CHAPTER

1

Mathematical Preliminaries and Error Analysis

Introduction

In beginning chemistry courses, we see the *ideal gas law*,

$$PV = NRT,$$

which relates the pressure P, volume V, temperature T, and number of moles N of an "ideal" gas. In this equation, R is a constant that depends on the measurement system.

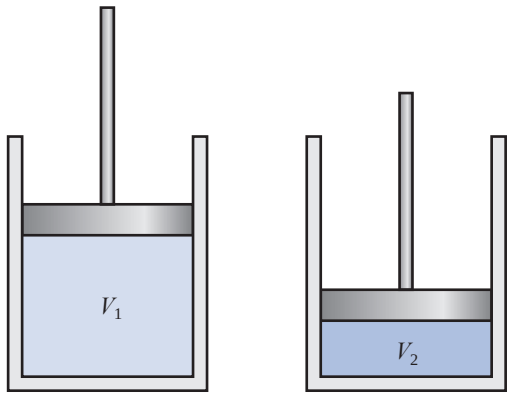
Suppose two experiments are conducted to test this law, using the same gas in each case. In the first experiment,

$$P = 1.00 \text{ atm}, \qquad V = 0.100 \text{ m}^3,$$
$$N = 0.00420 \text{ mol}, \qquad R = 0.08206.$$

The ideal gas law predicts the temperature of the gas to be

$$T = \frac{PV}{NR} = \frac{(1.00)(0.100)}{(0.00420)(0.08206)} = 290.15 \text{ K} = 17^\circ\text{C}.$$

When we measure the temperature of the gas however, we find that the true temperature is 15°C.



We then repeat the experiment using the same values of R and N, but increase the pressure by a factor of two and reduce the volume by the same factor. The product PV remains the same, so the predicted temperature is still 17°C. But now we find that the actual temperature of the gas is 19°C.

Clearly, the ideal gas law is suspect, but before concluding that the law is invalid in this situation, we should examine the data to see whether the error could be attributed to the experimental results. If so, we might be able to determine how much more accurate our experimental results would need to be to ensure that an error of this magnitude did not occur.

Analysis of the error involved in calculations is an important topic in numerical analysis and is introduced in Section 1.2. This particular application is considered in Exercise 28 of that section.

This chapter contains a short review of those topics from single-variable calculus that will be needed in later chapters. A solid knowledge of calculus is essential for an understanding of the analysis of numerical techniques, and more thorough review might be needed if you have been away from this subject for a while. In addition there is an introduction to convergence, error analysis, the machine representation of numbers, and some techniques for categorizing and minimizing computational error.

1.1 Review of Calculus

Limits and Continuity

The concepts of *limit* and *continuity* of a function are fundamental to the study of calculus, and form the basis for the analysis of numerical techniques.

Definition 1.1 A function f defined on a set X of real numbers has the **limit** L at x_0, written

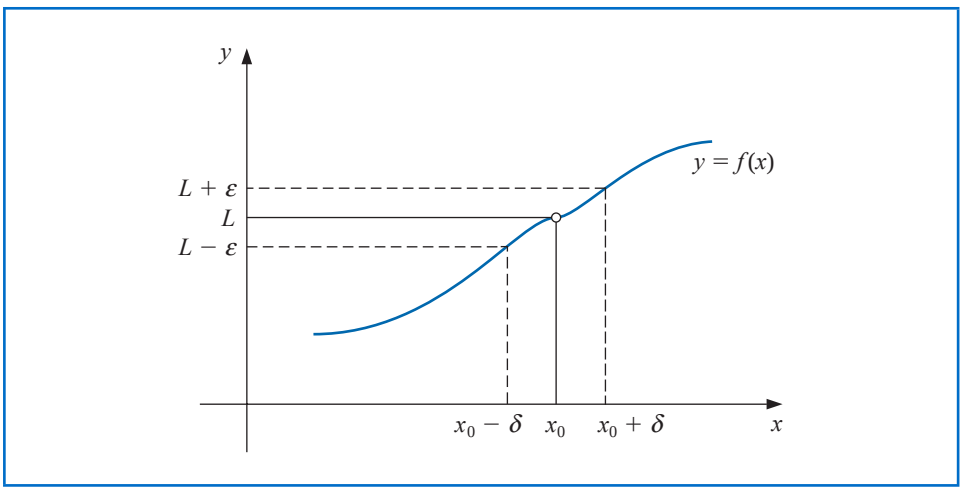
$$\lim_{x \to x_0} f(x) = L,$$

if, given any real number $\varepsilon > 0$, there exists a real number $\delta > 0$ such that

$$|f(x) - L| < \varepsilon, \quad \text{whenever} \quad x \in X \quad \text{and} \quad 0 < |x - x_0| < \delta.$$

(See Figure 1.1.) ■

Figure 1.1



Definition 1.2 Let f be a function defined on a set X of real numbers and $x_0 \in X$. Then f is **continuous** at x_0 if

$$\lim_{x \to x_0} f(x) = f(x_0).$$

The function f is **continuous on the set** X if it is continuous at each number in X. ■

The basic concepts of calculus and its applications were developed in the late 17th and early 18th centuries, but the mathematically precise concepts of limits and continuity were not described until the time of Augustin Louis Cauchy (1789–1857), Heinrich Eduard Heine (1821–1881), and Karl Weierstrass (1815 –1897) in the latter portion of the 19th century.

The set of all functions that are continuous on the set X is denoted $C(X)$. When X is an interval of the real line, the parentheses in this notation are omitted. For example, the set of all functions continuous on the closed interval $[a, b]$ is denoted $C[a, b]$. The symbol $\mathbb{R}$ denotes the set of all real numbers, which also has the interval notation $(-\infty, \infty)$. So the set of all functions that are continuous at every real number is denoted by $C(\mathbb{R})$ or by $C(-\infty, \infty)$.

The *limit of a sequence* of real or complex numbers is defined in a similar manner.

Definition 1.3 Let $\{x_n\}_{n=1}^{\infty}$ be an infinite sequence of real numbers. This sequence has the **limit** x (**converges to** x) if, for any $\varepsilon > 0$ there exists a positive integer $N(\varepsilon)$ such that $|x_n - x| < \varepsilon$, whenever $n > N(\varepsilon)$. The notation

$$\lim_{n \to \infty} x_n = x, \quad \text{or} \quad x_n \to x \quad \text{as} \quad n \to \infty,$$

means that the sequence $\{x_n\}_{n=1}^{\infty}$ converges to x. ■

Theorem 1.4 If f is a function defined on a set X of real numbers and $x_0 \in X$, then the following statements are equivalent:

a. f is continuous at x_0;

b. If $\{x_n\}_{n=1}^{\infty}$ is any sequence in X converging to x_0, then $\lim_{n \to \infty} f(x_n) = f(x_0)$. ■

The functions we will consider when discussing numerical methods will be assumed to be continuous because this is a minimal requirement for predictable behavior. Functions that are not continuous can skip over points of interest, which can cause difficulties when attempting to approximate a solution to a problem.

Differentiability

More sophisticated assumptions about a function generally lead to better approximation results. For example, a function with a smooth graph will normally behave more predictably than one with numerous jagged features. The smoothness condition relies on the concept of the derivative.

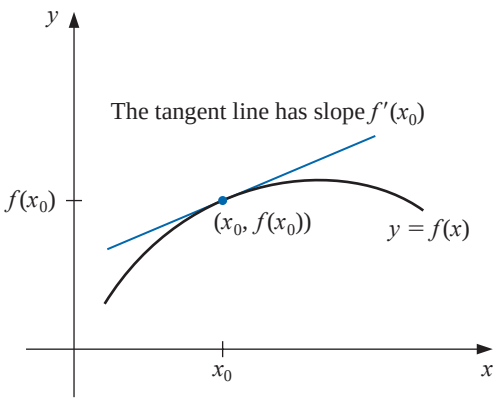
Definition 1.5 Let f be a function defined in an open interval containing x_0. The function f is **differentiable** at x_0 if

$$f'(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}$$

exists. The number $f'(x_0)$ is called the **derivative** of f at x_0. A function that has a derivative at each number in a set X is **differentiable on** X. ■

The derivative of f at x_0 is the slope of the tangent line to the graph of f at $(x_0, f(x_0))$, as shown in Figure 1.2.

Figure 1.2



Theorem 1.6 If the function f is differentiable at x_0, then f is continuous at x_0. ■

The theorem attributed to Michel Rolle (1652–1719) appeared in 1691 in a little-known treatise entitled *Méthode pour résoundre les égalites*. Rolle originally criticized the calculus that was developed by Isaac Newton and Gottfried Leibniz, but later became one of its proponents.

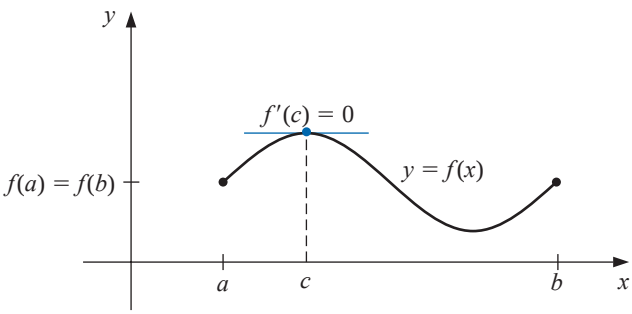
The next theorems are of fundamental importance in deriving methods for error estimation. The proofs of these theorems and the other unreferenced results in this section can be found in any standard calculus text.

The set of all functions that have n continuous derivatives on X is denoted $C^n(X)$, and the set of functions that have derivatives of all orders on X is denoted $C^\infty(X)$. Polynomial, rational, trigonometric, exponential, and logarithmic functions are in $C^\infty(X)$, where X consists of all numbers for which the functions are defined. When X is an interval of the real line, we will again omit the parentheses in this notation.

Theorem 1.7 **(Rolle's Theorem)**

Suppose $f \in C[a, b]$ and f is differentiable on (a, b). If $f(a) = f(b)$, then a number c in (a, b) exists with $f'(c) = 0$. (See Figure 1.3.) ■

Figure 1.3

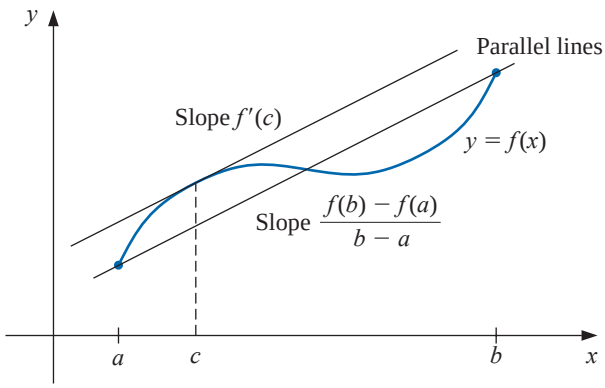


Theorem 1.8 **(Mean Value Theorem)**

If $f \in C[a, b]$ and f is differentiable on (a, b), then a number c in (a, b) exists with (See Figure 1.4.)

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$ ■

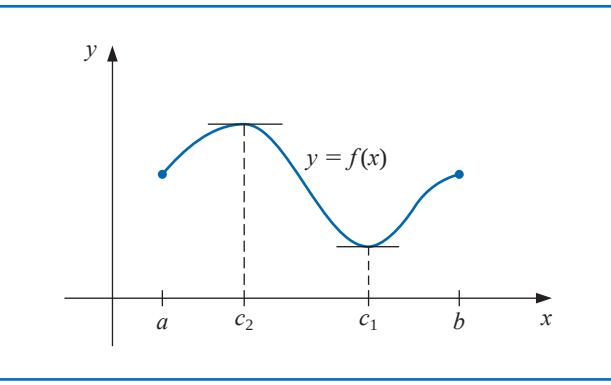
Figure 1.4



Theorem 1.9 **(Extreme Value Theorem)**

If $f \in C[a, b]$, then $c_1, c_2 \in [a, b]$ exist with $f(c_1) \leq f(x) \leq f(c_2)$, for all $x \in [a, b]$. In addition, if f is differentiable on (a, b), then the numbers c_1 and c_2 occur either at the endpoints of $[a, b]$ or where f' is zero. (See Figure 1.5.) ■

Figure 1.5



Research work on the design of algorithms and systems for performing symbolic mathematics began in the 1960s. The first system to be operational, in the 1970s, was a LISP-based system called MACSYMA.

As mentioned in the preface, we will use the computer algebra system Maple whenever appropriate. Computer algebra systems are particularly useful for symbolic differentiation and plotting graphs. Both techniques are illustrated in Example 1.

Example 1 Use Maple to find the absolute minimum and absolute maximum values of

$$f(x) = 5 \cos 2x - 2x \sin 2x f(x)$$

on the intervals **(a)** $[1, 2]$, and **(b)** $[0.5, 1]$

Solution There is a choice of Text input or Math input under the Maple C 2D Math option. The Text input is used to document worksheets by adding standard text information in the document. The Math input option is used to execute Maple commands. Maple input

The Maple development project began at the University of Waterloo in late 1980. Its goal was to be accessible to researchers in mathematics, engineering, and science, but additionally to students for educational purposes. To be effective it needed to be portable, as well as space and time efficient. Demonstrations of the system were presented in 1982, and the major paper setting out the design criteria for the MAPLE system was presented in 1983 [CGGG].

can either be typed or selected from the pallets at the left of the Maple screen. We will show the input as typed because it is easier to accurately describe the commands. For pallet input instructions you should consult the Maple tutorials. In our presentation, Maple input commands appear in *italic* type, and Maple responses appear in cyan type.

To ensure that the variables we use have not been previously assigned, we first issue the command.

restart

to clear the Maple memory. We first illustrate the graphing capabilities of Maple. To access the graphing package, enter the command

with(*plots*)

to load the plots subpackage. Maple responds with a list of available commands in the package. This list can be suppressed by placing a colon after the *with*(*plots*) command.

The following command defines $f(x) = 5\cos 2x - 2x\sin 2x$ as a function of x.

$f := x \to 5\cos(2x) - 2x \cdot \sin(2x)$

and Maple responds with

$$x \to 5\cos(2x) - 2x\sin(2x)$$

We can plot the graph of f on the interval $[0.5, 2]$ with the command

plot(f, 0.5 . . 2)

Figure 1.6 shows the screen that results from this command after doing a mouse click on the graph. This click tells Maple to enter its graph mode, which presents options for various views of the graph. We can determine the coordinates of a point of the graph by moving the mouse cursor to the point. The coordinates appear in the box above the left of the *plot*(f, 0.5 . . 2) command. This feature is useful for estimating the axis intercepts and extrema of functions.

The absolute maximum and minimum values of $f(x)$ on the interval $[a, b]$ can occur only at the endpoints, or at a critical point.

(a) When the interval is $[1, 2]$ we have

$$f(1) = 5\cos 2 - 2\sin 2 = -3.899329036 \quad \text{and} \quad f(2) = 5\cos 4 - 4\sin 4 = -0.241008123.$$

A critical point occurs when $f'(x) = 0$. To use Maple to find this point, we first define a function *fp* to represent f' with the command

$fp := x \to diff(f(x), x)$

and Maple responds with

$$x \to \frac{d}{dx} f(x)$$

To find the explicit representation of $f'(x)$ we enter the command

fp(x)

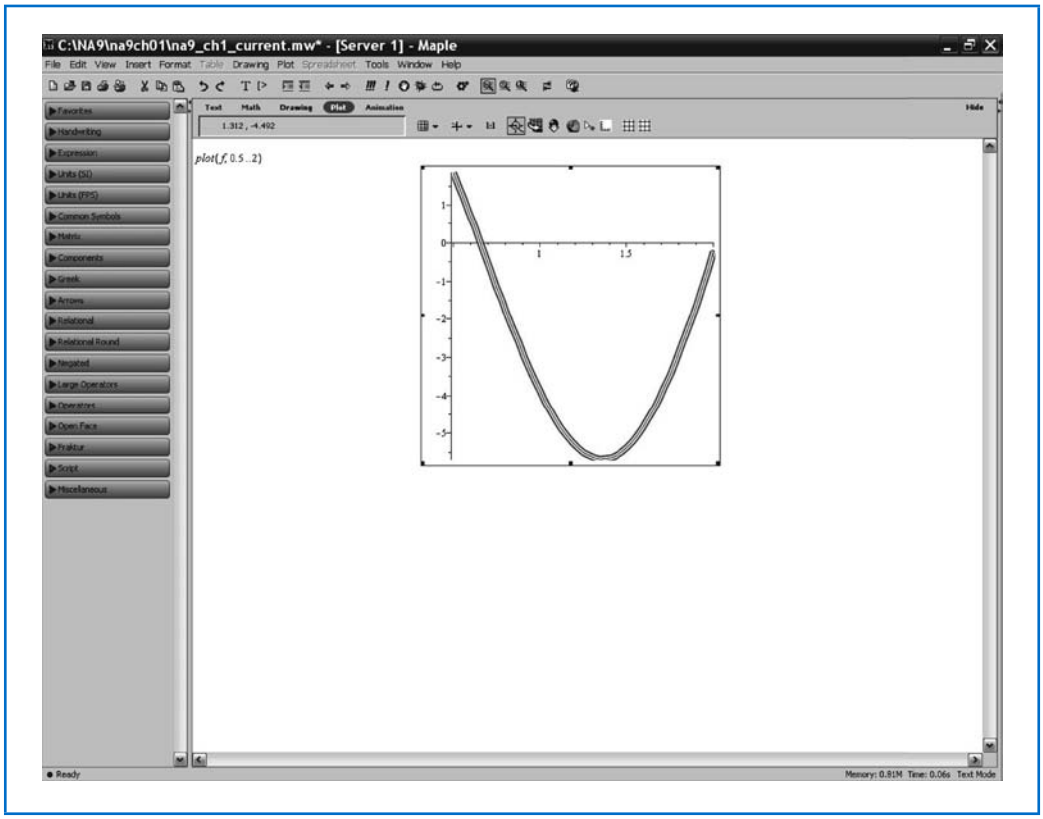
and Maple gives the derivative as

$$-12\sin(2x) - 4x\cos(2x)$$

To determine the critical point we use the command

fsolve(*fp*(x), x, 1 . . 2)

Figure 1.6



and Maple tells us that $f'(x) = fp(x) = 0$ for x in $[1, 2]$ when x is

1.358229874

We evaluate $f(x)$ at this point with the command

$f(\%)$

The % is interpreted as the last Maple response. The value of f at the critical point is

−5.675301338

As a consequence, the absolute maximum value of $f(x)$ in $[1, 2]$ is $f(2) = -0.241008123$ and the absolute minimum value is $f(1.358229874) = -5.675301338$, accurate at least to the places listed.

(b) When the interval is $[0.5, 1]$ we have the values at the endpoints given by

$$f(0.5) = 5\cos 1 - 1\sin 1 = 1.860040545 \quad \text{and} \quad f(1) = 5\cos 2 - 2\sin 2 = -3.899329036.$$

However, when we attempt to determine the critical point in the interval $[0.5, 1]$ with the command

$fsolve(fp(x), x, 0.5\ldots 1)$

Maple gives the response

$$fsolve(-12\sin(2x) - 4x\cos(2x), x, .5 \ldots 1)$$

This indicates that Maple is unable to determine the solution. The reason is obvious once the graph in Figure 1.6 is considered. The function f is always decreasing on this interval, so no solution exists. Be suspicious when Maple returns the same response it is given; it is as if it was questioning your request.

In summary, on $[0.5, 1]$ the absolute maximum value is $f(0.5) = 1.86004545$ and the absolute minimum value is $f(1) = -3.899329036$, accurate at least to the places listed. ■

The following theorem is not generally presented in a basic calculus course, but is derived by applying Rolle's Theorem successively to f, f', ..., and, finally, to $f^{(n-1)}$. This result is considered in Exercise 23.

Theorem 1.10 **(Generalized Rolle's Theorem)**

Suppose $f \in C[a, b]$ is n times differentiable on (a, b). If $f(x) = 0$ at the $n + 1$ distinct numbers $a \le x_0 < x_1 < \ldots < x_n \le b$, then a number c in (x_0, x_n), and hence in (a, b), exists with $f^{(n)}(c) = 0$. ■

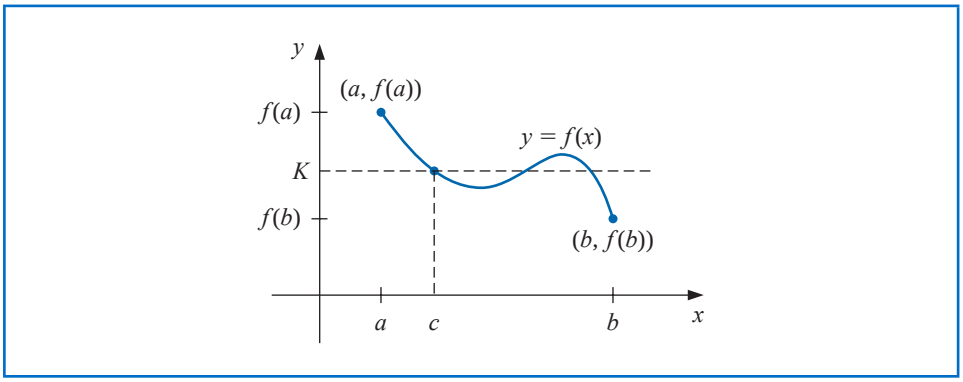
We will also make frequent use of the Intermediate Value Theorem. Although its statement seems reasonable, its proof is beyond the scope of the usual calculus course. It can, however, be found in most analysis texts.

Theorem 1.11 **(Intermediate Value Theorem)**

If $f \in C[a, b]$ and K is any number between $f(a)$ and $f(b)$, then there exists a number c in (a, b) for which $f(c) = K$. ■

Figure 1.7 shows one choice for the number that is guaranteed by the Intermediate Value Theorem. In this example there are two other possibilities.

Figure 1.7



Example 2 Show that $x^5 - 2x^3 + 3x^2 - 1 = 0$ has a solution in the interval $[0, 1]$.

Solution Consider the function defined by $f(x) = x^5 - 2x^3 + 3x^2 - 1$. The function f is continuous on $[0, 1]$. In addition,

$$f(0) = -1 < 0 \qquad \text{and} \qquad 0 < 1 = f(1).$$

The Intermediate Value Theorem implies that a number x exists, with $0 < x < 1$, for which $x^5 - 2x^3 + 3x^2 - 1 = 0$. ■

As seen in Example 2, the Intermediate Value Theorem is used to determine when solutions to certain problems exist. It does not, however, give an efficient means for finding these solutions. This topic is considered in Chapter 2.

Integration

The other basic concept of calculus that will be used extensively is the Riemann integral.

Definition 1.12

The **Riemann integral** of the function f on the interval $[a, b]$ is the following limit, provided it exists:

$$\int_a^b f(x)\,dx = \lim_{\max \Delta x_i \to 0} \sum_{i=1}^{n} f(z_i)\,\Delta x_i,$$

where the numbers $x_0, x_1, \ldots, x_n$ satisfy $a = x_0 \le x_1 \le \cdots \le x_n = b$, where $\Delta x_i = x_i - x_{i-1}$, for each $i = 1, 2, \ldots, n$, and z_i is arbitrarily chosen in the interval $[x_{i-1}, x_i]$. ■

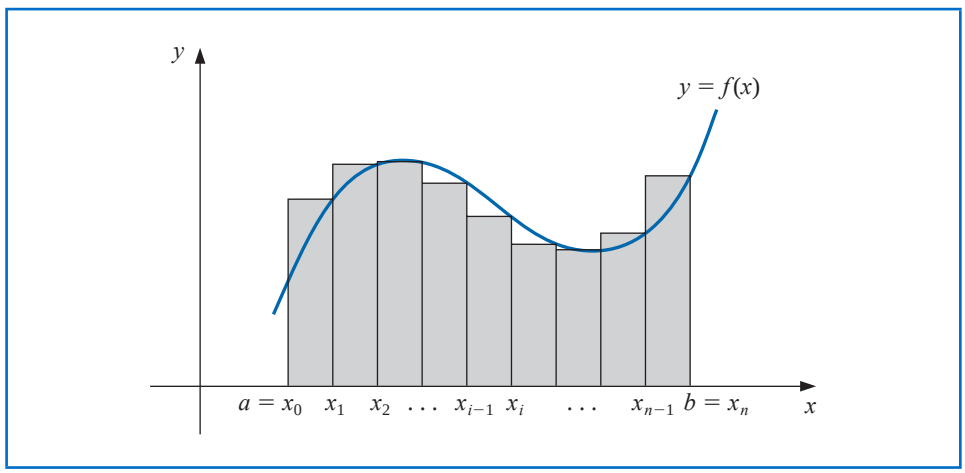
George Fredrich Berhard Riemann (1826–1866) made many of the important discoveries classifying the functions that have integrals. He also did fundamental work in geometry and complex function theory, and is regarded as one of the profound mathematicians of the nineteenth century.

A function f that is continuous on an interval $[a, b]$ is also Riemann integrable on $[a, b]$. This permits us to choose, for computational convenience, the points x_i to be equally spaced in $[a, b]$, and for each $i = 1, 2, \ldots, n$, to choose $z_i = x_i$. In this case,

$$\int_a^b f(x)\,dx = \lim_{n \to \infty} \frac{b-a}{n} \sum_{i=1}^{n} f(x_i),$$

where the numbers shown in Figure 1.8 as x_i are $x_i = a + i(b-a)/n$.

Figure 1.8



Two other results will be needed in our study of numerical analysis. The first is a generalization of the usual Mean Value Theorem for Integrals.

Theorem 1.13 **(Weighted Mean Value Theorem for Integrals)**

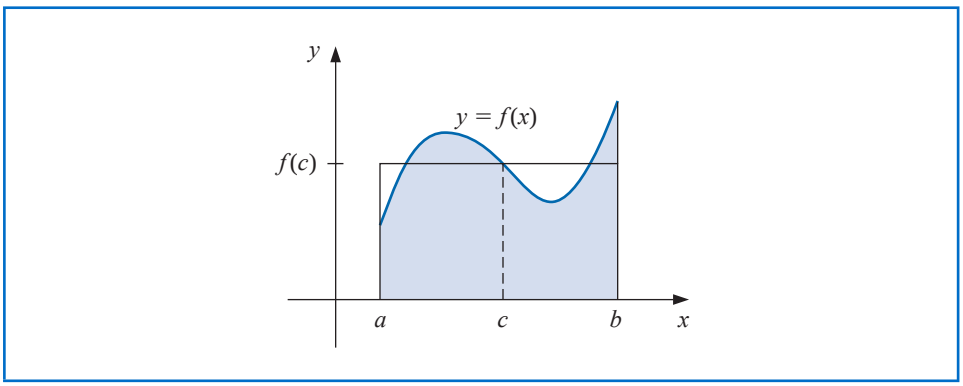
Suppose $f \in C[a, b]$, the Riemann integral of g exists on $[a, b]$, and $g(x)$ does not change sign on $[a, b]$. Then there exists a number c in (a, b) with

$$\int_a^b f(x)g(x)\,dx = f(c)\int_a^b g(x)\,dx.$$ ■

When $g(x) \equiv 1$, Theorem 1.13 is the usual Mean Value Theorem for Integrals. It gives the **average value** of the function f over the interval $[a, b]$ as (See Figure 1.9.)

$$f(c) = \frac{1}{b-a}\int_a^b f(x)\,dx.$$

Figure 1.9



The proof of Theorem 1.13 is not generally given in a basic calculus course but can be found in most analysis texts (see, for example, [Fu], p. 162).

Taylor Polynomials and Series

The final theorem in this review from calculus describes the Taylor polynomials. These polynomials are used extensively in numerical analysis.

Theorem 1.14 **(Taylor's Theorem)**

Brook Taylor (1685–1731) described this series in 1715 in the paper *Methodus incrementorum directa et inversa*. Special cases of the result, and likely the result itself, had been previously known to Isaac Newton, James Gregory, and others.

Suppose $f \in C^n[a, b]$, that $f^{(n+1)}$ exists on $[a, b]$, and $x_0 \in [a, b]$. For every $x \in [a, b]$, there exists a number $\xi(x)$ between x_0 and x with

$$f(x) = P_n(x) + R_n(x),$$

where

$$P_n(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n$$

$$= \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k$$

and

$$R_n(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x - x_0)^{n+1}. \quad \blacksquare$$

Colin Maclaurin (1698–1746) is best known as the defender of the calculus of Newton when it came under bitter attack by the Irish philosopher, the Bishop George Berkeley.

Maclaurin did not discover the series that bears his name; it was known to 17th century mathematicians before he was born. However, he did devise a method for solving a system of linear equations that is known as Cramer's rule, which Cramer did not publish until 1750.

Here $P_n(x)$ is called the ***n*th Taylor polynomial** for f about x_0, and $R_n(x)$ is called the **remainder term** (or **truncation error**) associated with $P_n(x)$. Since the number $\xi(x)$ in the truncation error $R_n(x)$ depends on the value of x at which the polynomial $P_n(x)$ is being evaluated, it is a function of the variable x. However, we should not expect to be able to explicitly determine the function $\xi(x)$. Taylor's Theorem simply ensures that such a function exists, and that its value lies between x and x_0. In fact, one of the common problems in numerical methods is to try to determine a realistic bound for the value of $f^{(n+1)}(\xi(x))$ when x is in some specified interval.

The infinite series obtained by taking the limit of $P_n(x)$ as $n \to \infty$ is called the **Taylor series** for f about x_0. In the case $x_0 = 0$, the Taylor polynomial is often called a **Maclaurin polynomial**, and the Taylor series is often called a **Maclaurin series**.

The term **truncation error** in the Taylor polynomial refers to the error involved in using a truncated, or finite, summation to approximate the sum of an infinite series.

Example 3 Let $f(x) = \cos x$ and $x_0 = 0$. Determine

(a) the second Taylor polynomial for f about x_0; and

(b) the third Taylor polynomial for f about x_0.

Solution Since $f \in C^\infty(\mathbb{R})$, Taylor's Theorem can be applied for any $n \geq 0$. Also,

$$f'(x) = -\sin x, \quad f''(x) = -\cos x, \quad f'''(x) = \sin x, \quad \text{and} \quad f^{(4)}(x) = \cos x,$$

so

$$f(0) = 1, \quad f'(0) = 0, \quad f''(0) = -1, \quad \text{and} \quad f'''(0) = 0.$$

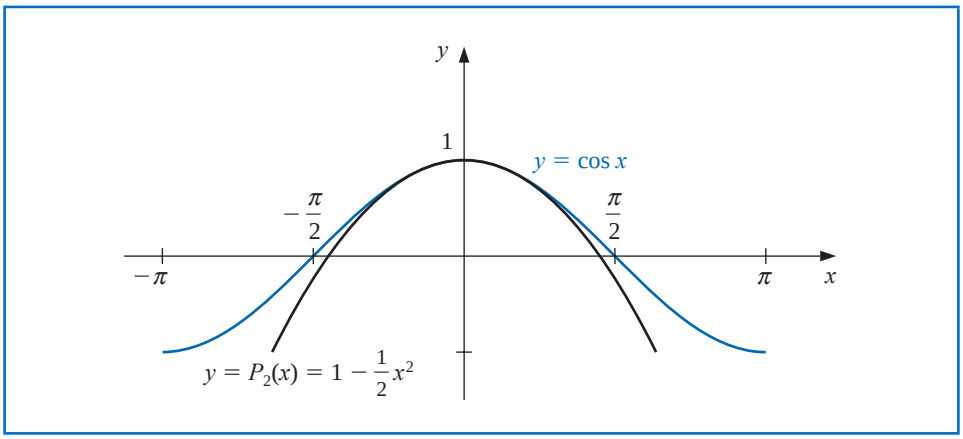
(a) For $n = 2$ and $x_0 = 0$, we have

$$\cos x = f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \frac{f'''(\xi(x))}{3!}x^3$$

$$= 1 - \frac{1}{2}x^2 + \frac{1}{6}x^3 \sin \xi(x),$$

where $\xi(x)$ is some (generally unknown) number between 0 and x. (See Figure 1.10.)

Figure 1.10



When $x = 0.01$, this becomes

$$\cos 0.01 = 1 - \frac{1}{2}(0.01)^2 + \frac{1}{6}(0.01)^3 \sin \xi(0.01) = 0.99995 + \frac{10^{-6}}{6} \sin \xi(0.01).$$

The approximation to $\cos 0.01$ given by the Taylor polynomial is therefore 0.99995. The truncation error, or remainder term, associated with this approximation is

$$\frac{10^{-6}}{6} \sin \xi(0.01) = 0.1\overline{6} \times 10^{-6} \sin \xi(0.01),$$

where the bar over the 6 in $0.1\overline{6}$ is used to indicate that this digit repeats indefinitely. Although we have no way of determining $\sin \xi(0.01)$, we know that all values of the sine lie in the interval $[-1, 1]$, so the error occurring if we use the approximation 0.99995 for the value of $\cos 0.01$ is bounded by

$$|\cos(0.01) - 0.99995| = 0.1\overline{6} \times 10^{-6} |\sin \xi(0.01)| \leq 0.1\overline{6} \times 10^{-6}.$$

Hence the approximation 0.99995 matches at least the first five digits of $\cos 0.01$, and

$$\begin{aligned} 0.9999483 < 0.99995 - 1.\overline{6} \times 10^{-6} &\leq \cos 0.01 \\ &\leq 0.99995 + 1.\overline{6} \times 10^{-6} < 0.9999517. \end{aligned}$$

The error bound is much larger than the actual error. This is due in part to the poor bound we used for $|\sin \xi(x)|$. It is shown in Exercise 24 that for all values of x, we have $|\sin x| \leq |x|$. Since $0 \leq \xi < 0.01$, we could have used the fact that $|\sin \xi(x)| \leq 0.01$ in the error formula, producing the bound $0.1\overline{6} \times 10^{-8}$.

(b) Since $f'''(0) = 0$, the third Taylor polynomial with remainder term about $x_0 = 0$ is

$$\cos x = 1 - \frac{1}{2}x^2 + \frac{1}{24}x^4 \cos \tilde{\xi}(x),$$

where $0 < \tilde{\xi}(x) < 0.01$. The approximating polynomial remains the same, and the approximation is still 0.99995, but we now have much better accuracy assurance. Since $|\cos \tilde{\xi}(x)| \leq 1$ for all x, we have

$$\left| \frac{1}{24}x^4 \cos \tilde{\xi}(x) \right| \leq \frac{1}{24}(0.01)^4(1) \approx 4.2 \times 10^{-10}.$$

So

$$|\cos 0.01 - 0.99995| \leq 4.2 \times 10^{-10},$$

and

$$\begin{aligned} 0.99994999958 &= 0.99995 - 4.2 \times 10^{-10} \\ &\leq \cos 0.01 \leq 0.99995 + 4.2 \times 10^{-10} = 0.99995000042. \end{aligned}$$

■

Example 3 illustrates the two objectives of numerical analysis:

(i) Find an approximation to the solution of a given problem.

(ii) Determine a bound for the accuracy of the approximation.

The Taylor polynomials in both parts provide the same answer to (i), but the third Taylor polynomial gave a much better answer to (ii) than the second Taylor polynomial.

We can also use the Taylor polynomials to give us approximations to integrals.

Illustration We can use the third Taylor polynomial and its remainder term found in Example 3 to approximate $\int_0^{0.1} \cos x\, dx$. We have

$$\begin{aligned}\int_0^{0.1} \cos x\, dx &= \int_0^{0.1} \left(1 - \frac{1}{2}x^2\right) dx + \frac{1}{24}\int_0^{0.1} x^4 \cos\tilde{\xi}(x)\, dx \\ &= \left[x - \frac{1}{6}x^3\right]_0^{0.1} + \frac{1}{24}\int_0^{0.1} x^4 \cos\tilde{\xi}(x)\, dx \\ &= 0.1 - \frac{1}{6}(0.1)^3 + \frac{1}{24}\int_0^{0.1} x^4 \cos\tilde{\xi}(x)\, dx.\end{aligned}$$

Therefore

$$\int_0^{0.1} \cos x\, dx \approx 0.1 - \frac{1}{6}(0.1)^3 = 0.0998\overline{3}.$$

A bound for the error in this approximation is determined from the integral of the Taylor remainder term and the fact that $|\cos\tilde{\xi}(x)| \le 1$ for all x:

$$\begin{aligned}\frac{1}{24}\left|\int_0^{0.1} x^4 \cos\tilde{\xi}(x)\, dx\right| &\le \frac{1}{24}\int_0^{0.1} x^4 |\cos\tilde{\xi}(x)|\, dx \\ &\le \frac{1}{24}\int_0^{0.1} x^4\, dx = \frac{(0.1)^5}{120} = 8.\overline{3} \times 10^{-8}.\end{aligned}$$

The true value of this integral is

$$\int_0^{0.1} \cos x\, dx = \sin x\Big]_0^{0.1} = \sin 0.1 \approx 0.099833416647,$$

so the actual error for this approximation is 8.3314×10^{-8}, which is within the error bound. □

We can also use Maple to obtain these results. Define f by

$f := \cos(x)$

Maple allows us to place multiple statements on a line separated by either a semicolon or a colon. A semicolon will produce all the output, and a colon suppresses all but the final Maple response. For example, the third Taylor polynomial is given by

$s3 := taylor(f, x = 0, 4) :\ p3 := convert(s3, polynom)$

$$1 - \frac{1}{2}x^2$$

The first statement $s3 := taylor(f, x = 0, 4)$ determines the Taylor polynomial about $x_0 = 0$ with four terms (degree 3) and an indication of its remainder. The second $p3 := convert(s3, polynom)$ converts the series $s3$ to the polynomial $p3$ by dropping the remainder term.

Maple normally displays 10 decimal digits for approximations. To instead obtain the 11 digits we want for this illustration, enter

$Digits := 11$

and evaluate $f(0.01)$ and $P_3(0.01)$ with

$y1 := evalf(subs(x = 0.01, f)); y2 := evalf(subs(x = 0.01, p3))$

This produces

$$0.99995000042$$

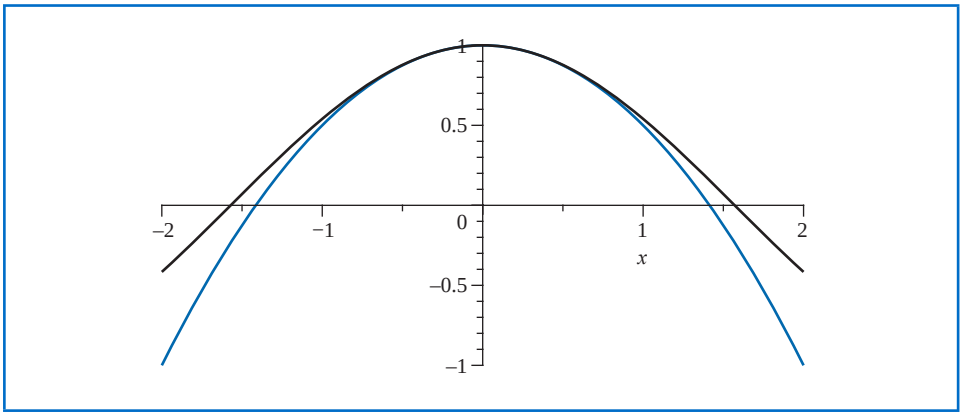
$$0.99995000000$$

To show both the function (in black) and the polynomial (in cyan) near $x_0 = 0$, we enter

$plot((f, p3), x = -2 \ldots 2)$

and obtain the Maple plot shown in Figure 1.11.

Figure 1.11



The integrals of f and the polynomial are given by

$q1 := int(f, x = 0 \ldots 0.1); q2 := int(p3, x = 0 \ldots 0.1)$

$$0.099833416647$$

$$0.099833333333$$

We assigned the names $q1$ and $q2$ to these values so that we could easily determine the error with the command

$err := |q1 - q2|$

$$8.3314\ 10^{-8}$$

There is an alternate method for generating the Taylor polynomials within the *NumericalAnalysis* subpackage of Maple's *Student* package. This subpackage will be discussed in Chapter 2.

EXERCISE SET 1.1

1. Show that the following equations have at least one solution in the given intervals.

a. $x \cos x - 2x^2 + 3x - 1 = 0, \quad [0.2, 0.3]$ and $[1.2, 1.3]$

b. $(x - 2)^2 - \ln x = 0, \quad [1, 2]$ and $[e, 4]$

c. $2x\cos(2x) - (x-2)^2 = 0, \quad [2,3]$ and $[3,4]$

d. $x - (\ln x)^x = 0, \quad [4,5]$

2. Find intervals containing solutions to the following equations.

a. $x - 3^{-x} = 0$

b. $4x^2 - e^x = 0$

c. $x^3 - 2x^2 - 4x + 2 = 0$

d. $x^3 + 4.001x^2 + 4.002x + 1.101 = 0$

3. Show that $f'(x)$ is 0 at least once in the given intervals.

a. $f(x) = 1 - e^x + (e-1)\sin((\pi/2)x), \quad [0,1]$

b. $f(x) = (x-1)\tan x + x\sin \pi x, \quad [0,1]$

c. $f(x) = x\sin \pi x - (x-2)\ln x, \quad [1,2]$

d. $f(x) = (x-2)\sin x \ln(x+2), \quad [-1,3]$

4. Find $\max_{a\le x\le b} |f(x)|$ for the following functions and intervals.

a. $f(x) = (2 - e^x + 2x)/3, \quad [0,1]$

b. $f(x) = (4x-3)/(x^2 - 2x), \quad [0.5,1]$

c. $f(x) = 2x\cos(2x) - (x-2)^2, \quad [2,4]$

d. $f(x) = 1 + e^{-\cos(x-1)}, \quad [1,2]$

5. Use the Intermediate Value Theorem 1.11 and Rolle's Theorem 1.7 to show that the graph of $f(x) = x^3 + 2x + k$ crosses the x-axis exactly once, regardless of the value of the constant k.

6. Suppose $f \in C[a,b]$ and $f'(x)$ exists on (a,b). Show that if $f'(x) \neq 0$ for all x in (a,b), then there can exist at most one number p in $[a,b]$ with $f(p) = 0$.

7. Let $f(x) = x^3$.

a. Find the second Taylor polynomial $P_2(x)$ about $x_0 = 0$.

b. Find $R_2(0.5)$ and the actual error in using $P_2(0.5)$ to approximate $f(0.5)$.

c. Repeat part (a) using $x_0 = 1$.

d. Repeat part (b) using the polynomial from part (c).

8. Find the third Taylor polynomial $P_3(x)$ for the function $f(x) = \sqrt{x+1}$ about $x_0 = 0$. Approximate $\sqrt{0.5}$, $\sqrt{0.75}$, $\sqrt{1.25}$, and $\sqrt{1.5}$ using $P_3(x)$, and find the actual errors.

9. Find the second Taylor polynomial $P_2(x)$ for the function $f(x) = e^x \cos x$ about $x_0 = 0$.

a. Use $P_2(0.5)$ to approximate $f(0.5)$. Find an upper bound for error $|f(0.5) - P_2(0.5)|$ using the error formula, and compare it to the actual error.

b. Find a bound for the error $|f(x) - P_2(x)|$ in using $P_2(x)$ to approximate $f(x)$ on the interval $[0,1]$.

c. Approximate $\int_0^1 f(x)\,dx$ using $\int_0^1 P_2(x)\,dx$.

d. Find an upper bound for the error in (c) using $\int_0^1 |R_2(x)\,dx|$, and compare the bound to the actual error.

10. Repeat Exercise 9 using $x_0 = \pi/6$.

11. Find the third Taylor polynomial $P_3(x)$ for the function $f(x) = (x-1)\ln x$ about $x_0 = 1$.

a. Use $P_3(0.5)$ to approximate $f(0.5)$. Find an upper bound for error $|f(0.5) - P_3(0.5)|$ using the error formula, and compare it to the actual error.

b. Find a bound for the error $|f(x) - P_3(x)|$ in using $P_3(x)$ to approximate $f(x)$ on the interval $[0.5, 1.5]$.

c. Approximate $\int_{0.5}^{1.5} f(x)\,dx$ using $\int_{0.5}^{1.5} P_3(x)\,dx$.

d. Find an upper bound for the error in (c) using $\int_{0.5}^{1.5} |R_3(x)\,dx|$, and compare the bound to the actual error.

12. Let $f(x) = 2x\cos(2x) - (x-2)^2$ and $x_0 = 0$.

a. Find the third Taylor polynomial $P_3(x)$, and use it to approximate $f(0.4)$.

b. Use the error formula in Taylor's Theorem to find an upper bound for the error $|f(0.4) - P_3(0.4)|$. Compute the actual error.

c. Find the fourth Taylor polynomial $P_4(x)$, and use it to approximate $f(0.4)$.

d. Use the error formula in Taylor's Theorem to find an upper bound for the error $|f(0.4) - P_4(0.4)|$. Compute the actual error.

