum x_i\right)a_1 + \left(\sum x_i^2\right)a_2 &= \sum y_i\\ \left(\sum x_i\right)a_0 + \left(\sum x_i^2\right)a_1 + \left(\sum x_i^3\right)a_2 &= \sum x_i y_i\\ \left(\sum x_i^2\right)a_0 + \left(\sum x_i^3\right)a_1 + \left(\sum x_i^4\right)a_2 &= \sum x_i^2 y_i\end{aligned} \tag{17.19}$$

where all summations are from $i = 1$ through n. Note that the above three equations are linear and have three unknowns: a_0, a_1, and a_2. The coefficients of the unknowns can be calculated directly from the observed data.

For this case, we see that the problem of determining a least-squares second-order polynomial is equivalent to solving a system of three simultaneous linear equations. Techniques to solve such equations were discussed in Part Three.

The two-dimensional case can be easily extended to an mth-order polynomial as

$$y = a_0 + a_1 x + a_2 x^2 + \cdots + a_m x^m + e$$

The foregoing analysis can be easily extended to this more general case. Thus, we can recognize that determining the coefficients of an mth-order polynomial is equivalent to solving a system of $m + 1$ simultaneous linear equations. For this case, the standard error is formulated as

$$s_{y/x} = \sqrt{\frac{S_r}{n - (m+1)}} \tag{17.20}$$

This quantity is divided by $n - (m + 1)$ because $(m + 1)$ data-derived coefficients—$a_0, a_1, \ldots, a_m$—were used to compute S_r; thus, we have lost $m + 1$ degrees of freedom. In addition to the standard error, a coefficient of determination can also be computed for polynomial regression with Eq. (17.10).

EXAMPLE 17.5 Polynomial Regression

Problem Statement. Fit a second-order polynomial to the data in the first two columns of Table 17.4.

Solution. From the given data,

$$\begin{aligned} m &= 2 & \sum x_i &= 15 & \sum x_i^4 &= 979\\ n &= 6 & \sum y_i &= 152.6 & \sum x_i y_i &= 585.6\\ \bar{x} &= 2.5 & \sum x_i^2 &= 55 & \sum x_i^2 y_i &= 2488.8\\ \bar{y} &= 25.433 & \sum x_i^3 &= 225 \end{aligned}$$

TABLE 17.4 Computations for an error analysis of the quadratic least-squares fit.

x_i	y_i	$(y_i - \bar{y})^2$	$(y_i - a_0 - a_1 x_i - a_2 x_i^2)^2$
0	2.1	544.44	0.14332
1	7.7	314.47	1.00286
2	13.6	140.03	1.08158
3	27.2	3.12	0.80491
4	40.9	239.22	0.61951
5	61.1	1272.11	0.09439
Σ	152.6	2513.39	3.74657

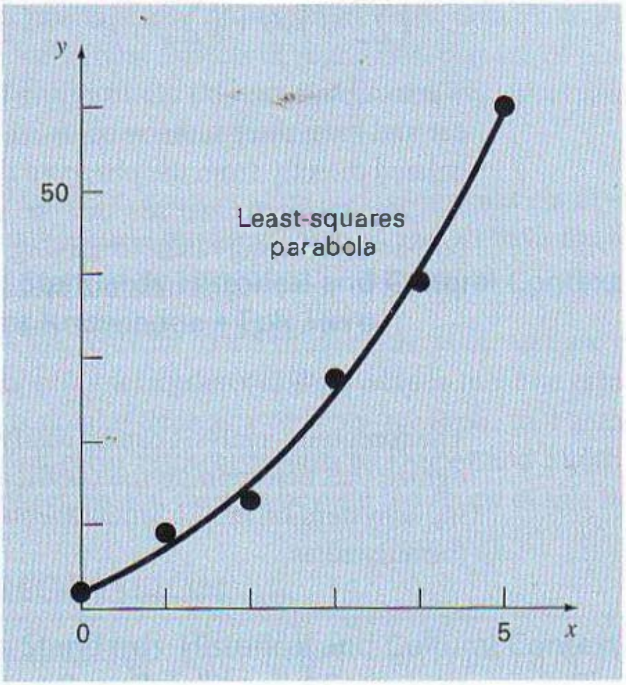


FIGURE 17.11
Fit of a second-order polynomial.

Therefore, the simultaneous linear equations are

$$\begin{bmatrix} 6 & 15 & 55 \\ 15 & 55 & 225 \\ 55 & 225 & 979 \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \\ a_2 \end{Bmatrix} = \begin{Bmatrix} 152.6 \\ 585.6 \\ 2488.8 \end{Bmatrix}$$

Solving these equations through a technique such as Gauss elimination gives a_0 $a_1 = 2.35929$, and $a_2 = 1.86071$. Therefore, the least-squares quadratic equat case is

$$y = 2.47857 + 2.35929x + 1.86071x^2$$

The standard error of the estimate based on the regression polynomial is [Eq. (

$$s_{y/x} = \sqrt{\frac{3.74657}{6-3}} = 1.12$$

The coefficient of determination is

$$r^2 = \frac{2513.39 - 3.74657}{2513.39} = 0.99851$$

and the correlation coefficient is $r = 0.99925$.

These results indicate that 99.851 percent of the original uncertainty has been explained by the model. This result supports the conclusion that the quadratic equation represents an excellent fit, as is also evident from Fig. 17.11.

17.2.1 Algorithm for Polynomial Regression

An algorithm for polynomial regression is delineated in Fig. 17.12. Note that the primary task is the generation of the coefficients of the normal equations [Eq. (17.19)]. (Pseudocode for accomplishing this is presented in Fig. 17.13.) Then, techniques from Part Three can be applied to solve these simultaneous equations for the coefficients.