13. Find the fourth Taylor polynomial $P_4(x)$ for the function $f(x) = xe^{x^2}$ about $x_0 = 0$.

a. Find an upper bound for $|f(x) - P_4(x)|$, for $0 \le x \le 0.4$.

b. Approximate $\int_0^{0.4} f(x)\,dx$ using $\int_0^{0.4} P_4(x)\,dx$.

c. Find an upper bound for the error in (b) using $\int_0^{0.4} P_4(x)\,dx$.

d. Approximate $f'(0.2)$ using $P_4'(0.2)$, and find the error.

14. Use the error term of a Taylor polynomial to estimate the error involved in using $\sin x \approx x$ to approximate $\sin 1°$.

15. Use a Taylor polynomial about $\pi/4$ to approximate $\cos 42°$ to an accuracy of 10^{-6}.

16. Let $f(x) = e^{x/2}\sin(x/3)$. Use Maple to determine the following.

a. The third Maclaurin polynomial $P_3(x)$.

b. $f^{(4)}(x)$ and a bound for the error $|f(x) - P_3(x)|$ on $[0, 1]$.

17. Let $f(x) = \ln(x^2 + 2)$. Use Maple to determine the following.

a. The Taylor polynomial $P_3(x)$ for f expanded about $x_0 = 1$.

b. The maximum error $|f(x) - P_3(x)|$, for $0 \le x \le 1$.

c. The Maclaurin polynomial $\tilde{P}_3(x)$ for f.

d. The maximum error $|f(x) - \tilde{P}_3(x)|$, for $0 \le x \le 1$.

e. Does $P_3(0)$ approximate $f(0)$ better than $\tilde{P}_3(1)$ approximates $f(1)$?

18. Let $f(x) = (1 - x)^{-1}$ and $x_0 = 0$. Find the nth Taylor polynomial $P_n(x)$ for $f(x)$ about x_0. Find a value of n necessary for $P_n(x)$ to approximate $f(x)$ to within 10^{-6} on $[0, 0.5]$.

19. Let $f(x) = e^x$ and $x_0 = 0$. Find the nth Taylor polynomial $P_n(x)$ for $f(x)$ about x_0. Find a value of n necessary for $P_n(x)$ to approximate $f(x)$ to within 10^{-6} on $[0, 0.5]$.

20. Find the nth Maclaurin polynomial $P_n(x)$ for $f(x) = \arctan x$.

21. The polynomial $P_2(x) = 1 - \frac{1}{2}x^2$ is to be used to approximate $f(x) = \cos x$ in $[-\frac{1}{2}, \frac{1}{2}]$. Find a bound for the maximum error.

22. The nth Taylor polynomial for a function f at x_0 is sometimes referred to as the polynomial of degree at most n that "best" approximates f near x_0.

a. Explain why this description is accurate.

b. Find the quadratic polynomial that best approximates a function f near $x_0 = 1$ if the tangent line at $x_0 = 1$ has equation $y = 4x - 1$, and if $f''(1) = 6$.

23. Prove the Generalized Rolle's Theorem, Theorem 1.10, by verifying the following.

a. Use Rolle's Theorem to show that $f'(z_i) = 0$ for $n - 1$ numbers in $[a, b]$ with $a < z_1 < z_2 < \cdots < z_{n-1} < b$.

b. Use Rolle's Theorem to show that $f''(w_i) = 0$ for $n - 2$ numbers in $[a, b]$ with $z_1 < w_1 < z_2 < w_2 \cdots w_{n-2} < z_{n-1} < b$.

c. Continue the arguments in **a.** and **b.** to show that for each $j = 1, 2, \ldots, n - 1$ there are $n - j$ distinct numbers in $[a, b]$ where $f^{(j)}$ is 0.

d. Show that part **c.** implies the conclusion of the theorem.

24. In Example 3 it is stated that for all x we have $|\sin x| \le |x|$. Use the following to verify this statement.

a. Show that for all $x \ge 0$ we have $f(x) = x - \sin x$ is non-decreasing, which implies that $\sin x \le x$ with equality only when $x = 0$.

b. Use the fact that the sine function is odd to reach the conclusion.

25. A Maclaurin polynomial for e^x is used to give the approximation 2.5 to e. The error bound in this approximation is established to be $E = \frac{1}{6}$. Find a bound for the error in E.

26. The *error function* defined by

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\,dt$$

gives the probability that any one of a series of trials will lie within x units of the mean, assuming that the trials have a normal distribution with mean 0 and standard deviation $\sqrt{2}/2$. This integral cannot be evaluated in terms of elementary functions, so an approximating technique must be used.

a. Integrate the Maclaurin series for e^{-x^2} to show that

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k+1}}{(2k+1)k!}.$$

b. The error function can also be expressed in the form

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} e^{-x^2} \sum_{k=0}^{\infty} \frac{2^k x^{2k+1}}{1 \cdot 3 \cdot 5 \cdots (2k+1)}.$$

Verify that the two series agree for $k = 1, 2, 3$, and 4. [*Hint:* Use the Maclaurin series for e^{-x^2}.]

c. Use the series in part (a) to approximate erf(1) to within 10^{-7}.

d. Use the same number of terms as in part (c) to approximate erf(1) with the series in part (b).

e. Explain why difficulties occur using the series in part (b) to approximate erf(x).

27. A function $f : [a, b] \to \mathbb{R}$ is said to satisfy a *Lipschitz condition* with Lipschitz constant L on $[a, b]$ if, for every $x, y \in [a, b]$, we have $|f(x) - f(y)| \le L|x - y|$.

a. Show that if f satisfies a Lipschitz condition with Lipschitz constant L on an interval $[a, b]$, then $f \in C[a, b]$.

b. Show that if f has a derivative that is bounded on $[a, b]$ by L, then f satisfies a Lipschitz condition with Lipschitz constant L on $[a, b]$.

c. Give an example of a function that is continuous on a closed interval but does not satisfy a Lipschitz condition on the interval.

28. Suppose $f \in C[a, b]$, that x_1 and x_2 are in $[a, b]$.

a. Show that a number ξ exists between x_1 and x_2 with

$$f(\xi) = \frac{f(x_1) + f(x_2)}{2} = \frac{1}{2} f(x_1) + \frac{1}{2} f(x_2).$$

b. Suppose that c_1 and c_2 are positive constants. Show that a number ξ exists between x_1 and x_2 with

$$f(\xi) = \frac{c_1 f(x_1) + c_2 f(x_2)}{c_1 + c_2}.$$

c. Give an example to show that the result in part **b.** does not necessarily hold when c_1 and c_2 have opposite signs with $c_1 \ne -c_2$.

29. Let $f \in C[a, b]$, and let p be in the open interval (a, b).

a. Suppose $f(p) \ne 0$. Show that a $\delta > 0$ exists with $f(x) \ne 0$, for all x in $[p - \delta, p + \delta]$, with $[p - \delta, p + \delta]$ a subset of $[a, b]$.

b. Suppose $f(p) = 0$ and $k > 0$ is given. Show that a $\delta > 0$ exists with $|f(x)| \le k$, for all x in $[p - \delta, p + \delta]$, with $[p - \delta, p + \delta]$ a subset of $[a, b]$.

1.2 Round-off Errors and Computer Arithmetic

The arithmetic performed by a calculator or computer is different from the arithmetic in algebra and calculus courses. You would likely expect that we always have as true statements things such as $2+2 = 4$, $4 \cdot 8 = 32$, and $(\sqrt{3})^2 = 3$. However, with *computer* arithmetic we expect exact results for $2 + 2 = 4$ and $4 \cdot 8 = 32$, but we will not have precisely $(\sqrt{3})^2 = 3$. To understand why this is true we must explore the world of finite-digit arithmetic.

In our traditional mathematical world we permit numbers with an infinite number of digits. The arithmetic we use in this world *defines* $\sqrt{3}$ as that unique positive number that when multiplied by itself produces the integer 3. In the computational world, however, each representable number has only a fixed and finite number of digits. This means, for example, that only rational numbers—and not even all of these—can be represented exactly. Since $\sqrt{3}$ is not rational, it is given an approximate representation, one whose square will not be precisely 3, although it will likely be sufficiently close to 3 to be acceptable in most situations. In most cases, then, this machine arithmetic is satisfactory and passes without notice or concern, but at times problems arise because of this discrepancy.

Error due to rounding should be expected whenever computations are performed using numbers that are not powers of 2. Keeping this error under control is extremely important when the number of calculations is large.

The error that is produced when a calculator or computer is used to perform real-number calculations is called **round-off error**. It occurs because the arithmetic performed in a machine involves numbers with only a finite number of digits, with the result that calculations are performed with only approximate representations of the actual numbers. In a computer, only a relatively small subset of the real number system is used for the representation of all the real numbers. This subset contains only rational numbers, both positive and negative, and stores the fractional part, together with an exponential part.

Binary Machine Numbers

In 1985, the IEEE (Institute for Electrical and Electronic Engineers) published a report called *Binary Floating Point Arithmetic Standard 754–1985*. An updated version was published in 2008 as *IEEE 754-2008*. This provides standards for binary and decimal floating point numbers, formats for data interchange, algorithms for rounding arithmetic operations, and for the handling of exceptions. Formats are specified for single, double, and extended precisions, and these standards are generally followed by all microcomputer manufacturers using floating-point hardware.

A 64-bit (binary digit) representation is used for a real number. The first bit is a sign indicator, denoted s. This is followed by an 11-bit exponent, c, called the **characteristic**, and a 52-bit binary fraction, f, called the **mantissa**. The base for the exponent is 2.

Since 52 binary digits correspond to between 16 and 17 decimal digits, we can assume that a number represented in this system has at least 16 decimal digits of precision. The exponent of 11 binary digits gives a range of 0 to $2^{11} - 1 = 2047$. However, using only positive integers for the exponent would not permit an adequate representation of numbers with small magnitude. To ensure that numbers with small magnitude are equally representable, 1023 is subtracted from the characteristic, so the range of the exponent is actually from -1023 to 1024.

To save storage and provide a unique representation for each floating-point number, a normalization is imposed. Using this system gives a floating-point number of the form

$$(-1)^s 2^{c-1023}(1 + f).$$

Illustration Consider the machine number

0 10000000011 1011100100010000000000000000000000000000000000000000000000.

The leftmost bit is $s = 0$, which indicates that the number is positive. The next 11 bits, 10000000011, give the characteristic and are equivalent to the decimal number

$$c = 1 \cdot 2^{10} + 0 \cdot 2^9 + \cdots + 0 \cdot 2^2 + 1 \cdot 2^1 + 1 \cdot 2^0 = 1024 + 2 + 1 = 1027.$$

The exponential part of the number is, therefore, $2^{1027-1023} = 2^4$. The final 52 bits specify that the mantissa is

$$f = 1 \cdot \left(\frac{1}{2}\right)^1 + 1 \cdot \left(\frac{1}{2}\right)^3 + 1 \cdot \left(\frac{1}{2}\right)^4 + 1 \cdot \left(\frac{1}{2}\right)^5 + 1 \cdot \left(\frac{1}{2}\right)^8 + 1 \cdot \left(\frac{1}{2}\right)^{12}.$$

As a consequence, this machine number precisely represents the decimal number

$$\begin{aligned}(-1)^s 2^{c-1023}(1+f) &= (-1)^0 \cdot 2^{1027-1023}\left(1 + \left(\frac{1}{2} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32} + \frac{1}{256} + \frac{1}{4096}\right)\right) \\ &= 27.56640625.\end{aligned}$$

However, the next smallest machine number is

0 10000000011 1011100100001111111111111111111111111111111111111111111111,

and the next largest machine number is

0 10000000011 1011100100010000000000000000000000000000000000000000000001.

This means that our original machine number represents not only 27.56640625, but also half of the real numbers that are between 27.56640625 and the next smallest machine number, as well as half the numbers between 27.56640625 and the next largest machine number. To be precise, it represents any real number in the interval

$$\begin{aligned}&[27.566406249999998223643160599749535322189331054687 5,\\ &27.566406250000001776356839400250464677810668945312 5).\end{aligned}$$ □

The smallest normalized positive number that can be represented has $s = 0$, $c = 1$, and $f = 0$ and is equivalent to

$$2^{-1022} \cdot (1 + 0) \approx 0.22251 \times 10^{-307},$$

and the largest has $s = 0$, $c = 2046$, and $f = 1 - 2^{-52}$ and is equivalent to

$$2^{1023} \cdot (2 - 2^{-52}) \approx 0.17977 \times 10^{309}.$$

Numbers occurring in calculations that have a magnitude less than

$$2^{-1022} \cdot (1 + 0)$$

result in **underflow** and are generally set to zero. Numbers greater than

$$2^{1023} \cdot (2 - 2^{-52})$$

result in **overflow** and typically cause the computations to stop (unless the program has been designed to detect this occurrence). Note that there are two representations for the number zero; a positive 0 when $s = 0$, $c = 0$ and $f = 0$, and a negative 0 when $s = 1$, $c = 0$ and $f = 0$.

Decimal Machine Numbers

The use of binary digits tends to conceal the computational difficulties that occur when a finite collection of machine numbers is used to represent all the real numbers. To examine these problems, we will use more familiar decimal numbers instead of binary representation. Specifically, we assume that machine numbers are represented in the normalized *decimal* floating-point form

$$\pm 0.d_1d_2 \dots d_k \times 10^n, \quad 1 \le d_1 \le 9, \quad \text{and} \quad 0 \le d_i \le 9,$$

for each $i = 2, \dots, k$. Numbers of this form are called k-digit *decimal machine numbers*.

Any positive real number within the numerical range of the machine can be normalized to the form

$$y = 0.d_1d_2 \dots d_kd_{k+1}d_{k+2} \dots \times 10^n.$$

The error that results from replacing a number with its floating-point form is called **round-off error** regardless of whether the rounding or chopping method is used.

The floating-point form of y, denoted $fl(y)$, is obtained by terminating the mantissa of y at k decimal digits. There are two common ways of performing this termination. One method, called **chopping**, is to simply chop off the digits $d_{k+1}d_{k+2} \dots$. This produces the floating-point form

$$fl(y) = 0.d_1d_2 \dots d_k \times 10^n.$$

The other method, called **rounding**, adds $5 \times 10^{n-(k+1)}$ to y and then chops the result to obtain a number of the form

$$fl(y) = 0.\delta_1\delta_2 \dots \delta_k \times 10^n.$$

For rounding, when $d_{k+1} \ge 5$, we add 1 to d_k to obtain $fl(y)$; that is, we *round up*. When $d_{k+1} < 5$, we simply chop off all but the first k digits; so we *round down*. If we round down, then $\delta_i = d_i$, for each $i = 1, 2, \dots, k$. However, if we round up, the digits (and even the exponent) might change.

Example 1 Determine the five-digit **(a)** chopping and **(b)** rounding values of the irrational number π.

Solution The number π has an infinite decimal expansion of the form $\pi = 3.14159265\dots$. Written in normalized decimal form, we have

$$\pi = 0.314159265 \dots \times 10^1.$$

The relative error is generally a better measure of accuracy than the absolute error because it takes into consideration the size of the number being approximated.

(a) The floating-point form of π using five-digit chopping is

$$fl(\pi) = 0.31415 \times 10^1 = 3.1415.$$

(b) The sixth digit of the decimal expansion of π is a 9, so the floating-point form of π using five-digit rounding is

$$fl(\pi) = (0.31415 + 0.00001) \times 10^1 = 3.1416.$$ ■

The following definition describes two methods for measuring approximation errors.

Definition 1.15 Suppose that p^* is an approximation to p. The **absolute error** is $|p - p^*|$, and the **relative error** is $\dfrac{|p - p^*|}{|p|}$, provided that $p \ne 0$. ■

Consider the absolute and relative errors in representing p by p^* in the following example.

Example 2 Determine the absolute and relative errors when approximating p by p^* when

(a) $p = 0.3000 \times 10^1$ and $p^* = 0.3100 \times 10^1$;

(b) $p = 0.3000 \times 10^{-3}$ and $p^* = 0.3100 \times 10^{-3}$;

(c) $p = 0.3000 \times 10^4$ and $p^* = 0.3100 \times 10^4$.

Solution

We often cannot find an accurate value for the true error in an approximation. Instead we find a bound for the error, which gives us a "worst-case" error.

(a) For $p = 0.3000 \times 10^1$ and $p^* = 0.3100 \times 10^1$ the absolute error is 0.1, and the relative error is $0.333\overline{3} \times 10^{-1}$.

(b) For $p = 0.3000 \times 10^{-3}$ and $p^* = 0.3100 \times 10^{-3}$ the absolute error is 0.1×10^{-4}, and the relative error is $0.333\overline{3} \times 10^{-1}$.

(c) For $p = 0.3000 \times 10^4$ and $p^* = 0.3100 \times 10^4$, the absolute error is 0.1×10^3, and the relative error is again $0.333\overline{3} \times 10^{-1}$.

This example shows that the same relative error, $0.333\overline{3} \times 10^{-1}$, occurs for widely varying absolute errors. As a measure of accuracy, the absolute error can be misleading and the relative error more meaningful, because the relative error takes into consideration the size of the value. ■

The following definition uses relative error to give a measure of significant digits of accuracy for an approximation.

Definition 1.16

The term significant digits is often used to loosely describe the number of decimal digits that appear to be accurate. The definition is more precise, and provides a continuous concept.

The number p^* is said to approximate p to t **significant digits** (or figures) if t is the largest nonnegative integer for which

$$\frac{|p - p^*|}{|p|} \leq 5 \times 10^{-t}.$$ ■

Table 1.1 illustrates the continuous nature of significant digits by listing, for the various values of p, the least upper bound of $|p - p^*|$, denoted max $|p - p^*|$, when p^* agrees with p to four significant digits.

Table 1.1

p	0.1	0.5	100	1000	5000	9990	10000
max $\lvert p - p^* \rvert$	0.00005	0.00025	0.05	0.5	2.5	4.995	5.

Returning to the machine representation of numbers, we see that the floating-point representation $fl(y)$ for the number y has the relative error

$$\left| \frac{y - fl(y)}{y} \right|.$$

If k decimal digits and chopping are used for the machine representation of

$$y = 0.d_1 d_2 \ldots d_k d_{k+1} \ldots \times 10^n,$$

then

$$\left|\frac{y - fl(y)}{y}\right| = \left|\frac{0.d_1d_2 \ldots d_kd_{k+1} \ldots \times 10^n - 0.d_1d_2 \ldots d_k \times 10^n}{0.d_1d_2 \ldots \times 10^n}\right|$$

$$= \left|\frac{0.d_{k+1}d_{k+2} \ldots \times 10^{n-k}}{0.d_1d_2 \ldots \times 10^n}\right| = \left|\frac{0.d_{k+1}d_{k+2} \ldots}{0.d_1d_2 \ldots}\right| \times 10^{-k}.$$

Since $d_1 \neq 0$, the minimal value of the denominator is 0.1. The numerator is bounded above by 1. As a consequence,

$$\left|\frac{y - fl(y)}{y}\right| \leq \frac{1}{0.1} \times 10^{-k} = 10^{-k+1}.$$

In a similar manner, a bound for the relative error when using k-digit rounding arithmetic is $0.5 \times 10^{-k+1}$. (See Exercise 24.)

Note that the bounds for the relative error using k-digit arithmetic are independent of the number being represented. This result is due to the manner in which the machine numbers are distributed along the real line. Because of the exponential form of the characteristic, the same number of decimal machine numbers is used to represent each of the intervals $[0.1, 1]$, $[1, 10]$, and $[10, 100]$. In fact, within the limits of the machine, the number of decimal machine numbers in $[10^n, 10^{n+1}]$ is constant for all integers n.

Finite-Digit Arithmetic

In addition to inaccurate representation of numbers, the arithmetic performed in a computer is not exact. The arithmetic involves manipulating binary digits by various shifting, or logical, operations. Since the actual mechanics of these operations are not pertinent to this presentation, we shall devise our own approximation to computer arithmetic. Although our arithmetic will not give the exact picture, it suffices to explain the problems that occur. (For an explanation of the manipulations actually involved, the reader is urged to consult more technically oriented computer science texts, such as [Ma], *Computer System Architecture*.)

Assume that the floating-point representations $fl(x)$ and $fl(y)$ are given for the real numbers x and y and that the symbols $\oplus$, $\ominus$, $\otimes$, $\oslash$ represent machine addition, subtraction, multiplication, and division operations, respectively. We will assume a finite-digit arithmetic given by

$$x \oplus y = fl(fl(x) + fl(y)), \quad x \otimes y = fl(fl(x) \times fl(y)),$$
$$x \ominus y = fl(fl(x) - fl(y)), \quad x \oslash y = fl(fl(x) \div fl(y)).$$

This arithmetic corresponds to performing exact arithmetic on the floating-point representations of x and y and then converting the exact result to its finite-digit floating-point representation.

Rounding arithmetic is easily implemented in Maple. For example, the command

Digits := 5

causes all arithmetic to be rounded to 5 digits. To ensure that Maple uses approximate rather than exact arithmetic we use the *evalf*. For example, if $x = \pi$ and $y = \sqrt{2}$ then

evalf(*x*); *evalf*(*y*)

produces 3.1416 and 1.4142, respectively. Then $fl(fl(x) + fl(y))$ is performed using 5-digit rounding arithmetic with

evalf(*evalf*(*x*) + *evalf*(*y*))

which gives 4.5558. Implementing finite-digit chopping arithmetic is more difficult and requires a sequence of steps or a procedure. Exercise 27 explores this problem.

Example 3 Suppose that $x = \frac{5}{7}$ and $y = \frac{1}{3}$. Use five-digit chopping for calculating $x + y$, $x - y$, $x \times y$, and $x \div y$.

Solution Note that

$$x = \frac{5}{7} = 0.\overline{714285} \quad \text{and} \quad y = \frac{1}{3} = 0.\overline{3}$$

implies that the five-digit chopping values of x and y are

$$fl(x) = 0.71428 \times 10^0 \quad \text{and} \quad fl(y) = 0.33333 \times 10^0.$$

Thus

$$\begin{aligned} x \oplus y &= fl(fl(x) + fl(y)) = fl\left(0.71428 \times 10^0 + 0.33333 \times 10^0\right) \\ &= fl\left(1.04761 \times 10^0\right) = 0.10476 \times 10^1. \end{aligned}$$

The true value is $x + y = \frac{5}{7} + \frac{1}{3} = \frac{22}{21}$, so we have

$$\text{Absolute Error} = \left|\frac{22}{21} - 0.10476 \times 10^1\right| = 0.190 \times 10^{-4}$$

and

$$\text{Relative Error} = \left|\frac{0.190 \times 10^{-4}}{22/21}\right| = 0.182 \times 10^{-4}.$$

Table 1.2 lists the values of this and the other calculations. ■

Table 1.2

Operation	Result	Actual value	Absolute error	Relative error
$x \oplus y$	0.10476×10^1	$22/21$	0.190×10^{-4}	0.182×10^{-4}
$x \ominus y$	0.38095×10^0	$8/21$	0.238×10^{-5}	0.625×10^{-5}
$x \otimes y$	0.23809×10^0	$5/21$	0.524×10^{-5}	0.220×10^{-4}
$x \oslash y$	0.21428×10^1	$15/7$	0.571×10^{-4}	0.267×10^{-4}

The maximum relative error for the operations in Example 3 is 0.267×10^{-4}, so the arithmetic produces satisfactory five-digit results. This is not the case in the following example.

Example 4 Suppose that in addition to $x = \frac{5}{7}$ and $y = \frac{1}{3}$ we have

$$u = 0.714251, \quad v = 98765.9, \quad \text{and} \quad w = 0.111111 \times 10^{-4},$$

so that

$$fl(u) = 0.71425 \times 10^0, \quad fl(v) = 0.98765 \times 10^5, \quad \text{and} \quad fl(w) = 0.11111 \times 10^{-4}.$$

Determine the five-digit chopping values of $x \ominus u$, $(x \ominus u) \oslash w$, $(x \ominus u) \otimes v$, and $u \oplus v$.

Solution These numbers were chosen to illustrate some problems that can arise with finite-digit arithmetic. Because x and u are nearly the same, their difference is small. The absolute error for $x \ominus u$ is

$$\begin{aligned}|(x-u)-(x \ominus u)| &= |(x-u)-(fl(fl(x)-fl(u)))| \\ &= \left|\left(\frac{5}{7}-0.714251\right)-\left(fl\left(0.71428\times 10^{0}-0.71425\times 10^{0}\right)\right)\right| \\ &= \left|0.347143\times 10^{-4}-fl\left(0.00003\times 10^{0}\right)\right| = 0.47143\times 10^{-5}.\end{aligned}$$

This approximation has a small absolute error, but a large relative error

$$\left|\frac{0.47143\times 10^{-5}}{0.347143\times 10^{-4}}\right| \leq 0.136.$$

The subsequent division by the small number w or multiplication by the large number v magnifies the absolute error without modifying the relative error. The addition of the large and small numbers u and v produces large absolute error but not large relative error. These calculations are shown in Table 1.3. ■

Table 1.3

Operation	Result	Actual value	Absolute error	Relative error
$x \ominus u$	0.30000×10^{-4}	0.34714×10^{-4}	0.471×10^{-5}	0.136
$(x \ominus u) \oplus w$	0.27000×10^{1}	0.31242×10^{1}	0.424	0.136
$(x \ominus u) \otimes v$	0.29629×10^{1}	0.34285×10^{1}	0.465	0.136
$u \oplus v$	0.98765×10^{5}	0.98766×10^{5}	0.161×10^{1}	0.163×10^{-4}

One of the most common error-producing calculations involves the cancelation of significant digits due to the subtraction of nearly equal numbers. Suppose two nearly equal numbers x and y, with $x > y$, have the k-digit representations

$$fl(x) = 0.d_1d_2\ldots d_p\alpha_{p+1}\alpha_{p+2}\ldots\alpha_k\times 10^n,$$

and

$$fl(y) = 0.d_1d_2\ldots d_p\beta_{p+1}\beta_{p+2}\ldots\beta_k\times 10^n.$$

The floating-point form of $x-y$ is

$$fl(fl(x)-fl(y)) = 0.\sigma_{p+1}\sigma_{p+2}\ldots\sigma_k\times 10^{n-p},$$

where

$$0.\sigma_{p+1}\sigma_{p+2}\ldots\sigma_k = 0.\alpha_{p+1}\alpha_{p+2}\ldots\alpha_k - 0.\beta_{p+1}\beta_{p+2}\ldots\beta_k.$$

The floating-point number used to represent $x-y$ has at most $k-p$ digits of significance. However, in most calculation devices, $x-y$ will be assigned k digits, with the last p being either zero or randomly assigned. Any further calculations involving $x-y$ retain the problem of having only $k-p$ digits of significance, since a chain of calculations is no more accurate than its weakest portion.

If a finite-digit representation or calculation introduces an error, further enlargement of the error occurs when dividing by a number with small magnitude (or, equivalently, when

multiplying by a number with large magnitude). Suppose, for example, that the number z has the finite-digit approximation $z + \delta$, where the error δ is introduced by representation or by previous calculation. Now divide by $\varepsilon = 10^{-n}$, where $n > 0$. Then

$$\frac{z}{\varepsilon} \approx fl\left(\frac{fl(z)}{fl(\varepsilon)}\right) = (z + \delta) \times 10^n.$$

The absolute error in this approximation, $|\delta| \times 10^n$, is the original absolute error, $|\delta|$, multiplied by the factor 10^n.

Example 5 Let $p = 0.54617$ and $q = 0.54601$. Use four-digit arithmetic to approximate $p - q$ and determine the absolute and relative errors using **(a)** rounding and **(b)** chopping.

Solution The exact value of $r = p - q$ is $r = 0.00016$.

(a) Suppose the subtraction is performed using four-digit rounding arithmetic. Rounding p and q to four digits gives $p^* = 0.5462$ and $q^* = 0.5460$, respectively, and $r^* = p^* - q^* = 0.0002$ is the four-digit approximation to r. Since

$$\frac{|r - r^*|}{|r|} = \frac{|0.00016 - 0.0002|}{|0.00016|} = 0.25,$$

the result has only one significant digit, whereas p^* and q^* were accurate to four and five significant digits, respectively.

(b) If chopping is used to obtain the four digits, the four-digit approximations to p, q, and r are $p^* = 0.5461$, $q^* = 0.5460$, and $r^* = p^* - q^* = 0.0001$. This gives

$$\frac{|r - r^*|}{|r|} = \frac{|0.00016 - 0.0001|}{|0.00016|} = 0.375,$$

which also results in only one significant digit of accuracy. ■

The loss of accuracy due to round-off error can often be avoided by a reformulation of the calculations, as illustrated in the next example.

Illustration The quadratic formula states that the roots of $ax^2 + bx + c = 0$, when $a \neq 0$, are

$$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \quad \text{and} \quad x_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}. \tag{1.1}$$

Consider this formula applied to the equation $x^2 + 62.10x + 1 = 0$, whose roots are approximately

$$x_1 = -0.01610723 \quad \text{and} \quad x_2 = -62.08390.$$

The roots x_1 and x_2 of a general quadratic equation are related to the coefficients by the fact that

$$x_1 + x_2 = -\frac{b}{a}$$

and

$$x_1 x_2 = \frac{c}{a}.$$

This is a special case of Viète's Formulas for the coefficients of polynomials.

We will again use four-digit rounding arithmetic in the calculations to determine the root. In this equation, b^2 is much larger than $4ac$, so the numerator in the calculation for x_1 involves the *subtraction* of nearly equal numbers. Because

$$\begin{aligned}\sqrt{b^2 - 4ac} &= \sqrt{(62.10)^2 - (4.000)(1.000)(1.000)} \\ &= \sqrt{3856. - 4.000} = \sqrt{3852.} = 62.06,\end{aligned}$$

we have

$$fl(x_1) = \frac{-62.10 + 62.06}{2.000} = \frac{-0.04000}{2.000} = -0.02000,$$

a poor approximation to $x_1 = -0.01611$, with the large relative error

$$\frac{|-0.01611 + 0.02000|}{|-0.01611|} \approx 2.4 \times 10^{-1}.$$

On the other hand, the calculation for x_2 involves the *addition* of the nearly equal numbers $-b$ and $-\sqrt{b^2 - 4ac}$. This presents no problem since

$$fl(x_2) = \frac{-62.10 - 62.06}{2.000} = \frac{-124.2}{2.000} = -62.10$$

has the small relative error

$$\frac{|-62.08 + 62.10|}{|-62.08|} \approx 3.2 \times 10^{-4}.$$

To obtain a more accurate four-digit rounding approximation for x_1, we change the form of the quadratic formula by *rationalizing the numerator*:

$$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a}\left(\frac{-b - \sqrt{b^2 - 4ac}}{-b - \sqrt{b^2 - 4ac}}\right) = \frac{b^2 - (b^2 - 4ac)}{2a(-b - \sqrt{b^2 - 4ac})},$$

which simplifies to an alternate quadratic formula

$$x_1 = \frac{-2c}{b + \sqrt{b^2 - 4ac}}. \tag{1.2}$$

Using (1.2) gives

$$fl(x_1) = \frac{-2.000}{62.10 + 62.06} = \frac{-2.000}{124.2} = -0.01610,$$

which has the small relative error 6.2×10^{-4}.

The rationalization technique can also be applied to give the following alternative quadratic formula for x_2:

$$x_2 = \frac{-2c}{b - \sqrt{b^2 - 4ac}}. \tag{1.3}$$

This is the form to use if b is a negative number. In the Illustration, however, the mistaken use of this formula for x_2 would result in not only the subtraction of nearly equal numbers, but also the division by the small result of this subtraction. The inaccuracy that this combination produces,

$$fl(x_2) = \frac{-2c}{b - \sqrt{b^2 - 4ac}} = \frac{-2.000}{62.10 - 62.06} = \frac{-2.000}{0.04000} = -50.00,$$

has the large relative error 1.9×10^{-1}. □

- The lesson: Think before you compute!

Nested Arithmetic

Accuracy loss due to round-off error can also be reduced by rearranging calculations, as shown in the next example.

Example 6 Evaluate $f(x) = x^3 - 6.1x^2 + 3.2x + 1.5$ at $x = 4.71$ using three-digit arithmetic.

Solution Table 1.4 gives the intermediate results in the calculations.

Table 1.4

	x	x^2	x^3	$6.1x^2$	$3.2x$
Exact	4.71	22.1841	104.487111	135.32301	15.072
Three-digit (chopping)	4.71	22.1	104.	134.	15.0
Three-digit (rounding)	4.71	22.2	105.	135.	15.1

To illustrate the calculations, let us look at those involved with finding x^3 using three-digit rounding arithmetic. First we find

$$x^2 = 4.71^2 = 22.1841 \quad \text{which rounds to } 22.2.$$

Then we use this value of x^2 to find

$$x^3 = x^2 \cdot x = 22.2 \cdot 4.71 = 104.562 \quad \text{which rounds to } 105.$$

Also,

$$6.1x^2 = 6.1(22.2) = 135.42 \quad \text{which rounds to } 135,$$

and

$$3.2x = 3.2(4.71) = 15.072 \quad \text{which rounds to } 15.1.$$

The exact result of the evaluation is

$$\text{Exact:} \quad f(4.71) = 104.487111 - 135.32301 + 15.072 + 1.5 = -14.263899.$$

Using finite-digit arithmetic, the way in which we add the results can effect the final result. Suppose that we add left to right. Then for chopping arithmetic we have

$$\text{Three-digit (chopping):} \quad f(4.71) = ((104. - 134.) + 15.0) + 1.5 = -13.5,$$

and for rounding arithmetic we have

$$\text{Three-digit (rounding):} \quad f(4.71) = ((105. - 135.) + 15.1) + 1.5 = -13.4.$$

(You should carefully verify these results to be sure that your notion of finite-digit arithmetic is correct.) Note that the three-digit chopping values simply retain the leading three digits, with no rounding involved, and differ significantly from the three-digit rounding values.

The relative errors for the three-digit methods are

$$\text{Chopping:} \left|\frac{-14.263899 + 13.5}{-14.263899}\right| \approx 0.05, \quad \text{and Rounding:} \left|\frac{-14.263899 + 13.4}{-14.263899}\right| \approx 0.06.$$

■

Illustration

Remember that chopping (or rounding) is performed after each calculation.

As an alternative approach, the polynomial $f(x)$ in Example 6 can be written in a **nested** manner as

$$f(x) = x^3 - 6.1x^2 + 3.2x + 1.5 = ((x - 6.1)x + 3.2)x + 1.5.$$

Using three-digit chopping arithmetic now produces

$$\begin{aligned} f(4.71) &= ((4.71 - 6.1)4.71 + 3.2)4.71 + 1.5 = ((-1.39)(4.71) + 3.2)4.71 + 1.5 \\ &= (-6.54 + 3.2)4.71 + 1.5 = (-3.34)4.71 + 1.5 = -15.7 + 1.5 = -14.2. \end{aligned}$$

In a similar manner, we now obtain a three-digit rounding answer of -14.3. The new relative errors are

$$\text{Three-digit (chopping):} \quad \left|\frac{-14.263899 + 14.2}{-14.263899}\right| \approx 0.0045;$$

$$\text{Three-digit (rounding):} \quad \left|\frac{-14.263899 + 14.3}{-14.263899}\right| \approx 0.0025.$$

Nesting has reduced the relative error for the chopping approximation to less than 10% of that obtained initially. For the rounding approximation the improvement has been even more dramatic; the error in this case has been reduced by more than 95%. □

Polynomials should *always* be expressed in nested form before performing an evaluation, because this form minimizes the number of arithmetic calculations. The decreased error in the Illustration is due to the reduction in computations from four multiplications and three additions to two multiplications and three additions. One way to reduce round-off error is to reduce the number of computations.

EXERCISE SET 1.2

1. Compute the absolute error and relative error in approximations of p by p^*.

a. $p = \pi, p^* = 22/7$
b. $p = \pi, p^* = 3.1416$
c. $p = e, p^* = 2.718$
d. $p = \sqrt{2}, p^* = 1.414$
e. $p = e^{10}, p^* = 22000$
f. $p = 10^{\pi}, p^* = 1400$
g. $p = 8!, p^* = 39900$
h. $p = 9!, p^* = \sqrt{18\pi}(9/e)^9$

2. Find the largest interval in which p^* must lie to approximate p with relative error at most 10^{-4} for each value of p.

a. π
b. e
c. $\sqrt{2}$
d. $\sqrt[3]{7}$

3. Suppose p^* must approximate p with relative error at most 10^{-3}. Find the largest interval in which p^* must lie for each value of p.

a. 150
b. 900
c. 1500
d. 90

4. Perform the following computations (i) exactly, (ii) using three-digit chopping arithmetic, and (iii) using three-digit rounding arithmetic. (iv) Compute the relative errors in parts (ii) and (iii).

a. $\frac{4}{5} + \frac{1}{3}$
b. $\frac{4}{5} \cdot \frac{1}{3}$
c. $\left(\frac{1}{3} - \frac{3}{11}\right) + \frac{3}{20}$
d. $\left(\frac{1}{3} + \frac{3}{11}\right) - \frac{3}{20}$

5. Use three-digit rounding arithmetic to perform the following calculations. Compute the absolute error and relative error with the exact value determined to at least five digits.

a. $133 + 0.921$
b. $133 - 0.499$
c. $(121 - 0.327) - 119$
d. $(121 - 119) - 0.327$
e. $\dfrac{\frac{13}{14} - \frac{6}{7}}{2e - 5.4}$
f. $-10\pi + 6e - \frac{3}{62}$
g. $\left(\frac{2}{9}\right) \cdot \left(\frac{9}{7}\right)$
h. $\dfrac{\pi - \frac{22}{7}}{\frac{1}{17}}$

6. Repeat Exercise 5 using four-digit rounding arithmetic.

7. Repeat Exercise 5 using three-digit chopping arithmetic.

8. Repeat Exercise 5 using four-digit chopping arithmetic.

9. The first three nonzero terms of the Maclaurin series for the arctangent function are $x - (1/3)x^3 + (1/5)x^5$. Compute the absolute error and relative error in the following approximations of π using the polynomial in place of the arctangent:

a. $4\left[\arctan\left(\frac{1}{2}\right) + \arctan\left(\frac{1}{3}\right)\right]$

b. $16\arctan\left(\frac{1}{5}\right) - 4\arctan\left(\frac{1}{239}\right)$

10. The number e can be defined by $e = \sum_{n=0}^{\infty}(1/n!)$, where $n! = n(n-1)\cdots 2\cdot 1$ for $n \neq 0$ and $0! = 1$. Compute the absolute error and relative error in the following approximations of e:

a. $\sum_{n=0}^{5}\frac{1}{n!}$ **b.** $\sum_{n=0}^{10}\frac{1}{n!}$

11. Let

$$f(x) = \frac{x\cos x - \sin x}{x - \sin x}.$$

a. Find $\lim_{x\to 0} f(x)$.

b. Use four-digit rounding arithmetic to evaluate $f(0.1)$.

c. Replace each trigonometric function with its third Maclaurin polynomial, and repeat part (b).

d. The actual value is $f(0.1) = -1.99899998$. Find the relative error for the values obtained in parts (b) and (c).

12. Let

$$f(x) = \frac{e^x - e^{-x}}{x}.$$

a. Find $\lim_{x\to 0}(e^x - e^{-x})/x$.

b. Use three-digit rounding arithmetic to evaluate $f(0.1)$.

c. Replace each exponential function with its third Maclaurin polynomial, and repeat part (b).

d. The actual value is $f(0.1) = 2.003335000$. Find the relative error for the values obtained in parts (b) and (c).

13. Use four-digit rounding arithmetic and the formulas (1.1), (1.2), and (1.3) to find the most accurate approximations to the roots of the following quadratic equations. Compute the absolute errors and relative errors.

a. $\frac{1}{3}x^2 - \frac{123}{4}x + \frac{1}{6} = 0$

b. $\frac{1}{3}x^2 + \frac{123}{4}x - \frac{1}{6} = 0$

c. $1.002x^2 - 11.01x + 0.01265 = 0$

d. $1.002x^2 + 11.01x + 0.01265 = 0$

14. Repeat Exercise 13 using four-digit chopping arithmetic.

15. Use the 64-bit long real format to find the decimal equivalent of the following floating-point machine numbers.

a. 0 10000001010 1001001100000000000000000000000000000000000000000000

b. 1 10000001010 1001001100000000000000000000000000000000000000000000

c. 0 01111111111 0101001100000000000000000000000000000000000000000000

d. 0 01111111111 0101001100000000000000000000000000000000000000000001

16. Find the next largest and smallest machine numbers in decimal form for the numbers given in Exercise 15.

17. Suppose two points (x_0, y_0) and (x_1, y_1) are on a straight line with $y_1 \neq y_0$. Two formulas are available to find the x-intercept of the line:

$$x = \frac{x_0y_1 - x_1y_0}{y_1 - y_0} \quad \text{and} \quad x = x_0 - \frac{(x_1 - x_0)y_0}{y_1 - y_0}.$$

a. Show that both formulas are algebraically correct.

b. Use the data $(x_0, y_0) = (1.31, 3.24)$ and $(x_1, y_1) = (1.93, 4.76)$ and three-digit rounding arithmetic to compute the x-intercept both ways. Which method is better and why?

18. The Taylor polynomial of degree n for $f(x) = e^x$ is $\sum_{i=0}^{n}(x^i/i!)$. Use the Taylor polynomial of degree nine and three-digit chopping arithmetic to find an approximation to e^{-5} by each of the following methods.

a. $e^{-5} \approx \sum_{i=0}^{9} \frac{(-5)^i}{i!} = \sum_{i=0}^{9} \frac{(-1)^i 5^i}{i!}$

b. $e^{-5} = \frac{1}{e^5} \approx \frac{1}{\sum_{i=0}^{9} \frac{5^i}{i!}}$.

c. An approximate value of e^{-5} correct to three digits is 6.74×10^{-3}. Which formula, (a) or (b), gives the most accuracy, and why?

19. The two-by-two linear system

$$ax + by = e,$$
$$cx + dy = f,$$

where a, b, c, d, e, f are given, can be solved for x and y as follows:

$$\text{set } m = \frac{c}{a}, \quad \text{provided } a \neq 0;$$
$$d_1 = d - mb;$$
$$f_1 = f - me;$$
$$y = \frac{f_1}{d_1};$$
$$x = \frac{(e - by)}{a}.$$

Solve the following linear systems using four-digit rounding arithmetic.

a. $1.130x - 6.990y = 14.20$
$1.013x - 6.099y = 14.22$

b. $8.110x + 12.20y = -0.1370$
$-18.11x + 112.2y = -0.1376$

20. Repeat Exercise 19 using four-digit chopping arithmetic.

21. **a.** Show that the polynomial nesting technique described in Example 6 can also be applied to the evaluation of

$$f(x) = 1.01e^{4x} - 4.62e^{3x} - 3.11e^{2x} + 12.2e^x - 1.99.$$

b. Use three-digit rounding arithmetic, the assumption that $e^{1.53} = 4.62$, and the fact that $e^{nx} = (e^x)^n$ to evaluate $f(1.53)$ as given in part (a).

c. Redo the calculation in part (b) by first nesting the calculations.

d. Compare the approximations in parts (b) and (c) to the true three-digit result $f(1.53) = -7.61$.

22. A rectangular parallelepiped has sides of length 3 cm, 4 cm, and 5 cm, measured to the nearest centimeter. What are the best upper and lower bounds for the volume of this parallelepiped? What are the best upper and lower bounds for the surface area?

23. Let $P_n(x)$ be the Maclaurin polynomial of degree n for the arctangent function. Use Maple carrying 75 decimal digits to find the value of n required to approximate π to within 10^{-25} using the following formulas.

a. $4\left[P_n\left(\frac{1}{2}\right) + P_n\left(\frac{1}{3}\right)\right]$

b. $16P_n\left(\frac{1}{5}\right) - 4P_n\left(\frac{1}{239}\right)$

24. Suppose that $fl(y)$ is a k-digit rounding approximation to y. Show that

$$\left|\frac{y - fl(y)}{y}\right| \leq 0.5 \times 10^{-k+1}.$$

[*Hint:* If $d_{k+1} < 5$, then $fl(y) = 0.d_1d_2 \ldots d_k \times 10^n$. If $d_{k+1} \geq 5$, then $fl(y) = 0.d_1d_2 \ldots d_k \times 10^n + 10^{n-k}$.]

25. The binomial coefficient

$$\binom{m}{k} = \frac{m!}{k!\,(m-k)!}$$

describes the number of ways of choosing a subset of k objects from a set of m elements.

a. Suppose decimal machine numbers are of the form

$$\pm 0.d_1d_2d_3d_4 \times 10^n, \quad \text{with } 1 \leq d_1 \leq 9,\ 0 \leq d_i \leq 9, \text{ if } i = 2, 3, 4 \quad \text{and} \quad |n| \leq 15.$$

What is the largest value of m for which the binomial coefficient $\binom{m}{k}$ can be computed for all k by the definition without causing overflow?

b. Show that $\binom{m}{k}$ can also be computed by

$$\binom{m}{k} = \left(\frac{m}{k}\right)\left(\frac{m-1}{k-1}\right)\cdots\left(\frac{m-k+1}{1}\right).$$

c. What is the largest value of m for which the binomial coefficient $\binom{m}{3}$ can be computed by the formula in part (b) without causing overflow?

d. Use the equation in (b) and four-digit chopping arithmetic to compute the number of possible 5-card hands in a 52-card deck. Compute the actual and relative errors.

26. Let $f \in C[a, b]$ be a function whose derivative exists on (a, b). Suppose f is to be evaluated at x_0 in (a, b), but instead of computing the actual value $f(x_0)$, the approximate value, $\tilde{f}(x_0)$, is the actual value of f at $x_0 + \epsilon$, that is, $\tilde{f}(x_0) = f(x_0 + \epsilon)$.

a. Use the Mean Value Theorem 1.8 to estimate the absolute error $|f(x_0) - \tilde{f}(x_0)|$ and the relative error $|f(x_0) - \tilde{f}(x_0)|/|f(x_0)|$, assuming $f(x_0) \neq 0$.

b. If $\epsilon = 5 \times 10^{-6}$ and $x_0 = 1$, find bounds for the absolute and relative errors for

i. $f(x) = e^x$

ii. $f(x) = \sin x$

c. Repeat part (b) with $\epsilon = (5 \times 10^{-6})x_0$ and $x_0 = 10$.

27. The following Maple procedure chops a floating-point number x to t digits. (Use the Shift and Enter keys at the end of each line when creating the procedure.)

```
chop := proc(x, t);
    local e, x2;
    if x = 0 then 0
    else
        e := ceil(evalf(log10(abs(x))));
        x2 := evalf(trunc(x · 10^(t−e)) · 10^(e−t));
    end if
end;
```

Verify the procedure works for the following values.

a. $x = 124.031,\ t = 5$ **b.** $x = 124.036,\ t = 5$

c. $x = -124.031,\ t = 5$ **d.** $x = -124.036,\ t = 5$

e. $x = 0.00653,\ t = 2$ **f.** $x = 0.00656,\ t = 2$

g. $x = -0.00653,\ t = 2$ **h.** $x = -0.00656,\ t = 2$

28. The opening example to this chapter described a physical experiment involving the temperature of a gas under pressure. In this application, we were given $P = 1.00$ atm, $V = 0.100$ m^3, $N = 0.00420$ mol, and $R = 0.08206$. Solving for T in the ideal gas law gives

$$T = \frac{PV}{NR} = \frac{(1.00)(0.100)}{(0.00420)(0.08206)} = 290.15 \text{ K} = 17°\text{C}.$$

In the laboratory, it was found that T was 15°C under these conditions, and when the pressure was doubled and the volume halved, T was 19°C. Assume that the data are rounded values accurate to the places given, and show that both laboratory figures are within the bounds of accuracy for the ideal gas law.

1.3 Algorithms and Convergence

Throughout the text we will be examining approximation procedures, called *algorithms*, involving sequences of calculations. An **algorithm** is a procedure that describes, in an unambiguous manner, a finite sequence of steps to be performed in a specified order. The object of the algorithm is to implement a procedure to solve a problem or approximate a solution to the problem.

The use of an algorithm is as old as formal mathematics, but the name derives from the Arabic mathematician Muhammad ibn-Mŝâ al-Khwarârizmî (c. 780–850). The Latin translation of his works begins with the words "Dixit Algorismi" meaning "al-Khwarârizmî says."

We use a **pseudocode** to describe the algorithms. This pseudocode specifies the form of the input to be supplied and the form of the desired output. Not all numerical procedures give satisfactory output for arbitrarily chosen input. As a consequence, a stopping technique independent of the numerical technique is incorporated into each algorithm to avoid infinite loops.

Two punctuation symbols are used in the algorithms:

- a period (.) indicates the termination of a step,
- a semicolon (;) separates tasks within a step.

Indentation is used to indicate that groups of statements are to be treated as a single entity.

Looping techniques in the algorithms are either counter-controlled, such as,

For $i = 1, 2, \ldots, n$

Set $x_i = a + i \cdot h$

or condition-controlled, such as

While $i < N$ do Steps 3–6.

To allow for conditional execution, we use the standard

If . . . then or If . . . then

else

constructions.

The steps in the algorithms follow the rules of structured program construction. They have been arranged so that there should be minimal difficulty translating pseudocode into any programming language suitable for scientific applications.

The algorithms are liberally laced with comments. These are written in italics and contained within parentheses to distinguish them from the algorithmic statements.

Illustration The following algorithm computes $x_1 + x_2 + \cdots + x_N = \sum_{i=1}^{N} x_i$, given N and the numbers $x_1, x_2, \ldots, x_N$.

INPUT $N, x_1, x_2, \ldots, x_n$.

OUTPUT $SUM = \sum_{i=1}^{N} x_i$.

Step 1 Set $SUM = 0$. (*Initialize accumulator.*)

Step 2 For $i = 1, 2, \ldots, N$ do
set $SUM = SUM + x_i$. (*Add the next term.*)

Step 3 OUTPUT (SUM);
STOP. □

Example 1 The Nth Taylor polynomial for $f(x) = \ln x$ expanded about $x_0 = 1$ is

$$P_N(x) = \sum_{i=1}^{N} \frac{(-1)^{i+1}}{i}(x-1)^i,$$

and the value of $\ln 1.5$ to eight decimal places is 0.40546511. Construct an algorithm to determine the minimal value of N required for

$$|\ln 1.5 - P_N(1.5)| < 10^{-5},$$

without using the Taylor polynomial remainder term.

Solution From calculus we know that if $\sum_{n=1}^{\infty} a_n$ is an alternating series with limit A whose terms decrease in magnitude, then A and the Nth partial sum $A_N = \sum_{n=1}^{N} a_n$ differ by less than the magnitude of the $(N+1)$st term; that is,

$$|A - A_N| \leq |a_{N+1}|.$$

The following algorithm uses this bound.

INPUT value x, tolerance TOL, maximum number of iterations M.
OUTPUT degree N of the polynomial or a message of failure.

Step 1 Set $N = 1$;
$y = x - 1$;
$SUM = 0$;
$POWER = y$;
$TERM = y$;
$SIGN = -1$. (*Used to implement alternation of signs.*)

Step 2 While $N \leq M$ do Steps 3–5.

Step 3 Set $SIGN = -SIGN$; (*Alternate the signs.*)
$SUM = SUM + SIGN \cdot TERM$; (*Accumulate the terms.*)
$POWER = POWER \cdot y$;
$TERM = POWER/(N+1)$. (*Calculate the next term.*)

Step 4 If $|TERM| < TOL$ then (*Test for accuracy.*)
OUTPUT (N);
STOP. (*The procedure was successful.*)

Step 5 Set $N = N + 1$. (*Prepare for the next iteration.*)

Step 6 OUTPUT ('Method Failed'); (*The procedure was unsuccessful.*)
STOP.

The input for our problem is $x = 1.5$, $TOL = 10^{-5}$, and perhaps $M = 15$. This choice of M provides an upper bound for the number of calculations we are willing to perform, recognizing that the algorithm is likely to fail if this bound is exceeded. Whether the output is a value for N or the failure message depends on the precision of the computational device. ■

Characterizing Algorithms

We will be considering a variety of approximation problems throughout the text, and in each case we need to determine approximation methods that produce dependably accurate results for a wide class of problems. Because of the differing ways in which the approximation methods are derived, we need a variety of conditions to categorize their accuracy. Not all of these conditions will be appropriate for any particular problem.

One criterion we will impose on an algorithm whenever possible is that small changes in the initial data produce correspondingly small changes in the final results. An algorithm that satisfies this property is called **stable**; otherwise it is **unstable**. Some algorithms are stable only for certain choices of initial data, and are called **conditionally stable**. We will characterize the stability properties of algorithms whenever possible.

The word *stable* has the same root as the words stand and standard. In mathematics, the term stable applied to a problem indicates that a small change in initial data or conditions does not result in a dramatic change in the solution to the problem.

To further consider the subject of round-off error growth and its connection to algorithm stability, suppose an error with magnitude $E_0 > 0$ is introduced at some stage in the calculations and that the magnitude of the error after n subsequent operations is denoted by E_n. The two cases that arise most often in practice are defined as follows.

Definition 1.17 Suppose that $E_0 > 0$ denotes an error introduced at some stage in the calculations and E_n represents the magnitude of the error after n subsequent operations.

- If $E_n \approx CnE_0$, where C is a constant independent of n, then the growth of error is said to be **linear**.
- If $E_n \approx C^nE_0$, for some $C > 1$, then the growth of error is called **exponential**. ■

Linear growth of error is usually unavoidable, and when C and E_0 are small the results are generally acceptable. Exponential growth of error should be avoided, because the term C^n becomes large for even relatively small values of n. This leads to unacceptable inaccuracies, regardless of the size of E_0. As a consequence, an algorithm that exhibits linear growth of error is stable, whereas an algorithm exhibiting exponential error growth is unstable. (See Figure 1.12.)

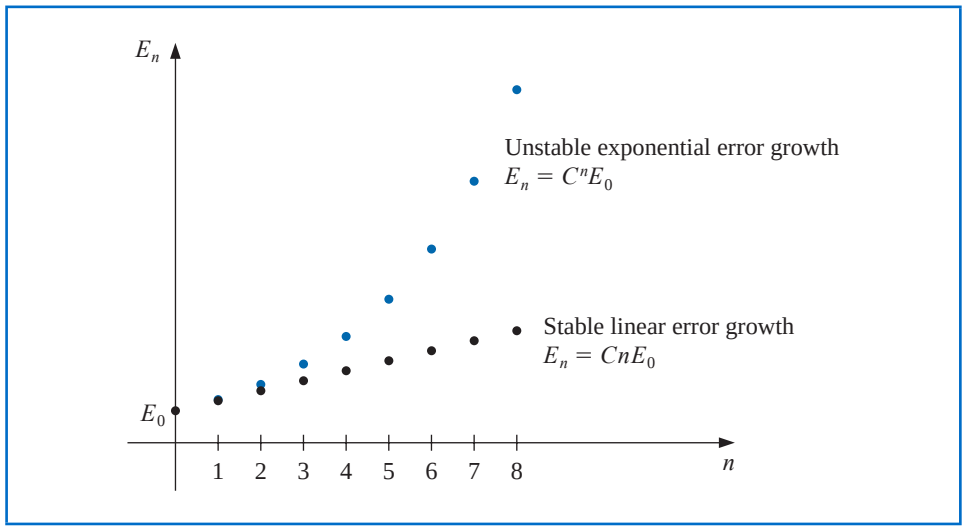
Illustration For any constants c_1 and c_2,

$$p_n = c_1\left(\frac{1}{3}\right)^n + c_2 3^n,$$

is a solution to the recursive equation

$$p_n = \frac{10}{3}p_{n-1} - p_{n-2}, \quad \text{for } n = 2, 3, \ldots.$$

Figure 1.12



This can be seen by noting that

$$\begin{aligned}\frac{10}{3}p_{n-1} - p_{n-2} &= \frac{10}{3}\left[c_1\left(\frac{1}{3}\right)^{n-1} + c_2 3^{n-1}\right] - \left[c_1\left(\frac{1}{3}\right)^{n-2} + c_2 3^{n-2}\right] \\ &= c_1\left(\frac{1}{3}\right)^{n-2}\left[\frac{10}{3}\cdot\frac{1}{3} - 1\right] + c_2 3^{n-2}\left[\frac{10}{3}\cdot 3 - 1\right] \\ &= c_1\left(\frac{1}{3}\right)^{n-2}\left(\frac{1}{9}\right) + c_2 3^{n-2}(9) = c_1\left(\frac{1}{3}\right)^{n} + c_2 3^{n} = p_n.\end{aligned}$$

Suppose that we are given $p_0 = 1$ and $p_1 = \frac{1}{3}$. This determines unique values for the constants as $c_1 = 1$ and $c_2 = 0$. So $p_n = \left(\frac{1}{3}\right)^n$ for all n.

If five-digit rounding arithmetic is used to compute the terms of the sequence given by this equation, then $\hat{p}_0 = 1.0000$ and $\hat{p}_1 = 0.33333$, which requires modifying the constants to $\hat{c}_1 = 1.0000$ and $\hat{c}_2 = -0.12500 \times 10^{-5}$. The sequence $\{\hat{p}_n\}_{n=0}^{\infty}$ generated is then given by

$$\hat{p}_n = 1.0000\left(\frac{1}{3}\right)^n - 0.12500 \times 10^{-5}(3)^n,$$

which has round-off error,

$$p_n - \hat{p}_n = 0.12500 \times 10^{-5}(3^n),$$

This procedure is unstable because the error grows *exponentially* with n, which is reflected in the extreme inaccuracies after the first few terms, as shown in Table 1.5 on page 36.

Now consider this recursive equation:

$$p_n = 2p_{n-1} - p_{n-2}, \quad \text{for } n = 2, 3, \ldots.$$

It has the solution $p_n = c_1 + c_2 n$ for any constants c_1 and c_2, because

$$\begin{aligned}2p_{n-1} - p_{n-2} &= 2(c_1 + c_2(n-1)) - (c_1 + c_2(n-2)) \\ &= c_1(2-1) + c_2(2n - 2 - n + 2) = c_1 + c_2 n = p_n.\end{aligned}$$

Table 1.5

n	Computed $\hat{p}_n$	Correct p_n	Relative Error
0	0.10000×10^{1}	0.10000×10^{1}	
1	0.33333×10^{0}	0.33333×10^{0}	
2	0.11110×10^{0}	0.11111×10^{0}	9×10^{-5}
3	0.37000×10^{-1}	0.37037×10^{-1}	1×10^{-3}
4	0.12230×10^{-1}	0.12346×10^{-1}	9×10^{-3}
5	0.37660×10^{-2}	0.41152×10^{-2}	8×10^{-2}
6	0.32300×10^{-3}	0.13717×10^{-2}	8×10^{-1}
7	-0.26893×10^{-2}	0.45725×10^{-3}	7×10^{0}
8	-0.92872×10^{-2}	0.15242×10^{-3}	6×10^{1}

If we are given $p_0 = 1$ and $p_1 = \frac{1}{3}$, then constants in this equation are uniquely determined to be $c_1 = 1$ and $c_2 = -\frac{2}{3}$. This implies that $p_n = 1 - \frac{2}{3}n$.

If five-digit rounding arithmetic is used to compute the terms of the sequence given by this equation, then $\hat{p}_0 = 1.0000$ and $\hat{p}_1 = 0.33333$. As a consequence, the five-digit rounding constants are $\hat{c}_1 = 1.0000$ and $\hat{c}_2 = -0.66667$. Thus

$$\hat{p}_n = 1.0000 - 0.66667n,$$

which has round-off error

$$p_n - \hat{p}_n = \left(0.66667 - \frac{2}{3}\right) n.$$

This procedure is stable because the error grows grows *linearly* with n, which is reflected in the approximations shown in Table 1.6. □

Table 1.6

n	Computed $\hat{p}_n$	Correct p_n	Relative Error
0	0.10000×10^{1}	0.10000×10^{1}	
1	0.33333×10^{0}	0.33333×10^{0}	
2	-0.33330×10^{0}	-0.33333×10^{0}	9×10^{-5}
3	-0.10000×10^{1}	-0.10000×10^{1}	0
4	-0.16667×10^{1}	-0.16667×10^{1}	0
5	-0.23334×10^{1}	-0.23333×10^{1}	4×10^{-5}
6	-0.30000×10^{1}	-0.30000×10^{1}	0
7	-0.36667×10^{1}	-0.36667×10^{1}	0
8	-0.43334×10^{1}	-0.43333×10^{1}	2×10^{-5}

The effects of round-off error can be reduced by using high-order-digit arithmetic such as the double- or multiple-precision option available on most computers. Disadvantages in using double-precision arithmetic are that it takes more computation time and the growth of round-off error is not entirely eliminated.

One approach to estimating round-off error is to use interval arithmetic (that is, to retain the largest and smallest possible values at each step), so that, in the end, we obtain

an interval that contains the true value. Unfortunately, a very small interval may be needed for reasonable implementation.

Rates of Convergence

Since iterative techniques involving sequences are often used, this section concludes with a brief discussion of some terminology used to describe the rate at which convergence occurs. In general, we would like the technique to converge as rapidly as possible. The following definition is used to compare the convergence rates of sequences.

Definition 1.18 Suppose $\{\beta_n\}_{n=1}^{\infty}$ is a sequence known to converge to zero, and $\{\alpha_n\}_{n=1}^{\infty}$ converges to a number α. If a positive constant K exists with

$$|\alpha_n - \alpha| \le K|\beta_n|, \quad \text{for large } n,$$

then we say that $\{\alpha_n\}_{n=1}^{\infty}$ converges to α with **rate, or order, of convergence** $O(\beta_n)$. (This expression is read "big oh of β_n".) It is indicated by writing $\alpha_n = \alpha + O(\beta_n)$. ■

Although Definition 1.18 permits $\{\alpha_n\}_{n=1}^{\infty}$ to be compared with an arbitrary sequence $\{\beta_n\}_{n=1}^{\infty}$, in nearly every situation we use

$$\beta_n = \frac{1}{n^p},$$

for some number $p > 0$. We are generally interested in the largest value of p with $\alpha_n = \alpha + O(1/n^p)$.

Example 2 Suppose that, for $n \ge 1$,

$$\alpha_n = \frac{n+1}{n^2} \quad \text{and} \quad \hat{\alpha}_n = \frac{n+3}{n^3}.$$

Both $\lim_{n\to\infty} \alpha_n = 0$ and $\lim_{n\to\infty} \hat{\alpha}_n = 0$, but the sequence $\{\hat{\alpha}_n\}$ converges to this limit much faster than the sequence $\{\alpha_n\}$. Using five-digit rounding arithmetic we have the values shown in Table 1.7. Determine rates of convergence for these two sequences.

Table 1.7

n	1	2	3	4	5	6	7
α_n	2.00000	0.75000	0.44444	0.31250	0.24000	0.19444	0.16327
$\hat{\alpha}_n$	4.00000	0.62500	0.22222	0.10938	0.064000	0.041667	0.029155

There are numerous other ways of describing the growth of sequences and functions, some of which require bounds both above and below the sequence or function under consideration. Any good book that analyzes algorithms, for example [CLRS], will include this information.

Solution Define the sequences $\beta_n = 1/n$ and $\hat{\beta}_n = 1/n^2$. Then

$$|\alpha_n - 0| = \frac{n+1}{n^2} \le \frac{n+n}{n^2} = 2 \cdot \frac{1}{n} = 2\beta_n$$

and

$$|\hat{\alpha}_n - 0| = \frac{n+3}{n^3} \le \frac{n+3n}{n^3} = 4 \cdot \frac{1}{n^2} = 4\hat{\beta}_n.$$

Hence the rate of convergence of $\{\alpha_n\}$ to zero is similar to the convergence of $\{1/n\}$ to zero, whereas $\{\hat{\alpha}_n\}$ converges to zero at a rate similar to the more rapidly convergent sequence $\{1/n^2\}$. We express this by writing

$$\alpha_n = 0 + O\left(\frac{1}{n}\right) \quad \text{and} \quad \hat{\alpha}_n = 0 + O\left(\frac{1}{n^2}\right). \qquad \blacksquare$$

We also use the O (*big oh*) notation to describe the rate at which functions converge.

Definition 1.19 Suppose that $\lim_{h\to 0} G(h) = 0$ and $\lim_{h\to 0} F(h) = L$. If a positive constant K exists with

$$|F(h) - L| \le K|G(h)|, \quad \text{for sufficiently small } h,$$

then we write $F(h) = L + O(G(h))$. $\blacksquare$

The functions we use for comparison generally have the form $G(h) = h^p$, where $p > 0$. We are interested in the largest value of p for which $F(h) = L + O(h^p)$.

Example 3 Use the third Taylor polynomial about $h = 0$ to show that $\cos h + \frac{1}{2}h^2 = 1 + O(h^4)$.

Solution In Example 3(b) of Section 1.1 we found that this polynomial is

$$\cos h = 1 - \frac{1}{2}h^2 + \frac{1}{24}h^4 \cos\tilde{\xi}(h),$$

for some number $\tilde{\xi}(h)$ between zero and h. This implies that

$$\cos h + \frac{1}{2}h^2 = 1 + \frac{1}{24}h^4 \cos\tilde{\xi}(h).$$

Hence

$$\left|\left(\cos h + \frac{1}{2}h^2\right) - 1\right| = \left|\frac{1}{24}\cos\tilde{\xi}(h)\right| h^4 \le \frac{1}{24}h^4,$$

so as $h \to 0$, $\cos h + \frac{1}{2}h^2$ converges to its limit, 1, about as fast as h^4 converges to 0. That is,

$$\cos h + \frac{1}{2}h^2 = 1 + O(h^4). \qquad \blacksquare$$

Maple uses the O notation to indicate the form of the error in Taylor polynomials and in other situations. For example, at the end of Section 1.1 the third Taylor polynomial for $f(x) = \cos(x)$ was found by first defining

$f := \cos(x)$

and then calling the third Taylor polynomial with

$taylor(f, x = 0, 4)$

Maple responds with

$$1 - \frac{1}{2}x^2 + O(x^4)$$

to indicate that the lowest term in the truncation error is x^4.

EXERCISE SET 1.3

1. **a.** Use three-digit chopping arithmetic to compute the sum $\sum_{i=1}^{10}(1/i^2)$ first by $\frac{1}{1}+\frac{1}{4}+\cdots+\frac{1}{100}$ and then by $\frac{1}{100}+\frac{1}{81}+\cdots+\frac{1}{1}$. Which method is more accurate, and why?

b. Write an algorithm to sum the finite series $\sum_{i=1}^{N} x_i$ in reverse order.

2. The number e is defined by $e=\sum_{n=0}^{\infty}(1/n!)$, where $n!=n(n-1)\cdots 2\cdot 1$ for $n\neq 0$ and $0!=1$. Use four-digit chopping arithmetic to compute the following approximations to e, and determine the absolute and relative errors.

a. $e\approx\sum_{n=0}^{5}\frac{1}{n!}$ **b.** $e\approx\sum_{j=0}^{5}\frac{1}{(5-j)!}$

c. $e\approx\sum_{n=0}^{10}\frac{1}{n!}$ **d.** $e\approx\sum_{j=0}^{10}\frac{1}{(10-j)!}$

3. The Maclaurin series for the arctangent function converges for $-1<x\leq 1$ and is given by

$$\arctan x=\lim_{n\to\infty}P_n(x)=\lim_{n\to\infty}\sum_{i=1}^{n}(-1)^{i+1}\frac{x^{2i-1}}{2i-1}.$$

a. Use the fact that $\tan\pi/4=1$ to determine the number of n terms of the series that need to be summed to ensure that $|4P_n(1)-\pi|<10^{-3}$.

b. The C++ programming language requires the value of π to be within 10^{-10}. How many terms of the series would we need to sum to obtain this degree of accuracy?

4. Exercise 3 details a rather inefficient means of obtaining an approximation to π. The method can be improved substantially by observing that $\pi/4=\arctan\frac{1}{2}+\arctan\frac{1}{3}$ and evaluating the series for the arctangent at $\frac{1}{2}$ and at $\frac{1}{3}$. Determine the number of terms that must be summed to ensure an approximation to π to within 10^{-3}.

5. Another formula for computing π can be deduced from the identity $\pi/4=4\arctan\frac{1}{5}-\arctan\frac{1}{239}$. Determine the number of terms that must be summed to ensure an approximation to π to within 10^{-3}.

6. Find the rates of convergence of the following sequences as $n\to\infty$.

a. $\lim_{n\to\infty}\sin\frac{1}{n}=0$ **b.** $\lim_{n\to\infty}\sin\frac{1}{n^2}=0$

c. $\lim_{n\to\infty}\left(\sin\frac{1}{n}\right)^2=0$ **d.** $\lim_{n\to\infty}[\ln(n+1)-\ln(n)]=0$

7. Find the rates of convergence of the following functions as $h\to 0$.

a. $\lim_{h\to 0}\frac{\sin h}{h}=1$ **b.** $\lim_{h\to 0}\frac{1-\cos h}{h}=0$

c. $\lim_{h\to 0}\frac{\sin h-h\cos h}{h}=0$ **d.** $\lim_{h\to 0}\frac{1-e^h}{h}=-1$

8. **a.** How many multiplications and additions are required to determine a sum of the form

$$\sum_{i=1}^{n}\sum_{j=1}^{i}a_ib_j?$$

b. Modify the sum in part (a) to an equivalent form that reduces the number of computations.

9. Let $P(x)=a_nx^n+a_{n-1}x^{n-1}+\cdots+a_1x+a_0$ be a polynomial, and let x_0 be given. Construct an algorithm to evaluate $P(x_0)$ using nested multiplication.

10. Equations (1.2) and (1.3) in Section 1.2 give alternative formulas for the roots x_1 and x_2 of $ax^2+bx+c=0$. Construct an algorithm with input a,b,c and output x_1, x_2 that computes the roots x_1 and x_2 (which may be equal or be complex conjugates) using the best formula for each root.

11. Construct an algorithm that has as input an integer $n\geq 1$, numbers $x_0,x_1,\ldots,x_n$, and a number x and that produces as output the product $(x-x_0)(x-x_1)\cdots(x-x_n)$.

12. Assume that

$$\frac{1-2x}{1-x+x^2}+\frac{2x-4x^3}{1-x^2+x^4}+\frac{4x^3-8x^7}{1-x^4+x^8}+\cdots=\frac{1+2x}{1+x+x^2},$$

for $x < 1$, and let $x = 0.25$. Write and execute an algorithm that determines the number of terms needed on the left side of the equation so that the left side differs from the right side by less than 10^{-6}.

13. **a.** Suppose that $0 < q < p$ and that $\alpha_n = \alpha + O\left(n^{-p}\right)$. Show that $\alpha_n = \alpha + O\left(n^{-q}\right)$.

b. Make a table listing $1/n$, $1/n^2$, $1/n^3$, and $1/n^4$ for $n = 5, 10, 100$, and 1000, and discuss the varying rates of convergence of these sequences as n becomes large.

14. **a.** Suppose that $0 < q < p$ and that $F(h) = L + O(h^p)$. Show that $F(h) = L + O(h^q)$.

b. Make a table listing h, h^2, h^3, and h^4 for $h = 0.5, 0.1, 0.01$, and 0.001, and discuss the varying rates of convergence of these powers of h as h approaches zero.

15. Suppose that as x approaches zero,

$$F_1(x) = L_1 + O(x^\alpha) \quad \text{and} \quad F_2(x) = L_2 + O(x^\beta).$$

Let c_1 and c_2 be nonzero constants, and define

$$F(x) = c_1F_1(x) + c_2F_2(x) \quad \text{and}$$
$$G(x) = F_1(c_1x) + F_2(c_2x).$$

Show that if γ = minimum $\{\alpha, \beta\}$, then as x approaches zero,

a. $F(x) = c_1L_1 + c_2L_2 + O(x^\gamma)$

b. $G(x) = L_1 + L_2 + O(x^\gamma)$.

16. The sequence $\{F_n\}$ described by $F_0 = 1, F_1 = 1$, and $F_{n+2} = F_n + F_{n+1}$, if $n \geq 0$, is called a *Fibonacci sequence*. Its terms occur naturally in many botanical species, particularly those with petals or scales arranged in the form of a logarithmic spiral. Consider the sequence $\{x_n\}$, where $x_n = F_{n+1}/F_n$. Assuming that $\lim_{n\to\infty} x_n = x$ exists, show that $x = (1 + \sqrt{5})/2$. This number is called the *golden ratio.*

17. The Fibonacci sequence also satisfies the equation

$$F_n \equiv \tilde{F}_n = \frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^n - \left(\frac{1-\sqrt{5}}{2}\right)^n\right].$$

a. Write a Maple procedure to calculate F_{100}.

b. Use Maple with the default value of *Digits* followed by *evalf* to calculate $\tilde{F}_{100}$.

c. Why is the result from part (a) more accurate than the result from part (b)?

d. Why is the result from part (b) obtained more rapidly than the result from part (a)?

e. What results when you use the command *simplify* instead of *evalf* to compute $\tilde{F}_{100}$?

18. The harmonic series $1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots$ diverges, but the sequence $\gamma_n = 1 + \frac{1}{2} + \cdots + \frac{1}{n} - \ln n$ converges, since $\{\gamma_n\}$ is a bounded, nonincreasing sequence. The limit $\gamma = 0.5772156649\ldots$ of the sequence $\{\gamma_n\}$ is called Euler's constant.

a. Use the default value of *Digits* in Maple to determine the value of n for γ_n to be within 10^{-2} of γ.

b. Use the default value of *Digits* in Maple to determine the value of n for γ_n to be within 10^{-3} of γ.

c. What happens if you use the default value of *Digits* in Maple to determine the value of n for γ_n to be within 10^{-4} of γ?

1.4 Numerical Software

Computer software packages for approximating the numerical solutions to problems are available in many forms. On our web site for the book

```
http://www.math.ysu.edu/~faires/Numerical-Analysis/Programs.html
```

we have provided programs written in C, FORTRAN, Maple, Mathematica, MATLAB, and Pascal, as well as JAVA applets. These can be used to solve the problems given in the examples and exercises, and will give satisfactory results for most problems that you may need to solve. However, they are what we call *special-purpose* programs. We use this term to distinguish these programs from those available in the standard mathematical subroutine libraries. The programs in these packages will be called *general purpose*.

The programs in general-purpose software packages differ in their intent from the algorithms and programs provided with this book. General-purpose software packages consider ways to reduce errors due to machine rounding, underflow, and overflow. They also describe the range of input that will lead to results of a certain specified accuracy. These are machine-dependent characteristics, so general-purpose software packages use parameters that describe the floating-point characteristics of the machine being used for computations.

Illustration To illustrate some differences between programs included in a general-purpose package and a program that we would provide for use in this book, let us consider an algorithm that computes the Euclidean norm of an n-dimensional vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)^t$. This norm is often required within larger programs and is defined by

$$||\mathbf{x}||_2 = \left[\sum_{i=1}^{n} x_i^2\right]^{1/2}.$$

The norm gives a measure for the distance from the vector $\mathbf{x}$ to the vector $\mathbf{0}$. For example, the vector $\mathbf{x} = (2, 1, 3, -2, -1)^t$ has

$$||\mathbf{x}||_2 = [2^2 + 1^2 + 3^2 + (-2)^2 + (-1)^2]^{1/2} = \sqrt{19},$$

so its distance from $\mathbf{0} = (0, 0, 0, 0, 0)^t$ is $\sqrt{19} \approx 4.36$.

An algorithm of the type we would present for this problem is given here. It includes no machine-dependent parameters and provides no accuracy assurances, but it will give accurate results "most of the time."

INPUT $n, x_1, x_2, \ldots, x_n$.

OUTPUT $NORM$.

Step 1 Set $SUM = 0$.

Step 2 For $i = 1, 2, \ldots, n$ set $SUM = SUM + x_i^2$.

Step 3 Set $NORM = SUM^{1/2}$.

Step 4 OUTPUT ($NORM$);
STOP. □

A program based on our algorithm is easy to write and understand. However, the program could fail to give sufficient accuracy for a number of reasons. For example, the magnitude of some of the numbers might be too large or too small to be accurately represented in

the floating-point system of the computer. Also, this order for performing the calculations might not produce the most accurate results, or the standard software square-root routine might not be the best available for the problem. Matters of this type are considered by algorithm designers when writing programs for general-purpose software. These programs are often used as subprograms for solving larger problems, so they must incorporate controls that we will not need.

General Purpose Algorithms

Let us now consider an algorithm for a general-purpose software program for computing the Euclidean norm. First, it is possible that although a component x_i of the vector is within the range of the machine, the square of the component is not. This can occur when some $|x_i|$ is so small that x_i^2 causes underflow or when some $|x_i|$ is so large that x_i^2 causes overflow. It is also possible for all these terms to be within the range of the machine, but overflow occurs from the addition of a square of one of the terms to the previously computed sum.

Accuracy criteria depend on the machine on which the calculations are being performed, so machine-dependent parameters are incorporated into the algorithm. Suppose we are working on a hypothetical computer with base 10, having $t \geq 4$ digits of precision, a minimum exponent *emin*, and a maximum exponent *emax*. Then the set of floating-point numbers in this machine consists of 0 and the numbers of the form

$$x = f \cdot 10^e, \quad \text{where} \quad f = \pm(f_1 10^{-1} + f_2 10^{-2} + \cdots + f_t 10^{-t}),$$

where $1 \leq f_1 \leq 9$ and $0 \leq f_i \leq 9$, for each $i = 2, \ldots, t$, and where $emin \leq e \leq emax$. These constraints imply that the smallest positive number represented in the machine is $\sigma = 10^{emin-1}$, so any computed number x with $|x| < \sigma$ causes underflow and results in x being set to 0. The largest positive number is $\lambda = (1 - 10^{-t})10^{emax}$, and any computed number x with $|x| > \lambda$ causes overflow. When underflow occurs, the program will continue, often without a significant loss of accuracy. If overflow occurs, the program will fail.

The algorithm assumes that the floating-point characteristics of the machine are described using parameters N, s, S, y, and Y. The maximum number of entries that can be summed with at least $t/2$ digits of accuracy is given by N. This implies the algorithm will proceed to find the norm of a vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)^t$ only if $n \leq N$. To resolve the underflow-overflow problem, the nonzero floating-point numbers are partitioned into three groups:

- small-magnitude numbers x, those satisfying $0 < |x| < y$;
- medium-magnitude numbers x, where $y \leq |x| < Y$;
- large-magnitude numbers x, where $Y \leq |x|$.

The parameters y and Y are chosen so that there will be no underflow-overflow problem in squaring and summing the medium-magnitude numbers. Squaring small-magnitude numbers can cause underflow, so a scale factor S much greater than 1 is used with the result that $(Sx)^2$ avoids the underflow even when x^2 does not. Summing and squaring numbers having a large magnitude can cause overflow. So in this case, a positive scale factor s much smaller than 1 is used to ensure that $(sx)^2$ does not cause overflow when calculated or incorporated into a sum, even though x^2 would.

To avoid unnecessary scaling, y and Y are chosen so that the range of medium-magnitude numbers is as large as possible. The algorithm that follows is a modification of one described in [Brow, W], p. 471. It incorporates a procedure for adding scaled components of the vector that are small in magnitude until a component with medium magnitude

is encountered. It then unscales the previous sum and continues by squaring and summing small and medium numbers until a component with a large magnitude is encountered. Once a component with large magnitude appears, the algorithm scales the previous sum and proceeds to scale, square, and sum the remaining numbers.

The algorithm assumes that, in transition from small to medium numbers, unscaled small numbers are negligible when compared to medium numbers. Similarly, in transition from medium to large numbers, unscaled medium numbers are negligible when compared to large numbers. Thus, the choices of the scaling parameters must be made so that numbers are equated to 0 only when they are truly negligible. Typical relationships between the machine characteristics as described by t, σ, λ, *emin*, *emax*, and the algorithm parameters N, s, S, y, and Y are given after the algorithm.

The algorithm uses three flags to indicate the various stages in the summation process. These flags are given initial values in Step 3 of the algorithm. FLAG 1 is 1 until a medium or large component is encountered; then it is changed to 0. FLAG 2 is 0 while small numbers are being summed, changes to 1 when a medium number is first encountered, and changes back to 0 when a large number is found. FLAG 3 is initially 0 and changes to 1 when a large number is first encountered. Step 3 also introduces the flag DONE, which is 0 until the calculations are complete, and then changes to 1.

INPUT $N, s, S, y, Y, \lambda, n, x_1, x_2, \ldots, x_n$.

OUTPUT *NORM* or an appropriate error message.

Step 1 If $n \leq 0$ then OUTPUT ('The integer n must be positive.');
STOP.

Step 2 If $n \geq N$ then OUTPUT ('The integer n is too large.');
STOP.

Step 3 Set $SUM = 0$;
$FLAG1 = 1$; (*The small numbers are being summed.*)
$FLAG2 = 0$;
$FLAG3 = 0$;
$DONE = 0$;
$i = 1$.

Step 4 While ($i \leq n$ and $FLAG1 = 1$) do Step 5.

Step 5 If $|x_i| < y$ then set $SUM = SUM + (Sx_i)^2$;
$i = i + 1$
else set $FLAG1 = 0$. (*A non-small number encountered.*)

Step 6 If $i > n$ then set $NORM = (SUM)^{1/2}/S$;
$DONE = 1$
else set $SUM = (SUM/S)/S$; (*Scale for larger numbers.*)
$FLAG2 = 1$.

Step 7 While ($i \leq n$ and $FLAG2 = 1$) do Step 8. (*Sum the medium-sized numbers.*)

Step 8 If $|x_i| < Y$ then set $SUM = SUM + x_i^2$;
$i = i + 1$
else set $FLAG2 = 0$. (*A large number has been encountered.*)

Step 9 If $DONE = 0$ then
if $i > n$ then set $NORM = (SUM)^{1/2}$;
$DONE = 1$
else set $SUM = ((SUM)s)s$; (*Scale the large numbers.*)
$FLAG3 = 1$.

Step 10 While ($i \leq n$ and $FLAG3 = 1$) do Step 11.

Step 11 Set $SUM = SUM + (sx_i)^2$; (*Sum the large numbers.*)
$i = i + 1$.

Step 12 If $DONE = 0$ then
if $SUM^{1/2} < \lambda s$ then set $NORM = (SUM)^{1/2}/s$;
$DONE = 1$
else set $SUM = \lambda$. (*The norm is too large.*)

Step 13 If $DONE = 1$ then OUTPUT ('Norm is', *NORM*)
else OUTPUT ('Norm $\geq$', *NORM*, 'overflow occurred').

Step 14 STOP.

The relationships between the machine characteristics t, σ, λ, *emin*, *emax*, and the algorithm parameters N, s, S, y, and Y were chosen in [Brow, W], p. 471, as:

$N = 10^{e_N}$, where $e_N = \lfloor (t-2)/2 \rfloor$, the greatest integer less than or equal to $(t-2)/2$;

$s = 10^{e_s}$, where $e_s = \lfloor -(emax + e_N)/2 \rfloor$;

$S = 10^{e_S}$, where $e_S = \lceil (1 - emin)/2 \rceil$, the smallest integer greater than or equal to $(1 - emin)/2$;

$y = 10^{e_y}$, where $e_y = \lceil (emin + t - 2)/2 \rceil$;

$Y = 10^{e_Y}$, where $e_Y = \lfloor (emax - e_N)/2 \rfloor$.

The first portable computer was the Osborne I, produced in 1981, although it was much larger and heaver than we would currently think of as portable.

The system FORTRAN (FORmula TRANslator) was the original general-purpose scientific programming language. It is still in wide use in situations that require intensive scientific computations.

The EISPACK project was the first large-scale numerical software package to be made available in the public domain and led the way for many packages to follow.

The reliability built into this algorithm has greatly increased the complexity compared to the algorithm given earlier in the section. In the majority of cases the special-purpose and general-purpose algorithms give identical results. The advantage of the general-purpose algorithm is that it provides security for its results.

Many forms of general-purpose numerical software are available commercially and in the public domain. Most of the early software was written for mainframe computers, and a good reference for this is *Sources and Development of Mathematical Software*, edited by Wayne Cowell [Co].

Now that personal computers are sufficiently powerful, standard numerical software is available for them. Most of this numerical software is written in FORTRAN, although some packages are written in C, C++, and FORTRAN90.

ALGOL procedures were presented for matrix computations in 1971 in [WR]. A package of FORTRAN subroutines based mainly on the ALGOL procedures was then developed into the EISPACK routines. These routines are documented in the manuals published by Springer-Verlag as part of their Lecture Notes in Computer Science series [Sm,B] and [Gar]. The FORTRAN subroutines are used to compute eigenvalues and eigenvectors for a variety of different types of matrices.

LINPACK is a package of FORTRAN subroutines for analyzing and solving systems of linear equations and solving linear least squares problems. The documentation for this package is contained in [DBMS]. A step-by-step introduction to LINPACK, EISPACK, and BLAS (Basic Linear Algebra Subprograms) is given in [CV].

The LAPACK package, first available in 1992, is a library of FORTRAN subroutines that supercedes LINPACK and EISPACK by integrating these two sets of algorithms into a unified and updated package. The software has been restructured to achieve greater efficiency on vector processors and other high-performance or shared-memory multiprocessors. LAPACK is expanded in depth and breadth in version 3.0, which is available in FORTRAN, FORTRAN90, C, C++, and JAVA. C, and JAVA are only available as language interfaces

or translations of the FORTRAN libraries of LAPACK. The package BLAS is not a part of LAPACK, but the code for BLAS is distributed with LAPACK.

Other packages for solving specific types of problems are available in the public domain. As an alternative to netlib, you can use Xnetlib to search the database and retrieve software. More information can be found in the article *Software Distribution using Netlib* by Dongarra, Roman, and Wade [DRW].

Software engineering was established as a laboratory discipline during the 1970s and 1980s. EISPACK was developed at Argonne Labs and LINPACK there shortly thereafter. By the early 1980s, Argonne was internationally recognized as a world leader in symbolic and numerical computation.

These software packages are highly efficient, accurate, and reliable. They are thoroughly tested, and documentation is readily available. Although the packages are portable, it is a good idea to investigate the machine dependence and read the documentation thoroughly. The programs test for almost all special contingencies that might result in error and failures. At the end of each chapter we will discuss some of the appropriate general-purpose packages.

Commercially available packages also represent the state of the art in numerical methods. Their contents are often based on the public-domain packages but include methods in libraries for almost every type of problem.

In 1970 IMSL became the first large-scale scientific library for mainframes. Since that time, the libraries have been made available for computer systems ranging from supercomputers to personal computers.

IMSL (International Mathematical and Statistical Libraries) consists of the libraries MATH, STAT, and SFUN for numerical mathematics, statistics, and special functions, respectively. These libraries contain more than 900 subroutines originally available in FORTRAN 77 and now available in C, FORTRAN90, and JAVA. These subroutines solve the most common numerical analysis problems. The libraries are available commercially from Visual Numerics.

The packages are delivered in compiled form with extensive documentation. There is an example program for each routine as well as background reference information. IMSL contains methods for linear systems, eigensystem analysis, interpolation and approximation, integration and differentiation, differential equations, transforms, nonlinear equations, optimization, and basic matrix/vector operations. The library also contains extensive statistical routines.

The Numerical Algorithms Group (NAG) was instituted in the UK in 1971 and developed the first mathematical software library. It now has over 10,000 users world-wide and contains over 1000 mathematical and statistical functions ranging from statistical, symbolic, visualisation, and numerical simulation software, to compilers and application development tools.

The Numerical Algorithms Group (NAG) has been in existence in the United Kingdom since 1970. NAG offers more than 1000 subroutines in a FORTRAN 77 library, about 400 subroutines in a C library, more than 200 subroutines in a FORTRAN 90 library, and an MPI FORTRAN numerical library for parallel machines and clusters of workstations or personal computers. A useful introduction to the NAG routines is [Ph]. The NAG library contains routines to perform most standard numerical analysis tasks in a manner similar to those in the IMSL. It also includes some statistical routines and a set of graphic routines.

The IMSL and NAG packages are designed for the mathematician, scientist, or engineer who wishes to call high-quality C, Java, or FORTRAN subroutines from within a program. The documentation available with the commercial packages illustrates the typical driver program required to use the library routines. The next three software packages are stand-alone environments. When activated, the user enters commands to cause the package to solve a problem. However, each package allows programming within the command language.

MATLAB was originally written to provide easy access to matrix software developed in the LINPACK and EISPACK projects. The first version was written in the late 1970s for use in courses in matrix theory, linear algebra, and numerical analysis. There are currently more than 500,000 users of MATLAB in more than 100 countries.

MATLAB is a matrix laboratory that was originally a Fortran program published by Cleve Moler [Mo] in the 1980s. The laboratory is based mainly on the EISPACK and LINPACK subroutines, although functions such as nonlinear systems, numerical integration, cubic splines, curve fitting, optimization, ordinary differential equations, and graphical tools have been incorporated. MATLAB is currently written in C and assembler, and the PC version of this package requires a numeric coprocessor. The basic structure is to perform matrix operations, such as finding the eigenvalues of a matrix entered from the command line or from an external file via function calls. This is a powerful self-contained system that is especially useful for instruction in an applied linear algebra course.