A potential problem associated with implementing polynomial regression on the computer is that the normal equations are sometimes ill-conditioned. This is particularly true for higher-order versions. For these cases, the computed coefficients may be highly susceptible to round-off error, and consequently, the results can be inaccurate. Among other things, this problem is related to the structure of the normal equations and to the fact that for higher-order polynomials the normal equations can have very large and very small coefficients. This is because the coefficients are summations of the data raised to powers.

Although the strategies for mitigating round-off error discussed in Part Three, such as pivoting, can help to partially remedy this problem, a simpler alternative is to use a computer with higher precision. Fortunately, most practical problems are limited to lower-order polynomials for which round-off is usually negligible. In situations where higher-order versions are required, other alternatives are available for certain types of data. However, these techniques (such as orthogonal polynomials) are beyond the scope of this book. The reader should consult texts on regression, such as Draper and Smith (1981), for additional information regarding the problem and possible alternatives.

FIGURE 17.12
Algorithm for implementation of polynomial and multiple linear regression.

Step 1: Input order of polynomial to be fit, m.
Step 2: Input number of data points, n.
Step 3: If $n < m + 1$, print out an error message that regression is impossible and terminate the process. If $n \geq m + 1$, continue.
Step 4: Compute the elements of the normal equation in the form of an augmented matrix.
Step 5: Solve the augmented matrix for the coefficients $a_0, a_1, a_2, \ldots, a_m$, using an elimination method.
Step 6: Print out the coefficients.

```
DOFOR i = 1, order + 1
  DOFOR j = 1, i
    k = i + j - 2
    sum = 0
    DOFOR ℓ = 1, n
      sum = sum + x_ℓ^k
    END DO
    a_i,j = sum
    a_j,i = sum
  END DO
  sum = 0
  DOFOR ℓ = 1, n
    sum = sum + y_ℓ · x_ℓ^(i-1)
  END DO
  a_i,order+2 = sum
END DO
```

FIGURE 17.13
Pseudocode to assemble the elements of the normal equations for polynomial regression.

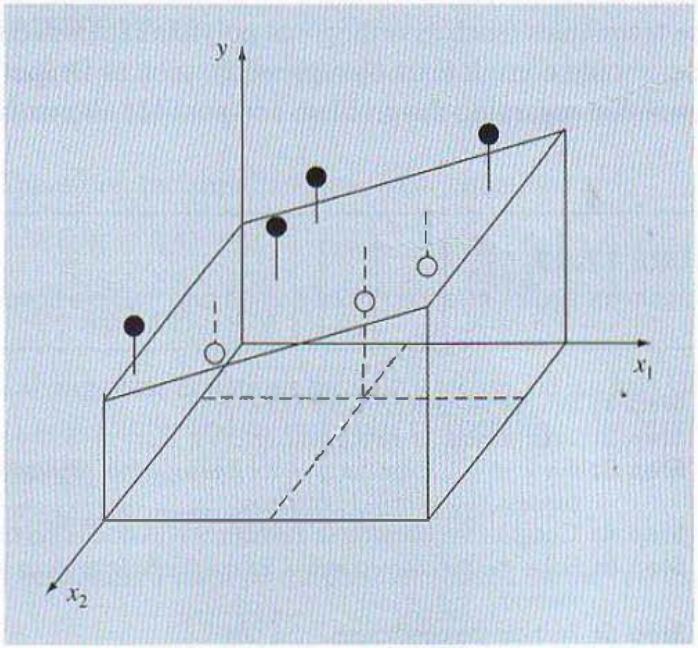
17.3 MULTIPLE LINEAR REGRESSION

A useful extension of linear regression is the case where y is a linear functio more independent variables. For example, y might be a linear function of x_1 a

$$y = a_0 + a_1 x_1 + a_2 x_2 + e$$

Such an equation is particularly useful when fitting experimental data where being studied is often a function of two other variables. For this two-dimensior regression "line" becomes a "plane" (Fig. 17.14).

FIGURE 17.14
Graphical depiction of multiple linear regression where y is a linear function of x_1 and x_2.



As with the previous cases, the "best" values of the coefficients are determined by setting up the sum of the squares of the residuals,

$$S_r = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_{1i} - a_2 x_{2i})^2 \tag{17.21}$$

and differentiating with respect to each of the unknown coefficients,

$$\frac{\partial S_r}{\partial a_0} = -2 \sum (y_i - a_0 - a_1 x_{1i} - a_2 x_{2i})$$

$$\frac{\partial S_r}{\partial a_1} = -2 \sum x_{1i} (y_i - a_0 - a_1 x_{1i} - a_2 x_{2i})$$

$$\frac{\partial S_r}{\partial a_2} = -2 \sum x_{2i} (y_i - a_0 - a_1 x_{1i} - a_2 x_{2i})$$

The coefficients yielding the minimum sum of the squares of the residuals are obtained by setting the partial derivatives equal to zero and expressing the result in matrix form as

$$\begin{bmatrix} n & \Sigma x_{1i} & \Sigma x_{2i} \\ \Sigma x_{1i} & \Sigma x_{1i}^2 & \Sigma x_{1i} x_{2i} \\ \Sigma x_{2i} & \Sigma x_{1i} x_{2i} & \Sigma x_{2i}^2 \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \\ a_2 \end{Bmatrix} = \begin{Bmatrix} \Sigma y_i \\ \Sigma x_{1i} y_i \\ \Sigma x_{2i} y_i \end{Bmatrix} \tag{17.22}$$

EXAMPLE 17.6 Multiple Linear Regression

Problem Statement. The following data was calculated from the equation $y = 5 + 4x_1 - 3x_2$:

x_1	x_2	y
0	0	5
2	1	10
2.5	2	9
1	3	0
4	6	3
7	2	27

Use multiple linear regression to fit this data.