The second package is GAUSS, a mathematical and statistical system produced by Lee E. Ediefson and Samuel D. Jones in 1985. It is coded mainly in assembler and based primarily

on EISPACK and LINPACK. As in the case of MATLAB, integration/differentiation, nonlinear systems, fast Fourier transforms, and graphics are available. GAUSS is oriented less toward instruction in linear algebra and more toward statistical analysis of data. This package also uses a numeric coprocessor if one is available.

The third package is Maple, a computer algebra system developed in 1980 by the Symbolic Computational Group at the University of Waterloo. The design for the original Maple system is presented in the paper by B.W. Char, K.O. Geddes, W.M. Gentlemen, and G.H. Gonnet [CGGG].

The NAG routines are compatible with Maple beginning with version 9.0.

Maple, which is written in C, has the ability to manipulate information in a symbolic manner. This symbolic manipulation allows the user to obtain exact answers instead of numerical values. Maple can give exact answers to mathematical problems such as integrals, differential equations, and linear systems. It contains a programming structure and permits text, as well as commands, to be saved in its worksheet files. These worksheets can then be loaded into Maple and the commands executed. Because of the properties of symbolic computation, numerical computation, and worksheets, Maple is the language of choice for this text. Throughout the book Maple commands, particularly from the *NumericalAnalysis* package, will be included in the text.

Although we have chosen Maple as our standard computer algebra system, the equally popular Mathematica, released in 1988, can also be used for this purpose.

Numerous packages are available that can be classified as supercalculator packages for the PC. These should not be confused, however, with the general-purpose software listed here. If you have an interest in one of these packages, you should read *Supercalculators on the PC* by B. Simon and R. M. Wilson [SW].

Additional information about software and software libraries can be found in the books by Cody and Waite [CW] and by Kockler [Ko], and in the 1995 article by Dongarra and Walker [DW]. More information about floating-point computation can be found in the book by Chaitini-Chatelin and Frayse [CF] and the article by Goldberg [Go].

Books that address the application of numerical techniques on parallel computers include those by Schendell [Sche], Phillips and Freeman [PF], Ortega [Or1], and Golub and Ortega [GO].

CHAPTER

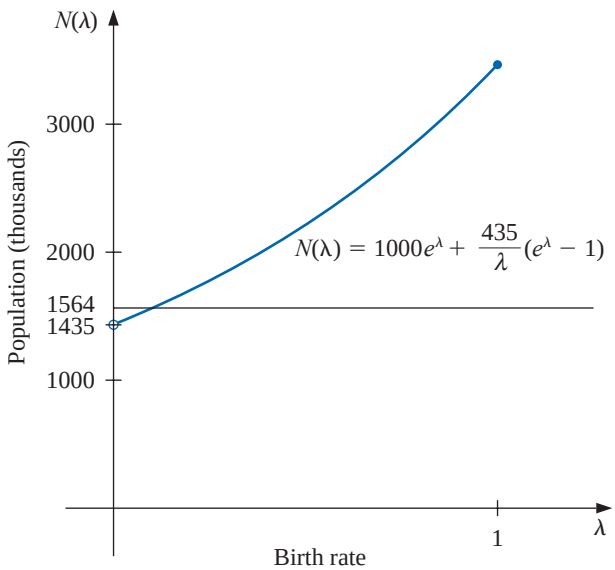
2 Solutions of Equations in One Variable

Introduction

The growth of a population can often be modeled over short periods of time by assuming that the population grows continuously with time at a rate proportional to the number present at that time. Suppose that $N(t)$ denotes the number in the population at time t and λ denotes the constant birth rate of the population. Then the population satisfies the differential equation

$$\frac{dN(t)}{dt} = \lambda N(t),$$

whose solution is $N(t) = N_0 e^{\lambda t}$, where N_0 denotes the initial population.



This exponential model is valid only when the population is isolated, with no immigration. If immigration is permitted at a constant rate v, then the differential equation becomes

$$\frac{dN(t)}{dt} = \lambda N(t) + v,$$

whose solution is

$$N(t) = N_0 e^{\lambda t} + \frac{v}{\lambda}(e^{\lambda t} - 1).$$

Suppose a certain population contains $N(0) = 1{,}000{,}000$ individuals initially, that 435,000 individuals immigrate into the community in the first year, and that $N(1) = 1{,}564{,}000$ individuals are present at the end of one year. To determine the birth rate of this population, we need to find λ in the equation

$$1{,}564{,}000 = 1{,}000{,}000e^{\lambda} + \frac{435{,}000}{\lambda}(e^{\lambda} - 1).$$

It is not possible to solve explicitly for λ in this equation, but numerical methods discussed in this chapter can be used to approximate solutions of equations of this type to an arbitrarily high accuracy. The solution to this particular problem is considered in Exercise 24 of Section 2.3.

2.1 The Bisection Method

In this chapter we consider one of the most basic problems of numerical approximation, the **root-finding problem**. This process involves finding a **root**, or solution, of an equation of the form $f(x) = 0$, for a given function f. A root of this equation is also called a **zero** of the function f.

The problem of finding an approximation to the root of an equation can be traced back at least to 1700 B.C.E. A cuneiform table in the Yale Babylonian Collection dating from that period gives a sexigesimal (base-60) number equivalent to 1.414222 as an approximation to $\sqrt{2}$, a result that is accurate to within 10^{-5}. This approximation can be found by applying a technique described in Exercise 19 of Section 2.2.

Bisection Technique

In computer science, the process of dividing a set continually in half to search for the solution to a problem, as the bisection method does, is known as a *binary search* procedure.

The first technique, based on the Intermediate Value Theorem, is called the **Bisection**, or **Binary-search, method**.

Suppose f is a continuous function defined on the interval $[a, b]$, with $f(a)$ and $f(b)$ of opposite sign. The Intermediate Value Theorem implies that a number p exists in (a, b) with $f(p) = 0$. Although the procedure will work when there is more than one root in the interval (a, b), we assume for simplicity that the root in this interval is unique. The method calls for a repeated halving (or bisecting) of subintervals of $[a, b]$ and, at each step, locating the half containing p.

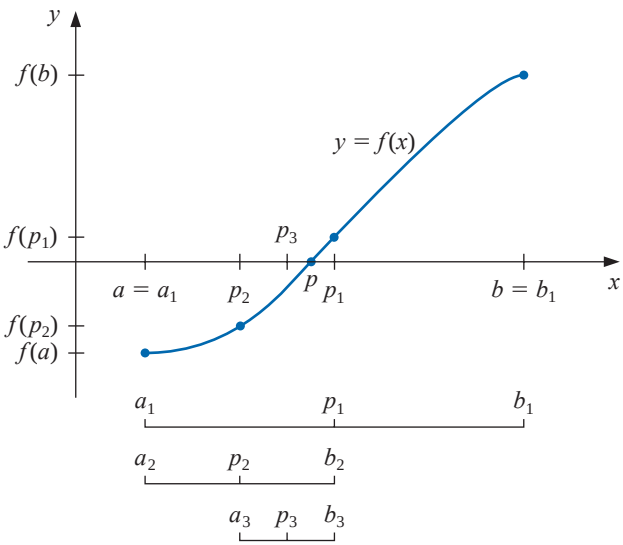
To begin, set $a_1 = a$ and $b_1 = b$, and let p_1 be the midpoint of $[a, b]$; that is,

$$p_1 = a_1 + \frac{b_1 - a_1}{2} = \frac{a_1 + b_1}{2}.$$

- If $f(p_1) = 0$, then $p = p_1$, and we are done.
- If $f(p_1) \neq 0$, then $f(p_1)$ has the same sign as either $f(a_1)$ or $f(b_1)$.
  - If $f(p_1)$ and $f(a_1)$ have the same sign, $p \in (p_1, b_1)$. Set $a_2 = p_1$ and $b_2 = b_1$.
  - If $f(p_1)$ and $f(a_1)$ have opposite signs, $p \in (a_1, p_1)$. Set $a_2 = a_1$ and $b_2 = p_1$.

Then reapply the process to the interval $[a_2, b_2]$. This produces the method described in Algorithm 2.1. (See Figure 2.1.)

Figure 2.1



ALGORITHM
2.1

Bisection

To find a solution to $f(x) = 0$ given the continuous function f on the interval $[a, b]$, where $f(a)$ and $f(b)$ have opposite signs:

INPUT endpoints a, b; tolerance TOL; maximum number of iterations N_0.

OUTPUT approximate solution p or message of failure.

Step 1 Set $i = 1$;
$FA = f(a)$.

Step 2 While $i \le N_0$ do Steps 3–6.

Step 3 Set $p = a + (b - a)/2$; (*Compute* p_i.)
$FP = f(p)$.

Step 4 If $FP = 0$ or $(b - a)/2 < TOL$ then
OUTPUT (p); (*Procedure completed successfully.*)
STOP.

Step 5 Set $i = i + 1$.

Step 6 If $FA \cdot FP > 0$ then set $a = p$; (*Compute* a_i, b_i.)
$FA = FP$
else set $b = p$. (*FA is unchanged.*)

Step 7 OUTPUT ('Method failed after N_0 iterations, $N_0 =$', N_0);
(*The procedure was unsuccessful.*)
STOP.

■

Other stopping procedures can be applied in Step 4 of Algorithm 2.1 or in any of the iterative techniques in this chapter. For example, we can select a tolerance $\varepsilon > 0$ and generate $p_1, \ldots, p_N$ until one of the following conditions is met:

$$|p_N - p_{N-1}| < \varepsilon, \tag{2.1}$$

$$\frac{|p_N - p_{N-1}|}{|p_N|} < \varepsilon, \quad p_N \neq 0, \quad \text{or} \tag{2.2}$$

$$|f(p_N)| < \varepsilon. \tag{2.3}$$

Unfortunately, difficulties can arise using any of these stopping criteria. For example, there are sequences $\{p_n\}_{n=0}^{\infty}$ with the property that the differences $p_n - p_{n-1}$ converge to zero while the sequence itself diverges. (See Exercise 17.) It is also possible for $f(p_n)$ to be close to zero while p_n differs significantly from p. (See Exercise 16.) Without additional knowledge about f or p, Inequality (2.2) is the best stopping criterion to apply because it comes closest to testing relative error.

When using a computer to generate approximations, it is good practice to set an upper bound on the number of iterations. This eliminates the possibility of entering an infinite loop, a situation that can arise when the sequence diverges (and also when the program is incorrectly coded). This was done in Step 2 of Algorithm 2.1 where the bound N_0 was set and the procedure terminated if $i > N_0$.

Note that to start the Bisection Algorithm, an interval $[a, b]$ must be found with $f(a) \cdot f(b) < 0$. At each step the length of the interval known to contain a zero of f is reduced by a factor of 2; hence it is advantageous to choose the initial interval $[a, b]$ as small as possible. For example, if $f(x) = 2x^3 - x^2 + x - 1$, we have both

$$f(-4) \cdot f(4) < 0 \quad \text{and} \quad f(0) \cdot f(1) < 0,$$

so the Bisection Algorithm could be used on $[-4, 4]$ or on $[0, 1]$. Starting the Bisection Algorithm on $[0, 1]$ instead of $[-4, 4]$ will reduce by 3 the number of iterations required to achieve a specified accuracy.

The following example illustrates the Bisection Algorithm. The iteration in this example is terminated when a bound for the relative error is less than 0.0001. This is ensured by having

$$\frac{|p - p_n|}{\min\{|a_n|, |b_n|\}} < 10^{-4}.$$

Example 1 Show that $f(x) = x^3 + 4x^2 - 10 = 0$ has a root in $[1, 2]$, and use the Bisection method to determine an approximation to the root that is accurate to at least within 10^{-4}.

Solution Because $f(1) = -5$ and $f(2) = 14$ the Intermediate Value Theorem 1.11 ensures that this continuous function has a root in $[1, 2]$.

For the first iteration of the Bisection method we use the fact that at the midpoint of $[1, 2]$ we have $f(1.5) = 2.375 > 0$. This indicates that we should select the interval $[1, 1.5]$ for our second iteration. Then we find that $f(1.25) = -1.796875$ so our new interval becomes $[1.25, 1.5]$, whose midpoint is 1.375. Continuing in this manner gives the values in Table 2.1. After 13 iterations, $p_{13} = 1.365112305$ approximates the root p with an error

$$|p - p_{13}| < |b_{14} - a_{14}| = |1.365234375 - 1.365112305| = 0.000122070.$$

Since $|a_{14}| < |p|$, we have

$$\frac{|p - p_{13}|}{|p|} < \frac{|b_{14} - a_{14}|}{|a_{14}|} \leq 9.0 \times 10^{-5},$$

Table 2.1

n	a_n	b_n	p_n	$f(p_n)$
1	1.0	2.0	1.5	2.375
2	1.0	1.5	1.25	−1.79687
3	1.25	1.5	1.375	0.16211
4	1.25	1.375	1.3125	−0.84839
5	1.3125	1.375	1.34375	−0.35098
6	1.34375	1.375	1.359375	−0.09641
7	1.359375	1.375	1.3671875	0.03236
8	1.359375	1.3671875	1.36328125	−0.03215
9	1.36328125	1.3671875	1.365234375	0.000072
10	1.36328125	1.365234375	1.364257813	−0.01605
11	1.364257813	1.365234375	1.364746094	−0.00799
12	1.364746094	1.365234375	1.364990235	−0.00396
13	1.364990235	1.365234375	1.365112305	−0.00194

so the approximation is correct to at least within 10^{-4}. The correct value of p to nine decimal places is $p = 1.365230013$. Note that p_9 is closer to p than is the final approximation p_{13}. You might suspect this is true because $|f(p_9)| < |f(p_{13})|$, but we cannot be sure of this unless the true answer is known. ■

The Bisection method, though conceptually clear, has significant drawbacks. It is relatively slow to converge (that is, N may become quite large before $|p - p_N|$ is sufficiently small), and a good intermediate approximation might be inadvertently discarded. However, the method has the important property that it always converges to a solution, and for that reason it is often used as a starter for the more efficient methods we will see later in this chapter.

Theorem 2.1 Suppose that $f \in C[a, b]$ and $f(a) \cdot f(b) < 0$. The Bisection method generates a sequence $\{p_n\}_{n=1}^{\infty}$ approximating a zero p of f with

$$|p_n - p| \leq \frac{b-a}{2^n}, \quad \text{when} \quad n \geq 1.$$ ■

Proof For each $n \geq 1$, we have

$$b_n - a_n = \frac{1}{2^{n-1}}(b-a) \quad \text{and} \quad p \in (a_n, b_n).$$

Since $p_n = \frac{1}{2}(a_n + b_n)$ for all $n \geq 1$, it follows that

$$|p_n - p| \leq \frac{1}{2}(b_n - a_n) = \frac{b-a}{2^n}.$$ ■ ■ ■

Because

$$|p_n - p| \leq (b-a)\frac{1}{2^n},$$

the sequence $\{p_n\}_{n=1}^{\infty}$ converges to p with rate of convergence $O\left(\frac{1}{2^n}\right)$; that is,

$$p_n = p + O\left(\frac{1}{2^n}\right).$$

It is important to realize that Theorem 2.1 gives only a bound for approximation error and that this bound might be quite conservative. For example, this bound applied to the problem in Example 1 ensures only that

$$|p - p_9| \le \frac{2-1}{2^9} \approx 2 \times 10^{-3},$$

but the actual error is much smaller:

$$|p - p_9| = |1.365230013 - 1.365234375| \approx 4.4 \times 10^{-6}.$$

Example 2 Determine the number of iterations necessary to solve $f(x) = x^3 + 4x^2 - 10 = 0$ with accuracy 10^{-3} using $a_1 = 1$ and $b_1 = 2$.

Solution We we will use logarithms to find an integer N that satisfies

$$|p_N - p| \le 2^{-N}(b-a) = 2^{-N} < 10^{-3}.$$

Logarithms to any base would suffice, but we will use base-10 logarithms because the tolerance is given as a power of 10. Since $2^{-N} < 10^{-3}$ implies that $\log_{10} 2^{-N} < \log_{10} 10^{-3} = -3$, we have

$$-N \log_{10} 2 < -3 \quad \text{and} \quad N > \frac{3}{\log_{10} 2} \approx 9.96.$$

Hence, ten iterations will ensure an approximation accurate to within 10^{-3}.

Table 2.1 shows that the value of $p_9 = 1.365234375$ is accurate to within 10^{-4}. Again, it is important to keep in mind that the error analysis gives only a bound for the number of iterations. In many cases this bound is much larger than the actual number required. ■

Maple has a *NumericalAnalysis* package that implements many of the techniques we will discuss, and the presentation and examples in the package are closely aligned with this text. The Bisection method in this package has a number of options, some of which we will now consider. In what follows, Maple code is given in *black italic* type and Maple response in cyan.

Load the *NumericalAnalysis* package with the command

with(*Student*[*NumericalAnalysis*])

which gives access to the procedures in the package. Define the function with

$f := x^3 + 4x^2 - 10$

and use

Bisection $(f, x = [1, 2], tolerance = 0.005)$

Maple returns

1.363281250

Note that the value that is output is the same as p_8 in Table 2.1.

The sequence of bisection intervals can be output with the command

Bisection $(f, x = [1, 2], tolerance = 0.005, output = sequence)$

and Maple returns the intervals containing the solution together with the solution

[1., 2.], [1., 1.500000000], [1.250000000, 1.500000000], [1.250000000, 1.375000000],

[1.312500000, 1.375000000], [1.343750000, 1.375000000], [1.359375000, 1.375000000],

[1.359375000, 1.367187500], 1.363281250

The stopping criterion can also be based on relative error by choosing the option

Bisection $(f, x = [1, 2], tolerance = 0.005, stoppingcriterion = relative)$

Now Maple returns

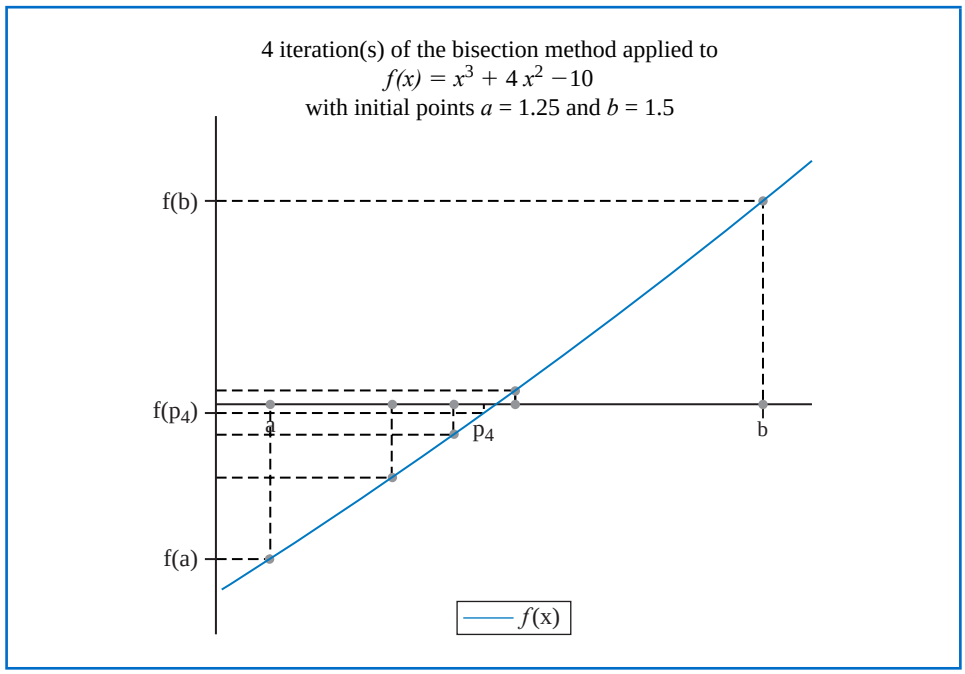
$$1.363281250$$

The option *output* = *plot* given in

$$Bisection\,(f, x = [1.25, 1.5], output = plot, tolerance = 0.02)$$

produces the plot shown in Figure 2.2.

Figure 2.2



We can also set the maximum number of iterations with the option *maxiterations* = . An error message will be displayed if the stated tolerance is not met within the specified number of iterations.

The results from Bisection method can also be obtained using the command Roots. For example,

$$Roots\left(f, x = [1.0, 2.0], method = bisection, tolerance = \frac{1}{100}, output = information\right)$$

uses the Bisection method to produce the information

n	a_n	b_n	p_n	$f\,(p_n)$	*relative error*
1	1.0	2.0	1.500000000	2.37500000	0.3333333333
2	1.0	1.500000000	1.250000000	−1.796875000	0.2000000000
3	1.250000000	1.500000000	1.375000000	0.16210938	0.09090909091
4	1.250000000	1.375000000	1.312500000	−0.848388672	0.04761904762
5	1.312500000	1.375000000	1.343750000	−0.350982668	0.02325581395
6	1.343750000	1.375000000	1.359375000	−0.096408842	0.01149425287
7	1.359375000	1.375000000	1.367187500	0.03235578	0.005714285714

The bound for the number of iterations for the Bisection method assumes that the calculations are performed using infinite-digit arithmetic. When implementing the method on a computer, we need to consider the effects of round-off error. For example, the computation of the midpoint of the interval $[a_n, b_n]$ should be found from the equation

$$p_n = a_n + \frac{b_n - a_n}{2} \quad \text{instead of} \quad p_n = \frac{a_n + b_n}{2}.$$

The first equation adds a small correction, $(b_n - a_n)/2$, to the known value a_n. When $b_n - a_n$ is near the maximum precision of the machine, this correction might be in error, but the error would not significantly affect the computed value of p_n. However, when $b_n - a_n$ is near the maximum precision of the machine, it is possible for $(a_n + b_n)/2$ to return a midpoint that is not even in the interval $[a_n, b_n]$.

The Latin word *signum* means "token" or "sign". So the signum function quite naturally returns the sign of a number (unless the number is 0).

As a final remark, to determine which subinterval of $[a_n, b_n]$ contains a root of f, it is better to make use of the **signum** function, which is defined as

$$\operatorname{sgn}(x) = \begin{cases} -1, & \text{if } x < 0, \\ 0, & \text{if } x = 0, \\ 1, & \text{if } x > 0. \end{cases}$$

The test

$$\operatorname{sgn}\,(f(a_n))\,\operatorname{sgn}\,(f(p_n)) < 0 \qquad \text{instead of} \qquad f(a_n)f(p_n) < 0$$

gives the same result but avoids the possibility of overflow or underflow in the multiplication of $f(a_n)$ and $f(p_n)$.

EXERCISE SET 2.1

1. Use the Bisection method to find p_3 for $f(x) = \sqrt{x} - \cos x$ on $[0, 1]$.

2. Let $f(x) = 3(x + 1)(x - \frac{1}{2})(x - 1)$. Use the Bisection method on the following intervals to find p_3.
a. $[-2, 1.5]$ **b.** $[-1.25, 2.5]$

3. Use the Bisection method to find solutions accurate to within 10^{-2} for $x^3 - 7x^2 + 14x - 6 = 0$ on each interval.
a. $[0, 1]$ **b.** $[1, 3.2]$ **c.** $[3.2, 4]$

4. Use the Bisection method to find solutions accurate to within 10^{-2} for $x^4 - 2x^3 - 4x^2 + 4x + 4 = 0$ on each interval.
a. $[-2, -1]$ **b.** $[0, 2]$ **c.** $[2, 3]$ **d.** $[-1, 0]$

5. Use the Bisection method to find solutions accurate to within 10^{-5} for the following problems.
a. $x - 2^{-x} = 0$ for $0 \le x \le 1$
b. $e^x - x^2 + 3x - 2 = 0$ for $0 \le x \le 1$
c. $2x\cos(2x) - (x + 1)^2 = 0$ for $-3 \le x \le -2$ and $-1 \le x \le 0$
d. $x\cos x - 2x^2 + 3x - 1 = 0$ for $0.2 \le x \le 0.3$ and $1.2 \le x \le 1.3$

6. Use the Bisection method to find solutions, accurate to within 10^{-5} for the following problems.
a. $3x - e^x = 0$ for $1 \le x \le 2$
b. $2x + 3\cos x - e^x = 0$ for $0 \le x \le 1$
c. $x^2 - 4x + 4 - \ln x = 0$ for $1 \le x \le 2$ and $2 \le x \le 4$
d. $x + 1 - 2\sin \pi x = 0$ for $0 \le x \le 0.5$ and $0.5 \le x \le 1$

7. **a.** Sketch the graphs of $y = x$ and $y = 2\sin x$.

b. Use the Bisection method to find an approximation to within 10^{-5} to the first positive value of x with $x = 2\sin x$.

8. **a.** Sketch the graphs of $y = x$ and $y = \tan x$.

b. Use the Bisection method to find an approximation to within 10^{-5} to the first positive value of x with $x = \tan x$.

9. **a.** Sketch the graphs of $y = e^x - 2$ and $y = \cos(e^x - 2)$.

b. Use the Bisection method to find an approximation to within 10^{-5} to a value in $[0.5, 1.5]$ with $e^x - 2 = \cos(e^x - 2)$.

10. Let $f(x) = (x+2)(x+1)^2x(x-1)^3(x-2)$. To which zero of f does the Bisection method converge when applied on the following intervals?

a. $[-1.5, 2.5]$ **b.** $[-0.5, 2.4]$ **c.** $[-0.5, 3]$ **d.** $[-3, -0.5]$

11. Let $f(x) = (x+2)(x+1)x(x-1)^3(x-2)$. To which zero of f does the Bisection method converge when applied on the following intervals?

a. $[-3, 2.5]$ **b.** $[-2.5, 3]$ **c.** $[-1.75, 1.5]$ **d.** $[-1.5, 1.75]$

12. Find an approximation to $\sqrt{3}$ correct to within 10^{-4} using the Bisection Algorithm. [*Hint:* Consider $f(x) = x^2 - 3$.]

13. Find an approximation to $\sqrt[3]{25}$ correct to within 10^{-4} using the Bisection Algorithm.

14. Use Theorem 2.1 to find a bound for the number of iterations needed to achieve an approximation with accuracy 10^{-3} to the solution of $x^3 + x - 4 = 0$ lying in the interval $[1, 4]$. Find an approximation to the root with this degree of accuracy.

15. Use Theorem 2.1 to find a bound for the number of iterations needed to achieve an approximation with accuracy 10^{-4} to the solution of $x^3 - x - 1 = 0$ lying in the interval $[1, 2]$. Find an approximation to the root with this degree of accuracy.

16. Let $f(x) = (x-1)^{10}$, $p = 1$, and $p_n = 1 + 1/n$. Show that $|f(p_n)| < 10^{-3}$ whenever $n > 1$ but that $|p - p_n| < 10^{-3}$ requires that $n > 1000$.

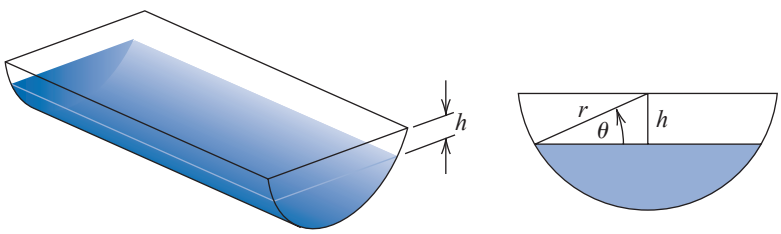
17. Let $\{p_n\}$ be the sequence defined by $p_n = \sum_{k=1}^{n} \frac{1}{k}$. Show that $\{p_n\}$ diverges even though $\lim_{n\to\infty}(p_n - p_{n-1}) = 0$.

18. The function defined by $f(x) = \sin \pi x$ has zeros at every integer. Show that when $-1 < a < 0$ and $2 < b < 3$, the Bisection method converges to

a. 0, if $a + b < 2$ **b.** 2, if $a + b > 2$ **c.** 1, if $a + b = 2$

19. A trough of length L has a cross section in the shape of a semicircle with radius r. (See the accompanying figure.) When filled with water to within a distance h of the top, the volume V of water is

$$V = L\left[0.5\pi r^2 - r^2 \arcsin(h/r) - h(r^2 - h^2)^{1/2}\right].$$



Suppose $L = 10$ ft, $r = 1$ ft, and $V = 12.4$ ft^3. Find the depth of water in the trough to within 0.01 ft.

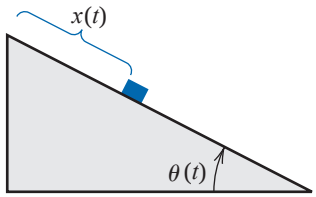
20. A particle starts at rest on a smooth inclined plane whose angle θ is changing at a constant rate

$$\frac{d\theta}{dt} = \omega < 0.$$

At the end of t seconds, the position of the object is given by

$$x(t) = -\frac{g}{2\omega^2}\left(\frac{e^{wt} - e^{-wt}}{2} - \sin \omega t\right).$$

Suppose the particle has moved 1.7 ft in 1 s. Find, to within 10^{-5}, the rate ω at which θ changes. Assume that $g = 32.17$ ft/s^2.



2.2 Fixed-Point Iteration

A *fixed point* for a function is a number at which the value of the function does not change when the function is applied.

Definition 2.2 The number p is a **fixed point** for a given function g if $g(p) = p$. ■

Fixed-point results occur in many areas of mathematics, and are a major tool of economists for proving results concerning equilibria. Although the idea behind the technique is old, the terminology was first used by the Dutch mathematician L. E. J. Brouwer (1882–1962) in the early 1900s.

In this section we consider the problem of finding solutions to fixed-point problems and the connection between the fixed-point problems and the root-finding problems we wish to solve. Root-finding problems and fixed-point problems are equivalent classes in the following sense:

- Given a root-finding problem $f(p) = 0$, we can define functions g with a fixed point at p in a number of ways, for example, as

$$g(x) = x - f(x) \quad \text{or as} \quad g(x) = x + 3f(x).$$

- Conversely, if the function g has a fixed point at p, then the function defined by

$$f(x) = x - g(x)$$

has a zero at p.

Although the problems we wish to solve are in the root-finding form, the fixed-point form is easier to analyze, and certain fixed-point choices lead to very powerful root-finding techniques.

We first need to become comfortable with this new type of problem, and to decide when a function has a fixed point and how the fixed points can be approximated to within a specified accuracy.

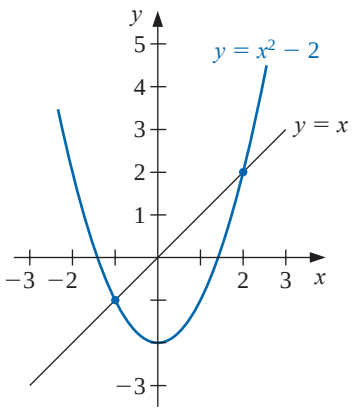
Example 1 Determine any fixed points of the function $g(x) = x^2 - 2$.

Solution A fixed point p for g has the property that

$$p = g(p) = p^2 - 2 \quad \text{which implies that} \quad 0 = p^2 - p - 2 = (p + 1)(p - 2).$$

A fixed point for g occurs precisely when the graph of $y = g(x)$ intersects the graph of $y = x$, so g has two fixed points, one at $p = -1$ and the other at $p = 2$. These are shown in Figure 2.3. ■

Figure 2.3



The following theorem gives sufficient conditions for the existence and uniqueness of a fixed point.

Theorem 2.3

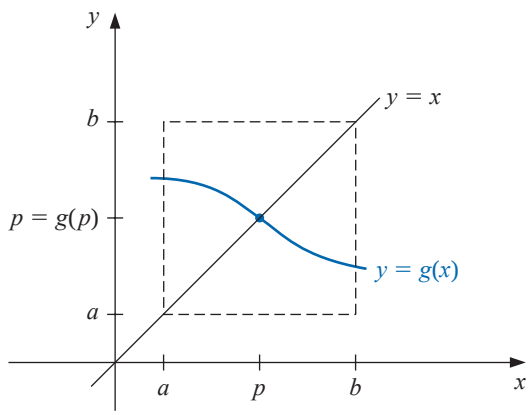
(i) If $g \in C[a, b]$ and $g(x) \in [a, b]$ for all $x \in [a, b]$, then g has at least one fixed point in $[a, b]$.

(ii) If, in addition, $g'(x)$ exists on (a, b) and a positive constant $k < 1$ exists with

$$|g'(x)| \leq k, \quad \text{for all } x \in (a, b),$$

then there is exactly one fixed point in $[a, b]$. (See Figure 2.4.) ■

Figure 2.4



Proof

(i) If $g(a) = a$ or $g(b) = b$, then g has a fixed point at an endpoint. If not, then $g(a) > a$ and $g(b) < b$. The function $h(x) = g(x) - x$ is continuous on $[a, b]$, with

$$h(a) = g(a) - a > 0 \quad \text{and} \quad h(b) = g(b) - b < 0.$$

The Intermediate Value Theorem implies that there exists $p \in (a, b)$ for which $h(p) = 0$. This number p is a fixed point for g because

$$0 = h(p) = g(p) - p \quad \text{implies that} \quad g(p) = p.$$

(ii) Suppose, in addition, that $|g'(x)| \leq k < 1$ and that p and q are both fixed points in $[a, b]$. If $p \neq q$, then the Mean Value Theorem implies that a number ξ exists between p and q, and hence in $[a, b]$, with

$$\frac{g(p) - g(q)}{p - q} = g'(\xi).$$

Thus

$$|p - q| = |g(p) - g(q)| = |g'(\xi)||p - q| \leq k|p - q| < |p - q|,$$

which is a contradiction. This contradiction must come from the only supposition, $p \neq q$. Hence, $p = q$ and the fixed point in $[a, b]$ is unique. ■ ■ ■

Example 2 Show that $g(x) = (x^2 - 1)/3$ has a unique fixed point on the interval $[-1, 1]$.

Solution The maximum and minimum values of $g(x)$ for x in $[-1, 1]$ must occur either when x is an endpoint of the interval or when the derivative is 0. Since $g'(x) = 2x/3$, the function g is continuous and $g'(x)$ exists on $[-1, 1]$. The maximum and minimum values of $g(x)$ occur at $x = -1$, $x = 0$, or $x = 1$. But $g(-1) = 0$, $g(1) = 0$, and $g(0) = -1/3$, so an absolute maximum for $g(x)$ on $[-1, 1]$ occurs at $x = -1$ and $x = 1$, and an absolute minimum at $x = 0$.

Moreover

$$|g'(x)| = \left|\frac{2x}{3}\right| \leq \frac{2}{3}, \quad \text{for all } x \in (-1, 1).$$

So g satisfies all the hypotheses of Theorem 2.3 and has a unique fixed point in $[-1, 1]$. ■

For the function in Example 2, the unique fixed point p in the interval $[-1, 1]$ can be determined algebraically. If

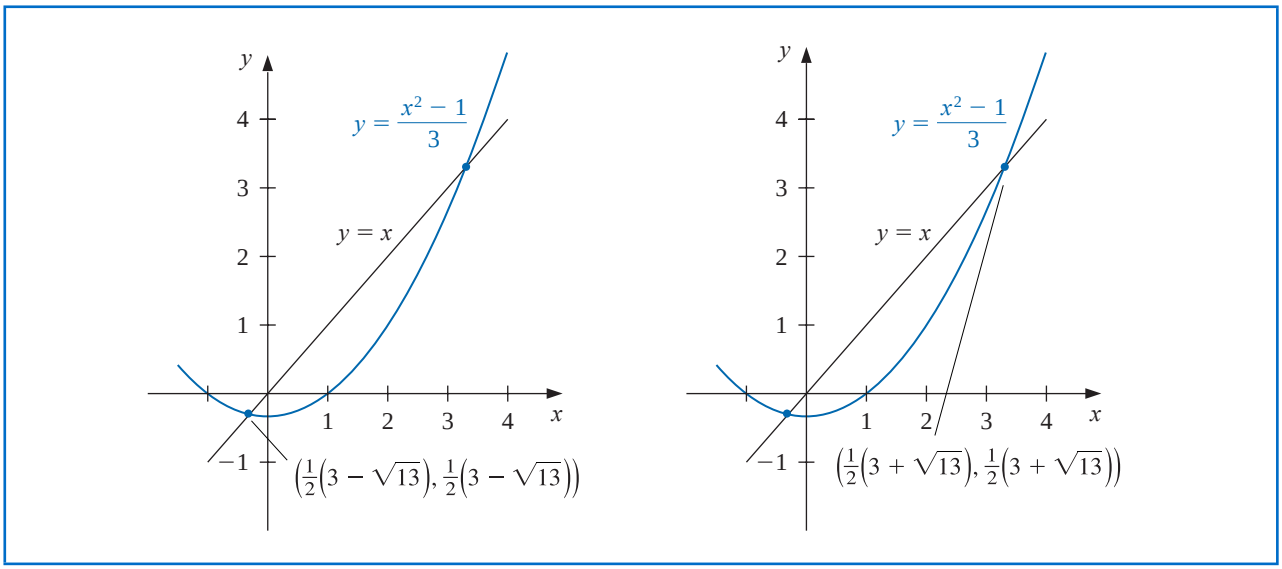
$$p = g(p) = \frac{p^2 - 1}{3}, \quad \text{then} \quad p^2 - 3p - 1 = 0,$$

which, by the quadratic formula, implies, as shown on the left graph in Figure 2.4, that

$$p = \frac{1}{2}(3 - \sqrt{13}).$$

Note that g also has a unique fixed point $p = \frac{1}{2}(3 + \sqrt{13})$ for the interval $[3, 4]$. However, $g(4) = 5$ and $g'(4) = \frac{8}{3} > 1$, so g does not satisfy the hypotheses of Theorem 2.3 on $[3, 4]$. This demonstrates that the hypotheses of Theorem 2.3 are sufficient to guarantee a unique fixed point but are not necessary. (See the graph on the right in Figure 2.5.)

Figure 2.5



Example 3 Show that Theorem 2.3 does not ensure a unique fixed point of $g(x) = 3^{-x}$ on the interval $[0, 1]$, even though a unique fixed point on this interval does exist.

Solution $g'(x) = -3^{-x} \ln 3 < 0$ on $[0, 1]$, the function g is strictly decreasing on $[0, 1]$. So

$$g(1) = \frac{1}{3} \le g(x) \le 1 = g(0), \quad \text{for} \quad 0 \le x \le 1.$$

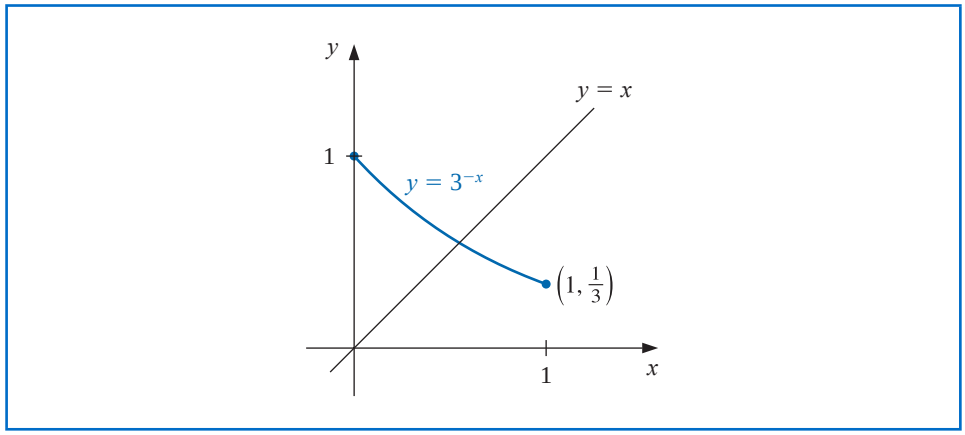
Thus, for $x \in [0, 1]$, we have $g(x) \in [0, 1]$. The first part of Theorem 2.3 ensures that there is at least one fixed point in $[0, 1]$.

However,

$$g'(0) = -\ln 3 = -1.098612289,$$

so $|g'(x)| \not\le 1$ on $(0, 1)$, and Theorem 2.3 cannot be used to determine uniqueness. But g is always decreasing, and it is clear from Figure 2.6 that the fixed point must be unique. ■

Figure 2.6



Fixed-Point Iteration

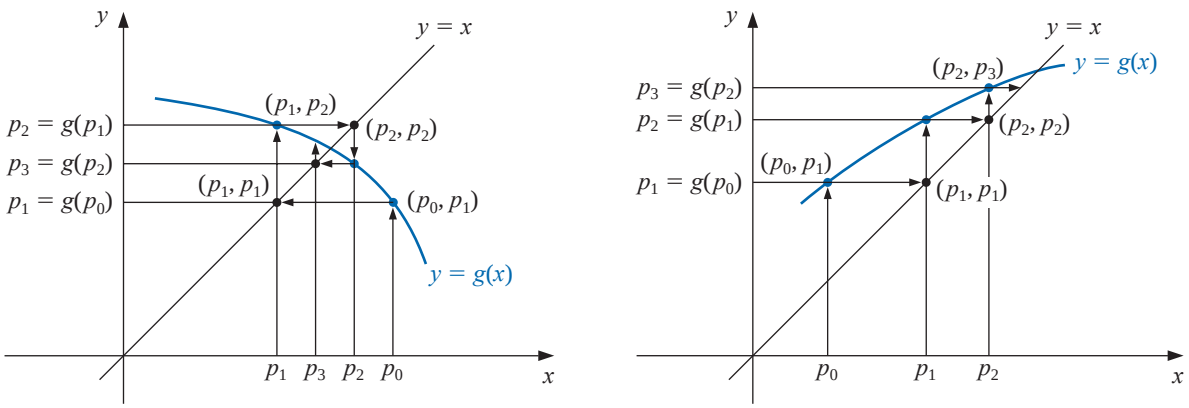
We cannot explicitly determine the fixed point in Example 3 because we have no way to solve for p in the equation $p = g(p) = 3^{-p}$. We can, however, determine approximations to this fixed point to any specified degree of accuracy. We will now consider how this can be done.

To approximate the fixed point of a function g, we choose an initial approximation p_0 and generate the sequence $\{p_n\}_{n=0}^{\infty}$ by letting $p_n = g(p_{n-1})$, for each $n \geq 1$. If the sequence converges to p and g is continuous, then

$$p = \lim_{n\to\infty} p_n = \lim_{n\to\infty} g(p_{n-1}) = g\left(\lim_{n\to\infty} p_{n-1}\right) = g(p),$$

and a solution to $x = g(x)$ is obtained. This technique is called **fixed-point**, or **functional iteration**. The procedure is illustrated in Figure 2.7 and detailed in Algorithm 2.2.

Figure 2.7



ALGORITHM
2.2

Fixed-Point Iteration

To find a solution to $p = g(p)$ given an initial approximation p_0:

INPUT initial approximation p_0; tolerance TOL; maximum number of iterations N_0.

OUTPUT approximate solution p or message of failure.

Step 1 Set $i = 1$.

Step 2 While $i \leq N_0$ do Steps 3–6.

Step 3 Set $p = g(p_0)$. (*Compute* p_i.)

Step 4 If $|p - p_0| < TOL$ then
OUTPUT (p); (*The procedure was successful.*)
STOP.

Step 5 Set $i = i + 1$.

Step 6 Set $p_0 = p$. (*Update* p_0.)

Step 7 OUTPUT ('The method failed after N_0 iterations, $N_0 =$', N_0);
(*The procedure was unsuccessful.*)
STOP. ■

The following illustrates some features of functional iteration.

Illustration The equation $x^3 + 4x^2 - 10 = 0$ has a unique root in $[1, 2]$. There are many ways to change the equation to the fixed-point form $x = g(x)$ using simple algebraic manipulation. For example, to obtain the function g described in part (c), we can manipulate the equation $x^3 + 4x^2 - 10 = 0$ as follows:

$$4x^2 = 10 - x^3, \quad \text{so} \quad x^2 = \frac{1}{4}(10 - x^3), \quad \text{and} \quad x = \pm\frac{1}{2}(10 - x^3)^{1/2}.$$

To obtain a positive solution, $g_3(x)$ is chosen. It is not important for you to derive the functions shown here, but you should verify that the fixed point of each is actually a solution to the original equation, $x^3 + 4x^2 - 10 = 0$.

(a) $x = g_1(x) = x - x^3 - 4x^2 + 10$

(b) $x = g_2(x) = \left(\frac{10}{x} - 4x\right)^{1/2}$

(c) $x = g_3(x) = \frac{1}{2}(10 - x^3)^{1/2}$

(d) $x = g_4(x) = \left(\frac{10}{4 + x}\right)^{1/2}$

(e) $x = g_5(x) = x - \frac{x^3 + 4x^2 - 10}{3x^2 + 8x}$

With $p_0 = 1.5$, Table 2.2 lists the results of the fixed-point iteration for all five choices of g.

Table 2.2

n	(a)	(b)	(c)	(d)	(e)
0	1.5	1.5	1.5	1.5	1.5
1	−0.875	0.8165	1.286953768	1.348399725	1.373333333
2	6.732	2.9969	1.402540804	1.367376372	1.365262015
3	−469.7	$(-8.65)^{1/2}$	1.345458374	1.364957015	1.365230014
4	1.03×10^8		1.375170253	1.365264748	1.365230013
5			1.360094193	1.365225594	
6			1.367846968	1.365230576	
7			1.363887004	1.365229942	
8			1.365916734	1.365230022	
9			1.364878217	1.365230012	
10			1.365410062	1.365230014	
15			1.365223680	1.365230013	
20			1.365230236		
25			1.365230006		
30			1.365230013		

The actual root is 1.365230013, as was noted in Example 1 of Section 2.1. Comparing the results to the Bisection Algorithm given in that example, it can be seen that excellent results have been obtained for choices (c), (d), and (e) (the Bisection method requires 27 iterations for this accuracy). It is interesting to note that choice (a) was divergent and that (b) became undefined because it involved the square root of a negative number. □

Although the various functions we have given are fixed-point problems for the same root-finding problem, they differ vastly as techniques for approximating the solution to the root-finding problem. Their purpose is to illustrate what needs to be answered:

- Question: How can we find a fixed-point problem that produces a sequence that reliably and rapidly converges to a solution to a given root-finding problem?

The following theorem and its corollary give us some clues concerning the paths we should pursue and, perhaps more importantly, some we should reject.

Theorem 2.4 (Fixed-Point Theorem)

Let $g \in C[a, b]$ be such that $g(x) \in [a, b]$, for all x in $[a, b]$. Suppose, in addition, that g' exists on (a, b) and that a constant $0 < k < 1$ exists with

$$|g'(x)| \leq k, \quad \text{for all } x \in (a, b).$$

Then for any number p_0 in $[a, b]$, the sequence defined by

$$p_n = g(p_{n-1}), \quad n \geq 1,$$

converges to the unique fixed point p in $[a, b]$. ■

Proof Theorem 2.3 implies that a unique point p exists in $[a, b]$ with $g(p) = p$. Since g maps $[a, b]$ into itself, the sequence $\{p_n\}_{n=0}^{\infty}$ is defined for all $n \geq 0$, and $p_n \in [a, b]$ for all n. Using the fact that $|g'(x)| \leq k$ and the Mean Value Theorem 1.8, we have, for each n,

$$|p_n - p| = |g(p_{n-1}) - g(p)| = |g'(\xi_n)||p_{n-1} - p| \leq k|p_{n-1} - p|,$$

where $\xi_n \in (a, b)$. Applying this inequality inductively gives

$$|p_n - p| \leq k|p_{n-1} - p| \leq k^2|p_{n-2} - p| \leq \cdots \leq k^n|p_0 - p|. \tag{2.4}$$

Since $0 < k < 1$, we have $\lim_{n\to\infty} k^n = 0$ and

$$\lim_{n\to\infty} |p_n - p| \leq \lim_{n\to\infty} k^n|p_0 - p| = 0.$$

Hence $\{p_n\}_{n=0}^{\infty}$ converges to p. ■ ■ ■

Corollary 2.5 If g satisfies the hypotheses of Theorem 2.4, then bounds for the error involved in using p_n to approximate p are given by

$$|p_n - p| \leq k^n \max\{p_0 - a, b - p_0\} \tag{2.5}$$

and

$$|p_n - p| \leq \frac{k^n}{1 - k}|p_1 - p_0|, \quad \text{for all} \quad n \geq 1. \tag{2.6}$$

■

Proof Because $p \in [a, b]$, the first bound follows from Inequality (2.4):

$$|p_n - p| \leq k^n|p_0 - p| \leq k^n \max\{p_0 - a, b - p_0\}.$$

For $n \geq 1$, the procedure used in the proof of Theorem 2.4 implies that

$$|p_{n+1} - p_n| = |g(p_n) - g(p_{n-1})| \leq k|p_n - p_{n-1}| \leq \cdots \leq k^n|p_1 - p_0|.$$

Thus for $m > n \geq 1$,

$$\begin{aligned}|p_m - p_n| &= |p_m - p_{m-1} + p_{m-1} - \cdots + p_{n+1} - p_n| \\ &\leq |p_m - p_{m-1}| + |p_{m-1} - p_{m-2}| + \cdots + |p_{n+1} - p_n| \\ &\leq k^{m-1}|p_1 - p_0| + k^{m-2}|p_1 - p_0| + \cdots + k^n|p_1 - p_0| \\ &= k^n|p_1 - p_0|\left(1 + k + k^2 + \cdots + k^{m-n-1}\right).\end{aligned}$$

By Theorem 2.3, $\lim_{m\to\infty} p_m = p$, so

$$|p - p_n| = \lim_{m\to\infty} |p_m - p_n| \leq \lim_{m\to\infty} k^n|p_1 - p_0| \sum_{i=0}^{m-n-1} k^i \leq k^n|p_1 - p_0| \sum_{i=0}^{\infty} k^i.$$

But $\sum_{i=0}^{\infty} k^i$ is a geometric series with ratio k and $0 < k < 1$. This sequence converges to $1/(1-k)$, which gives the second bound:

$$|p - p_n| \leq \frac{k^n}{1-k}|p_1 - p_0|. \qquad \blacksquare\ \blacksquare\ \blacksquare$$

Both inequalities in the corollary relate the rate at which $\{p_n\}_{n=0}^{\infty}$ converges to the bound k on the first derivative. The rate of convergence depends on the factor k^n. The smaller the value of k, the faster the convergence, which may be very slow if k is close to 1.

Illustration Let us reconsider the various fixed-point schemes described in the preceding illustration in light of the Fixed-point Theorem 2.4 and its Corollary 2.5.

(a) For $g_1(x) = x - x^3 - 4x^2 + 10$, we have $g_1(1) = 6$ and $g_1(2) = -12$, so g_1 does not map $[1, 2]$ into itself. Moreover, $g_1'(x) = 1 - 3x^2 - 8x$, so $|g_1'(x)| > 1$ for all x in $[1, 2]$. Although Theorem 2.4 does not guarantee that the method must fail for this choice of g, there is no reason to expect convergence.

(b) With $g_2(x) = [(10/x) - 4x]^{1/2}$, we can see that g_2 does not map $[1, 2]$ into $[1, 2]$, and the sequence $\{p_n\}_{n=0}^{\infty}$ is not defined when $p_0 = 1.5$. Moreover, there is no interval containing $p \approx 1.365$ such that $|g_2'(x)| < 1$, because $|g_2'(p)| \approx 3.4$. There is no reason to expect that this method will converge.

(c) For the function $g_3(x) = \frac{1}{2}(10 - x^3)^{1/2}$, we have

$$g_3'(x) = -\frac{3}{4}x^2(10 - x^3)^{-1/2} < 0 \quad \text{on } [1, 2],$$

so g_3 is strictly decreasing on $[1, 2]$. However, $|g_3'(2)| \approx 2.12$, so the condition $|g_3'(x)| \leq k < 1$ fails on $[1, 2]$. A closer examination of the sequence $\{p_n\}_{n=0}^{\infty}$ starting with $p_0 = 1.5$ shows that it suffices to consider the interval $[1, 1.5]$ instead of $[1, 2]$. On this interval it is still true that $g_3'(x) < 0$ and g_3 is strictly decreasing, but, additionally,

$$1 < 1.28 \approx g_3(1.5) \leq g_3(x) \leq g_3(1) = 1.5,$$

for all $x \in [1, 1.5]$. This shows that g_3 maps the interval $[1, 1.5]$ into itself. It is also true that $|g_3'(x)| \leq |g_3'(1.5)| \approx 0.66$ on this interval, so Theorem 2.4 confirms the convergence of which we were already aware.

(d) For $g_4(x) = (10/(4 + x))^{1/2}$, we have

$$|g_4'(x)| = \left|\frac{-5}{\sqrt{10}(4 + x)^{3/2}}\right| \leq \frac{5}{\sqrt{10}(5)^{3/2}} < 0.15, \quad \text{for all} \quad x \in [1, 2].$$

The bound on the magnitude of $g_4'(x)$ is much smaller than the bound (found in (**c**)) on the magnitude of $g_3'(x)$, which explains the more rapid convergence using g_4.

(e) The sequence defined by

$$g_5(x) = x - \frac{x^3 + 4x^2 - 10}{3x^2 + 8x}$$

converges much more rapidly than our other choices. In the next sections we will see where this choice came from and why it is so effective. □

From what we have seen,

- Question: How can we find a fixed-point problem that produces a sequence that reliably and rapidly converges to a solution to a given root-finding problem?

might have

- Answer: Manipulate the root-finding problem into a fixed point problem that satisfies the conditions of Fixed-Point Theorem 2.4 and has a derivative that is as small as possible near the fixed point.

In the next sections we will examine this in more detail.

Maple has the fixed-point algorithm implemented in its *NumericalAnalysis* package. The options for the Bisection method are also available for fixed-point iteration. We will show only one option. After accessing the package using *with*(*Student*[*NumericalAnalysis*]): we enter the function

$$g := x - \frac{(x^3 + 4x^2 - 10)}{3x^2 + 8x}$$

and Maple returns

$$x - \frac{x^3 + 4x^2 - 10}{3x^2 + 8x}$$

Enter the command

FixedPointIteration(*fixedpointiterator* = *g*, $x = 1.5$, tolerance $= 10^{-8}$, *output* = *sequence*, *maxiterations* $= 20$)

and Maple returns

$$1.5, 1.373333333, 1.365262015, 1.365230014, 1.365230013$$

EXERCISE SET 2.2

1. Use algebraic manipulation to show that each of the following functions has a fixed point at p precisely when $f(p) = 0$, where $f(x) = x^4 + 2x^2 - x - 3$.

a. $g_1(x) = \left(3 + x - 2x^2\right)^{1/4}$ **b.** $g_2(x) = \left(\frac{x + 3 - x^4}{2}\right)^{1/2}$

c. $g_3(x) = \left(\dfrac{x+3}{x^2+2}\right)^{1/2}$ **d.** $g_4(x) = \dfrac{3x^4+2x^2+3}{4x^3+4x-1}$

2. **a.** Perform four iterations, if possible, on each of the functions g defined in Exercise 1. Let $p_0 = 1$ and $p_{n+1} = g(p_n)$, for $n = 0, 1, 2, 3$.

b. Which function do you think gives the best approximation to the solution?

3. The following four methods are proposed to compute $21^{1/3}$. Rank them in order, based on their apparent speed of convergence, assuming $p_0 = 1$.

a. $p_n = \dfrac{20p_{n-1} + 21/p_{n-1}^2}{21}$ **b.** $p_n = p_{n-1} - \dfrac{p_{n-1}^3 - 21}{3p_{n-1}^2}$

c. $p_n = p_{n-1} - \dfrac{p_{n-1}^4 - 21p_{n-1}}{p_{n-1}^2 - 21}$ **d.** $p_n = \left(\dfrac{21}{p_{n-1}}\right)^{1/2}$

4. The following four methods are proposed to compute $7^{1/5}$. Rank them in order, based on their apparent speed of convergence, assuming $p_0 = 1$.

a. $p_n = p_{n-1}\left(1 + \dfrac{7 - p_{n-1}^5}{p_{n-1}^2}\right)^3$ **b.** $p_n = p_{n-1} - \dfrac{p_{n-1}^5 - 7}{p_{n-1}^2}$

c. $p_n = p_{n-1} - \dfrac{p_{n-1}^5 - 7}{5p_{n-1}^4}$ **d.** $p_n = p_{n-1} - \dfrac{p_{n-1}^5 - 7}{12}$

5. Use a fixed-point iteration method to determine a solution accurate to within 10^{-2} for $x^4 - 3x^2 - 3 = 0$ on $[1, 2]$. Use $p_0 = 1$.

6. Use a fixed-point iteration method to determine a solution accurate to within 10^{-2} for $x^3 - x - 1 = 0$ on $[1, 2]$. Use $p_0 = 1$.

7. Use Theorem 2.3 to show that $g(x) = \pi + 0.5\sin(x/2)$ has a unique fixed point on $[0, 2\pi]$. Use fixed-point iteration to find an approximation to the fixed point that is accurate to within 10^{-2}. Use Corollary 2.5 to estimate the number of iterations required to achieve 10^{-2} accuracy, and compare this theoretical estimate to the number actually needed.

8. Use Theorem 2.3 to show that $g(x) = 2^{-x}$ has a unique fixed point on $[\frac{1}{3}, 1]$. Use fixed-point iteration to find an approximation to the fixed point accurate to within 10^{-4}. Use Corollary 2.5 to estimate the number of iterations required to achieve 10^{-4} accuracy, and compare this theoretical estimate to the number actually needed.

9. Use a fixed-point iteration method to find an approximation to $\sqrt{3}$ that is accurate to within 10^{-4}. Compare your result and the number of iterations required with the answer obtained in Exercise 12 of Section 2.1.

10. Use a fixed-point iteration method to find an approximation to $\sqrt[3]{25}$ that is accurate to within 10^{-4}. Compare your result and the number of iterations required with the answer obtained in Exercise 13 of Section 2.1.

11. For each of the following equations, determine an interval $[a, b]$ on which fixed-point iteration will converge. Estimate the number of iterations necessary to obtain approximations accurate to within 10^{-5}, and perform the calculations.

a. $x = \dfrac{2 - e^x + x^2}{3}$ **b.** $x = \dfrac{5}{x^2} + 2$

c. $x = (e^x/3)^{1/2}$ **d.** $x = 5^{-x}$

e. $x = 6^{-x}$ **f.** $x = 0.5(\sin x + \cos x)$

12. For each of the following equations, use the given interval or determine an interval $[a, b]$ on which fixed-point iteration will converge. Estimate the number of iterations necessary to obtain approximations accurate to within 10^{-5}, and perform the calculations.

a. $2 + \sin x - x = 0$ use $[2, 3]$ **b.** $x^3 - 2x - 5 = 0$ use $[2, 3]$

c. $3x^2 - e^x = 0$ **d.** $x - \cos x = 0$

13. Find all the zeros of $f(x) = x^2 + 10\cos x$ by using the fixed-point iteration method for an appropriate iteration function g. Find the zeros accurate to within 10^{-4}.

14. Use a fixed-point iteration method to determine a solution accurate to within 10^{-4} for $x = \tan x$, for x in $[4, 5]$.

15. Use a fixed-point iteration method to determine a solution accurate to within 10^{-2} for $2 \sin \pi x + x = 0$ on $[1, 2]$. Use $p_0 = 1$.

16. Let A be a given positive constant and $g(x) = 2x - Ax^2$.

a. Show that if fixed-point iteration converges to a nonzero limit, then the limit is $p = 1/A$, so the inverse of a number can be found using only multiplications and subtractions.

b. Find an interval about $1/A$ for which fixed-point iteration converges, provided p_0 is in that interval.

17. Find a function g defined on $[0, 1]$ that satisfies none of the hypotheses of Theorem 2.3 but still has a unique fixed point on $[0, 1]$.

18. **a.** Show that Theorem 2.2 is true if the inequality $|g'(x)| \leq k$ is replaced by $g'(x) \leq k$, for all $x \in (a, b)$. [*Hint:* Only uniqueness is in question.]

b. Show that Theorem 2.3 may not hold if inequality $|g'(x)| \leq k$ is replaced by $g'(x) \leq k$. [*Hint:* Show that $g(x) = 1 - x^2$, for x in $[0, 1]$, provides a counterexample.]

19. **a.** Use Theorem 2.4 to show that the sequence defined by

$$x_n = \frac{1}{2}x_{n-1} + \frac{1}{x_{n-1}}, \quad \text{for } n \geq 1,$$

converges to $\sqrt{2}$ whenever $x_0 > \sqrt{2}$.

b. Use the fact that $0 < (x_0 - \sqrt{2})^2$ whenever $x_0 \neq \sqrt{2}$ to show that if $0 < x_0 < \sqrt{2}$, then $x_1 > \sqrt{2}$.

c. Use the results of parts (a) and (b) to show that the sequence in (a) converges to $\sqrt{2}$ whenever $x_0 > 0$.

20. **a.** Show that if A is any positive number, then the sequence defined by

$$x_n = \frac{1}{2}x_{n-1} + \frac{A}{2x_{n-1}}, \quad \text{for } n \geq 1,$$

converges to $\sqrt{A}$ whenever $x_0 > 0$.

b. What happens if $x_0 < 0$?

21. Replace the assumption in Theorem 2.4 that "a positive number $k < 1$ exists with $|g'(x)| \leq k$" with "g satisfies a Lipschitz condition on the interval $[a, b]$ with Lipschitz constant $L < 1$." (See Exercise 27, Section 1.1.) Show that the conclusions of this theorem are still valid.

22. Suppose that g is continuously differentiable on some interval (c, d) that contains the fixed point p of g. Show that if $|g'(p)| < 1$, then there exists a $\delta > 0$ such that if $|p_0 - p| \leq \delta$, then the fixed-point iteration converges.

23. An object falling vertically through the air is subjected to viscous resistance as well as to the force of gravity. Assume that an object with mass m is dropped from a height s_0 and that the height of the object after t seconds is

$$s(t) = s_0 - \frac{mg}{k}t + \frac{m^2 g}{k^2}(1 - e^{-kt/m}),$$

where $g = 32.17$ ft/s^2 and k represents the coefficient of air resistance in lb-s/ft. Suppose $s_0 = 300$ ft, $m = 0.25$ lb, and $k = 0.1$ lb-s/ft. Find, to within 0.01 s, the time it takes this quarter-pounder to hit the ground.

24. Let $g \in C^1[a, b]$ and p be in (a, b) with $g(p) = p$ and $|g'(p)| > 1$. Show that there exists a $\delta > 0$ such that if $0 < |p_0 - p| < \delta$, then $|p_0 - p| < |p_1 - p|$. Thus, no matter how close the initial approximation p_0 is to p, the next iterate p_1 is farther away, so the fixed-point iteration does not converge if $p_0 \neq p$.

2.3 Newton's Method and Its Extensions

Isaac Newton (1641–1727) was one of the most brilliant scientists of all time. The late 17th century was a vibrant period for science and mathematics and Newton's work touched nearly every aspect of mathematics. His method for solving was introduced to find a root of the equation $y^3 - 2y - 5 = 0$. Although he demonstrated the method only for polynomials, it is clear that he realized its broader applications.

Newton's (or the *Newton-Raphson*) **method** is one of the most powerful and well-known numerical methods for solving a root-finding problem. There are many ways of introducing Newton's method.

Newton's Method

If we only want an algorithm, we can consider the technique graphically, as is often done in calculus. Another possibility is to derive Newton's method as a technique to obtain faster convergence than offered by other types of functional iteration, as is done in Section 2.4. A third means of introducing Newton's method, which is discussed next, is based on Taylor polynomials. We will see there that this particular derivation produces not only the method, but also a bound for the error of the approximation.

Suppose that $f \in C^2[a, b]$. Let $p_0 \in [a, b]$ be an approximation to p such that $f'(p_0) \neq 0$ and $|p - p_0|$ is "small." Consider the first Taylor polynomial for $f(x)$ expanded about p_0 and evaluated at $x = p$.

$$f(p) = f(p_0) + (p - p_0)f'(p_0) + \frac{(p - p_0)^2}{2} f''(\xi(p)),$$

where $\xi(p)$ lies between p and p_0. Since $f(p) = 0$, this equation gives

$$0 = f(p_0) + (p - p_0)f'(p_0) + \frac{(p - p_0)^2}{2} f''(\xi(p)).$$

Newton's method is derived by assuming that since $|p - p_0|$ is small, the term involving $(p - p_0)^2$ is much smaller, so

$$0 \approx f(p_0) + (p - p_0)f'(p_0).$$

Solving for p gives

$$p \approx p_0 - \frac{f(p_0)}{f'(p_0)} \equiv p_1.$$

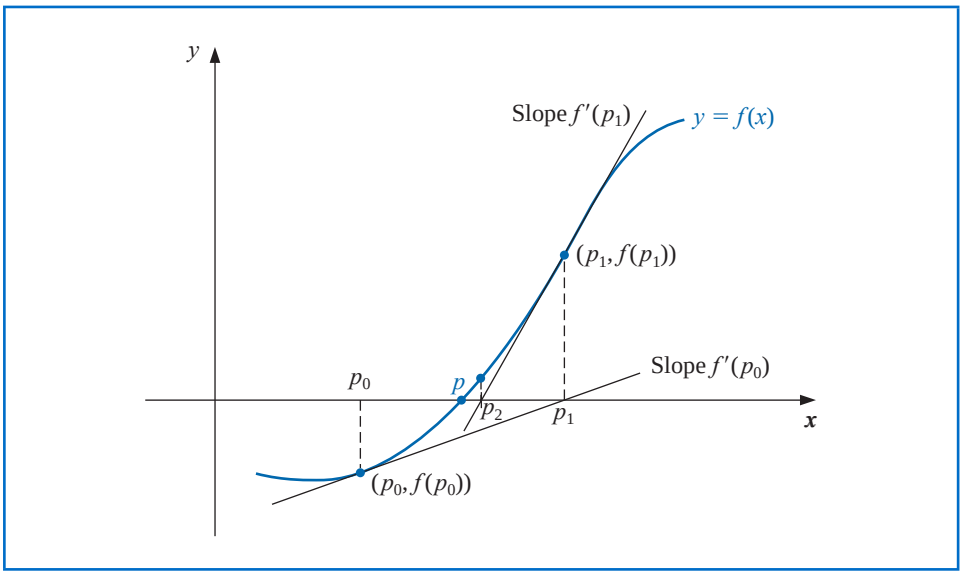
Joseph Raphson (1648–1715) gave a description of the method attributed to Isaac Newton in 1690, acknowledging Newton as the source of the discovery. Neither Newton nor Raphson explicitly used the derivative in their description since both considered only polynomials. Other mathematicians, particularly James Gregory (1636–1675), were aware of the underlying process at or before this time.

This sets the stage for Newton's method, which starts with an initial approximation p_0 and generates the sequence $\{p_n\}_{n=0}^{\infty}$, by

$$p_n = p_{n-1} - \frac{f(p_{n-1})}{f'(p_{n-1})}, \quad \text{for } n \geq 1. \tag{2.7}$$

Figure 2.8 on page 68 illustrates how the approximations are obtained using successive tangents. (Also see Exercise 15.) Starting with the initial approximation p_0, the approximation p_1 is the x-intercept of the tangent line to the graph of f at $(p_0, f(p_0))$. The approximation p_2 is the x-intercept of the tangent line to the graph of f at $(p_1, f(p_1))$ and so on. Algorithm 2.3 follows this procedure.

Figure 2.8



ALGORITHM
2.3

Newton's

To find a solution to $f(x) = 0$ given an initial approximation p_0:

INPUT initial approximation p_0; tolerance TOL; maximum number of iterations N_0.

OUTPUT approximate solution p or message of failure.

Step 1 Set $i = 1$.

Step 2 While $i \le N_0$ do Steps 3–6.

Step 3 Set $p = p_0 - f(p_0)/f'(p_0)$. (*Compute* p_i.)

Step 4 If $|p - p_0| < TOL$ then
OUTPUT (p); (*The procedure was successful.*)
STOP.

Step 5 Set $i = i + 1$.

Step 6 Set $p_0 = p$. (*Update* p_0.)

Step 7 OUTPUT ('The method failed after N_0 iterations, $N_0 =$', N_0);
(*The procedure was unsuccessful.*)
STOP. ■

The stopping-technique inequalities given with the Bisection method are applicable to Newton's method. That is, select a tolerance $\varepsilon > 0$, and construct $p_1, \ldots p_N$ until

$$|p_N - p_{N-1}| < \varepsilon, \tag{2.8}$$

$$\frac{|p_N - p_{N-1}|}{|p_N|} < \varepsilon, \quad p_N \neq 0, \tag{2.9}$$

or

$$|f(p_N)| < \varepsilon. \tag{2.10}$$

A form of Inequality (2.8) is used in Step 4 of Algorithm 2.3. Note that none of the inequalities (2.8), (2.9), or (2.10) give precise information about the actual error $|p_N - p|$. (See Exercises 16 and 17 in Section 2.1.)

Newton's method is a functional iteration technique with $p_n = g(p_{n-1})$, for which

$$g(p_{n-1}) = p_{n-1} - \frac{f(p_{n-1})}{f'(p_{n-1})}, \quad \text{for } n \geq 1. \tag{2.11}$$

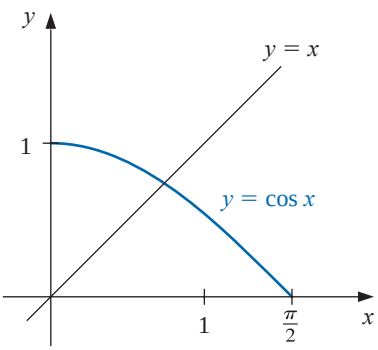
In fact, this is the functional iteration technique that was used to give the rapid convergence we saw in column (**e**) of Table 2.2 in Section 2.2.

It is clear from Equation (2.7) that Newton's method cannot be continued if $f'(p_{n-1}) = 0$ for some n. In fact, we will see that the method is most effective when f' is bounded away from zero near p.

Example 1 Consider the function $f(x) = \cos x - x = 0$. Approximate a root of f using **(a)** a fixed-point method, and **(b)** Newton's Method

Solution **(a)** A solution to this root-finding problem is also a solution to the fixed-point problem $x = \cos x$, and the graph in Figure 2.9 implies that a single fixed-point p lies in $[0, \pi/2]$.

Figure 2.9



Note that the variable in the trigonometric function is in radian measure, not degrees. This will always be the case unless specified otherwise.

Table 2.3

n	p_n
0	0.7853981635
1	0.7071067810
2	0.7602445972
3	0.7246674808
4	0.7487198858
5	0.7325608446
6	0.7434642113
7	0.7361282565

Table 2.4

	Newton's Method
n	p_n
0	0.7853981635
1	0.7395361337
2	0.7390851781
3	0.7390851332
4	0.7390851332

Table 2.3 shows the results of fixed-point iteration with $p_0 = \pi/4$. The best we could conclude from these results is that $p \approx 0.74$.

(b) To apply Newton's method to this problem we need $f'(x) = -\sin x - 1$. Starting again with $p_0 = \pi/4$, we generate the sequence defined, for $n \geq 1$, by

$$p_n = p_{n-1} - \frac{f(p_{n-1})}{f(p'_{n-1})} = p_{n-1} - \frac{\cos p_{n-1} - p_{n-1}}{-\sin p_{n-1} - 1}.$$

This gives the approximations in Table 2.4. An excellent approximation is obtained with $n = 3$. Because of the agreement of p_3 and p_4 we could reasonably expect this result to be accurate to the places listed. ■

Convergence using Newton's Method

Example 1 shows that Newton's method can provide extremely accurate approximations with very few iterations. For that example, only one iteration of Newton's method was needed to give better accuracy than 7 iterations of the fixed-point method. It is now time to examine Newton's method more carefully to discover why it is so effective.

The Taylor series derivation of Newton's method at the beginning of the section points out the importance of an accurate initial approximation. The crucial assumption is that the term involving $(p - p_0)^2$ is, by comparison with $|p - p_0|$, so small that it can be deleted. This will clearly be false unless p_0 is a good approximation to p. If p_0 is not sufficiently close to the actual root, there is little reason to suspect that Newton's method will converge to the root. However, in some instances, even poor initial approximations will produce convergence. (Exercises 20 and 21 illustrate some of these possibilities.)

The following convergence theorem for Newton's method illustrates the theoretical importance of the choice of p_0.

Theorem 2.6 Let $f \in C^2[a, b]$. If $p \in (a, b)$ is such that $f(p) = 0$ and $f'(p) \neq 0$, then there exists a $\delta > 0$ such that Newton's method generates a sequence $\{p_n\}_{n=1}^{\infty}$ converging to p for any initial approximation $p_0 \in [p - \delta, p + \delta]$. ■

Proof The proof is based on analyzing Newton's method as the functional iteration scheme $p_n = g(p_{n-1})$, for $n \geq 1$, with

$$g(x) = x - \frac{f(x)}{f'(x)}.$$

Let k be in $(0, 1)$. We first find an interval $[p - \delta, p + \delta]$ that g maps into itself and for which $|g'(x)| \leq k$, for all $x \in (p - \delta, p + \delta)$.

Since f' is continuous and $f'(p) \neq 0$, part (**a**) of Exercise 29 in Section 1.1 implies that there exists a $\delta_1 > 0$, such that $f'(x) \neq 0$ for $x \in [p - \delta_1, p + \delta_1] \subseteq [a, b]$. Thus g is defined and continuous on $[p - \delta_1, p + \delta_1]$. Also

$$g'(x) = 1 - \frac{f'(x)f'(x) - f(x)f''(x)}{[f'(x)]^2} = \frac{f(x)f''(x)}{[f'(x)]^2},$$

for $x \in [p - \delta_1, p + \delta_1]$, and, since $f \in C^2[a, b]$, we have $g \in C^1[p - \delta_1, p + \delta_1]$.

By assumption, $f(p) = 0$, so

$$g'(p) = \frac{f(p)f''(p)}{[f'(p)]^2} = 0.$$

Since g' is continuous and $0 < k < 1$, part (**b**) of Exercise 29 in Section 1.1 implies that there exists a δ, with $0 < \delta < \delta_1$, and

$$|g'(x)| \leq k, \quad \text{for all } x \in [p - \delta, p + \delta].$$

It remains to show that g maps $[p - \delta, p + \delta]$ into $[p - \delta, p + \delta]$. If $x \in [p - \delta, p + \delta]$, the Mean Value Theorem implies that for some number ξ between x and p, $|g(x) - g(p)| = |g'(\xi)||x - p|$. So

$$|g(x) - p| = |g(x) - g(p)| = |g'(\xi)||x - p| \leq k|x - p| < |x - p|.$$

Since $x \in [p - \delta, p + \delta]$, it follows that $|x - p| < \delta$ and that $|g(x) - p| < \delta$. Hence, g maps $[p - \delta, p + \delta]$ into $[p - \delta, p + \delta]$.

All the hypotheses of the Fixed-Point Theorem 2.4 are now satisfied, so the sequence $\{p_n\}_{n=1}^{\infty}$, defined by

$$p_n = g(p_{n-1}) = p_{n-1} - \frac{f(p_{n-1})}{f'(p_{n-1})}, \quad \text{for } n \geq 1,$$

converges to p for any $p_0 \in [p - \delta, p + \delta]$. ■ ■ ■

Theorem 2.6 states that, under reasonable assumptions, Newton's method converges provided a sufficiently accurate initial approximation is chosen. It also implies that the constant k that bounds the derivative of g, and, consequently, indicates the speed of convergence of the method, decreases to 0 as the procedure continues. This result is important for the theory of Newton's method, but it is seldom applied in practice because it does not tell us how to determine δ.

In a practical application, an initial approximation is selected and successive approximations are generated by Newton's method. These will generally either converge quickly to the root, or it will be clear that convergence is unlikely.

The Secant Method

Newton's method is an extremely powerful technique, but it has a major weakness: the need to know the value of the derivative of f at each approximation. Frequently, $f'(x)$ is far more difficult and needs more arithmetic operations to calculate than $f(x)$.

To circumvent the problem of the derivative evaluation in Newton's method, we introduce a slight variation. By definition,

$$f'(p_{n-1}) = \lim_{x \to p_{n-1}} \frac{f(x) - f(p_{n-1})}{x - p_{n-1}}.$$

If p_{n-2} is close to p_{n-1}, then

$$f'(p_{n-1}) \approx \frac{f(p_{n-2}) - f(p_{n-1})}{p_{n-2} - p_{n-1}} = \frac{f(p_{n-1}) - f(p_{n-2})}{p_{n-1} - p_{n-2}}.$$

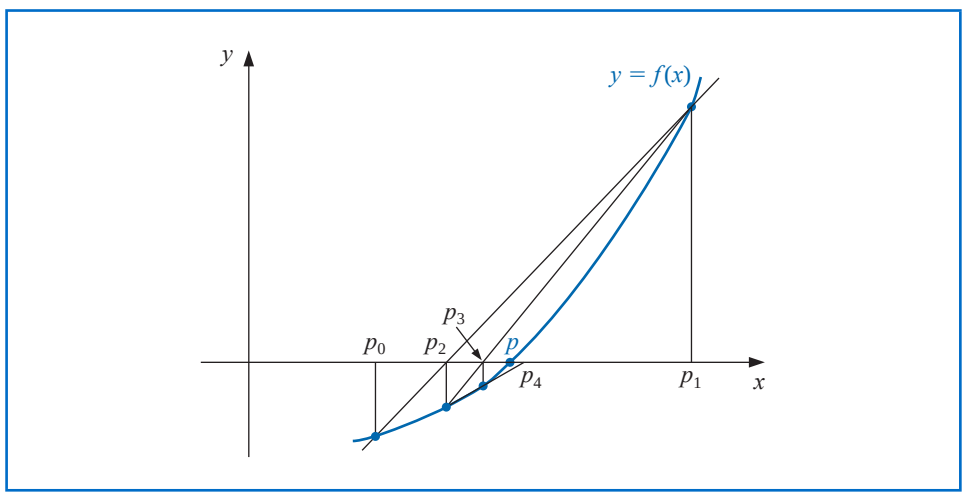
Using this approximation for $f'(p_{n-1})$ in Newton's formula gives

$$p_n = p_{n-1} - \frac{f(p_{n-1})(p_{n-1} - p_{n-2})}{f(p_{n-1}) - f(p_{n-2})}. \tag{2.12}$$

The word secant is derived from the Latin word *secan*, which means to cut. The secant method uses a secant line, a line joining two points that cut the curve, to approximate a root.

This technique is called the **Secant method** and is presented in Algorithm 2.4. (See Figure 2.10.) Starting with the two initial approximations p_0 and p_1, the approximation p_2 is the x-intercept of the line joining $(p_0, f(p_0))$ and $(p_1, f(p_1))$. The approximation p_3 is the x-intercept of the line joining $(p_1, f(p_1))$ and $(p_2, f(p_2))$, and so on. Note that only one function evaluation is needed per step for the Secant method after p_2 has been determined. In contrast, each step of Newton's method requires an evaluation of both the function and its derivative.

Figure 2.10



ALGORITHM
2.4

Secant

To find a solution to $f(x) = 0$ given initial approximations p_0 and p_1:

INPUT initial approximations p_0, p_1; tolerance TOL; maximum number of iterations N_0.

OUTPUT approximate solution p or message of failure.

Step 1 Set $i = 2$;
$q_0 = f(p_0)$;
$q_1 = f(p_1)$.

Step 2 While $i \le N_0$ do Steps 3–6.

Step 3 Set $p = p_1 - q_1(p_1 - p_0)/(q_1 - q_0)$. (*Compute* p_i.)

Step 4 If $|p - p_1| < TOL$ then
OUTPUT (p); (*The procedure was successful.*)
STOP.

Step 5 Set $i = i + 1$.

Step 6 Set $p_0 = p_1$; (*Update* p_0, q_0, p_1, q_1.)
$q_0 = q_1$;
$p_1 = p$;
$q_1 = f(p)$.

Step 7 OUTPUT ('The method failed after N_0 iterations, $N_0 =$', N_0);
(*The procedure was unsuccessful.*)
STOP. ■

The next example involves a problem considered in Example 1, where we used Newton's method with $p_0 = \pi/4$.

Example 2 Use the Secant method to find a solution to $x = \cos x$, and compare the approximations with those given in Example 1 which applied Newton's method.

Solution In Example 1 we compared fixed-point iteration and Newton's method starting with the initial approximation $p_0 = \pi/4$. For the Secant method we need two initial approximations. Suppose we use $p_0 = 0.5$ and $p_1 = \pi/4$. Succeeding approximations are generated by the formula

$$p_n = p_{n-1} - \frac{(p_{n-1} - p_{n-2})(\cos p_{n-1} - p_{n-1})}{(\cos p_{n-1} - p_{n-1}) - (\cos p_{n-2} - p_{n-2})}, \quad \text{for } n \ge 2.$$

These give the results in Table 2.5. ■

Table 2.5

	Secant
n	p_n
0	0.5
1	0.7853981635
2	0.7363841388
3	0.7390581392
4	0.7390851493
5	0.7390851332

	Newton
n	p_n
0	0.7853981635
1	0.7395361337
2	0.7390851781
3	0.7390851332
4	0.7390851332

Comparing the results in Table 2.5 from the Secant method and Newton's method, we see that the Secant method approximation p_5 is accurate to the tenth decimal place, whereas Newton's method obtained this accuracy by p_3. For this example, the convergence of the Secant method is much faster than functional iteration but slightly slower than Newton's method. This is generally the case. (See Exercise 14 of Section 2.4.)

Newton's method or the Secant method is often used to refine an answer obtained by another technique, such as the Bisection method, since these methods require good first approximations but generally give rapid convergence.

The Method of False Position

Each successive pair of approximations in the Bisection method brackets a root p of the equation; that is, for each positive integer n, a root lies between a_n and b_n. This implies that, for each n, the Bisection method iterations satisfy

$$|p_n - p| < \frac{1}{2}|a_n - b_n|,$$

which provides an easily calculated error bound for the approximations.

Root bracketing is not guaranteed for either Newton's method or the Secant method. In Example 1, Newton's method was applied to $f(x) = \cos x - x$, and an approximate root was found to be 0.7390851332. Table 2.5 shows that this root is not bracketed by either p_0 and p_1 or p_1 and p_2. The Secant method approximations for this problem are also given in Table 2.5. In this case the initial approximations p_0 and p_1 bracket the root, but the pair of approximations p_3 and p_4 fail to do so.

The term *Regula Falsi*, literally a false rule or false position, refers to a technique that uses results that are known to be false, but in some specific manner, to obtain convergence to a true result. False position problems can be found on the Rhind papyrus, which dates from about 1650 B.C.E.

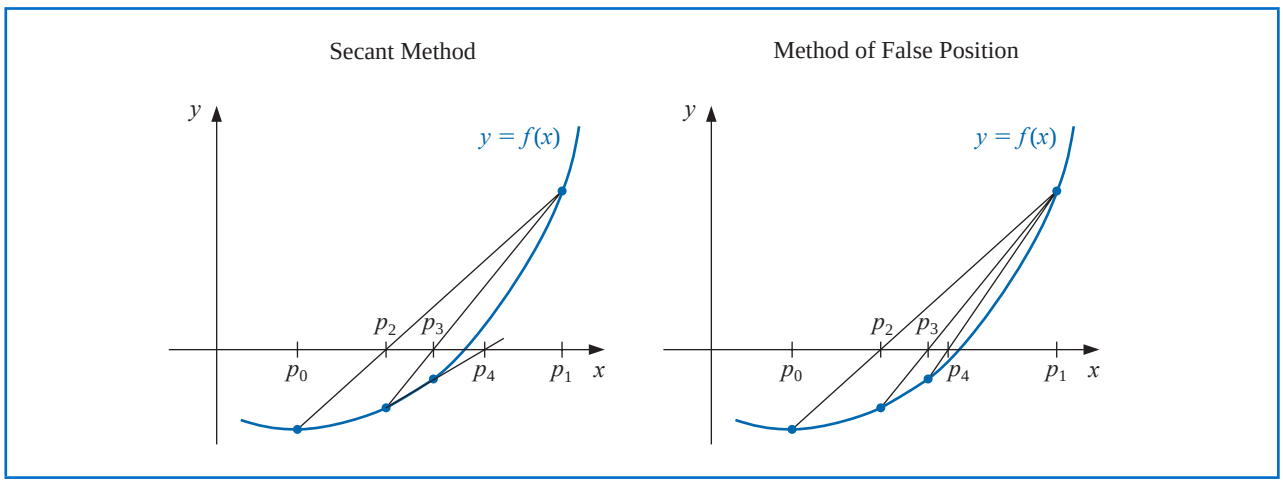
The **method of False Position** (also called *Regula Falsi*) generates approximations in the same manner as the Secant method, but it includes a test to ensure that the root is always bracketed between successive iterations. Although it is not a method we generally recommend, it illustrates how bracketing can be incorporated.

First choose initial approximations p_0 and p_1 with $f(p_0) \cdot f(p_1) < 0$. The approximation p_2 is chosen in the same manner as in the Secant method, as the x-intercept of the line joining $(p_0, f(p_0))$ and $(p_1, f(p_1))$. To decide which secant line to use to compute p_3, consider $f(p_2) \cdot f(p_1)$, or more correctly $\operatorname{sgn} f(p_2) \cdot \operatorname{sgn} f(p_1)$.

- If $\operatorname{sgn} f(p_2) \cdot \operatorname{sgn} f(p_1) < 0$, then p_1 and p_2 bracket a root. Choose p_3 as the x-intercept of the line joining $(p_1, f(p_1))$ and $(p_2, f(p_2))$.
- If not, choose p_3 as the x-intercept of the line joining $(p_0, f(p_0))$ and $(p_2, f(p_2))$, and then interchange the indices on p_0 and p_1.

In a similar manner, once p_3 is found, the sign of $f(p_3) \cdot f(p_2)$ determines whether we use p_2 and p_3 or p_3 and p_1 to compute p_4. In the latter case a relabeling of p_2 and p_1 is performed. The relabeling ensures that the root is bracketed between successive iterations. The process is described in Algorithm 2.5, and Figure 2.11 shows how the iterations can differ from those of the Secant method. In this illustration, the first three approximations are the same, but the fourth approximations differ.

Figure 2.11



ALGORITHM
2.5

False Position

To find a solution to $f(x) = 0$ given the continuous function f on the interval $[p_0, p_1]$ where $f(p_0)$ and $f(p_1)$ have opposite signs:

INPUT initial approximations p_0, p_1; tolerance TOL; maximum number of iterations N_0.

OUTPUT approximate solution p or message of failure.

Step 1 Set $i = 2$;
$q_0 = f(p_0)$;
$q_1 = f(p_1)$.

Step 2 While $i \le N_0$ do Steps 3–7.

Step 3 Set $p = p_1 - q_1(p_1 - p_0)/(q_1 - q_0)$. (*Compute* p_i.)

Step 4 If $|p - p_1| < TOL$ then
OUTPUT (p); (*The procedure was successful.*)
STOP.

Step 5 Set $i = i + 1$;
$q = f(p)$.

Step 6 If $q \cdot q_1 < 0$ then set $p_0 = p_1$;
$q_0 = q_1$.

Step 7 Set $p_1 = p$;
$q_1 = q$.

Step 8 OUTPUT ('Method failed after N_0 iterations, $N_0 =$', N_0);
(*The procedure unsuccessful.*)
STOP. ■

Example 3 Use the method of False Position to find a solution to $x = \cos x$, and compare the approximations with those given in Example 1 which applied fixed-point iteration and Newton's method, and to those found in Example 2 which applied the Secant method.

Solution To make a reasonable comparison we will use the same initial approximations as in the Secant method, that is, $p_0 = 0.5$ and $p_1 = \pi/4$. Table 2.6 shows the results of the method of False Position applied to $f(x) = \cos x - x$ together with those we obtained using the Secant and Newton's methods. Notice that the False Position and Secant approximations agree through p_3 and that the method of False Position requires an additional iteration to obtain the same accuracy as the Secant method. ■

Table 2.6

	False Position	Secant	Newton
n	p_n	p_n	p_n
0	0.5	0.5	0.7853981635
1	0.7853981635	0.7853981635	0.7395361337
2	0.7363841388	0.7363841388	0.7390851781
3	0.7390581392	0.7390581392	0.7390851332
4	0.7390848638	0.7390851493	0.7390851332
5	0.7390851305	0.7390851332	
6	0.7390851332		

The added insurance of the method of False Position commonly requires more calculation than the Secant method, just as the simplification that the Secant method provides over Newton's method usually comes at the expense of additional iterations. Further examples of the positive and negative features of these methods can be seen by working Exercises 17 and 18.

Maple has Newton's method, the Secant method, and the method of False Position implemented in its *NumericalAnalysis* package. The options that were available for the Bisection method are also available for these techniques. For example, to generate the results in Tables 2.4, 2.5, and 2.6 we could use the commands

with(*Student*[*NumericalAnalysis*])

$f := \cos(x) - x$

$Newton\left(f, x = \frac{\pi}{4.0}, tolerance = 10^{-8}, output = sequence, maxiterations = 20\right)$

$Secant\left(f, x = \left[0.5, \frac{\pi}{4.0}\right], tolerance = 10^{-8}, output = sequence, maxiterations = 20\right)$

and

$FalsePosition\left(f, x = \left[0.5, \frac{\pi}{4.0}\right], tolerance = 10^{-8}, output = sequence, maxiterations = 20\right)$

EXERCISE SET 2.3

1. Let $f(x) = x^2 - 6$ and $p_0 = 1$. Use Newton's method to find p_2.

2. Let $f(x) = -x^3 - \cos x$ and $p_0 = -1$. Use Newton's method to find p_2. Could $p_0 = 0$ be used?

3. Let $f(x) = x^2 - 6$. With $p_0 = 3$ and $p_1 = 2$, find p_3.

a. Use the Secant method.

b. Use the method of False Position.

c. Which of **a.** or **b.** is closer to $\sqrt{6}$?