Solution. The summations required to develop Eq. (17.22) are computed in Table 17.5. The result is

$$\begin{bmatrix} 6 & 16.5 & 14 \\ 16.5 & 76.25 & 48 \\ 14 & 48 & 54 \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \\ a_2 \end{Bmatrix} = \begin{Bmatrix} 54 \\ 243.5 \\ 100 \end{Bmatrix}$$

which can be solved using a method such as Gauss elimination for

$$a_0 = 5 \qquad a_1 = 4 \qquad a_2 = -3$$

which is consistent with the original equation from which the data was derived.

TABLE 17.5 Computations required to develop the normal equations for Example 17.

	y	x_1	x_2	x_1^2	x_2^2	x_1x_2	x_1y
	5	0	0	0	0	0	0
	10	2	1	4	1	2	20
	9	2.5	2	6.25	4	5	22.5
	0	1	3	1	9	3	0
	3	4	6	16	36	24	12
	27	7	2	49	4	14	189
Σ	54	16.5	14	76.25	54	48	243.5

The foregoing two-dimensional case can be easily extended to m dimensions.

$$y = a_0 + a_1x_1 + a_2x_2 + \cdots + a_mx_m + e$$

where the standard error is formulated as

$$s_{y/x} = \sqrt{\frac{S_r}{n - (m + 1)}}$$

and the coefficient of determination is computed as in Eq. (17.10). An algorithm to solve the normal equations is listed in Fig. 17.15.

Although there may be certain cases where a variable is linearly related to two or more other variables, multiple linear regression has additional utility in the derivation of power equations of the general form

$$y = a_0x_1^{a_1}x_2^{a_2} \cdots x_m^{a_m}$$

FIGURE 17.15
Pseudocode to assemble the elements of the normal equations for multiple regression. Note that aside from storing the independent variables in $x_{1,i}$, $x_{2,i}$, etc., 1's must be stored in $x_{0,i}$ for the algorithm to work.

```
DOFOR i = 1, order + 1
  DOFOR j = 1, i
    sum = 0
    DOFOR ℓ = 1, n
      sum = sum + x(i-1,ℓ) · x(j-1,ℓ)
    END DO
    a(i,j) = sum
    a(j,i) = sum
  END DO
  sum = 0
  DOFOR ℓ = 1, n
    sum = sum + y(ℓ) · x(i-1,ℓ)
  END DO
  a(i,order+2) = sum
END DO
```

Such equations are extremely useful when fitting experimental data. To use multiple linear regression, the equation is transformed by taking its logarithm to yield

$$\log y = \log a_0 + a_1 \log x_1 + a_2 \log x_2 + \cdots + a_m \log x_m$$

This transformation is similar in spirit to the one used in Sec. 17.1.5 and Example 17.4 to fit a power equation when y was a function of a single variable x. Section 20.4 provides an example of such an application for two independent variables.

17.4 GENERAL LINEAR LEAST SQUARES

To this point, we have focused on the mechanics of obtaining least-squares fits of some simple functions to data. Before turning to nonlinear regression, there are several issues that we would like to discuss to enrich your understanding of the preceding material.

17.4.1 General Matrix Formulation for Linear Least Squares

In the preceding pages, we have introduced three types of regression: simple linear, polynomial, and multiple linear. In fact, all three belong to the following general linear least-squares model:

$$y = a_0 z_0 + a_1 z_1 + a_2 z_2 + \cdots + a_m z_m + e \tag{17.23}$$

where $z_0, z_1, \ldots, z_m$ are $m + 1$ basis functions. It can easily be seen how simple and multiple linear regression fall within this model—that is, $z_0 = 1$, $z_1 = x_1$, $z_2 = x_2$, $\ldots$, $z_m = x_m$. Further, polynomial regression is also included if the basis functions are simple monomials as in $z_0 = x^0 = 1$, $z_1 = x$, $z_2 = x^2$, $\ldots$, $z_m = x^m$.

Note that the terminology "linear" refers only to the model's dependence on its parameters—that is, the a's. As in the case of polynomial regression, the functions themselves can be highly nonlinear. For example, the z's can be sinusoids, as in

$$y = a_0 + a_1 \cos(\omega t) + a_2 \sin(\omega t)$$

Such a format is the basis of Fourier analysis described in Chap. 19.

On the other hand, a simple-looking model like

$$f(x) = a_0(1 - e^{-a_1 x})$$

is truly nonlinear because it cannot be manipulated into the format of Eq. (17.23). We will turn to such models at the end of this chapter.