4. Let $f(x) = -x^3 - \cos x$. With $p_0 = -1$ and $p_1 = 0$, find p_3.

a. Use the Secant method.

b. Use the method of False Position.

5. Use Newton's method to find solutions accurate to within 10^{-4} for the following problems.

a. $x^3 - 2x^2 - 5 = 0, \quad [1, 4]$

b. $x^3 + 3x^2 - 1 = 0, \quad [-3, -2]$

c. $x - \cos x = 0, \quad [0, \pi/2]$

d. $x - 0.8 - 0.2 \sin x = 0, \quad [0, \pi/2]$

6. Use Newton's method to find solutions accurate to within 10^{-5} for the following problems.

a. $e^x + 2^{-x} + 2\cos x - 6 = 0 \quad$ for $1 \le x \le 2$

b. $\ln(x - 1) + \cos(x - 1) = 0 \quad$ for $1.3 \le x \le 2$

c. $2x \cos 2x - (x - 2)^2 = 0 \quad$ for $2 \le x \le 3$ and $3 \le x \le 4$

d. $(x - 2)^2 - \ln x = 0 \quad$ for $1 \le x \le 2$ and $e \le x \le 4$

e. $e^x - 3x^2 = 0 \quad$ for $0 \le x \le 1$ and $3 \le x \le 5$

f. $\sin x - e^{-x} = 0 \quad$ for $0 \le x \le 1 \;\; 3 \le x \le 4$ and $6 \le x \le 7$

7. Repeat Exercise 5 using the Secant method.

8. Repeat Exercise 6 using the Secant method.

9. Repeat Exercise 5 using the method of False Position.

10. Repeat Exercise 6 using the method of False Position.

11. Use all three methods in this Section to find solutions to within 10^{-5} for the following problems.

a. $3xe^x = 0 \quad$ for $1 \le x \le 2$

b. $2x + 3\cos x - e^x = 0 \quad$ for $0 \le x \le 1$

12. Use all three methods in this Section to find solutions to within 10^{-7} for the following problems.

a. $x^2 - 4x + 4 - \ln x = 0$ for $1 \le x \le 2$ and for $2 \le x \le 4$

b. $x + 1 - 2\sin \pi x = 0$ for $0 \le x \le 1/2$ and for $1/2 \le x \le 1$

13. Use Newton's method to approximate, to within 10^{-4}, the value of x that produces the point on the graph of $y = x^2$ that is closest to (1, 0). [*Hint:* Minimize $[d(x)]^2$, where $d(x)$ represents the distance from (x, x^2) to (1, 0).]

14. Use Newton's method to approximate, to within 10^{-4}, the value of x that produces the point on the graph of $y = 1/x$ that is closest to (2, 1).

15. The following describes Newton's method graphically: Suppose that $f'(x)$ exists on $[a, b]$ and that $f'(x) \neq 0$ on $[a, b]$. Further, suppose there exists one $p \in [a, b]$ such that $f(p) = 0$, and let $p_0 \in [a, b]$ be arbitrary. Let p_1 be the point at which the tangent line to f at $(p_0, f(p_0))$ crosses the x-axis. For each $n \ge 1$, let p_n be the x-intercept of the line tangent to f at $(p_{n-1}, f(p_{n-1}))$. Derive the formula describing this method.

16. Use Newton's method to solve the equation

$$0 = \frac{1}{2} + \frac{1}{4}x^2 - x\sin x - \frac{1}{2}\cos 2x, \quad \text{with } p_0 = \frac{\pi}{2}.$$

Iterate using Newton's method until an accuracy of 10^{-5} is obtained. Explain why the result seems unusual for Newton's method. Also, solve the equation with $p_0 = 5\pi$ and $p_0 = 10\pi$.

17. The fourth-degree polynomial

$$f(x) = 230x^4 + 18x^3 + 9x^2 - 221x - 9$$

has two real zeros, one in $[-1, 0]$ and the other in $[0, 1]$. Attempt to approximate these zeros to within 10^{-6} using the

a. Method of False Position

b. Secant method

c. Newton's method

Use the endpoints of each interval as the initial approximations in (a) and (b) and the midpoints as the initial approximation in (c).

18. The function $f(x) = \tan \pi x - 6$ has a zero at $(1/\pi)\arctan 6 \approx 0.447431543$. Let $p_0 = 0$ and $p_1 = 0.48$, and use ten iterations of each of the following methods to approximate this root. Which method is most successful and why?

a. Bisection method

b. Method of False Position

c. Secant method

19. The iteration equation for the Secant method can be written in the simpler form

$$p_n = \frac{f(p_{n-1})p_{n-2} - f(p_{n-2})p_{n-1}}{f(p_{n-1}) - f(p_{n-2})}.$$

Explain why, in general, this iteration equation is likely to be less accurate than the one given in Algorithm 2.4.

20. The equation $x^2 - 10\cos x = 0$ has two solutions, ± 1.3793646. Use Newton's method to approximate the solutions to within 10^{-5} with the following values of p_0.

a. $p_0 = -100$ **b.** $p_0 = -50$ **c.** $p_0 = -25$

d. $p_0 = 25$ **e.** $p_0 = 50$ **f.** $p_0 = 100$

21. The equation $4x^2 - e^x - e^{-x} = 0$ has two positive solutions x_1 and x_2. Use Newton's method to approximate the solution to within 10^{-5} with the following values of p_0.

a. $p_0 = -10$ **b.** $p_0 = -5$ **c.** $p_0 = -3$
d. $p_0 = -1$ **e.** $p_0 = 0$ **f.** $p_0 = 1$
g. $p_0 = 3$ **h.** $p_0 = 5$ **i.** $p_0 = 10$

22. Use Maple to determine how many iterations of Newton's method with $p_0 = \pi/4$ are needed to find a root of $f(x) = \cos x - x$ to within 10^{-100}.

23. The function described by $f(x) = \ln(x^2 + 1) - e^{0.4x} \cos \pi x$ has an infinite number of zeros.

a. Determine, within 10^{-6}, the only negative zero.

b. Determine, within 10^{-6}, the four smallest positive zeros.

c. Determine a reasonable initial approximation to find the nth smallest positive zero of f. [*Hint:* Sketch an approximate graph of f.]

d. Use part (c) to determine, within 10^{-6}, the 25th smallest positive zero of f.

24. Find an approximation for λ, accurate to within 10^{-4}, for the population equation

$$1{,}564{,}000 = 1{,}000{,}000e^{\lambda} + \frac{435{,}000}{\lambda}(e^{\lambda} - 1),$$

discussed in the introduction to this chapter. Use this value to predict the population at the end of the second year, assuming that the immigration rate during this year remains at 435,000 individuals per year.

25. The sum of two numbers is 20. If each number is added to its square root, the product of the two sums is 155.55. Determine the two numbers to within 10^{-4}.

26. The accumulated value of a savings account based on regular periodic payments can be determined from the *annuity due equation*,

$$A = \frac{P}{i}[(1 + i)^n - 1].$$

In this equation, A is the amount in the account, P is the amount regularly deposited, and i is the rate of interest per period for the n deposit periods. An engineer would like to have a savings account valued at \$750,000 upon retirement in 20 years and can afford to put \$1500 per month toward this goal. What is the minimal interest rate at which this amount can be invested, assuming that the interest is compounded monthly?

27. Problems involving the amount of money required to pay off a mortgage over a fixed period of time involve the formula

$$A = \frac{P}{i}[1 - (1 + i)^{-n}],$$

known as an *ordinary annuity equation.* In this equation, A is the amount of the mortgage, P is the amount of each payment, and i is the interest rate per period for the n payment periods. Suppose that a 30-year home mortgage in the amount of \$135,000 is needed and that the borrower can afford house payments of at most \$1000 per month. What is the maximal interest rate the borrower can afford to pay?

28. A drug administered to a patient produces a concentration in the blood stream given by $c(t) = Ate^{-t/3}$ milligrams per milliliter, t hours after A units have been injected. The maximum safe concentration is 1 mg/mL.

a. What amount should be injected to reach this maximum safe concentration, and when does this maximum occur?

b. An additional amount of this drug is to be administered to the patient after the concentration falls to 0.25 mg/mL. Determine, to the nearest minute, when this second injection should be given.

c. Assume that the concentration from consecutive injections is additive and that 75% of the amount originally injected is administered in the second injection. When is it time for the third injection?

29. Let $f(x) = 3^{3x+1} - 7 \cdot 5^{2x}$.

a. Use the Maple commands *solve* and *fsolve* to try to find all roots of f.

b. Plot $f(x)$ to find initial approximations to roots of f.

c. Use Newton's method to find roots of f to within 10^{-16}.

d. Find the exact solutions of $f(x) = 0$ without using Maple.

30. Repeat Exercise 29 using $f(x) = 2^{x^2} - 3 \cdot 7^{x+1}$.

31. The logistic population growth model is described by an equation of the form

$$P(t) = \frac{P_L}{1 - ce^{-kt}},$$

where P_L, c, and $k > 0$ are constants, and $P(t)$ is the population at time t. P_L represents the limiting value of the population since $\lim_{t\to\infty} P(t) = P_L$. Use the census data for the years 1950, 1960, and 1970 listed in the table on page 105 to determine the constants P_L, c, and k for a logistic growth model. Use the logistic model to predict the population of the United States in 1980 and in 2010, assuming $t = 0$ at 1950. Compare the 1980 prediction to the actual value.

32. The Gompertz population growth model is described by

$$P(t) = P_L e^{-ce^{-kt}},$$

where P_L, c, and $k > 0$ are constants, and $P(t)$ is the population at time t. Repeat Exercise 31 using the Gompertz growth model in place of the logistic model.

33. Player A will shut out (win by a score of 21–0) player B in a game of racquetball with probability

$$P = \frac{1+p}{2}\left(\frac{p}{1-p+p^2}\right)^{21},$$

where p denotes the probability A will win any specific rally (independent of the server). (See [Keller, J], p. 267.) Determine, to within 10^{-3}, the minimal value of p that will ensure that A will shut out B in at least half the matches they play.

34. In the design of all-terrain vehicles, it is necessary to consider the failure of the vehicle when attempting to negotiate two types of obstacles. One type of failure is called *hang-up failure* and occurs when the vehicle attempts to cross an obstacle that causes the bottom of the vehicle to touch the ground. The other type of failure is called *nose-in failure* and occurs when the vehicle descends into a ditch and its nose touches the ground.

The accompanying figure, adapted from [Bek], shows the components associated with the nose-in failure of a vehicle. In that reference it is shown that the maximum angle α that can be negotiated by a vehicle when β is the maximum angle at which hang-up failure does *not* occur satisfies the equation

$$A \sin\alpha \cos\alpha + B \sin^2\alpha - C \cos\alpha - E \sin\alpha = 0,$$

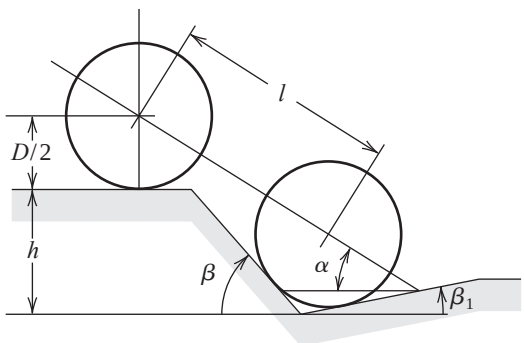
where

$$A = l \sin\beta_1, \quad B = l\cos\beta_1, \quad C = (h + 0.5D)\sin\beta_1 - 0.5D\tan\beta_1,$$

and $$E = (h + 0.5D)\cos\beta_1 - 0.5D.$$

a. It is stated that when $l = 89$ in., $h = 49$ in., $D = 55$ in., and $\beta_1 = 11.5°$, angle α is approximately 33°. Verify this result.

b. Find α for the situation when l, h, and β_1 are the same as in part (a) but $D = 30$ in.



2.4 Error Analysis for Iterative Methods

In this section we investigate the order of convergence of functional iteration schemes and, as a means of obtaining rapid convergence, rediscover Newton's method. We also consider ways of accelerating the convergence of Newton's method in special circumstances. First, however, we need a new procedure for measuring how rapidly a sequence converges.

Order of Convergence

Definition 2.7 Suppose $\{p_n\}_{n=0}^{\infty}$ is a sequence that converges to p, with $p_n \neq p$ for all n. If positive constants λ and α exist with

$$\lim_{n\to\infty} \frac{|p_{n+1} - p|}{|p_n - p|^{\alpha}} = \lambda,$$

then $\{p_n\}_{n=0}^{\infty}$ **converges to p of order α, with asymptotic error constant λ.** ■

An iterative technique of the form $p_n = g(p_{n-1})$ is said to be of *order* α if the sequence $\{p_n\}_{n=0}^{\infty}$ converges to the solution $p = g(p)$ of order α.

In general, a sequence with a high order of convergence converges more rapidly than a sequence with a lower order. The asymptotic constant affects the speed of convergence but not to the extent of the order. Two cases of order are given special attention.

(i) If $\alpha = 1$ (and $\lambda < 1$), the sequence is **linearly convergent**.

(ii) If $\alpha = 2$, the sequence is **quadratically convergent**.

The next illustration compares a linearly convergent sequence to one that is quadratically convergent. It shows why we try to find methods that produce higher-order convergent sequences.

Illustration Suppose that $\{p_n\}_{n=0}^{\infty}$ is linearly convergent to 0 with

$$\lim_{n\to\infty} \frac{|p_{n+1}|}{|p_n|} = 0.5$$

and that $\{\tilde{p}_n\}_{n=0}^{\infty}$ is quadratically convergent to 0 with the same asymptotic error constant,

$$\lim_{n\to\infty} \frac{|\tilde{p}_{n+1}|}{|\tilde{p}_n|^2} = 0.5.$$

For simplicity we assume that for each n we have

$$\frac{|p_{n+1}|}{|p_n|} \approx 0.5 \quad \text{and} \quad \frac{|\tilde{p}_{n+1}|}{|\tilde{p}_n|^2} \approx 0.5.$$

For the linearly convergent scheme, this means that

$$|p_n - 0| = |p_n| \approx 0.5|p_{n-1}| \approx (0.5)^2|p_{n-2}| \approx \cdots \approx (0.5)^n|p_0|,$$

whereas the quadratically convergent procedure has

$$\begin{aligned}|\tilde{p}_n - 0| = |\tilde{p}_n| &\approx 0.5|\tilde{p}_{n-1}|^2 \approx (0.5)[0.5|\tilde{p}_{n-2}|^2]^2 = (0.5)^3|\tilde{p}_{n-2}|^4 \\ &\approx (0.5)^3[(0.5)|\tilde{p}_{n-3}|^2]^4 = (0.5)^7|\tilde{p}_{n-3}|^8 \\ &\approx \cdots \approx (0.5)^{2^n-1}|\tilde{p}_0|^{2^n}.\end{aligned}$$

Table 2.7 illustrates the relative speed of convergence of the sequences to 0 if $|p_0| = |\tilde{p}_0| = 1$.

Table 2.7

n	Linear Convergence Sequence $\{p_n\}_{n=0}^{\infty}$ $(0.5)^n$	Quadratic Convergence Sequence $\{\tilde{p}_n\}_{n=0}^{\infty}$ $(0.5)^{2^n-1}$
1	5.0000×10^{-1}	5.0000×10^{-1}
2	2.5000×10^{-1}	1.2500×10^{-1}
3	1.2500×10^{-1}	7.8125×10^{-3}
4	6.2500×10^{-2}	3.0518×10^{-5}
5	3.1250×10^{-2}	4.6566×10^{-10}
6	1.5625×10^{-2}	1.0842×10^{-19}
7	7.8125×10^{-3}	5.8775×10^{-39}

The quadratically convergent sequence is within 10^{-38} of 0 by the seventh term. At least 126 terms are needed to ensure this accuracy for the linearly convergent sequence. □

Quadratically convergent sequences are expected to converge much quicker than those that converge only linearly, but the next result implies that an arbitrary technique that generates a convergent sequences does so only linearly.

Theorem 2.8 Let $g \in C[a, b]$ be such that $g(x) \in [a, b]$, for all $x \in [a, b]$. Suppose, in addition, that g' is continuous on (a, b) and a positive constant $k < 1$ exists with

$$|g'(x)| \leq k, \quad \text{for all } x \in (a, b).$$

If $g'(p) \neq 0$, then for any number $p_0 \neq p$ in $[a, b]$, the sequence

$$p_n = g(p_{n-1}), \quad \text{for } n \geq 1,$$

converges only linearly to the unique fixed point p in $[a, b]$. ■

Proof We know from the Fixed-Point Theorem 2.4 in Section 2.2 that the sequence converges to p. Since g' exists on (a, b), we can apply the Mean Value Theorem to g to show that for any n,

$$p_{n+1} - p = g(p_n) - g(p) = g'(\xi_n)(p_n - p),$$

where ξ_n is between p_n and p. Since $\{p_n\}_{n=0}^{\infty}$ converges to p, we also have $\{\xi_n\}_{n=0}^{\infty}$ converging to p. Since g' is continuous on (a, b), we have

$$\lim_{n\to\infty} g'(\xi_n) = g'(p).$$

Thus

$$\lim_{n\to\infty} \frac{p_{n+1} - p}{p_n - p} = \lim_{n\to\infty} g'(\xi_n) = g'(p) \quad \text{and} \quad \lim_{n\to\infty} \frac{|p_{n+1} - p|}{|p_n - p|} = |g'(p)|.$$

Hence, if $g'(p) \neq 0$, fixed-point iteration exhibits linear convergence with asymptotic error constant $|g'(p)|$. ■ ■ ■

Theorem 2.8 implies that higher-order convergence for fixed-point methods of the form $g(p) = p$ can occur only when $g'(p) = 0$. The next result describes additional conditions that ensure the quadratic convergence we seek.

Theorem 2.9 Let p be a solution of the equation $x = g(x)$. Suppose that $g'(p) = 0$ and g'' is continuous with $|g''(x)| < M$ on an open interval I containing p. Then there exists a $\delta > 0$ such that, for $p_0 \in [p - \delta, p + \delta]$, the sequence defined by $p_n = g(p_{n-1})$, when $n \geq 1$, converges at least quadratically to p. Moreover, for sufficiently large values of n,

$$|p_{n+1} - p| < \frac{M}{2}|p_n - p|^2.$$

■

Proof Choose k in $(0, 1)$ and $\delta > 0$ such that on the interval $[p-\delta, p+\delta]$, contained in I, we have $|g'(x)| \leq k$ and g'' continuous. Since $|g'(x)| \leq k < 1$, the argument used in the proof of Theorem 2.6 in Section 2.3 shows that the terms of the sequence $\{p_n\}_{n=0}^{\infty}$ are contained in $[p - \delta, p + \delta]$. Expanding $g(x)$ in a linear Taylor polynomial for $x \in [p - \delta, p + \delta]$ gives

$$g(x) = g(p) + g'(p)(x - p) + \frac{g''(\xi)}{2}(x - p)^2,$$

where ξ lies between x and p. The hypotheses $g(p) = p$ and $g'(p) = 0$ imply that

$$g(x) = p + \frac{g''(\xi)}{2}(x - p)^2.$$

In particular, when $x = p_n$,

$$p_{n+1} = g(p_n) = p + \frac{g''(\xi_n)}{2}(p_n - p)^2,$$

with ξ_n between p_n and p. Thus,

$$p_{n+1} - p = \frac{g''(\xi_n)}{2}(p_n - p)^2.$$

Since $|g'(x)| \leq k < 1$ on $[p - \delta, p + \delta]$ and g maps $[p - \delta, p + \delta]$ into itself, it follows from the Fixed-Point Theorem that $\{p_n\}_{n=0}^{\infty}$ converges to p. But ξ_n is between p and p_n for each n, so $\{\xi_n\}_{n=0}^{\infty}$ also converges to p, and

$$\lim_{n\to\infty} \frac{|p_{n+1} - p|}{|p_n - p|^2} = \frac{|g''(p)|}{2}.$$

This result implies that the sequence $\{p_n\}_{n=0}^{\infty}$ is quadratically convergent if $g''(p) \neq 0$ and of higher-order convergence if $g''(p) = 0$.

Because g'' is continuous and strictly bounded by M on the interval $[p - \delta, p + \delta]$, this also implies that, for sufficiently large values of n,

$$|p_{n+1} - p| < \frac{M}{2}|p_n - p|^2.$$

■ ■ ■

Theorems 2.8 and 2.9 tell us that our search for quadratically convergent fixed-point methods should point in the direction of functions whose derivatives are zero at the fixed point. That is:

- For a fixed point method to converge quadratically we need to have both $g(p) = p$, and $g'(p) = 0$.

The easiest way to construct a fixed-point problem associated with a root-finding problem $f(x) = 0$ is to add or subtract a multiple of $f(x)$ from x. Consider the sequence

$$p_n = g(p_{n-1}), \quad \text{for } n \geq 1,$$

for g in the form

$$g(x) = x - \phi(x) f(x),$$

where ϕ is a differentiable function that will be chosen later.

For the iterative procedure derived from g to be quadratically convergent, we need to have $g'(p) = 0$ when $f(p) = 0$. Because

$$g'(x) = 1 - \phi'(x) f(x) - f'(x)\phi(x),$$

and $f(p) = 0$, we have

$$g'(p) = 1 - \phi'(p) f(p) - f'(p)\phi(p) = 1 - \phi'(p) \cdot 0 - f'(p)\phi(p) = 1 - f'(p)\phi(p),$$

and $g'(p) = 0$ if and only if $\phi(p) = 1/f'(p)$.

If we let $\phi(x) = 1/f'(x)$, then we will ensure that $\phi(p) = 1/f'(p)$ and produce the quadratically convergent procedure

$$p_n = g(p_{n-1}) = p_{n-1} - \frac{f(p_{n-1})}{f'(p_{n-1})}.$$

This, of course, is simply Newton's method. Hence

- If $f(p) = 0$ and $f'(p) \neq 0$, then for starting values sufficiently close to p, Newton's method will converge at least quadratically.

Multiple Roots

In the preceding discussion, the restriction was made that $f'(p) \neq 0$, where p is the solution to $f(x) = 0$. In particular, Newton's method and the Secant method will generally give problems if $f'(p) = 0$ when $f(p) = 0$. To examine these difficulties in more detail, we make the following definition.

Definition 2.10 A solution p of $f(x) = 0$ is a **zero of multiplicity** m of f if for $x \neq p$, we can write $f(x) = (x - p)^m q(x)$, where $\lim_{x \to p} q(x) \neq 0$. ■

For polynomials, p is a zero of multiplicity m of f if $f(x) = (x - p)^m q(x)$, where $q(p) \neq 0$.

In essence, $q(x)$ represents that portion of $f(x)$ that does not contribute to the zero of f. The following result gives a means to easily identify **simple** zeros of a function, those that have multiplicity one.

Theorem 2.11 The function $f \in C^1[a, b]$ has a simple zero at p in (a, b) if and only if $f(p) = 0$, but $f'(p) \neq 0$. ■

Proof If f has a simple zero at p, then $f(p) = 0$ and $f(x) = (x - p)q(x)$, where $\lim_{x \to p} q(x) \neq 0$. Since $f \in C^1[a, b]$,

$$f'(p) = \lim_{x \to p} f'(x) = \lim_{x \to p} [q(x) + (x - p)q'(x)] = \lim_{x \to p} q(x) \neq 0.$$

Conversely, if $f(p) = 0$, but $f'(p) \neq 0$, expand f in a zeroth Taylor polynomial about p. Then

$$f(x) = f(p) + f'(\xi(x))(x - p) = (x - p) f'(\xi(x)),$$

where $\xi(x)$ is between x and p. Since $f \in C^1[a, b]$,

$$\lim_{x \to p} f'(\xi(x)) = f'\left(\lim_{x \to p} \xi(x)\right) = f'(p) \neq 0.$$

Letting $q = f' \circ \xi$ gives $f(x) = (x - p)q(x)$, where $\lim_{x \to p} q(x) \neq 0$. Thus f has a simple zero at p. ■ ■ ■

The following generalization of Theorem 2.11 is considered in Exercise 12.

Theorem 2.12 The function $f \in C^m[a, b]$ has a zero of multiplicity m at p in (a, b) if and only if

$$0 = f(p) = f'(p) = f''(p) = \cdots = f^{(m-1)}(p), \quad \text{but} \quad f^{(m)}(p) \neq 0. \qquad ■$$

The result in Theorem 2.12 implies that an interval about p exists where Newton's method converges quadratically to p for any initial approximation $p_0 = p$, provided that p is a simple zero. The following example shows that quadratic convergence might not occur if the zero is not simple.

Example 1 Let $f(x) = e^x - x - 1$. **(a)** Show that f has a zero of multiplicity 2 at $x = 0$. **(b)** Show that Newton's method with $p_0 = 1$ converges to this zero but not quadratically.

Solution **(a)** We have

$$f(x) = e^x - x - 1, \qquad f'(x) = e^x - 1 \qquad \text{and} \qquad f''(x) = e^x,$$

so

$$f(0) = e^0 - 0 - 1 = 0, \qquad f'(0) = e^0 - 1 = 0 \qquad \text{and} \qquad f''(0) = e^0 = 1.$$

Theorem 2.12 implies that f has a zero of multiplicity 2 at $x = 0$.

(b) The first two terms generated by Newton's method applied to f with $p_0 = 1$ are

$$p_1 = p_0 - \frac{f(p_0)}{f'(p_0)} = 1 - \frac{e - 2}{e - 1} \approx 0.58198,$$

and

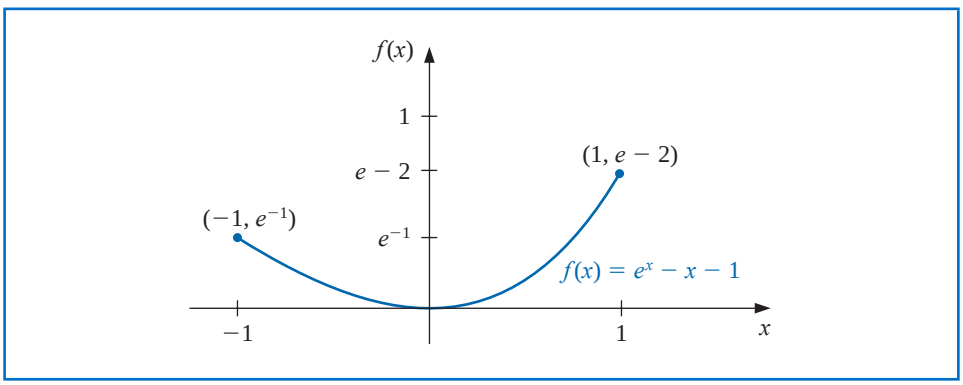
$$p_2 = p_1 - \frac{f(p_1)}{f'(p_1)} \approx 0.58198 - \frac{0.20760}{0.78957} \approx 0.31906.$$

The first sixteen terms of the sequence generated by Newton's method are shown in Table 2.8. The sequence is clearly converging to 0, but not quadratically. The graph of f is shown in Figure 2.12. ■

Table 2.8

n	p_n
0	1.0
1	0.58198
2	0.31906
3	0.16800
4	0.08635
5	0.04380
6	0.02206
7	0.01107
8	0.005545
9	2.7750×10^{-3}
10	1.3881×10^{-3}
11	6.9411×10^{-4}
12	3.4703×10^{-4}
13	1.7416×10^{-4}
14	8.8041×10^{-5}
15	4.2610×10^{-5}
16	1.9142×10^{-6}

Figure 2.12



One method of handling the problem of multiple roots of a function f is to define

$$\mu(x) = \frac{f(x)}{f'(x)}.$$

If p is a zero of f of multiplicity m with $f(x) = (x-p)^m q(x)$, then

$$\mu(x) = \frac{(x-p)^m q(x)}{m(x-p)^{m-1}q(x) + (x-p)^m q'(x)}$$
$$= (x-p)\frac{q(x)}{mq(x) + (x-p)q'(x)}$$

also has a zero at p. However, $q(p) \neq 0$, so

$$\frac{q(p)}{mq(p) + (p-p)q'(p)} = \frac{1}{m} \neq 0,$$

and p is a simple zero of $\mu(x)$. Newton's method can then be applied to $\mu(x)$ to give

$$g(x) = x - \frac{\mu(x)}{\mu'(x)} = x - \frac{f(x)/f'(x)}{\{[f'(x)]^2 - [f(x)][f''(x)]\}/[f'(x)]^2}$$

which simplifies to

$$g(x) = x - \frac{f(x)f'(x)}{[f'(x)]^2 - f(x)f''(x)}. \tag{2.13}$$

If g has the required continuity conditions, functional iteration applied to g will be quadratically convergent regardless of the multiplicity of the zero of f. Theoretically, the only drawback to this method is the additional calculation of $f''(x)$ and the more laborious procedure of calculating the iterates. In practice, however, multiple roots can cause serious round-off problems because the denominator of (2.13) consists of the difference of two numbers that are both close to 0.

Example 2 In Example 1 it was shown that $f(x) = e^x - x - 1$ has a zero of multiplicity 2 at $x = 0$ and that Newton's method with $p_0 = 1$ converges to this zero but not quadratically. Show that the modification of Newton's method as given in Eq. (2.13) improves the rate of convergence.

Solution Modified Newton's method gives

$$p_1 = p_0 - \frac{f(p_0)f'(p_0)}{f'(p_0)^2 - f(p_0)f''(p_0)} = 1 - \frac{(e-2)(e-1)}{(e-1)^2 - (e-2)e} \approx -2.3421061 \times 10^{-1}.$$

This is considerably closer to 0 than the first term using Newton's method, which was 0.58918. Table 2.9 lists the first five approximations to the double zero at $x = 0$. The results were obtained using a system with ten digits of precision. The relative lack of improvement in the last two entries is due to the fact that using this system both the numerator and the denominator approach 0. Consequently there is a loss of significant digits of accuracy as the approximations approach 0. ■

Table 2.9

n	p_n
1	$-2.3421061 \times 10^{-1}$
2	$-8.4582788 \times 10^{-3}$
3	$-1.1889524 \times 10^{-5}$
4	$-6.8638230 \times 10^{-6}$
5	$-2.8085217 \times 10^{-7}$

The following illustrates that the modified Newton's method converges quadratically even when in the case of a simple zero.

Illustration In Section 2.2 we found that a zero of $f(x) = x^3 + 4x^2 - 10 = 0$ is $p = 1.36523001$. Here we will compare convergence for a simple zero using both Newton's method and the modified Newton's method listed in Eq. (2.13). Let

(i) $p_n = p_{n-1} - \dfrac{p_{n-1}^3 + 4p_{n-1}^2 - 10}{3p_{n-1}^2 + 8p_{n-1}}$, from Newton's method

and, from the Modified Newton's method given by Eq. (2.13),

(ii) $p_n = p_{n-1} - \dfrac{(p_{n-1}^3 + 4p_{n-1}^2 - 10)(3p_{n-1}^2 + 8p_{n-1})}{(3p_{n-1}^2 + 8p_{n-1})^2 - (p_{n-1}^3 + 4p_{n-1}^2 - 10)(6p_{n-1} + 8)}$.

With $p_0 = 1.5$, we have

Newton's method

$$p_1 = 1.37333333, \quad p_2 = 1.36526201, \quad \text{and} \quad p_3 = 1.36523001.$$

Modified Newton's method

$$p_1 = 1.35689898, \quad p_2 = 1.36519585, \quad \text{and} \quad p_3 = 1.36523001.$$

Both methods are rapidly convergent to the actual zero, which is given by both methods as p_3. Note, however, that in the case of a simple zero the original Newton's method requires substantially less computation. □

Maple contains Modified Newton's method as described in Eq. (2.13) in its *Numerical-Analysis* package. The options for this command are the same as those for the Bisection method. To obtain results similar to those in Table 2.9 we can use

with(*Student*[*NumericalAnalysis*])

$f := e^x - x - 1$

ModifiedNewton $\left(f, x = 1.0, tolerance = 10^{-10}, output = sequence, maxiterations = 20\right)$

Remember that there is sensitivity to round-off error in these calculations, so you might need to reset *Digits* in Maple to get the exact values in Table 2.9.

EXERCISE SET 2.4

1. Use Newton's method to find solutions accurate to within 10^{-5} to the following problems.

a. $x^2 - 2xe^{-x} + e^{-2x} = 0, \quad \text{for } 0 \le x \le 1$

b. $\cos(x + \sqrt{2}) + x(x/2 + \sqrt{2}) = 0, \quad \text{for } -2 \le x \le -1$

c. $x^3 - 3x^2(2^{-x}) + 3x(4^{-x}) - 8^{-x} = 0, \quad \text{for } 0 \le x \le 1$

d. $e^{6x} + 3(\ln 2)^2 e^{2x} - (\ln 8)e^{4x} - (\ln 2)^3 = 0, \quad \text{for } -1 \le x \le 0$

2. Use Newton's method to find solutions accurate to within 10^{-5} to the following problems.

a. $1 - 4x\cos x + 2x^2 + \cos 2x = 0, \quad \text{for } 0 \le x \le 1$

b. $x^2 + 6x^5 + 9x^4 - 2x^3 - 6x^2 + 1 = 0, \quad \text{for } -3 \le x \le -2$

c. $\sin 3x + 3e^{-2x}\sin x - 3e^{-x}\sin 2x - e^{-3x} = 0, \quad \text{for } 3 \le x \le 4$

d. $e^{3x} - 27x^6 + 27x^4 e^x - 9x^2 e^{2x} = 0, \quad \text{for } 3 \le x \le 5$

3. Repeat Exercise 1 using the modified Newton's method described in Eq. (2.13). Is there an improvement in speed or accuracy over Exercise 1?

4. Repeat Exercise 2 using the modified Newton's method described in Eq. (2.13). Is there an improvement in speed or accuracy over Exercise 2?

5. Use Newton's method and the modified Newton's method described in Eq. (2.13) to find a solution accurate to within 10^{-5} to the problem

$$e^{6x} + 1.441e^{2x} - 2.079e^{4x} - 0.3330 = 0, \quad \text{for } -1 \le x \le 0.$$

This is the same problem as 1(d) with the coefficients replaced by their four-digit approximations. Compare the solutions to the results in 1(d) and 2(d).

6. Show that the following sequences converge linearly to $p = 0$. How large must n be before $|p_n - p| \le 5 \times 10^{-2}$?

a. $p_n = \dfrac{1}{n}, \quad n \ge 1$ **b.** $p_n = \dfrac{1}{n^2}, \quad n \ge 1$

7. **a.** Show that for any positive integer k, the sequence defined by $p_n = 1/n^k$ converges linearly to $p = 0$.

b. For each pair of integers k and m, determine a number N for which $1/N^k < 10^{-m}$.

8. **a.** Show that the sequence $p_n = 10^{-2^n}$ converges quadratically to 0.

b. Show that the sequence $p_n = 10^{-n^k}$ does not converge to 0 quadratically, regardless of the size of the exponent $k > 1$.

9. **a.** Construct a sequence that converges to 0 of order 3.

b. Suppose $\alpha > 1$. Construct a sequence that converges to 0 zero of order α.

10. Suppose p is a zero of multiplicity m of f, where $f^{(m)}$ is continuous on an open interval containing p. Show that the following fixed-point method has $g'(p) = 0$:

$$g(x) = x - \frac{mf(x)}{f'(x)}.$$

11. Show that the Bisection Algorithm 2.1 gives a sequence with an error bound that converges linearly to 0.

12. Suppose that f has m continuous derivatives. Modify the proof of Theorem 2.11 to show that f has a zero of multiplicity m at p if and only if

$$0 = f(p) = f'(p) = \cdots = f^{(m-1)}(p), \quad \text{but } f^{(m)}(p) \ne 0.$$

13. The iterative method to solve $f(x) = 0$, given by the fixed-point method $g(x) = x$, where

$$p_n = g(p_{n-1}) = p_{n-1} - \frac{f(p_{n-1})}{f'(p_{n-1})} - \frac{f''(p_{n-1})}{2f'(p_{n-1})}\left[\frac{f(p_{n-1})}{f'(p_{n-1})}\right]^2, \quad \text{for } n = 1, 2, 3, \ldots,$$

has $g'(p) = g''(p) = 0$. This will generally yield cubic ($\alpha = 3$) convergence. Expand the analysis of Example 1 to compare quadratic and cubic convergence.

14. It can be shown (see, for example, [DaB], pp. 228–229) that if $\{p_n\}_{n=0}^{\infty}$ are convergent Secant method approximations to p, the solution to $f(x) = 0$, then a constant C exists with $|p_{n+1} - p| \approx C\,|p_n - p|\,|p_{n-1} - p|$ for sufficiently large values of n. Assume $\{p_n\}$ converges to p of order α, and show that $\alpha = (1 + \sqrt{5})/2$. (*Note:* This implies that the order of convergence of the Secant method is approximately 1.62).

2.5 Accelerating Convergence

Theorem 2.8 indicates that it is rare to have the luxury of quadratic convergence. We now consider a technique called **Aitken's Δ^2 method** that can be used to accelerate the convergence of a sequence that is linearly convergent, regardless of its origin or application.

Aitken's Δ^2 Method

Alexander Aitken (1895-1967) used this technique in 1926 to accelerate the rate of convergence of a series in a paper on algebraic equations [Ai]. This process is similar to one used much earlier by the Japanese mathematician Takakazu Seki Kowa (1642-1708).

Suppose $\{p_n\}_{n=0}^{\infty}$ is a linearly convergent sequence with limit p. To motivate the construction of a sequence $\{\hat{p}_n\}_{n=0}^{\infty}$ that converges more rapidly to p than does $\{p_n\}_{n=0}^{\infty}$, let us first assume that the signs of $p_n - p$, $p_{n+1} - p$, and $p_{n+2} - p$ agree and that n is sufficiently large that

$$\frac{p_{n+1} - p}{p_n - p} \approx \frac{p_{n+2} - p}{p_{n+1} - p}.$$

Then

$$(p_{n+1} - p)^2 \approx (p_{n+2} - p)(p_n - p),$$

so

$$p_{n+1}^2 - 2p_{n+1}p + p^2 \approx p_{n+2}p_n - (p_n + p_{n+2})p + p^2$$

and

$$(p_{n+2} + p_n - 2p_{n+1})p \approx p_{n+2}p_n - p_{n+1}^2.$$

Solving for p gives

$$p \approx \frac{p_{n+2}p_n - p_{n+1}^2}{p_{n+2} - 2p_{n+1} + p_n}.$$

Adding and subtracting the terms p_n^2 and $2p_np_{n+1}$ in the numerator and grouping terms appropriately gives

$$\begin{aligned} p &\approx \frac{p_np_{n+2} - 2p_np_{n+1} + p_n^2 - p_{n+1}^2 + 2p_np_{n+1} - p_n^2}{p_{n+2} - 2p_{n+1} + p_n} \\ &= \frac{p_n(p_{n+2} - 2p_{n+1} + p_n) - (p_{n+1}^2 - 2p_np_{n+1} + p_n^2)}{p_{n+2} - 2p_{n+1} + p_n} \\ &= p_n - \frac{(p_{n+1} - p_n)^2}{p_{n+2} - 2p_{n+1} + p_n}. \end{aligned}$$

Aitken's Δ^2 method is based on the assumption that the sequence $\{\hat{p}_n\}_{n=0}^{\infty}$, defined by

$$\hat{p}_n = p_n - \frac{(p_{n+1} - p_n)^2}{p_{n+2} - 2p_{n+1} + p_n}, \tag{2.14}$$

converges more rapidly to p than does the original sequence $\{p_n\}_{n=0}^{\infty}$.

Table 2.10

n	p_n	$\hat{p}_n$
1	0.54030	0.96178
2	0.87758	0.98213
3	0.94496	0.98979
4	0.96891	0.99342
5	0.98007	0.99541
6	0.98614	
7	0.98981	

Example 1 The sequence $\{p_n\}_{n=1}^{\infty}$, where $p_n = \cos(1/n)$, converges linearly to $p = 1$. Determine the first five terms of the sequence given by Aitken's Δ^2 method.

Solution In order to determine a term $\hat{p}_n$ of the Aitken's Δ^2 method sequence we need to have the terms p_n, p_{n+1}, and p_{n+2} of the original sequence. So to determine $\hat{p}_5$ we need the first 7 terms of $\{p_n\}$. These are given in Table 2.10. It certainly appears that $\{\hat{p}_n\}_{n=1}^{\infty}$ converges more rapidly to $p = 1$ than does $\{p_n\}_{n=1}^{\infty}$. ■

The Δ notation associated with this technique has its origin in the following definition.

Definition 2.13 For a given sequence $\{p_n\}_{n=0}^{\infty}$, the **forward difference** Δp_n (read "delta p_n") is defined by

$$\Delta p_n = p_{n+1} - p_n, \quad \text{for } n \geq 0.$$

Higher powers of the operator Δ are defined recursively by

$$\Delta^k p_n = \Delta(\Delta^{k-1} p_n), \quad \text{for } k \geq 2.$$ ■

The definition implies that

$$\Delta^2 p_n = \Delta(p_{n+1} - p_n) = \Delta p_{n+1} - \Delta p_n = (p_{n+2} - p_{n+1}) - (p_{n+1} - p_n).$$

So $\Delta^2 p_n = p_{n+2} - 2p_{n+1} + p_n$, and the formula for $\hat{p}_n$ given in Eq. (2.14) can be written as

$$\hat{p}_n = p_n - \frac{(\Delta p_n)^2}{\Delta^2 p_n}, \quad \text{for } n \geq 0. \tag{2.15}$$

To this point in our discussion of Aitken's Δ^2 method, we have stated that the sequence $\{\hat{p}_n\}_{n=0}^{\infty}$, converges to p more rapidly than does the original sequence $\{p_n\}_{n=0}^{\infty}$, but we have not said what is meant by the term "more rapid" convergence. Theorem 2.14 explains and justifies this terminology. The proof of this theorem is considered in Exercise 16.

Theorem 2.14 Suppose that $\{p_n\}_{n=0}^{\infty}$ is a sequence that converges linearly to the limit p and that

$$\lim_{n\to\infty} \frac{p_{n+1} - p}{p_n - p} < 1.$$

Then the Aitken's Δ^2 sequence $\{\hat{p}_n\}_{n=0}^{\infty}$ converges to p faster than $\{p_n\}_{n=0}^{\infty}$ in the sense that

$$\lim_{n\to\infty} \frac{\hat{p}_n - p}{p_n - p} = 0.$$ ■

Steffensen's Method

Johan Frederik Steffensen (1873–1961) wrote an influential book entitled *Interpolation* in 1927.

By applying a modification of Aitken's Δ^2 method to a linearly convergent sequence obtained from fixed-point iteration, we can accelerate the convergence to quadratic. This procedure is known as Steffensen's method and differs slightly from applying Aitken's Δ^2 method directly to the linearly convergent fixed-point iteration sequence. Aitken's Δ^2 method constructs the terms in order:

$$p_0, \quad p_1 = g(p_0), \quad p_2 = g(p_1), \quad \hat{p}_0 = \{\Delta^2\}(p_0),$$
$$p_3 = g(p_2), \quad \hat{p}_1 = \{\Delta^2\}(p_1), \ldots,$$

where $\{\Delta^2\}$ indicates that Eq. (2.15) is used. Steffensen's method constructs the same first four terms, p_0, p_1, p_2, and $\hat{p}_0$. However, at this step we assume that $\hat{p}_0$ is a better approximation to p than is p_2 and apply fixed-point iteration to $\hat{p}_0$ instead of p_2. Using this notation, the sequence is

$$p_0^{(0)}, \quad p_1^{(0)} = g(p_0^{(0)}), \quad p_2^{(0)} = g(p_1^{(0)}), \quad p_0^{(1)} = \{\Delta^2\}(p_0^{(0)}), \quad p_1^{(1)} = g(p_0^{(1)}), \ldots.$$

Every third term of the Steffensen sequence is generated by Eq. (2.15); the others use fixed-point iteration on the previous term. The process is described in Algorithm 2.6.

ALGORITHM
2.6

Steffensen's

To find a solution to $p = g(p)$ given an initial approximation p_0:

INPUT initial approximation p_0; tolerance TOL; maximum number of iterations N_0.

OUTPUT approximate solution p or message of failure.

Step 1 Set $i = 1$.

Step 2 While $i \le N_0$ do Steps 3–6.

Step 3 Set $p_1 = g(p_0)$; (*Compute* $p_1^{(i-1)}$.)
$p_2 = g(p_1)$; (*Compute* $p_2^{(i-1)}$.)
$p = p_0 - (p_1 - p_0)^2/(p_2 - 2p_1 + p_0)$. (*Compute* $p_0^{(i)}$.)

Step 4 If $|p - p_0| < TOL$ then
OUTPUT (p); (*Procedure completed successfully.*)
STOP.

Step 5 Set $i = i + 1$.

Step 6 Set $p_0 = p$. (*Update* p_0.)

Step 7 OUTPUT ('Method failed after N_0 iterations, $N_0 =$', N_0);
(*Procedure completed unsuccessfully.*)
STOP. ■

Note that $\Delta^2 p_n$ might be 0, which would introduce a 0 in the denominator of the next iterate. If this occurs, we terminate the sequence and select $p_2^{(n-1)}$ as the best approximation.

Illustration To solve $x^3 + 4x^2 - 10 = 0$ using Steffensen's method, let $x^3 + 4x^2 = 10$, divide by $x + 4$, and solve for x. This procedure produces the fixed-point method

$$g(x) = \left(\frac{10}{x+4}\right)^{1/2}.$$

We considered this fixed-point method in Table 2.2 column **(d)** of Section 2.2.

Applying Steffensen's procedure with $p_0 = 1.5$ gives the values in Table 2.11. The iterate $p_0^{(2)} = 1.365230013$ is accurate to the ninth decimal place. In this example, Steffensen's method gave about the same accuracy as Newton's method applied to this polynomial. These results can be seen in the Illustration at the end of Section 2.4. □

Table 2.11

k	$p_0^{(k)}$	$p_1^{(k)}$	$p_2^{(k)}$
0	1.5	1.348399725	1.367376372
1	1.365265224	1.365225534	1.365230583
2	1.365230013		

From the Illustration, it appears that Steffensen's method gives quadratic convergence without evaluating a derivative, and Theorem 2.14 states that this is the case. The proof of this theorem can be found in [He2], pp. 90–92, or [IK], pp. 103–107.

Theorem 2.15 Suppose that $x = g(x)$ has the solution p with $g'(p) \neq 1$. If there exists a $\delta > 0$ such that $g \in C^3[p-\delta, p+\delta]$, then Steffensen's method gives quadratic convergence for any $p_0 \in [p-\delta, p+\delta]$. ■

Steffensen's method can be implemented in Maple with the *NumericalAnalysis* package. For example, after entering the function

$$g := \sqrt{\frac{10}{x+4}}$$

the Maple command

Steffensen(*fixedpointiterator* = $g, x = 1.5, tolerance = 10^{-8}$, *output* = *information*, *maxiterations* = 20)

produces the results in Table 2.11, as well as an indication that the final approximation has a relative error of approximately 7.32×10^{-10}.

EXERCISE SET 2.5

1. The following sequences are linearly convergent. Generate the first five terms of the sequence $\{\hat{p}_n\}$ using Aitken's Δ^2 method.

a. $p_0 = 0.5, \quad p_n = (2 - e^{p_{n-1}} + p_{n-1}^2)/3, \quad n \geq 1$

b. $p_0 = 0.75, \quad p_n = (e^{p_{n-1}}/3)^{1/2}, \quad n \geq 1$

c. $p_0 = 0.5, \quad p_n = 3^{-p_{n-1}}, \quad n \geq 1$

d. $p_0 = 0.5, \quad p_n = \cos p_{n-1}, \quad n \geq 1$

2. Consider the function $f(x) = e^{6x} + 3(\ln 2)^2 e^{2x} - (\ln 8)e^{4x} - (\ln 2)^3$. Use Newton's method with $p_0 = 0$ to approximate a zero of f. Generate terms until $|p_{n+1} - p_n| < 0.0002$. Construct the sequence $\{\hat{p}_n\}$. Is the convergence improved?

3. Let $g(x) = \cos(x-1)$ and $p_0^{(0)} = 2$. Use Steffensen's method to find $p_0^{(1)}$.

4. Let $g(x) = 1 + (\sin x)^2$ and $p_0^{(0)} = 1$. Use Steffensen's method to find $p_0^{(1)}$ and $p_0^{(2)}$.

5. Steffensen's method is applied to a function $g(x)$ using $p_0^{(0)} = 1$ and $p_2^{(0)} = 3$ to obtain $p_0^{(1)} = 0.75$. What is $p_1^{(0)}$?

6. Steffensen's method is applied to a function $g(x)$ using $p_0^{(0)} = 1$ and $p_1^{(0)} = \sqrt{2}$ to obtain $p_0^{(1)} = 2.7802$. What is $p_2^{(0)}$?

7. Use Steffensen's method to find, to an accuracy of 10^{-4}, the root of $x^3 - x - 1 = 0$ that lies in $[1, 2]$, and compare this to the results of Exercise 6 of Section 2.2.

8. Use Steffensen's method to find, to an accuracy of 10^{-4}, the root of $x - 2^{-x} = 0$ that lies in $[0, 1]$, and compare this to the results of Exercise 8 of Section 2.2.

9. Use Steffensen's method with $p_0 = 2$ to compute an approximation to $\sqrt{3}$ accurate to within 10^{-4}. Compare this result with those obtained in Exercise 9 of Section 2.2 and Exercise 12 of Section 2.1.

10. Use Steffensen's method with $p_0 = 3$ to compute an approximation to $\sqrt[3]{25}$ accurate to within 10^{-4}. Compare this result with those obtained in Exercise 10 of Section 2.2 and Exercise 13 of Section 2.1.

11. Use Steffensen's method to approximate the solutions of the following equations to within 10^{-5}.

a. $x = (2 - e^x + x^2)/3$, where g is the function in Exercise 11(a) of Section 2.2.

b. $x = 0.5(\sin x + \cos x)$, where g is the function in Exercise 11(f) of Section 2.2.

c. $x = (e^x/3)^{1/2}$, where g is the function in Exercise 11(c) of Section 2.2.

d. $x = 5^{-x}$, where g is the function in Exercise 11(d) of Section 2.2.

12. Use Steffensen's method to approximate the solutions of the following equations to within 10^{-5}.

a. $2 + \sin x - x = 0$, where g is the function in Exercise 12(a) of Section 2.2.

b. $x^3 - 2x - 5 = 0$, where g is the function in Exercise 12(b) of Section 2.2.

c. $3x^2 - e^x = 0$, where g is the function in Exercise 12(c) of Section 2.2.

d. $x - \cos x = 0$, where g is the function in Exercise 12(d) of Section 2.2.

13. The following sequences converge to 0. Use Aitken's Δ^2 method to generate $\{\hat{p}_n\}$ until $|\hat{p}_n| \leq 5 \times 10^{-2}$:

a. $p_n = \dfrac{1}{n}, \quad n \geq 1$ **b.** $p_n = \dfrac{1}{n^2}, \quad n \geq 1$

14. A sequence $\{p_n\}$ is said to be **superlinearly convergent** to p if

$$\lim_{n\to\infty} \frac{|p_{n+1} - p|}{|p_n - p|} = 0.$$

a. Show that if $p_n \to p$ of order α for $\alpha > 1$, then $\{p_n\}$ is superlinearly convergent to p.

b. Show that $p_n = \frac{1}{n^n}$ is superlinearly convergent to 0 but does not converge to 0 of order α for any $\alpha > 1$.

15. Suppose that $\{p_n\}$ is superlinearly convergent to p. Show that

$$\lim_{n\to\infty} \frac{|p_{n+1} - p_n|}{|p_n - p|} = 1.$$

16. Prove Theorem 2.14. [*Hint:* Let $\delta_n = (p_{n+1} - p)/(p_n - p) - \lambda$, and show that $\lim_{n\to\infty} \delta_n = 0$. Then express $(\hat{p}_{n+1} - p)/(p_n - p)$ in terms of δ_n, δ_{n+1}, and λ.]

17. Let $P_n(x)$ be the nth Taylor polynomial for $f(x) = e^x$ expanded about $x_0 = 0$.

a. For fixed x, show that $p_n = P_n(x)$ satisfies the hypotheses of Theorem 2.14.

b. Let $x = 1$, and use Aitken's Δ^2 method to generate the sequence $\hat{p}_0, \ldots, \hat{p}_8$.

c. Does Aitken's method accelerate convergence in this situation?

2.6 Zeros of Polynomials and Müller's Method

A *polynomial of degree n* has the form

$$P(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0,$$

where the a_i's, called the *coefficients* of P, are constants and $a_n \neq 0$. The zero function, $P(x) = 0$ for all values of x, is considered a polynomial but is assigned no degree.

Algebraic Polynomials

Theorem 2.16 **(Fundamental Theorem of Algebra)**

If $P(x)$ is a polynomial of degree $n \geq 1$ with real or complex coefficients, then $P(x) = 0$ has at least one (possibly complex) root. ■

Although the Fundamental Theorem of Algebra is basic to any study of elementary functions, the usual proof requires techniques from the study of complex function theory. The reader is referred to [SaS], p. 155, for the culmination of a systematic development of the topics needed to prove the Theorem.

Example 1 Determine all the zeros of the polynomial $P(x) = x^3 - 5x^2 + 17x - 13$.

Solution It is easily verified that $P(1) = 1 - 5 + 17 - 13 = 0$. so $x = 1$ is a zero of P and $(x - 1)$ is a factor of the polynomial. Dividing $P(x)$ by $x - 1$ gives

$$P(x) = (x - 1)(x^2 - 4x + 13).$$

Carl Friedrich Gauss (1777–1855), one of the greatest mathematicians of all time, proved the Fundamental Theorem of Algebra in his doctoral dissertation and published it in 1799. He published different proofs of this result throughout his lifetime, in 1815, 1816, and as late as 1848. The result had been stated, without proof, by Albert Girard (1595–1632), and partial proofs had been given by Jean d'Alembert (1717–1783), Euler, and Lagrange.

To determine the zeros of $x^2 - 4x + 13$ we use the quadratic formula in its standard form, which gives the complex zeros

$$\frac{-(-4) \pm \sqrt{(-4)^2 - 4(1)(13)}}{2(1)} = \frac{4 \pm \sqrt{-36}}{2} = 2 \pm 3i.$$

Hence the third-degree polynomial $P(x)$ has three zeros, $x_1 = 1$, $x_2 = 2 - 3i$, and $x_2 = 2 + 3i$. ■

In the preceding example we found that the third-degree polynomial had three distinct zeros. An important consequence of the Fundamental Theorem of Algebra is the following corollary. It states that this is always the case, provided that when the zeros are not distinct we count the number of zeros according to their multiplicities.

Corollary 2.17 If $P(x)$ is a polynomial of degree $n \geq 1$ with real or complex coefficients, then there exist unique constants $x_1, x_2, \ldots, x_k$, possibly complex, and unique positive integers $m_1, m_2, \ldots, m_k$, such that $\sum_{i=1}^{k} m_i = n$ and

$$P(x) = a_n(x - x_1)^{m_1}(x - x_2)^{m_2} \cdots (x - x_k)^{m_k}.$$ ■

By Corollary 2.17 the collection of zeros of a polynomial is unique and, if each zero x_i is counted as many times as its multiplicity m_i, a polynomial of degree n has exactly n zeros.

The following corollary of the Fundamental Theorem of Algebra is used often in this section and in later chapters.

Corollary 2.18 Let $P(x)$ and $Q(x)$ be polynomials of degree at most n. If $x_1, x_2, \ldots, x_k$, with $k > n$, are distinct numbers with $P(x_i) = Q(x_i)$ for $i = 1, 2, \ldots, k$, then $P(x) = Q(x)$ for all values of x. ■

This result implies that to show that two polynomials of degree less than or equal to n are the same, we only need to show that they agree at $n + 1$ values. This will be frequently used, particularly in Chapters 3 and 8.

Horner's Method

William Horner (1786–1837) was a child prodigy who became headmaster of a school in Bristol at age 18. Horner's method for solving algebraic equations was published in 1819 in the Philosophical Transactions of the Royal Society.

To use Newton's method to locate approximate zeros of a polynomial $P(x)$, we need to evaluate $P(x)$ and $P'(x)$ at specified values. Since $P(x)$ and $P'(x)$ are both polynomials, computational efficiency requires that the evaluation of these functions be done in the nested manner discussed in Section 1.2. Horner's method incorporates this nesting technique, and, as a consequence, requires only n multiplications and n additions to evaluate an arbitrary nth-degree polynomial.

Theorem 2.19 **(Horner's Method)**

Let

$$P(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0.$$

Define $b_n = a_n$ and

$$b_k = a_k + b_{k+1}x_0, \quad \text{for } k = n - 1, n - 2, \ldots, 1, 0.$$

Then $b_0 = P(x_0)$. Moreover, if

$$Q(x) = b_n x^{n-1} + b_{n-1}x^{n-2} + \cdots + b_2 x + b_1,$$

then

$$P(x) = (x - x_0)Q(x) + b_0.$$ ■

Paolo Ruffini (1765–1822) had described a similar method which won him the gold medal from the Italian Mathematical Society for Science. Neither Ruffini nor Horner was the first to discover this method; it was known in China at least 500 years earlier.

Proof By the definition of $Q(x)$,

$$\begin{aligned}(x - x_0)Q(x) + b_0 &= (x - x_0)(b_n x^{n-1} + \cdots + b_2 x + b_1) + b_0 \\ &= (b_n x^n + b_{n-1}x^{n-1} + \cdots + b_2 x^2 + b_1 x) \\ &\quad - (b_n x_0 x^{n-1} + \cdots + b_2 x_0 x + b_1 x_0) + b_0 \\ &= b_n x^n + (b_{n-1} - b_n x_0)x^{n-1} + \cdots + (b_1 - b_2 x_0)x + (b_0 - b_1 x_0).\end{aligned}$$

By the hypothesis, $b_n = a_n$ and $b_k - b_{k+1}x_0 = a_k$, so

$$(x - x_0)Q(x) + b_0 = P(x) \quad \text{and} \quad b_0 = P(x_0).$$ ■ ■ ■

Example 2 Use Horner's method to evaluate $P(x) = 2x^4 - 3x^2 + 3x - 4$ at $x_0 = -2$.

Solution When we use hand calculation in Horner's method, we first construct a table, which suggests the *synthetic division* name that is often applied to the technique. For this problem, the table appears as follows:

	Coefficient of x^4	Coefficient of x^3	Coefficient of x^2	Coefficient of x	Constant term
$x_0 = -2$	$a_4 = 2$	$a_3 = 0$ $b_4x_0 = -4$	$a_2 = -3$ $b_3x_0 = 8$	$a_1 = 3$ $b_2x_0 = -10$	$a_0 = -4$ $b_1x_0 = 14$
	$b_4 = 2$	$b_3 = -4$	$b_2 = 5$	$b_1 = -7$	$b_0 = 10$

So,

$$P(x) = (x + 2)(2x^3 - 4x^2 + 5x - 7) + 10.$$ ■

The word synthetic has its roots in various languages. In standard English it generally provides the sense of something that is "false" or "substituted". But in mathematics it takes the form of something that is "grouped together". Synthetic geometry treats shapes as whole, rather than as individual objects, which is the style in analytic geometry. In synthetic division of polynomials, the various powers of the variables are not explicitly given but kept grouped together.

An additional advantage of using the Horner (or synthetic-division) procedure is that, since

$$P(x) = (x - x_0)Q(x) + b_0,$$

where

$$Q(x) = b_n x^{n-1} + b_{n-1}x^{n-2} + \cdots + b_2 x + b_1,$$

differentiating with respect to x gives

$$P'(x) = Q(x) + (x - x_0)Q'(x) \quad \text{and} \quad P'(x_0) = Q(x_0). \tag{2.16}$$

When the Newton-Raphson method is being used to find an approximate zero of a polynomial, $P(x)$ and $P'(x)$ can be evaluated in the same manner.

Example 3 Find an approximation to a zero of

$$P(x) = 2x^4 - 3x^2 + 3x - 4,$$

using Newton's method with $x_0 = -2$ and synthetic division to evaluate $P(x_n)$ and $P'(x_n)$ for each iterate x_n.

Solution With $x_0 = -2$ as an initial approximation, we obtained $P(-2)$ in Example 1 by

$x_0 = -2$	2	0	−3	3	−4	
		−4	8	−10	14	
	2	−4	5	−7	10	$= P(-2).$

Using Theorem 2.19 and Eq. (2.16),

$$Q(x) = 2x^3 - 4x^2 + 5x - 7 \quad \text{and} \quad P'(-2) = Q(-2),$$

so $P'(-2)$ can be found by evaluating $Q(-2)$ in a similar manner:

$x_0 = -2$	2	−4	5	−7	
		−4	16	−42	
	2	−8	21	−49	$= Q(-2) = P'(-2)$

and

$$x_1 = x_0 - \frac{P(x_0)}{P'(x_0)} = x_0 - \frac{P(x_0)}{Q(x_0)} = -2 - \frac{10}{-49} \approx -1.796.$$

Repeating the procedure to find x_2 gives

−1.796	2	0	−3	3	−4	
		−3.592	6.451	−6.197	5.742	
	2	−3.592	3.451	−3.197	1.742	$= P(x_1)$
		−3.592	12.902	−29.368		
	2	−7.184	16.353	−32.565	$= Q(x_1)$	$= P'(x_1).$

So $P(-1.796) = 1.742$, $P'(-1.796) = Q(-1.796) = -32.565$, and

$$x_2 = -1.796 - \frac{1.742}{-32.565} \approx -1.7425.$$

In a similar manner, $x_3 = -1.73897$, and an actual zero to five decimal places is -1.73896.

Note that the polynomial $Q(x)$ depends on the approximation being used and changes from iterate to iterate. ■

Algorithm 2.7 computes $P(x_0)$ and $P'(x_0)$ using Horner's method.

ALGORITHM
2.7

Horner's

To evaluate the polynomial

$$P(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 = (x - x_0)Q(x) + b_0$$

and its derivative at x_0:

INPUT degree n; coefficients $a_0, a_1, \ldots, a_n$; x_0.

OUTPUT $y = P(x_0); z = P'(x_0)$.

Step 1 Set $y = a_n$; (*Compute* b_n *for* P.)
$z = a_n$. (*Compute* b_{n-1} *for* Q.)

Step 2 For $j = n - 1, n - 2, \ldots, 1$
set $y = x_0 y + a_j$; (*Compute* b_j *for* P.)
$z = x_0 z + y$. (*Compute* b_{j-1} *for* Q.)

Step 3 Set $y = x_0 y + a_0$. (*Compute* b_0 *for* P.)

Step 4 OUTPUT (y, z);
STOP. ■

If the Nth iterate, x_N, in Newton's method is an approximate zero for P, then

$$P(x) = (x - x_N)Q(x) + b_0 = (x - x_N)Q(x) + P(x_N) \approx (x - x_N)Q(x),$$

so $x - x_N$ is an approximate factor of $P(x)$. Letting $\hat{x}_1 = x_N$ be the approximate zero of P and $Q_1(x) \equiv Q(x)$ be the approximate factor gives

$$P(x) \approx (x - \hat{x}_1)Q_1(x).$$

We can find a second approximate zero of P by applying Newton's method to $Q_1(x)$.

If $P(x)$ is an nth-degree polynomial with n real zeros, this procedure applied repeatedly will eventually result in $(n-2)$ approximate zeros of P and an approximate quadratic factor $Q_{n-2}(x)$. At this stage, $Q_{n-2}(x) = 0$ can be solved by the quadratic formula to find the last two approximate zeros of P. Although this method can be used to find all the approximate zeros, it depends on repeated use of approximations and can lead to inaccurate results.

The procedure just described is called **deflation**. The accuracy difficulty with deflation is due to the fact that, when we obtain the approximate zeros of $P(x)$, Newton's method is used on the reduced polynomial $Q_k(x)$, that is, the polynomial having the property that

$$P(x) \approx (x - \hat{x}_1)(x - \hat{x}_2) \cdots (x - \hat{x}_k)Q_k(x).$$

An approximate zero $\hat{x}_{k+1}$ of Q_k will generally not approximate a root of $P(x) = 0$ as well as it does a root of the reduced equation $Q_k(x) = 0$, and inaccuracy increases as k increases. One way to eliminate this difficulty is to use the reduced equations to find approximations $\hat{x}_2$, $\hat{x}_3, \ldots, \hat{x}_k$ to the zeros of P, and then improve these approximations by applying Newton's method to the original polynomial $P(x)$.

Complex Zeros: Müller's Method

One problem with applying the Secant, False Position, or Newton's method to polynomials is the possibility of the polynomial having complex roots even when all the coefficients are

real numbers. If the initial approximation is a real number, all subsequent approximations will also be real numbers. One way to overcome this difficulty is to begin with a complex initial approximation and do all the computations using complex arithmetic. An alternative approach has its basis in the following theorem.

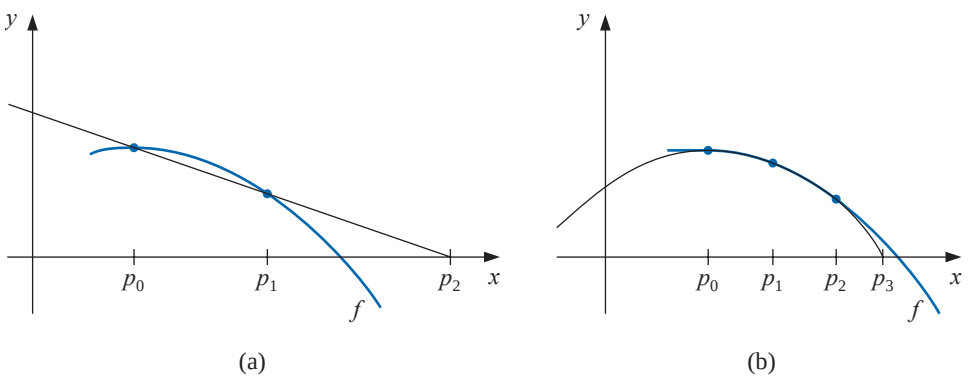
Theorem 2.20 If $z = a+bi$ is a complex zero of multiplicity m of the polynomial $P(x)$ with real coefficients, then $\bar{z} = a - bi$ is also a zero of multiplicity m of the polynomial $P(x)$, and $(x^2 - 2ax + a^2 + b^2)^m$ is a factor of $P(x)$. ■

A synthetic division involving quadratic polynomials can be devised to approximately factor the polynomial so that one term will be a quadratic polynomial whose complex roots are approximations to the roots of the original polynomial. This technique was described in some detail in our second edition [BFR]. Instead of proceeding along these lines, we will now consider a method first presented by D. E. Müller [Mu]. This technique can be used for any root-finding problem, but it is particularly useful for approximating the roots of polynomials.

Müller's method is similar to the Secant method. But whereas the Secant method uses a line through two points on the curve to approximate the root, Müller's method uses a parabola through three points on the curve for the approximation.

The Secant method begins with two initial approximations p_0 and p_1 and determines the next approximation p_2 as the intersection of the x-axis with the line through $(p_0, f(p_0))$ and $(p_1, f(p_1))$. (See Figure 2.13(a).) Müller's method uses three initial approximations, p_0, p_1, and p_2, and determines the next approximation p_3 by considering the intersection of the x-axis with the parabola through $(p_0, f(p_0))$, $(p_1, f(p_1))$, and $(p_2, f(p_2))$. (See Figure 2.13(b).)

Figure 2.13



The derivation of Müller's method begins by considering the quadratic polynomial

$$P(x) = a(x - p_2)^2 + b(x - p_2) + c$$

that passes through $(p_0, f(p_0))$, $(p_1, f(p_1))$, and $(p_2, f(p_2))$. The constants a, b, and c can be determined from the conditions

$$f(p_0) = a(p_0 - p_2)^2 + b(p_0 - p_2) + c, \tag{2.17}$$

$$f(p_1) = a(p_1 - p_2)^2 + b(p_1 - p_2) + c, \tag{2.18}$$

and

$$f(p_2) = a \cdot 0^2 + b \cdot 0 + c = c \tag{2.19}$$

to be

$$c = f(p_2), \tag{2.20}$$

$$b = \frac{(p_0 - p_2)^2[f(p_1) - f(p_2)] - (p_1 - p_2)^2[f(p_0) - f(p_2)]}{(p_0 - p_2)(p_1 - p_2)(p_0 - p_1)}, \tag{2.21}$$

and

$$a = \frac{(p_1 - p_2)[f(p_0) - f(p_2)] - (p_0 - p_2)[f(p_1) - f(p_2)]}{(p_0 - p_2)(p_1 - p_2)(p_0 - p_1)}. \tag{2.22}$$

To determine p_3, a zero of P, we apply the quadratic formula to $P(x) = 0$. However, because of round-off error problems caused by the subtraction of nearly equal numbers, we apply the formula in the manner prescribed in Eq (1.2) and (1.3) of Section 1.2:

$$p_3 - p_2 = \frac{-2c}{b \pm \sqrt{b^2 - 4ac}}.$$

This formula gives two possibilities for p_3, depending on the sign preceding the radical term. In Müller's method, the sign is chosen to agree with the sign of b. Chosen in this manner, the denominator will be the largest in magnitude and will result in p_3 being selected as the closest zero of P to p_2. Thus

$$p_3 = p_2 - \frac{2c}{b + \operatorname{sgn}(b)\sqrt{b^2 - 4ac}},$$

where a, b, and c are given in Eqs. (2.20) through (2.22).

Once p_3 is determined, the procedure is reinitialized using p_1, p_2, and p_3 in place of p_0, p_1, and p_2 to determine the next approximation, p_4. The method continues until a satisfactory conclusion is obtained. At each step, the method involves the radical $\sqrt{b^2 - 4ac}$, so the method gives approximate complex roots when $b^2 - 4ac < 0$. Algorithm 2.8 implements this procedure.

Müller's

To find a solution to $f(x) = 0$ given three approximations, p_0, p_1, and p_2:

INPUT p_0, p_1, p_2; tolerance TOL; maximum number of iterations N_0.

OUTPUT approximate solution p or message of failure.

Step 1 Set $h_1 = p_1 - p_0$;
$h_2 = p_2 - p_1$;
$\delta_1 = (f(p_1) - f(p_0))/h_1$;
$\delta_2 = (f(p_2) - f(p_1))/h_2$;
$d = (\delta_2 - \delta_1)/(h_2 + h_1)$;
$i = 3$.

Step 2 While $i \le N_0$ do Steps 3–7.

Step 3 $b = \delta_2 + h_2 d$;
$D = (b^2 - 4f(p_2)d)^{1/2}$. (*Note: May require complex arithmetic.*)

Step 4 If $|b - D| < |b + D|$ then set $E = b + D$
else set $E = b - D$.

Step 5 Set $h = -2f(p_2)/E$;
$p = p_2 + h$.

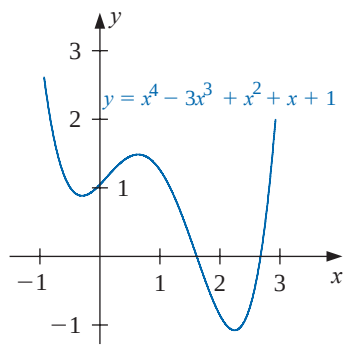
Step 6 If $|h| < TOL$ then
OUTPUT (p); (*The procedure was successful.*)
STOP.

Step 7 Set $p_0 = p_1$; (*Prepare for next iteration.*)
$p_1 = p_2$;
$p_2 = p$;
$h_1 = p_1 - p_0$;
$h_2 = p_2 - p_1$;
$\delta_1 = (f(p_1) - f(p_0))/h_1$;
$\delta_2 = (f(p_2) - f(p_1))/h_2$;
$d = (\delta_2 - \delta_1)/(h_2 + h_1)$;
$i = i + 1$.

Step 8 OUTPUT ('Method failed after N_0 iterations, $N_0 =$', N_0);
(*The procedure was unsuccessful.*)
STOP. ■

Illustration Consider the polynomial $f(x) = x^4 - 3x^3 + x^2 + x + 1$, part of whose graph is shown in Figure 2.14.

Figure 2.14



Three sets of three initial points will be used with Algorithm 2.8 and $TOL = 10^{-5}$ to approximate the zeros of f. The first set will use $p_0 = 0.5$, $p_1 = -0.5$, and $p_2 = 0$. The parabola passing through these points has complex roots because it does not intersect the x-axis. Table 2.12 gives approximations to the corresponding complex zeros of f.

Table 2.12

$p_0 = 0.5,\ p_1 = -0.5,\ p_2 = 0$		
i	p_i	$f(p_i)$
3	$-0.100000 + 0.888819i$	$-0.01120000 + 3.014875548i$
4	$-0.492146 + 0.447031i$	$-0.1691201 - 0.7367331502i$
5	$-0.352226 + 0.484132i$	$-0.1786004 + 0.0181872213i$
6	$-0.340229 + 0.443036i$	$0.01197670 - 0.0105562185i$
7	$-0.339095 + 0.446656i$	$-0.0010550 + 0.000387261i$
8	$-0.339093 + 0.446630i$	$0.000000 + 0.000000i$
9	$-0.339093 + 0.446630i$	$0.000000 + 0.000000i$

Table 2.13 gives the approximations to the two real zeros of f. The smallest of these uses $p_0 = 0.5$, $p_1 = 1.0$, and $p_2 = 1.5$, and the largest root is approximated when $p_0 = 1.5$, $p_1 = 2.0$, and $p_2 = 2.5$.

Table 2.13

$p_0 = 0.5$, $p_1 = 1.0$, $p_2 = 1.5$			$p_0 = 1.5$, $p_1 = 2.0$, $p_2 = 2.5$		
i	p_i	$f(p_i)$	i	p_i	$f(p_i)$
3	1.40637	−0.04851	3	2.24733	−0.24507
4	1.38878	0.00174	4	2.28652	−0.01446
5	1.38939	0.00000	5	2.28878	−0.00012
6	1.38939	0.00000	6	2.28880	0.00000
			7	2.28879	0.00000

The values in the tables are accurate approximations to the places listed. □

We used Maple to generate the results in Table 2.12. To find the first result in the table, define $f(x)$ with

$$f := x \to x^4 - 3x^3 + x^2 + x + 1$$

Then enter the initial approximations with

$$p0 := 0.5; p1 := -0.5; p2 := 0.0$$

and evaluate the function at these points with

$$f0 := f(p0); f1 := f(p1); f2 := f(p2)$$

To determine the coefficients a, b, c, and the approximate solution, enter

$$c := f2;$$

$$b := \frac{\left((p0 - p2)^2 \cdot (f1 - f2) - (p1 - p2)^2 \cdot (f0 - f2)\right)}{(p0 - p2) \cdot (p1 - p2) \cdot (p0 - p1)}$$

$$a := \frac{((p1 - p2) \cdot (f0 - f2) - (p0 - p2) \cdot (f1 - f2))}{(p0 - p2) \cdot (p1 - p2) \cdot (p0 - p1)}$$

$$p3 := p2 - \frac{2c}{b + \left(\frac{b}{abs(b)}\right)\sqrt{b^2 - 4a \cdot c}}$$

This produces the final Maple output

$$-0.1000000000 + 0.8888194418I$$

and evaluating at this approximation gives $f(p3)$ as

$$-0.0112000001 + 3.014875548I$$

This is our first approximation, as seen in Table 2.12.

The illustration shows that Müller's method can approximate the roots of polynomials with a variety of starting values. In fact, Müller's method generally converges to the root of a polynomial for any initial approximation choice, although problems can be constructed for

which convergence will not occur. For example, suppose that for some i we have $f(p_i) = f(p_{i+1}) = f(p_{i+2}) \neq 0$. The quadratic equation then reduces to a nonzero constant function and never intersects the x-axis. This is not usually the case, however, and general-purpose software packages using Müller's method request only one initial approximation per root and will even supply this approximation as an option.

EXERCISE SET 2.6

1. Find the approximations to within 10^{-4} to all the real zeros of the following polynomials using Newton's method.

a. $f(x) = x^3 - 2x^2 - 5$

b. $f(x) = x^3 + 3x^2 - 1$

c. $f(x) = x^3 - x - 1$

d. $f(x) = x^4 + 2x^2 - x - 3$

e. $f(x) = x^3 + 4.001x^2 + 4.002x + 1.101$

f. $f(x) = x^5 - x^4 + 2x^3 - 3x^2 + x - 4$

2. Find approximations to within 10^{-5} to all the zeros of each of the following polynomials by first finding the real zeros using Newton's method and then reducing to polynomials of lower degree to determine any complex zeros.

a. $f(x) = x^4 + 5x^3 - 9x^2 - 85x - 136$

b. $f(x) = x^4 - 2x^3 - 12x^2 + 16x - 40$

c. $f(x) = x^4 + x^3 + 3x^2 + 2x + 2$

d. $f(x) = x^5 + 11x^4 - 21x^3 - 10x^2 - 21x - 5$

e. $f(x) = 16x^4 + 88x^3 + 159x^2 + 76x - 240$

f. $f(x) = x^4 - 4x^2 - 3x + 5$

g. $f(x) = x^4 - 2x^3 - 4x^2 + 4x + 4$

h. $f(x) = x^3 - 7x^2 + 14x - 6$

3. Repeat Exercise 1 using Müller's method.

4. Repeat Exercise 2 using Müller's method.

5. Use Newton's method to find, within 10^{-3}, the zeros and critical points of the following functions. Use this information to sketch the graph of f.

a. $f(x) = x^3 - 9x^2 + 12$ **b.** $f(x) = x^4 - 2x^3 - 5x^2 + 12x - 5$

6. $f(x) = 10x^3 - 8.3x^2 + 2.295x - 0.21141 = 0$ has a root at $x = 0.29$. Use Newton's method with an initial approximation $x_0 = 0.28$ to attempt to find this root. Explain what happens.

7. Use Maple to find a real zero of the polynomial $f(x) = x^3 + 4x - 4$.

8. Use Maple to find a real zero of the polynomial $f(x) = x^3 - 2x - 5$.

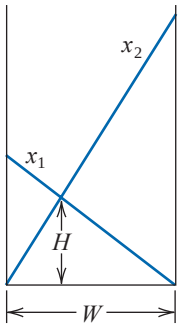
9. Use each of the following methods to find a solution in $[0.1, 1]$ accurate to within 10^{-4} for

$$600x^4 - 550x^3 + 200x^2 - 20x - 1 = 0.$$

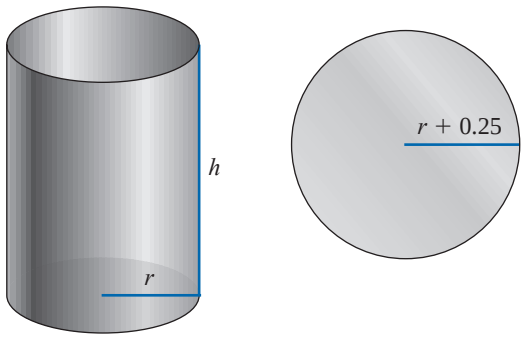
a. Bisection method **c.** Secant method **e.** Müller's method

b. Newton's method **d.** method of False Position

10. Two ladders crisscross an alley of width W. Each ladder reaches from the base of one wall to some point on the opposite wall. The ladders cross at a height H above the pavement. Find W given that the lengths of the ladders are $x_1 = 20$ ft and $x_2 = 30$ ft, and that $H = 8$ ft.



11. A can in the shape of a right circular cylinder is to be constructed to contain 1000 cm^3. The circular top and bottom of the can must have a radius of 0.25 cm more than the radius of the can so that the excess can be used to form a seal with the side. The sheet of material being formed into the side of the can must also be 0.25 cm longer than the circumference of the can so that a seal can be formed. Find, to within 10^{-4}, the minimal amount of material needed to construct the can.



12. In 1224, Leonardo of Pisa, better known as Fibonacci, answered a mathematical challenge of John of Palermo in the presence of Emperor Frederick II: find a root of the equation $x^3 + 2x^2 + 10x = 20$. He first showed that the equation had no rational roots and no Euclidean irrational root—that is, no root in any of the forms $a \pm \sqrt{b}$, $\sqrt{a} \pm \sqrt{b}$, $\sqrt{a \pm \sqrt{b}}$, or $\sqrt{\sqrt{a} \pm \sqrt{b}}$, where a and b are rational numbers. He then approximated the only real root, probably using an algebraic technique of Omar Khayyam involving the intersection of a circle and a parabola. His answer was given in the base-60 number system as

$$1 + 22\left(\frac{1}{60}\right) + 7\left(\frac{1}{60}\right)^2 + 42\left(\frac{1}{60}\right)^3 + 33\left(\frac{1}{60}\right)^4 + 4\left(\frac{1}{60}\right)^5 + 40\left(\frac{1}{60}\right)^6.$$

How accurate was his approximation?

2.7 Survey of Methods and Software

In this chapter we have considered the problem of solving the equation $f(x) = 0$, where f is a given continuous function. All the methods begin with initial approximations and generate a sequence that converges to a root of the equation, if the method is successful. If $[a, b]$ is an interval on which $f(a)$ and $f(b)$ are of opposite sign, then the Bisection method and the method of False Position will converge. However, the convergence of these methods might be slow. Faster convergence is generally obtained using the Secant method or Newton's method. Good initial approximations are required for these methods, two for the Secant method and one for Newton's method, so the root-bracketing techniques such as Bisection or the False Position method can be used as starter methods for the Secant or Newton's method.

Müller's method will give rapid convergence without a particularly good initial approximation. It is not quite as efficient as Newton's method; its order of convergence near a root is approximately $\alpha = 1.84$, compared to the quadratic, $\alpha = 2$, order of Newton's method. However, it is better than the Secant method, whose order is approximately $\alpha = 1.62$, and it has the added advantage of being able to approximate complex roots.

Deflation is generally used with Müller's method once an approximate root of a polynomial has been determined. After an approximation to the root of the deflated equation has been determined, use either Müller's method or Newton's method in the original polynomial with this root as the initial approximation. This procedure will ensure that the root being approximated is a solution to the true equation, not to the deflated equation. We recommended Müller's method for finding all the zeros of polynomials, real or complex. Müller's method can also be used for an arbitrary continuous function.

Other high-order methods are available for determining the roots of polynomials. If this topic is of particular interest, we recommend that consideration be given to Laguerre's method, which gives cubic convergence and also approximates complex roots (see [Ho], pp. 176–179 for a complete discussion), the Jenkins-Traub method (see [JT]), and Brent's method (see [Bre]).

Another method of interest, Cauchy's method, is similar to Müller's method but avoids the failure problem of Müller's method when $f(x_i) = f(x_{i+1}) = f(x_{i+2})$, for some i. For an interesting discussion of this method, as well as more detail on Müller's method, we recommend [YG], Sections 4.10, 4.11, and 5.4.

Given a specified function f and a tolerance, an efficient program should produce an approximation to one or more solutions of $f(x) = 0$, each having an absolute or relative error within the tolerance, and the results should be generated in a reasonable amount of time. If the program cannot accomplish this task, it should at least give meaningful explanations of why success was not obtained and an indication of how to remedy the cause of failure.

IMSL has subroutines that implement Müller's method with deflation. Also included in this package is a routine due to R. P. Brent that uses a combination of linear interpolation, an inverse quadratic interpolation similar to Müller's method, and the Bisection method. Laguerre's method is also used to find zeros of a real polynomial. Another routine for finding the zeros of real polynomials uses a method of Jenkins-Traub, which is also used to find zeros of a complex polynomial.

The NAG library has a subroutine that uses a combination of the Bisection method, linear interpolation, and extrapolation to approximate a real zero of a function on a given interval. NAG also supplies subroutines to approximate all zeros of a real polynomial or complex polynomial, respectively. Both subroutines use a modified Laguerre method.

The netlib library contains a subroutine that uses a combination of the Bisection and Secant method developed by T. J. Dekker to approximate a real zero of a function in the interval. It requires specifying an interval that contains a root and returns an interval with a width that is within a specified tolerance. Another subroutine uses a combination of the bisection method, interpolation, and extrapolation to find a real zero of the function on the interval.

MATLAB has a routine to compute all the roots, both real and complex, of a polynomial, and one that computes a zero near a specified initial approximation to within a specified tolerance.

Notice that in spite of the diversity of methods, the professionally written packages are based primarily on the methods and principles discussed in this chapter. You should be able to use these packages by reading the manuals accompanying the packages to better understand the parameters and the specifications of the results that are obtained.

There are three books that we consider to be classics on the solution of nonlinear equations: those by Traub [Tr], by Ostrowski [Os], and by Householder [Ho]. In addition, the book by Brent [Bre] served as the basis for many of the currently used root-finding methods.

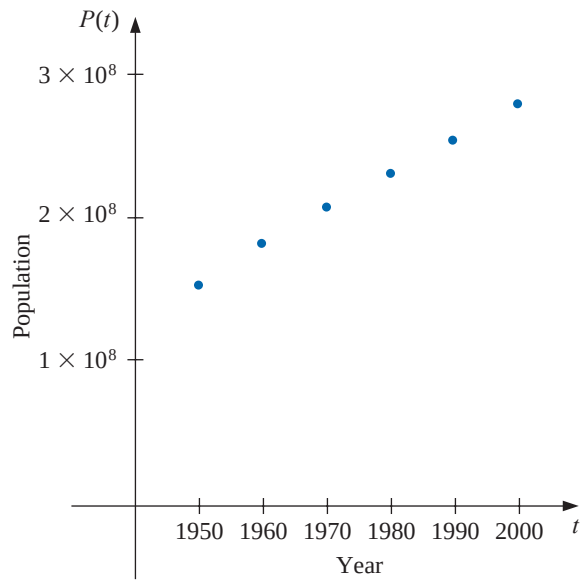
CHAPTER

3 Interpolation and Polynomial Approximation

Introduction

A census of the population of the United States is taken every 10 years. The following table lists the population, in thousands of people, from 1950 to 2000, and the data are also represented in the figure.

Year	1950	1960	1970	1980	1990	2000
Population (in thousands)	151,326	179,323	203,302	226,542	249,633	281,422



In reviewing these data, we might ask whether they could be used to provide a reasonable estimate of the population, say, in 1975 or even in the year 2020. Predictions of this type can be obtained by using a function that fits the given data. This process is called *interpolation* and is the subject of this chapter. This population problem is considered throughout the chapter and in Exercises 18 of Section 3.1, 18 of Section 3.3, and 28 of Section 3.5.

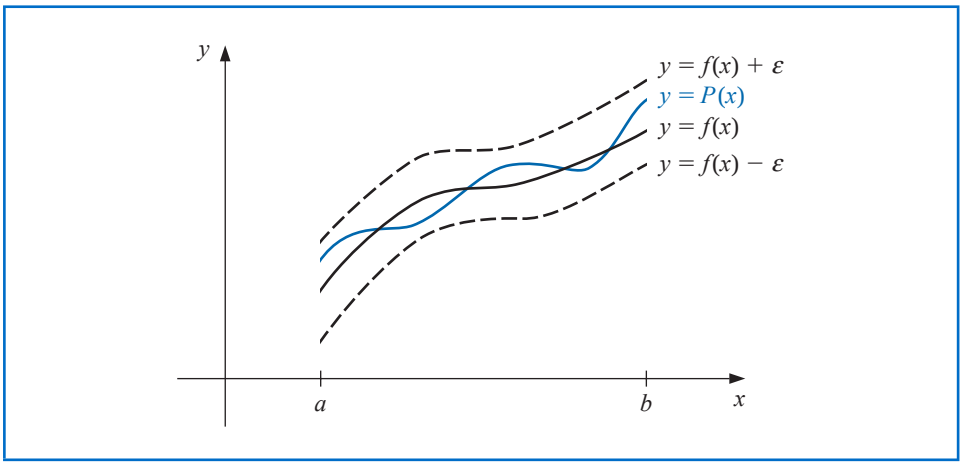
3.1 Interpolation and the Lagrange Polynomial

One of the most useful and well-known classes of functions mapping the set of real numbers into itself is the *algebraic polynomials*, the set of functions of the form

$$P_n(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0,$$

where n is a nonnegative integer and $a_0, \ldots, a_n$ are real constants. One reason for their importance is that they uniformly approximate continuous functions. By this we mean that given any function, defined and continuous on a closed and bounded interval, there exists a polynomial that is as "close" to the given function as desired. This result is expressed precisely in the Weierstrass Approximation Theorem. (See Figure 3.1.)

Figure 3.1



Theorem 3.1 **(Weierstrass Approximation Theorem)**

Suppose that f is defined and continuous on $[a, b]$. For each $\epsilon > 0$, there exists a polynomial $P(x)$, with the property that

$$|f(x) - P(x)| < \epsilon, \quad \text{for all } x \text{ in } [a, b]. \qquad \blacksquare$$

The proof of this theorem can be found in most elementary texts on real analysis (see, for example, [Bart], pp. 165–172).