For the time being, Eq. (17.23) can be expressed in matrix notation as

$$\{Y\} = [Z]\{A\} + \{E\} \tag{17.24}$$

where $[Z]$ is a matrix of the calculated values of the basis functions at the measured values of the independent variables,

$$[Z] = \begin{bmatrix} z_{01} & z_{11} & \cdots & z_{m1} \\ z_{02} & z_{12} & \cdots & z_{m2} \\ \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \\ z_{0n} & z_{1n} & \cdots & z_{mn} \end{bmatrix}$$

where m is the number of variables in the model and n is the number of data po
cause $n \geq m + 1$, you should recognize that most of the time, $[Z]$ is not a squar

The column vector $\{Y\}$ contains the observed values of the dependent variab

$$\{Y\}^T = \lfloor y_1 \quad y_2 \quad \cdots \quad y_n \rfloor$$

The column vector $\{A\}$ contains the unknown coefficients

$$\{A\}^T = \lfloor a_0 \quad a_1 \quad \cdots \quad a_m \rfloor$$

and the column vector $\{E\}$ contains the residuals

$$\{E\}^T = \lfloor e_1 \quad e_2 \quad \cdots \quad e_n \rfloor$$

As was done throughout this chapter, the sum of the squares of the residua
model can be defined as

$$S_r = \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{m} a_j z_{ji} \right)^2$$

This quantity can be minimized by taking its partial derivative with respect to ea
coefficients and setting the resulting equation equal to zero. The outcome of this p
the normal equations that can be expressed concisely in matrix form as

$$\left[[Z]^T [Z] \right] \{A\} = \left\{ [Z]^T \{Y\} \right\}$$

It can be shown that Eq. (17.25) is, in fact, equivalent to the normal equations d
previously for simple linear, polynomial, and multiple linear regression.

Our primary motivation for the foregoing has been to illustrate the unity an
three approaches and to show how they can all be expressed simply in the same m
tation. It also sets the stage for the next section where we will gain some insights
preferred strategies for solving Eq. (17.25). The matrix notation will also have r
when we turn to nonlinear regression in the last section of this chapter.

17.4.2 Solution Techniques

In previous discussions in this chapter, we have glossed over the issue of the spe
merical techniques to solve the normal equations. Now that we have established
among the various models, we can explore this question in more detail.

First, it should be clear that Gauss-Seidel cannot be employed because th
equations are not diagonally dominant. We are thus left with the elimination meth
the present purposes, we can divide these techniques into three categories: (1) *LU*
position methods including Gauss elimination, (2) Cholesky's method, and (3
inversion approaches. There are obviously overlaps involved in this breakdown
ample, Cholesky's method is, in fact, an *LU* decomposition, and all the approach
formulated so that they can generate the matrix inverse. However, this breakd
merit in that each category offers benefits regarding the solution of the normal e

LU Decomposition. If you are merely interested in applying a least-squares fit fo
where the appropriate model is known a priori, any of the *LU* decomposition ap
described in Chap. 9 is perfectly acceptable. In fact, the non-*LU*-decomposition for
of Gauss elimination can also be employed. It is a relatively straightforward prog

task to incorporate any of these into an algorithm for linear least squares. In fact, if a modular approach has been followed, it is almost trivial.

Cholesky's Method. Cholesky's decomposition algorithm has several advantages with regard to the solution of the general linear regression problem. First, it is expressly designed for solving symmetric matrices like the normal equations. Thus, it is fast and requires less storage space to solve such systems. Second, it is ideally suited for cases where the order of the model [that is, the value of m in Eq. (17.23)] is not known beforehand (see Ralston and Rabinowitz, 1978). A case in point would be polynomial regression. For this case, we might not know a priori whether a linear, quadratic, cubic, or higher-order polynomial is the "best" model to describe our data. Because of the way in which both the normal equations are constructed and the Cholesky algorithm proceeds (Fig. 11.3), we can develop successively higher-order models in an extremely efficient manner. At each step we could examine the residual sum of the squares error (and a plot!) to examine whether the inclusion of higher-order terms significantly improves the fit.

The analogous situation for multiple linear regression occurs when independent variables are added to the model one at a time. Suppose that the dependent variable of interest is a function of a number of independent variables, say, temperature, moisture content, pressure, etc. We could first perform a linear regression with temperature and compute a residual error. Next, we could include moisture content by performing a two-variable multiple regression and see whether the additional variable results in an improved fit. Cholesky's method makes this process efficient because the decomposition of the linear model would merely be supplemented to incorporate a new variable.

Matrix Inverse Approaches. From Eq. (PT3.6), recall that the matrix inverse can be employed to solve Eq. (17.25), as in

$$\{A\} = \left[[Z]^T[Z]\right]^{-1}\left\{[Z]^T\{Y\}\right\} \tag{17.26}$$

Each of the elimination methods can be used to determine the inverse and, thus, can be used to implement Eq. (17.26). However, as we have learned in Part Three, this is an inefficient approach for solving a set of simultaneous equations. Thus, if we were merely interested in solving for the regression coefficients, it is preferable to employ an LU decomposition approach without inversion. However, from a statistical perspective, there are a number of reasons why we might be interested in obtaining the inverse and examining its coefficients. These reasons will be discussed next.

17.4.3 Statistical Aspects of Least-Squares Theory

In Sec. PT5.2.1, we reviewed a number of descriptive statistics that can be used to describe a sample. These included the arithmetic mean, the standard deviation, and the variance.

Aside from yielding a solution for the regression coefficients, the matrix formulation of Eq. (17.26) provides estimates of their statistics. It can be shown (Draper and Smith, 1981) that the diagonal and off-diagonal terms of the matrix $[[Z]^T[Z]]^{-1}$ give, respectively, the variances and the covariances[1] of the a's. If the diagonal elements of

[1]The covariance is a statistic that measures the dependency of one variable on another. Thus, cov(x, y) indicates the dependency of x and y. For example, cov(x, y) $= 0$ would indicate that x and y are totally independent.

$[[Z]^T[Z]]^{-1}$ are designated as $z_{i,i}^{-1}$,

$$\text{var}(a_{i-1}) = z_{i,i}^{-1} s_{y/x}^2$$

and

$$\text{cov}(a_{i-1}, a_{j-1}) = z_{i,j}^{-1} s_{y/x}^2$$

These statistics have a number of important applications. For our present purp will illustrate how they can be used to develop confidence intervals for the inte slope.

Using an approach similar to that in Sec. PT5.2.3, it can be shown that lower a bounds on the intercept can be formulated as (see Milton and Arnold, 1995, fo

$$L = a_0 - t_{\alpha/2,n-2}\, s(a_0) \qquad U = a_0 + t_{\alpha/2,n-2}\, s(a_0)$$

where $s(a_j)$ = the standard error of coefficient $a_j = \sqrt{\text{var}(a_j)}$. In a similar manne and upper bounds on the slope can be formulated as

$$L = a_1 - t_{\alpha/2,n-2}\, s(a_1) \qquad U = a_1 + t_{\alpha/2,n-2}\, s(a_1)$$

The following example illustrates how these intervals can be used to make quanti ferences related to linear regression.

EXAMPLE 17.7 Confidence Intervals for Linear Regression

Problem Statement. In Example 17.3, we used regression to develop the follow tionship between measurements and model predictions:

$$y = -0.859 + 1.032x$$

where y = the model predictions and x = the measurements. We concluded that th good agreement between the two because the intercept was approximately equal to slope approximately equal to 1. Recompute the regression but use the matrix approx timate standard errors for the parameters. Then employ these errors to develop co intervals, and use these to make a probabilistic statement regarding the goodness

Solution. The data can be written in matrix format for simple linear regression

$$[Z] = \begin{bmatrix} 1 & 10 \\ 1 & 16.3 \\ 1 & 23 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & 50 \end{bmatrix} \qquad \{Y\} = \begin{Bmatrix} 8.953 \\ 16.405 \\ 22.607 \\ \cdot \\ \cdot \\ \cdot \\ 49.988 \end{Bmatrix}$$

Matrix transposition and multiplication can then be used to generate the normal eq

$$[[Z]^T[Z]] \qquad \{A\} = \{[Z]^T\{Y\}\}$$

$$\begin{bmatrix} 15 & 548.3 \\ 548.3 & 22191.21 \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \end{Bmatrix} = \begin{Bmatrix} 552.741 \\ 22421.43 \end{Bmatrix}$$

Matrix inversion can be used to obtain the slope and intercept as

$$\{A\} = [[Z]^T[Z]]^{-1} \{[Z]^T\{Y\}\}$$

$$= \begin{bmatrix} 0.688414 & -0.01701 \\ -0.01701 & 0.000465 \end{bmatrix} \begin{Bmatrix} 552.741 \\ 22421.43 \end{Bmatrix} = \begin{Bmatrix} -0.85872 \\ 1.031592 \end{Bmatrix}$$

Thus, the intercept and the slope are determined as $a_0 = -0.85872$ and $a_1 = 1.031592$, respectively. These values in turn can be used to compute the standard error of the estimate as $s_{y/x} = 0.863403$. This value can be used along with the diagonal elements of the matrix inverse to calculate the standard errors of the coefficients,

$$s(a_0) = \sqrt{z_{11}^{-1} s_{y/x}^2} = \sqrt{0.688414(0.863403)^2} = 0.716372$$

$$s(a_1) = \sqrt{z_{22}^{-1} s_{y/x}^2} = \sqrt{0.000465(0.863403)^2} = 0.018625$$

The statistic, $t_{\alpha/2,n-1}$ needed for a 95% confidence interval with $n - 2 = 15 - 2 = 13$ degrees of freedom can be determined from a statistics table or using software. We used an Excel function, TINV, to come up with the proper value, as in

$$= \text{TINV}(0.05, 13)$$

which yielded a value of 2.160368. Equations (17.29) and (17.30) can then be used to compute the confidence intervals as

$$a_0 = -0.85872 \pm 2.160368(0.716372)$$
$$= -0.85872 \pm 1.547627 = [-2.40634, 0.688912]$$
$$a_1 = 1.031592 \pm 2.160368(0.018625)$$
$$= 1.031592 \pm 0.040237 = [0.991355, 1.071828]$$

Notice that the desired values (0 for intercept and slope and 1 for the intercept) fall within the intervals. On the basis of this analysis we could make the following statement regarding the slope: We have strong grounds for believing that the slope of the true regression line lies within the interval from 0.991355 to 1.071828. Because 1 falls within this interval, we also have strong grounds for believing that the result supports the agreement between the measurements and the model. Because zero falls within the intercept interval, a similar statement can be made regarding the intercept.

The foregoing is a limited introduction to the rich topic of statistical inference and its relationship to regression. There are many subleties that are beyond the scope of this book. Our primary motivation has been to illustrate the power of the matrix approach to general linear least squares. You should consult some of the excellent books on the subject (for example, Draper and Smith 1981) for additional information. In addition, it should be noted that software packages and libraries can generate least-squares regression fits along with information relevant to inferential statistics. We will explore some of these capabilities when we describe these packages at the end of Chap. 19.

17.5 NONLINEAR REGRESSION

There are many cases in engineering where nonlinear models must be fit to d present context, these models are defined as those that have a nonlinear depe their parameters. For example,

$$f(x) = a_0(1 - e^{-a_1 x}) + e$$

This equation cannot be manipulated so that it conforms to the general form of E

As with linear least squares, nonlinear regression is based on determining of the parameters that minimize the sum of the squares of the residuals. Howe nonlinear case, the solution must proceed in an iterative fashion.

The *Gauss-Newton method* is one algorithm for minimizing the sum of the the residuals between data and nonlinear equations. The key concept underlyin nique is that a Taylor series expansion is used to express the original nonlinear e an approximate, linear form. Then, least-squares theory can be used to obtain mates of the parameters that move in the direction of minimizing the residual.

To illustrate how this is done, first the relationship between the nonlinear eq the data can be expressed generally as

$$y_i = f(x_i; a_0, a_1, \ldots, a_m) + e_i$$

where y_i = a measured value of the dependent variable, $f(x_i; a_0, a_1, \ldots, a_m)$ = th that is a function of the independent variable x_i and a nonlinear function of the p $a_0, a_1, \ldots, a_m$, and e_i = a random error. For convenience, this model can be ex abbreviated form by omitting the parameters,

$$y_i = f(x_i) + e_i$$

The nonlinear model can be expanded in a Taylor series around the param and curtailed after the first derivative. For example, for a two-parameter case,

$$f(x_i)_{j+1} = f(x_i)_j + \frac{\partial f(x_i)_j}{\partial a_0} \Delta a_0 + \frac{\partial f(x_i)_j}{\partial a_1} \Delta a_1$$

where j = the initial guess, $j + 1$ = the prediction, $\Delta a_0 = a_{0,j+1} - a_{0,j}$, and Δa_1 $a_{1,j}$. Thus, we have linearized the original model with respect to the parameters (17.33) can be substituted into Eq. (17.32) to yield

$$y_i - f(x_i)_j = \frac{\partial f(x_i)_j}{\partial a_0} \Delta a_0 + \frac{\partial f(x_i)_j}{\partial a_1} \Delta a_1 + e_i$$

or in matrix form [compare with Eq. (17.24)],

$$\{D\} = [Z_j]\{\Delta A\} + \{E\}$$

where $[Z_j]$ is the matrix of partial derivatives of the function evaluated at the ini

$$[Z_j] = \begin{bmatrix} \partial f_1/\partial a_0 & \partial f_1/\partial a_1 \\ \partial f_2/\partial a_0 & \partial f_2/\partial a_1 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ \partial f_n/\partial a_0 & \partial f_n/\partial a_1 \end{bmatrix}$$

where n = the number of data points and $\partial f_i/\partial a_k$ = the partial derivative of the function with respect to the kth parameter evaluated at the ith data point. The vector $\{D\}$ contains the differences between the measurements and the function values,

$$\{D\} = \begin{Bmatrix} y_1 - f(x_1) \\ y_2 - f(x_2) \\ \cdot \\ \cdot \\ \cdot \\ y_n - f(x_n) \end{Bmatrix}$$

and the vector $\{\Delta A\}$ contains the changes in the parameter values,

$$\{\Delta A\} = \begin{Bmatrix} \Delta a_0 \\ \Delta a_1 \\ \cdot \\ \cdot \\ \cdot \\ \Delta a_m \end{Bmatrix}$$

Applying linear least-squares theory to Eq. (17.34) results in the following normal equations [recall Eq. (17.25)]:

$$[[Z_j]^T[Z_j]]\{\Delta A\} = \{[Z_j]^T\{D\}\} \tag{17.35}$$

Thus, the approach consists of solving Eq. (17.35) for $\{\Delta A\}$, which can be employed to compute improved values for the parameters, as in

$$a_{0,j+1} = a_{0,j} + \Delta a_0$$

and

$$a_{1,j+1} = a_{1,j} + \Delta a_1$$

This procedure is repeated until the solution converges—that is, until

$$|\varepsilon_a|_k = \left|\frac{a_{k,j+1} - a_{k,j}}{a_{k,j+1}}\right| 100\% \tag{17.36}$$

falls below an acceptable stopping criterion.

EXAMPLE 17.9 Gauss-Newton Method

Problem Statement. Fit the function $f(x; a_0, a_1) = a_0(1 - e^{-a_1 x})$ to the data:

x	0.25	0.75	1.25	1.75	2.25
y	0.28	0.57	0.68	0.74	0.79

Use initial guesses of $a_0 = 1.0$ and $a_1 = 1.0$ for the parameters. Note that for these guesses, the initial sum of the squares of the residuals is 0.0248.

Solution. The partial derivatives of the function with respect to the parameters are

$$\frac{\partial f}{\partial a_0} = 1 - e^{-a_1 x} \tag{E17.9.1}$$

and

$$\frac{\partial f}{\partial a_1} = a_0 x e^{-a_1 x}$$

Equations (E17.9.1) and (E17.9.2) can be used to evaluate the matrix

$$[Z_0] = \begin{bmatrix} 0.2212 & 0.1947 \\ 0.5276 & 0.3543 \\ 0.7135 & 0.3581 \\ 0.8262 & 0.3041 \\ 0.8946 & 0.2371 \end{bmatrix}$$

This matrix multiplied by its transpose results in

$$[Z_0]^T[Z_0] = \begin{bmatrix} 2.3193 & 0.9489 \\ 0.9489 & 0.4404 \end{bmatrix}$$

which in turn can be inverted to yield

$$[[Z_0]^T[Z_0]]^{-1} = \begin{bmatrix} 3.6397 & -7.8421 \\ -7.8421 & 19.1678 \end{bmatrix}$$

The vector $\{D\}$ consists of the differences between the measurements and the mod dictions,

$$\{D\} = \begin{Bmatrix} 0.28 - 0.2212 \\ 0.57 - 0.5276 \\ 0.68 - 0.7135 \\ 0.74 - 0.8262 \\ 0.79 - 0.8946 \end{Bmatrix} = \begin{Bmatrix} 0.0588 \\ 0.0424 \\ -0.0335 \\ -0.0862 \\ -0.1046 \end{Bmatrix}$$

It is multiplied by $[Z_0]^T$ to give

$$[Z_0]^T\{D\} = \begin{bmatrix} -0.1533 \\ -0.0365 \end{bmatrix}$$

The vector $\{\Delta A\}$ is then calculated by solving Eq. (17.35) for

$$\Delta A = \begin{Bmatrix} -0.2714 \\ 0.5019 \end{Bmatrix}$$

which can be added to the initial parameter guesses to yield

$$\begin{Bmatrix} a_0 \\ a_1 \end{Bmatrix} = \begin{Bmatrix} 1.0 \\ 1.0 \end{Bmatrix} + \begin{Bmatrix} -0.2714 \\ 0.5019 \end{Bmatrix} = \begin{Bmatrix} 0.7286 \\ 1.5019 \end{Bmatrix}$$

Thus, the improved estimates of the parameters are $a_0 = 0.7286$ and $a_1 = 1.5019$. T parameters result in a sum of the squares of the residuals equal to 0.0242. Equation can be used to compute ε_0 and ε_1 equal to 37 and 33 percent, respectively. The comp would then be repeated until these values fell below the prescribed stopping criteri final result is $a_0 = 0.79186$ and $a_1 = 1.6751$. These coefficients give a sum of the squ the residuals of 0.000662.

A potential problem with the Gauss-Newton method as developed to this point is that the partial derivatives of the function may be difficult to evaluate. Consequently, many computer programs use difference equations to approximate the partial derivatives. One method is

$$\frac{\partial f_i}{\partial a_k} \cong \frac{f(x_i; a_0, \ldots, a_k + \delta a_k, \ldots, a_m) - f(x_i; a_0, \ldots, a_k, \ldots, a_m)}{\delta a_k} \tag{17.37}$$

where δ = a small fractional perturbation.

The Gauss-Newton method has a number of other possible shortcomings:

1. It may converge slowly.
2. It may oscillate widely, that is, continually change directions.
3. It may not converge at all.

Modifications of the method (Booth and Peterson, 1958; Hartley, 1961) have been developed to remedy the shortcomings.

In addition, although there are several approaches expressly designed for regression, a more general approach is to use nonlinear optimization routines as described in Part Four. To do this, a guess for the parameters is made, and the sum of the squares of the residuals is computed. For example, for Eq. (17.31) it would be computed as

$$S_r = \sum_{i=1}^{n} [y_i - a_0(1 - e^{-a_1 x_i})]^2 \tag{17.38}$$

Then, the parameters would be adjusted systematically to minimize S_r using search techniques of the type described previously in Chap. 14. We will illustrate how this is done when we describe software applications at the end of Chap. 19.

BLEMS

ata

).5	9.8	9.4	10.0
).1	9.2	11.3	9.4
).4	7.9	10.4	9.8
).5	8.9	8.8	10.6
).5	9.6	10.2	8.9

e mean, (b) the standard deviation, (c) the varificient of variation, and (e) the 95% confidence ean.

histogram from the data from Prob. 17.1. Use a 11.5 with intervals of 0.5.

ata

5	26.65	27.65	27.35	28.35	26.85
5	27.85	27.05	28.25	28.85	26.75
5	28.65	28.45	31.65	26.35	27.75
5	28.65	27.65	28.55	27.65	27.25

Determine (a) the mean, (b) the standard deviation, (c) the variance, (d) the coefficient of variation, and (e) the 90% confidence interval for the mean. (f) Construct a histogram. Use a range from 26 to 32 with increments of 0.5. (g) Assuming that the distribution is normal and that your estimate of the standard deviation is valid, compute the range (that is, the lower and the upper values) that encompasses 68% of the readings. Determine whether this is a valid estimate for the data in this problem.

17.4 Use least-squares regression to fit a straight line to

x	0	2	4	6	9	11	12	15	17	19
y	5	6	7	6	9	8	7	10	12	12

Along with the slope and intercept, compute the standard error of the estimate and the correlation coefficient. Plot the data and the regression line. Then repeat the problem, but regress x versus y—that is, switch the variables. Interpret your results.

17.5 Use least-squares regression to fit a straight line to

x	6	7	11	15	17	21	23	29	29	37	39
y	29	21	29	14	21	15	7	7	13	0	3

Along with the slope and the intercept, compute the standard error of the estimate and the correlation coefficient. Plot the data and the regression line. If someone made an additional measurement of $x = 10$, $y = 10$, would you suspect, based on a visual assessment and the standard error, that the measurement was valid or faulty? Justify your conclusion.

17.6 Using the same approach as was employed to derive Eqs. (17.15) and (17.16), derive the least-squares fit of the following model:

$$y = a_1 x + e$$

That is, determine the slope that results in the least-squares fit for a straight line with a zero intercept. Fit the following data with this model and display the result graphically:

x	2	4	6	7	10	11	14	17	20
y	1	2	5	2	8	7	6	9	12

17.7 Use least-squares regression to fit a straight line to

x	1	2	3	4	5	6	7	8	9
y	1	1.5	2	3	4	5	8	10	13

(a) Along with the slope and intercept, compute the standard error of the estimate and the correlation coefficient. Plot the data and the straight line. Assess the fit.

(b) Recompute (a), but use polynomial regression to fit a parabola to the data. Compare the results with those of (a).

17.8 Fit the following data with (a) a saturation-growth-rate model, (b) a power equation, and (c) a parabola. In each case, plot the data and the equation.

x	0.75	2	3	4	6	8	8.5
y	1.2	1.95	2	2.4	2.4	2.7	2.6

17.9 Fit the following data with the power model ($y = ax^b$). Use the resulting power equation to predict y at $x = 9$:

x	2.5	3.5	5	6	7.5	10	12.5	15	17.5	20
y	13	11	8.5	8.2	7	6.2	5.2	4.8	4.6	4.3

17.10 Fit an exponential model to

x	0.4	0.8	1.2	1.6	2	2.3
y	800	975	1500	1950	2900	3600

Plot the data and the equation on both standard and semi-log graph paper.

17.11 Rather than using the base-e exponential model (Eq. 17.20), a common alternative is to use a base-10 model,

$$y = \alpha_5 10^{\beta_5 x}$$

When used for curve fitting, this equation yields identical results to the base-e version, but the value of the exponent parameter will differ from that estimated with Eq. 17.22 (β_1). Use the base-10 version to solve Prob. 17.10. In addition, develop a formula to relate β_1 to β_5.

17.12 Beyond the examples in Fig. 17.10, there are other models that can be linearized using transformations. For example,

$$y = \alpha_4 x e^{\beta_4 x}$$

Linearize this model and use it to estimate α_4 and β_4 based on the following data. Develop a plot of your fit along with the data.

x	0.1	0.2	0.4	0.6	0.9	1.3	1.5	1.7
y	0.75	1.25	1.45	1.25	0.85	0.55	0.35	0.28

17.13 An investigator has reported the data tabulated below from an experiment to determine the growth rate of bacteria k (per day) as a function of oxygen concentration c (mg/L). It is known that the data can be modeled by the following equation:

$$k = \frac{k_{max} c^2}{c_s + c^2}$$

where c_s and k_{max} are parameters. Use a transformation to linearize this equation. Then use linear regression to estimate c_s and k_{max}. Finally, predict the growth rate at $c = 2$ mg/L.

c	0.5	0.8	1.5	2.5	4
k	1.1	2.4	5.3	7.6	8.9

17.14 Given the data

x	5	10	15	20	25	30	35	40	45
y	17	24	31	33	37	37	40	40	42

use least-squares regression to fit (a) a straight line, (b) a power equation, (c) a saturation-growth-rate equation, and (d) a parabola. Plot the data along with all the curves. Is any one of the curves superior? If so, justify.

17.15 Fit a cubic equation to the following data:

x	3	4	5	7	8	9	11	12
y	1.6	3.6	4.4	3.4	2.2	2.8	3.8	4.6

Along with the coefficients, determine r^2 and $s_{y/x}$.

ple linear regression to fit

1	2	2	3	3	4	4
2	1	2	1	2	1	2
12.7	25.6	20.5	35.1	29.7	45.4	40.2

fficients, the standard error of the estimate, and the cient.

ple linear regression to fit

1	2	0	1	2	2	1
2	4	4	6	6	2	1
11	12	23	23	14	6	11

fficients, the standard error of the estimate, and the cient.

near regression to fit a parabola to the following

0.8	1.2	1.7	2	2.3
1000	1200	2200	2650	3750

near regression to fit a saturation-growth-rate ta in Prob. 17.14.

e the regression fits from Probs. (a) 17.4, and (b) matrix approach. Estimate the standard errors and fidence intervals for the coefficients.

debug, and test a program in either a high-level o language of your choice to implement linear re- other things: (a) include statements to document determine the standard error and the coefficient of

is tested for cyclic fatigue failure whereby a applied to the material and the number of cycles failure is measured. The results are in the table g-log plot of stress versus cycles is generated, the a linear relationship. Use least-squares regression st-fit equation for this data.

10	100	1000	10,000	100,000	1,000,000
1000	925	800	625	550	420

ing data shows the relationship between the vis- oil and temperature. After taking the log of the gression to find the equation of the line that best ne r^2 value.

	26.67	93.33	148.89	315.56
m^2	1.35	0.085	0.012	0.00075

17.24 The data below represents the bacterial growth in a liquid culture over a number of days.

Day	0	4	8	12	16	20
Amount × 10^6	67	84	98	125	149	185

Find a best-fit equation to the data trend. Try several possibilities—linear, parabolic, and exponential. Use the software package of your choice to find the best equation to predict the amount of bacteria after 40 days.

17.25 The concentration of *E. coli* bacteria in a swimming area is monitored after a storm:

t (hrs)	4	8	12	16	20	24
c (CFU/100 mL)	1590	1320	1000	900	650	560

The time is measured in hours following the end of the storm and the unit CFU is a "colony forming unit." Use this data to estimate (a) the concentration at the end of the storm ($t = 0$) and (b) the time at which the concentration will reach 200 CFU/100 mL. Note that your choice of model should be consistent with the fact that negative concentrations are impossible and that the bacteria concentration always decreases with time.

17.26 An object is suspended in a wind tunnel and the force measured for various levels of wind velocity. The results are tabulated below.

v, m/s	10	20	30	40	50	60	70	80
F, N	25	70	380	550	610	1220	830	1450

Use least-squares regression to fit this data with (a) a straight line, (b) a power equation based on log transformations, and (c) a power model based on nonlinear regression. Display the results graphically.

17.27 Fit a power model to the data from Prob. 17.26, but use natural logarithms to perform the transformations.

17.28 Derive the least-squares fit of the following model:

$$y = a_1 x + a_2 x^2 + e$$

That is, determine the coefficients that results in the least-squares fit for a second-order polynomial with a zero intercept. Test the approach by using it to fit the data from Prob. 17.26.

17.29 In Prob. 17.12 we used transformations to linearize and fit the following model:

$$y = \alpha_4 x e^{\beta_4 x}$$

Use nonlinear regression to estimate α_4 and β_4 based on the following data. Develop a plot of your fit along with the data.

x	0.1	0.2	0.4	0.6	0.9	1.3	1.5	1.7	1.8
y	0.75	1.25	1.45	1.25	0.85	0.55	0.35	0.28	0.18