Another important reason for considering the class of polynomials in the approximation of functions is that the derivative and indefinite integral of a polynomial are easy to determine and are also polynomials. For these reasons, polynomials are often used for approximating continuous functions.

The Taylor polynomials were introduced in Section 1.1, where they were described as one of the fundamental building blocks of numerical analysis. Given this prominence, you might expect that polynomial interpolation would make heavy use of these functions. However this is not the case. The Taylor polynomials agree as closely as possible with a given function at a specific point, but they concentrate their accuracy near that point. A good interpolation polynomial needs to provide a relatively accurate approximation over an entire interval, and Taylor polynomials do not generally do this. For example, suppose we calculate the first six Taylor polynomials about $x_0 = 0$ for $f(x) = e^x$. Since the derivatives of $f(x)$ are all e^x, which evaluated at $x_0 = 0$ gives 1, the Taylor polynomials are

Karl Weierstrass (1815–1897) is often referred to as the father of modern analysis because of his insistence on rigor in the demonstration of mathematical results. He was instrumental in developing tests for convergence of series, and determining ways to rigorously define irrational numbers. He was the first to demonstrate that a function could be everywhere continuous but nowhere differentiable, a result that shocked some of his contemporaries.

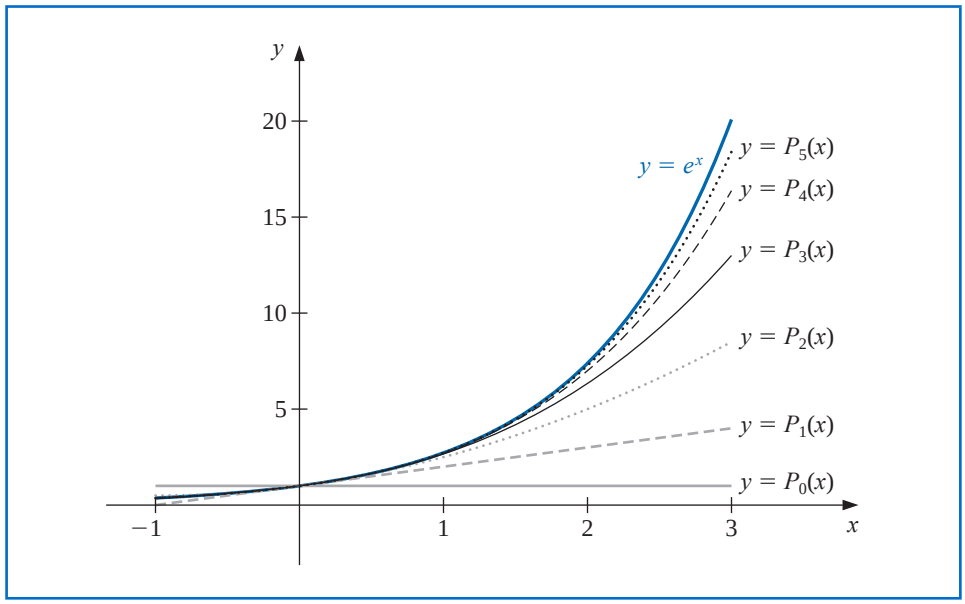
Very little of Weierstrass's work was published during his lifetime, but his lectures, particularly on the theory of functions, had significant influence on an entire generation of students.

$$P_0(x) = 1, \quad P_1(x) = 1 + x, \quad P_2(x) = 1 + x + \frac{x^2}{2}, \quad P_3(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{6},$$

$$P_4(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24}, \quad \text{and} \quad P_5(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \frac{x^5}{120}.$$

The graphs of the polynomials are shown in Figure 3.2. (Notice that even for the higher-degree polynomials, the error becomes progressively worse as we move away from zero.)

Figure 3.2



Although better approximations are obtained for $f(x) = e^x$ if higher-degree Taylor polynomials are used, this is not true for all functions. Consider, as an extreme example, using Taylor polynomials of various degrees for $f(x) = 1/x$ expanded about $x_0 = 1$ to approximate $f(3) = 1/3$. Since

$$f(x) = x^{-1}, \ f'(x) = -x^{-2}, \ f''(x) = (-1)^2 2 \cdot x^{-3},$$

and, in general,

$$f^{(k)}(x) = (-1)^k k! x^{-k-1},$$

the Taylor polynomials are

$$P_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(1)}{k!}(x-1)^k = \sum_{k=0}^{n} (-1)^k (x-1)^k.$$

To approximate $f(3) = 1/3$ by $P_n(3)$ for increasing values of n, we obtain the values in Table 3.1—rather a dramatic failure! When we approximate $f(3) = 1/3$ by $P_n(3)$ for larger values of n, the approximations become increasingly inaccurate.

Table 3.1

n	0	1	2	3	4	5	6	7
$P_n(3)$	1	−1	3	−5	11	−21	43	−85

For the Taylor polynomials all the information used in the approximation is concentrated at the single number x_0, so these polynomials will generally give inaccurate approximations as we move away from x_0. This limits Taylor polynomial approximation to the situation in which approximations are needed only at numbers close to x_0. For ordinary computational purposes it is more efficient to use methods that include information at various points. We consider this in the remainder of the chapter. The primary use of Taylor polynomials in numerical analysis is not for approximation purposes, but for the derivation of numerical techniques and error estimation.

Lagrange Interpolating Polynomials

The problem of determining a polynomial of degree one that passes through the distinct points (x_0, y_0) and (x_1, y_1) is the same as approximating a function f for which $f(x_0) = y_0$ and $f(x_1) = y_1$ by means of a first-degree polynomial **interpolating**, or agreeing with, the values of f at the given points. Using this polynomial for approximation within the interval given by the endpoints is called polynomial **interpolation**.

Define the functions

$$L_0(x) = \frac{x - x_1}{x_0 - x_1} \quad \text{and} \quad L_1(x) = \frac{x - x_0}{x_1 - x_0}.$$

The linear **Lagrange interpolating polynomial** through (x_0, y_0) and (x_1, y_1) is

$$P(x) = L_0(x)f(x_0) + L_1(x)f(x_1) = \frac{x - x_1}{x_0 - x_1}f(x_0) + \frac{x - x_0}{x_1 - x_0}f(x_1).$$

Note that

$$L_0(x_0) = 1, \quad L_0(x_1) = 0, \quad L_1(x_0) = 0, \quad \text{and} \quad L_1(x_1) = 1,$$

which implies that

$$P(x_0) = 1 \cdot f(x_0) + 0 \cdot f(x_1) = f(x_0) = y_0$$

and

$$P(x_1) = 0 \cdot f(x_0) + 1 \cdot f(x_1) = f(x_1) = y_1.$$

So P is the unique polynomial of degree at most one that passes through (x_0, y_0) and (x_1, y_1).

Example 1 Determine the linear Lagrange interpolating polynomial that passes through the points $(2, 4)$ and $(5, 1)$.

Solution In this case we have

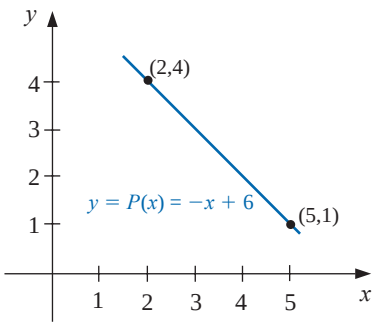
$$L_0(x) = \frac{x - 5}{2 - 5} = -\frac{1}{3}(x - 5) \quad \text{and} \quad L_1(x) = \frac{x - 2}{5 - 2} = \frac{1}{3}(x - 2),$$

so

$$P(x) = -\frac{1}{3}(x - 5) \cdot 4 + \frac{1}{3}(x - 2) \cdot 1 = -\frac{4}{3}x + \frac{20}{3} + \frac{1}{3}x - \frac{2}{3} = -x + 6.$$

The graph of $y = P(x)$ is shown in Figure 3.3. ■

Figure 3.3

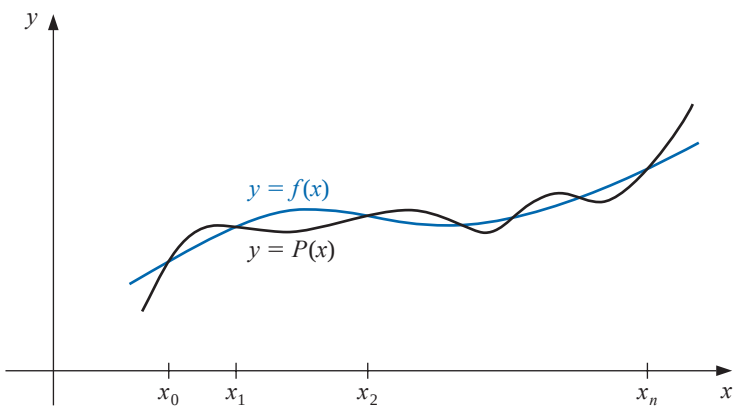


To generalize the concept of linear interpolation, consider the construction of a polynomial of degree at most n that passes through the $n + 1$ points

$$(x_0, f(x_0)),\ (x_1, f(x_1)), \ldots, (x_n, f(x_n)).$$

(See Figure 3.4.)

Figure 3.4



In this case we first construct, for each $k = 0, 1, \ldots, n$, a function $L_{n,k}(x)$ with the property that $L_{n,k}(x_i) = 0$ when $i \neq k$ and $L_{n,k}(x_k) = 1$. To satisfy $L_{n,k}(x_i) = 0$ for each $i \neq k$ requires that the numerator of $L_{n,k}(x)$ contain the term

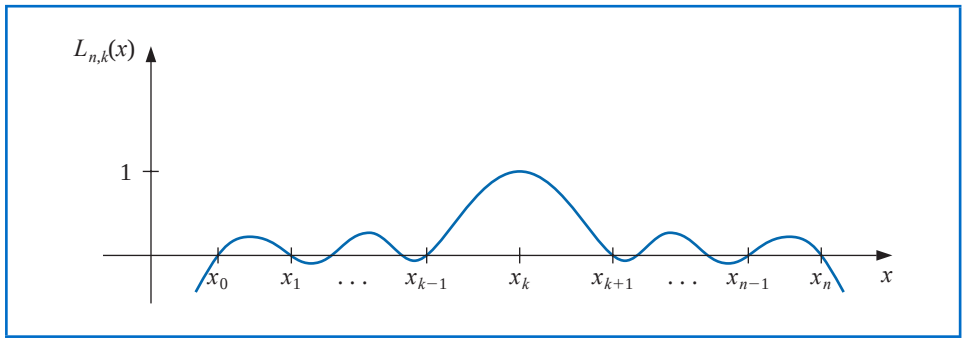
$$(x - x_0)(x - x_1) \cdots (x - x_{k-1})(x - x_{k+1}) \cdots (x - x_n).$$

To satisfy $L_{n,k}(x_k) = 1$, the denominator of $L_{n,k}(x)$ must be this same term but evaluated at $x = x_k$. Thus

$$L_{n,k}(x) = \frac{(x - x_0) \cdots (x - x_{k-1})(x - x_{k+1}) \cdots (x - x_n)}{(x_k - x_0) \cdots (x_k - x_{k-1})(x_k - x_{k+1}) \cdots (x_k - x_n)}.$$

A sketch of the graph of a typical $L_{n,k}$ (when n is even) is shown in Figure 3.5.

Figure 3.5



The interpolating polynomial is easily described once the form of $L_{n,k}$ is known. This polynomial, called the ***n*th Lagrange interpolating polynomial**, is defined in the following theorem.

Theorem 3.2 If $x_0, x_1, \ldots, x_n$ are $n+1$ distinct numbers and f is a function whose values are given at these numbers, then a unique polynomial $P(x)$ of degree at most n exists with

$$f(x_k) = P(x_k), \quad \text{for each } k = 0, 1, \ldots, n.$$

This polynomial is given by

$$P(x) = f(x_0)L_{n,0}(x) + \cdots + f(x_n)L_{n,n}(x) = \sum_{k=0}^{n} f(x_k)L_{n,k}(x), \tag{3.1}$$

where, for each $k = 0, 1, \ldots, n$,

$$L_{n,k}(x) = \frac{(x-x_0)(x-x_1)\cdots(x-x_{k-1})(x-x_{k+1})\cdots(x-x_n)}{(x_k-x_0)(x_k-x_1)\cdots(x_k-x_{k-1})(x_k-x_{k+1})\cdots(x_k-x_n)} \tag{3.2}$$

$$= \prod_{\substack{i=0 \\ i\neq k}}^{n} \frac{(x-x_i)}{(x_k-x_i)}.$$

■

The interpolation formula named for Joseph Louis Lagrange (1736–1813) was likely known by Isaac Newton around 1675, but it appears to first have been published in 1779 by Edward Waring (1736–1798). Lagrange wrote extensively on the subject of interpolation and his work had significant influence on later mathematicians. He published this result in 1795.

The symbol $\prod$ is used to write products compactly and parallels the symbol $\sum$, which is used for writing sums.

We will write $L_{n,k}(x)$ simply as $L_k(x)$ when there is no confusion as to its degree.

Example 2

(a) Use the numbers (called *nodes*) $x_0 = 2$, $x_1 = 2.75$, and $x_2 = 4$ to find the second Lagrange interpolating polynomial for $f(x) = 1/x$.

(b) Use this polynomial to approximate $f(3) = 1/3$.

Solution **(a)** We first determine the coefficient polynomials $L_0(x)$, $L_1(x)$, and $L_2(x)$. In nested form they are

$$L_0(x) = \frac{(x-2.75)(x-4)}{(2-2.5)(2-4)} = \frac{2}{3}(x-2.75)(x-4),$$

$$L_1(x) = \frac{(x-2)(x-4)}{(2.75-2)(2.75-4)} = -\frac{16}{15}(x-2)(x-4),$$

and

$$L_2(x) = \frac{(x-2)(x-2.75)}{(4-2)(4-2.5)} = \frac{2}{5}(x-2)(x-2.75).$$

Also, $f(x_0) = f(2) = 1/2$, $f(x_1) = f(2.75) = 4/11$, and $f(x_2) = f(4) = 1/4$, so

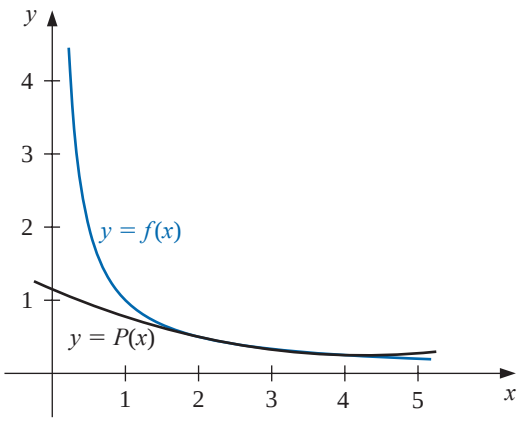
$$\begin{aligned} P(x) &= \sum_{k=0}^{2} f(x_k)L_k(x) \\ &= \frac{1}{3}(x - 2.75)(x - 4) - \frac{64}{165}(x - 2)(x - 4) + \frac{1}{10}(x - 2)(x - 2.75) \\ &= \frac{1}{22}x^2 - \frac{35}{88}x + \frac{49}{44}. \end{aligned}$$

(b) An approximation to $f(3) = 1/3$ (see Figure 3.6) is

$$f(3) \approx P(3) = \frac{9}{22} - \frac{105}{88} + \frac{49}{44} = \frac{29}{88} \approx 0.32955.$$

Recall that in the opening section of this chapter (see Table 3.1) we found that no Taylor polynomial expanded about $x_0 = 1$ could be used to reasonably approximate $f(x) = 1/x$ at $x = 3$. ■

Figure 3.6



The interpolating polynomial P of degree less than or equal to 3 is defined in Maple with

$P := x \rightarrow interp([2, 11/4, 4], [1/2, 4/11, 1/4], x)$

$$x \rightarrow interp\left(\left[2, \frac{11}{4}, 4\right], \left[\frac{1}{2}, \frac{4}{11}, \frac{1}{4}\right], x\right)$$

To see the polynomial, enter

$P(x)$

$$\frac{1}{22}x^2 - \frac{35}{88}x + \frac{49}{44}$$

Evaluating $P(3)$ as an approximation to $f(3) = 1/3$, is found with

$evalf(P(3))$

$$0.3295454545$$

The interpolating polynomial can also be defined in Maple using the *CurveFitting* package and the call *PolynomialInterpolation*.

The next step is to calculate a remainder term or bound for the error involved in approximating a function by an interpolating polynomial.

Theorem 3.3 Suppose $x_0, x_1, \ldots, x_n$ are distinct numbers in the interval $[a, b]$ and $f \in C^{n+1}[a, b]$. Then, for each x in $[a, b]$, a number $\xi(x)$ (generally unknown) between $x_0, x_1, \ldots, x_n$, and hence in (a, b), exists with

$$f(x) = P(x) + \frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x - x_0)(x - x_1)\cdots(x - x_n), \tag{3.3}$$

where $P(x)$ is the interpolating polynomial given in Eq. (3.1). ■

There are other ways that the error term for the Lagrange polynomial can be expressed, but this is the most useful form and the one that most closely agrees with the standard Taylor polynomial error form.

Proof Note first that if $x = x_k$, for any $k = 0, 1, \ldots, n$, then $f(x_k) = P(x_k)$, and choosing $\xi(x_k)$ arbitrarily in (a, b) yields Eq. (3.3).

If $x \neq x_k$, for all $k = 0, 1, \ldots, n$, define the function g for t in $[a, b]$ by

$$g(t) = f(t) - P(t) - [f(x) - P(x)]\frac{(t - x_0)(t - x_1)\cdots(t - x_n)}{(x - x_0)(x - x_1)\cdots(x - x_n)}$$

$$= f(t) - P(t) - [f(x) - P(x)]\prod_{i=0}^{n}\frac{(t - x_i)}{(x - x_i)}.$$

Since $f \in C^{n+1}[a, b]$, and $P \in C^{\infty}[a, b]$, it follows that $g \in C^{n+1}[a, b]$. For $t = x_k$, we have

$$g(x_k) = f(x_k) - P(x_k) - [f(x) - P(x)]\prod_{i=0}^{n}\frac{(x_k - x_i)}{(x - x_i)} = 0 - [f(x) - P(x)] \cdot 0 = 0.$$

Moreover,

$$g(x) = f(x) - P(x) - [f(x) - P(x)]\prod_{i=0}^{n}\frac{(x - x_i)}{(x - x_i)} = f(x) - P(x) - [f(x) - P(x)] = 0.$$

Thus $g \in C^{n+1}[a, b]$, and g is zero at the $n + 2$ distinct numbers $x, x_0, x_1, \ldots, x_n$. By Generalized Rolle's Theorem 1.10, there exists a number ξ in (a, b) for which $g^{(n+1)}(\xi) = 0$. So

$$0 = g^{(n+1)}(\xi) = f^{(n+1)}(\xi) - P^{(n+1)}(\xi) - [f(x) - P(x)]\frac{d^{n+1}}{dt^{n+1}}\left[\prod_{i=0}^{n}\frac{(t - x_i)}{(x - x_i)}\right]_{t=\xi}. \tag{3.4}$$

However $P(x)$ is a polynomial of degree at most n, so the $(n+1)$st derivative, $P^{(n+1)}(x)$, is identically zero. Also, $\prod_{i=0}^{n}[(t - x_i)/(x - x_i)]$ is a polynomial of degree $(n + 1)$, so

$$\prod_{i=0}^{n}\frac{(t - x_i)}{(x - x_i)} = \left[\frac{1}{\prod_{i=0}^{n}(x - x_i)}\right]t^{n+1} + (\text{lower-degree terms in } t),$$

and

$$\frac{d^{n+1}}{dt^{n+1}}\prod_{i=0}^{n}\frac{(t - x_i)}{(x - x_i)} = \frac{(n+1)!}{\prod_{i=0}^{n}(x - x_i)}.$$

Equation (3.4) now becomes

$$0 = f^{(n+1)}(\xi) - 0 - [f(x) - P(x)]\frac{(n+1)!}{\prod_{i=0}^{n}(x - x_i)},$$

and, upon solving for $f(x)$, we have

$$f(x) = P(x) + \frac{f^{(n+1)}(\xi)}{(n+1)!}\prod_{i=0}^{n}(x - x_i). \quad \blacksquare\ \blacksquare\ \blacksquare$$

The error formula in Theorem 3.3 is an important theoretical result because Lagrange polynomials are used extensively for deriving numerical differentiation and integration methods. Error bounds for these techniques are obtained from the Lagrange error formula.

Note that the error form for the Lagrange polynomial is quite similar to that for the Taylor polynomial. The nth Taylor polynomial about x_0 concentrates all the known information at x_0 and has an error term of the form

$$\frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x - x_0)^{n+1}.$$

The Lagrange polynomial of degree n uses information at the distinct numbers $x_0, x_1, \ldots, x_n$ and, in place of $(x - x_0)^n$, its error formula uses a product of the $n + 1$ terms $(x - x_0)$, $(x - x_1), \ldots, (x - x_n)$:

$$\frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x - x_0)(x - x_1)\cdots(x - x_n).$$

Example 3 In Example 2 we found the second Lagrange polynomial for $f(x) = 1/x$ on $[2, 4]$ using the nodes $x_0 = 2$, $x_1 = 2.75$, and $x_2 = 4$. Determine the error form for this polynomial, and the maximum error when the polynomial is used to approximate $f(x)$ for $x \in [2, 4]$.

Solution Because $f(x) = x^{-1}$, we have

$$f'(x) = -x^{-2}, \quad f''(x) = 2x^{-3}, \quad \text{and} \quad f'''(x) = -6x^{-4}.$$

As a consequence, the second Lagrange polynomial has the error form

$$\frac{f'''(\xi(x))}{3!}(x-x_0)(x-x_1)(x-x_2) = -(\xi(x))^{-4}(x-2)(x-2.75)(x-4), \quad \text{for } \xi(x) \text{ in } (2, 4).$$

The maximum value of $(\xi(x))^{-4}$ on the interval is $2^{-4} = 1/16$. We now need to determine the maximum value on this interval of the absolute value of the polynomial

$$g(x) = (x - 2)(x - 2.75)(x - 4) = x^3 - \frac{35}{4}x^2 + \frac{49}{2}x - 22.$$

Because

$$D_x\left(x^3 - \frac{35}{4}x^2 + \frac{49}{2}x - 22\right) = 3x^2 - \frac{35}{2}x + \frac{49}{2} = \frac{1}{2}(3x - 7)(2x - 7),$$

the critical points occur at

$$x = \frac{7}{3}, \text{ with } g\left(\frac{7}{3}\right) = \frac{25}{108}, \qquad \text{and} \qquad x = \frac{7}{2}, \text{ with } g\left(\frac{7}{2}\right) = -\frac{9}{16}.$$

Hence, the maximum error is

$$\frac{f'''(\xi(x))}{3!}|(x - x_0)(x - x_1)(x - x_2)| \le \frac{1}{16 \cdot 6}\left|-\frac{9}{16}\right| = \frac{3}{512} \approx 0.00586. \quad \blacksquare$$

The next example illustrates how the error formula can be used to prepare a table of data that will ensure a specified interpolation error within a specified bound.

Example 4 Suppose a table is to be prepared for the function $f(x) = e^x$, for x in $[0, 1]$. Assume the number of decimal places to be given per entry is $d \geq 8$ and that the difference between adjacent x-values, the step size, is h. What step size h will ensure that linear interpolation gives an absolute error of at most 10^{-6} for all x in $[0, 1]$?

Solution Let $x_0, x_1, \ldots$ be the numbers at which f is evaluated, x be in [0,1], and suppose j satisfies $x_j \leq x \leq x_{j+1}$. Eq. (3.3) implies that the error in linear interpolation is

$$|f(x) - P(x)| = \left| \frac{f^{(2)}(\xi)}{2!}(x - x_j)(x - x_{j+1}) \right| = \frac{|f^{(2)}(\xi)|}{2}|(x - x_j)||(x - x_{j+1})|.$$

The step size is h, so $x_j = jh$, $x_{j+1} = (j + 1)h$, and

$$|f(x) - P(x)| \leq \frac{|f^{(2)}(\xi)|}{2!}|(x - jh)(x - (j + 1)h)|.$$

Hence

$$\begin{aligned} |f(x) - P(x)| &\leq \frac{\max_{\xi \in [0,1]} e^{\xi}}{2} \max_{x_j \leq x \leq x_{j+1}} |(x - jh)(x - (j + 1)h)| \\ &\leq \frac{e}{2} \max_{x_j \leq x \leq x_{j+1}} |(x - jh)(x - (j + 1)h)|. \end{aligned}$$

Consider the function $g(x) = (x - jh)(x - (j + 1)h)$, for $jh \leq x \leq (j + 1)h$. Because

$$g'(x) = (x - (j + 1)h) + (x - jh) = 2\left(x - jh - \frac{h}{2}\right),$$

the only critical point for g is at $x = jh + h/2$, with $g(jh + h/2) = (h/2)^2 = h^2/4$.

Since $g(jh) = 0$ and $g((j + 1)h) = 0$, the maximum value of $|g'(x)|$ in $[jh, (j + 1)h]$ must occur at the critical point which implies that

$$|f(x) - P(x)| \leq \frac{e}{2} \max_{x_j \leq x \leq x_{j+1}} |g(x)| \leq \frac{e}{2} \cdot \frac{h^2}{4} = \frac{eh^2}{8}.$$

Consequently, to ensure that the the error in linear interpolation is bounded by 10^{-6}, it is sufficient for h to be chosen so that

$$\frac{eh^2}{8} \leq 10^{-6}. \quad \text{This implies that} \quad h < 1.72 \times 10^{-3}.$$

Because $n = (1 - 0)/h$ must be an integer, a reasonable choice for the step size is $h = 0.001$. ■

EXERCISE SET 3.1

1. For the given functions $f(x)$, let $x_0 = 0$, $x_1 = 0.6$, and $x_2 = 0.9$. Construct interpolation polynomials of degree at most one and at most two to approximate $f(0.45)$, and find the absolute error.

a. $f(x) = \cos x$

b. $f(x) = \sqrt{1 + x}$

c. $f(x) = \ln(x + 1)$

d. $f(x) = \tan x$

2. For the given functions $f(x)$, let $x_0 = 1, x_1 = 1.25$, and $x_2 = 1.6$. Construct interpolation polynomials of degree at most one and at most two to approximate $f(1.4)$, and find the absolute error.

a. $f(x) = \sin \pi x$

b. $f(x) = \sqrt[3]{x-1}$

c. $f(x) = \log_{10}(3x-1)$

d. $f(x) = e^{2x} - x$

3. Use Theorem 3.3 to find an error bound for the approximations in Exercise 1.

4. Use Theorem 3.3 to find an error bound for the approximations in Exercise 2.

5. Use appropriate Lagrange interpolating polynomials of degrees one, two, and three to approximate each of the following:

a. $f(8.4)$ if $f(8.1) = 16.94410$, $f(8.3) = 17.56492$, $f(8.6) = 18.50515$, $f(8.7) = 18.82091$

b. $f\left(-\frac{1}{3}\right)$ if $f(-0.75) = -0.07181250$, $f(-0.5) = -0.02475000$, $f(-0.25) = 0.33493750$, $f(0) = 1.10100000$

c. $f(0.25)$ if $f(0.1) = 0.62049958$, $f(0.2) = -0.28398668$, $f(0.3) = 0.00660095$, $f(0.4) = 0.24842440$

d. $f(0.9)$ if $f(0.6) = -0.17694460$, $f(0.7) = 0.01375227$, $f(0.8) = 0.22363362$, $f(1.0) = 0.65809197$

6. Use appropriate Lagrange interpolating polynomials of degrees one, two, and three to approximate each of the following:

a. $f(0.43)$ if $f(0) = 1$, $f(0.25) = 1.64872$, $f(0.5) = 2.71828$, $f(0.75) = 4.48169$

b. $f(0)$ if $f(-0.5) = 1.93750$, $f(-0.25) = 1.33203$, $f(0.25) = 0.800781$, $f(0.5) = 0.687500$

c. $f(0.18)$ if $f(0.1) = -0.29004986$, $f(0.2) = -0.56079734$, $f(0.3) = -0.81401972$, $f(0.4) = -1.0526302$

d. $f(0.25)$ if $f(-1) = 0.86199480$, $f(-0.5) = 0.95802009$, $f(0) = 1.0986123$, $f(0.5) = 1.2943767$

7. The data for Exercise 5 were generated using the following functions. Use the error formula to find a bound for the error, and compare the bound to the actual error for the cases $n = 1$ and $n = 2$.

a. $f(x) = x \ln x$

b. $f(x) = x^3 + 4.001x^2 + 4.002x + 1.101$

c. $f(x) = x \cos x - 2x^2 + 3x - 1$

d. $f(x) = \sin(e^x - 2)$

8. The data for Exercise 6 were generated using the following functions. Use the error formula to find a bound for the error, and compare the bound to the actual error for the cases $n = 1$ and $n = 2$.

a. $f(x) = e^{2x}$

b. $f(x) = x^4 - x^3 + x^2 - x + 1$

c. $f(x) = x^2 \cos x - 3x$

d. $f(x) = \ln(e^x + 2)$

9. Let $P_3(x)$ be the interpolating polynomial for the data $(0, 0)$, $(0.5, y)$, $(1, 3)$, and $(2, 2)$. The coefficient of x^3 in $P_3(x)$ is 6. Find y.

10. Let $f(x) = \sqrt{x - x^2}$ and $P_2(x)$ be the interpolation polynomial on $x_0 = 0$, x_1 and $x_2 = 1$. Find the largest value of x_1 in $(0, 1)$ for which $f(0.5) - P_2(0.5) = -0.25$.

11. Use the following values and four-digit rounding arithmetic to construct a third Lagrange polynomial approximation to $f(1.09)$. The function being approximated is $f(x) = \log_{10}(\tan x)$. Use this knowledge to find a bound for the error in the approximation.

$$f(1.00) = 0.1924 \quad f(1.05) = 0.2414 \quad f(1.10) = 0.2933 \quad f(1.15) = 0.3492$$

12. Use the Lagrange interpolating polynomial of degree three or less and four-digit chopping arithmetic to approximate $\cos 0.750$ using the following values. Find an error bound for the approximation.

$$\cos 0.698 = 0.7661 \quad \cos 0.733 = 0.7432 \quad \cos 0.768 = 0.7193 \quad \cos 0.803 = 0.6946$$

The actual value of $\cos 0.750$ is 0.7317 (to four decimal places). Explain the discrepancy between the actual error and the error bound.

13. Construct the Lagrange interpolating polynomials for the following functions, and find a bound for the absolute error on the interval $[x_0, x_n]$.

a. $f(x) = e^{2x}\cos 3x, \quad x_0 = 0, x_1 = 0.3, x_2 = 0.6, n = 2$

b. $f(x) = \sin(\ln x), \quad x_0 = 2.0, x_1 = 2.4, x_2 = 2.6, n = 2$

c. $f(x) = \ln x, \quad x_0 = 1, x_1 = 1.1, x_2 = 1.3, x_3 = 1.4, n = 3$

d. $f(x) = \cos x + \sin x, \quad x_0 = 0, x_1 = 0.25, x_2 = 0.5, x_3 = 1.0, n = 3$

14. Let $f(x) = e^x$, for $0 \le x \le 2$.

a. Approximate $f(0.25)$ using linear interpolation with $x_0 = 0$ and $x_1 = 0.5$.

b. Approximate $f(0.75)$ using linear interpolation with $x_0 = 0.5$ and $x_1 = 1$.

c. Approximate $f(0.25)$ and $f(0.75)$ by using the second interpolating polynomial with $x_0 = 0$, $x_1 = 1$, and $x_2 = 2$.

d. Which approximations are better and why?

15. Repeat Exercise 11 using Maple with *Digits* set to 10.

16. Repeat Exercise 12 using Maple with *Digits* set to 10.

17. Suppose you need to construct eight-decimal-place tables for the common, or base-10, logarithm function from $x = 1$ to $x = 10$ in such a way that linear interpolation is accurate to within 10^{-6}. Determine a bound for the step size for this table. What choice of step size would you make to ensure that $x = 10$ is included in the table?

18. **a.** The introduction to this chapter included a table listing the population of the United States from 1950 to 2000. Use Lagrange interpolation to approximate the population in the years 1940, 1975, and 2020.

b. The population in 1940 was approximately 132,165,000. How accurate do you think your 1975 and 2020 figures are?

19. It is suspected that the high amounts of tannin in mature oak leaves inhibit the growth of the winter moth (*Operophtera bromata L., Geometridae*) larvae that extensively damage these trees in certain years. The following table lists the average weight of two samples of larvae at times in the first 28 days after birth. The first sample was reared on young oak leaves, whereas the second sample was reared on mature leaves from the same tree.

a. Use Lagrange interpolation to approximate the average weight curve for each sample.

b. Find an approximate maximum average weight for each sample by determining the maximum of the interpolating polynomial.

Day	0	6	10	13	17	20	28
Sample 1 average weight (mg)	6.67	17.33	42.67	37.33	30.10	29.31	28.74
Sample 2 average weight (mg)	6.67	16.11	18.89	15.00	10.56	9.44	8.89

20. In Exercise 26 of Section 1.1 a Maclaurin series was integrated to approximate erf(1), where erf(x) is the normal distribution error function defined by

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt.$$

a. Use the Maclaurin series to construct a table for erf(x) that is accurate to within 10^{-4} for erf(x_i), where $x_i = 0.2i$, for $i = 0, 1, \ldots, 5$.

b. Use both linear interpolation and quadratic interpolation to obtain an approximation to erf($\frac{1}{3}$). Which approach seems most feasible?

21. Prove Taylor's Theorem 1.14 by following the procedure in the proof of Theorem 3.3. [*Hint:* Let

$$g(t) = f(t) - P(t) - [f(x) - P(x)] \cdot \frac{(t - x_0)^{n+1}}{(x - x_0)^{n+1}},$$

where P is the nth Taylor polynomial, and use the Generalized Rolle's Theorem 1.10.]

22. Show that $\max_{x_j \le x \le x_{j+1}} |g(x)| = h^2/4$, where $g(x) = (x - jh)(x - (j+1)h)$.

23. The Bernstein polynomial of degree n for $f \in C[0, 1]$ is given by

$$B_n(x) = \sum_{k=0}^{n} \binom{n}{k} f\left(\frac{k}{n}\right) x^k (1-x)^{n-k},$$

where $\binom{n}{k}$ denotes $n!/k!(n-k)!$. These polynomials can be used in a constructive proof of the Weierstrass Approximation Theorem 3.1 (see [Bart]) because $\lim_{n\to\infty} B_n(x) = f(x)$, for each $x \in [0, 1]$.

a. Find $B_3(x)$ for the functions

i. $f(x) = x$ **ii.** $f(x) = 1$

b. Show that for each $k \le n$,

$$\binom{n-1}{k-1} = \left(\frac{k}{n}\right)\binom{n}{k}.$$

c. Use part (b) and the fact, from (ii) in part (a), that

$$1 = \sum_{k=0}^{n} \binom{n}{k} x^k (1-x)^{n-k}, \quad \text{for each } n,$$

to show that, for $f(x) = x^2$,

$$B_n(x) = \left(\frac{n-1}{n}\right) x^2 + \frac{1}{n} x.$$

d. Use part (c) to estimate the value of n necessary for $\left|B_n(x) - x^2\right| \le 10^{-6}$ to hold for all x in $[0, 1]$.

3.2 Data Approximation and Neville's Method

In the previous section we found an explicit representation for Lagrange polynomials and their error when approximating a function on an interval. A frequent use of these polynomials involves the interpolation of tabulated data. In this case an explicit representation of the polynomial might not be needed, only the values of the polynomial at specified points. In this situation the function underlying the data might not be known so the explicit form of the error cannot be used. We will now illustrate a practical application of interpolation in such a situation.

Illustration Table 3.2 lists values of a function f at various points. The approximations to $f(1.5)$ obtained by various Lagrange polynomials that use this data will be compared to try and determine the accuracy of the approximation.

Table 3.2

x	$f(x)$
1.0	0.7651977
1.3	0.6200860
1.6	0.4554022
1.9	0.2818186
2.2	0.1103623

The most appropriate linear polynomial uses $x_0 = 1.3$ and $x_1 = 1.6$ because 1.5 is between 1.3 and 1.6. The value of the interpolating polynomial at 1.5 is

$$\begin{aligned} P_1(1.5) &= \frac{(1.5-1.6)}{(1.3-1.6)} f(1.3) + \frac{(1.5-1.3)}{(1.6-1.3)} f(1.6) \\ &= \frac{(1.5-1.6)}{(1.3-1.6)}(0.6200860) + \frac{(1.5-1.3)}{(1.6-1.3)}(0.4554022) = 0.5102968. \end{aligned}$$

Two polynomials of degree 2 can reasonably be used, one with $x_0 = 1.3$, $x_1 = 1.6$, and $x_2 = 1.9$, which gives

$$P_2(1.5) = \frac{(1.5-1.6)(1.5-1.9)}{(1.3-1.6)(1.3-1.9)}(0.6200860) + \frac{(1.5-1.3)(1.5-1.9)}{(1.6-1.3)(1.6-1.9)}(0.4554022)$$
$$+ \frac{(1.5-1.3)(1.5-1.6)}{(1.9-1.3)(1.9-1.6)}(0.2818186) = 0.5112857,$$

and one with $x_0 = 1.0$, $x_1 = 1.3$, and $x_2 = 1.6$, which gives $\hat{P}_2(1.5) = 0.5124715$.

In the third-degree case, there are also two reasonable choices for the polynomial. One with $x_0 = 1.3$, $x_1 = 1.6$, $x_2 = 1.9$, and $x_3 = 2.2$, which gives $P_3(1.5) = 0.5118302$.

The second third-degree approximation is obtained with $x_0 = 1.0$, $x_1 = 1.3$, $x_2 = 1.6$, and $x_3 = 1.9$, which gives $\hat{P}_3(1.5) = 0.5118127$. The fourth-degree Lagrange polynomial uses all the entries in the table. With $x_0 = 1.0$, $x_1 = 1.3$, $x_2 = 1.6$, $x_3 = 1.9$, and $x_4 = 2.2$, the approximation is $P_4(1.5) = 0.5118200$.

Because $P_3(1.5)$, $\hat{P}_3(1.5)$, and $P_4(1.5)$ all agree to within 2×10^{-5} units, we expect this degree of accuracy for these approximations. We also expect $P_4(1.5)$ to be the most accurate approximation, since it uses more of the given data.

The function we are approximating is actually the Bessel function of the first kind of order zero, whose value at 1.5 is known to be 0.5118277. Therefore, the true accuracies of the approximations are as follows:

$$|P_1(1.5) - f(1.5)| \approx 1.53 \times 10^{-3},$$
$$|P_2(1.5) - f(1.5)| \approx 5.42 \times 10^{-4},$$
$$|\hat{P}_2(1.5) - f(1.5)| \approx 6.44 \times 10^{-4},$$
$$|P_3(1.5) - f(1.5)| \approx 2.5 \times 10^{-6},$$
$$|\hat{P}_3(1.5) - f(1.5)| \approx 1.50 \times 10^{-5},$$
$$|P_4(1.5) - f(1.5)| \approx 7.7 \times 10^{-6}.$$

Although $P_3(1.5)$ is the most accurate approximation, if we had no knowledge of the actual value of $f(1.5)$, we would accept $P_4(1.5)$ as the best approximation since it includes the most data about the function. The Lagrange error term derived in Theorem 3.3 cannot be applied here because we have no knowledge of the fourth derivative of f. Unfortunately, this is generally the case. □

Neville's Method

A practical difficulty with Lagrange interpolation is that the error term is difficult to apply, so the degree of the polynomial needed for the desired accuracy is generally not known until computations have been performed. A common practice is to compute the results given from various polynomials until appropriate agreement is obtained, as was done in the previous Illustration. However, the work done in calculating the approximation by the second polynomial does not lessen the work needed to calculate the third approximation; nor is the fourth approximation easier to obtain once the third approximation is known, and so on. We will now derive these approximating polynomials in a manner that uses the previous calculations to greater advantage.

Definition 3.4 Let f be a function defined at $x_0, x_1, x_2, \ldots, x_n$, and suppose that $m_1, m_2, \ldots, m_k$ are k distinct integers, with $0 \le m_i \le n$ for each i. The Lagrange polynomial that agrees with $f(x)$ at the k points $x_{m_1}, x_{m_2}, \ldots, x_{m_k}$ is denoted $P_{m_1,m_2,\ldots,m_k}(x)$. ■

Example 1 Suppose that $x_0 = 1$, $x_1 = 2$, $x_2 = 3$, $x_3 = 4$, $x_4 = 6$, and $f(x) = e^x$. Determine the interpolating polynomial denoted $P_{1,2,4}(x)$, and use this polynomial to approximate $f(5)$.

Solution This is the Lagrange polynomial that agrees with $f(x)$ at $x_1 = 2$, $x_2 = 3$, and $x_4 = 6$. Hence

$$P_{1,2,4}(x) = \frac{(x-3)(x-6)}{(2-3)(2-6)}e^2 + \frac{(x-2)(x-6)}{(3-2)(3-6)}e^3 + \frac{(x-2)(x-3)}{(6-2)(6-3)}e^6.$$

So

$$\begin{aligned} f(5) \approx P(5) &= \frac{(5-3)(5-6)}{(2-3)(2-6)}e^2 + \frac{(5-2)(5-6)}{(3-2)(3-6)}e^3 + \frac{(5-2)(5-3)}{(6-2)(6-3)}e^6 \\ &= -\frac{1}{2}e^2 + e^3 + \frac{1}{2}e^6 \approx 218.105. \end{aligned}$$

■

The next result describes a method for recursively generating Lagrange polynomial approximations.

Theorem 3.5 Let f be defined at $x_0, x_1, \ldots, x_k$, and let x_j and x_i be two distinct numbers in this set. Then

$$P(x) = \frac{(x-x_j)P_{0,1,\ldots,j-1,j+1,\ldots,k}(x) - (x-x_i)P_{0,1,\ldots,i-1,i+1,\ldots,k}(x)}{(x_i - x_j)}$$

is the kth Lagrange polynomial that interpolates f at the $k+1$ points $x_0, x_1, \ldots, x_k$. ■

Proof For ease of notation, let $Q \equiv P_{0,1,\ldots,i-1,i+1,\ldots,k}$ and $\hat{Q} \equiv P_{0,1,\ldots,j-1,j+1,\ldots,k}$. Since $Q(x)$ and $\hat{Q}(x)$ are polynomials of degree $k-1$ or less, $P(x)$ is of degree at most k.

First note that $\hat{Q}(x_i) = f(x_i)$, implies that

$$P(x_i) = \frac{(x_i - x_j)\hat{Q}(x_i) - (x_i - x_i)Q(x_i)}{x_i - x_j} = \frac{(x_i - x_j)}{(x_i - x_j)}f(x_i) = f(x_i).$$

Similarly, since $Q(x_j) = f(x_j)$, we have $P(x_j) = f(x_j)$.

In addition, if $0 \le r \le k$ and r is neither i nor j, then $Q(x_r) = \hat{Q}(x_r) = f(x_r)$. So

$$P(x_r) = \frac{(x_r - x_j)\hat{Q}(x_r) - (x_r - x_i)Q(x_r)}{x_i - x_j} = \frac{(x_i - x_j)}{(x_i - x_j)}f(x_r) = f(x_r).$$

But, by definition, $P_{0,1,\ldots,k}(x)$ is the unique polynomial of degree at most k that agrees with f at $x_0, x_1, \ldots, x_k$. Thus, $P \equiv P_{0,1,\ldots,k}$. ■ ■ ■

Theorem 3.5 implies that the interpolating polynomials can be generated recursively. For example, we have

$$P_{0,1} = \frac{1}{x_1 - x_0}[(x-x_0)P_1 - (x-x_1)P_0], \qquad P_{1,2} = \frac{1}{x_2 - x_1}[(x-x_1)P_2 - (x-x_2)P_1],$$

$$P_{0,1,2} = \frac{1}{x_2 - x_0}[(x-x_0)P_{1,2} - (x-x_2)P_{0,1}],$$

and so on. They are generated in the manner shown in Table 3.3, where each row is completed before the succeeding rows are begun.

Table 3.3

x_0	P_0				
x_1	P_1	$P_{0,1}$			
x_2	P_2	$P_{1,2}$	$P_{0,1,2}$		
x_3	P_3	$P_{2,3}$	$P_{1,2,3}$	$P_{0,1,2,3}$	
x_4	P_4	$P_{3,4}$	$P_{2,3,4}$	$P_{1,2,3,4}$	$P_{0,1,2,3,4}$

The procedure that uses the result of Theorem 3.5 to recursively generate interpolating polynomial approximations is called **Neville's method**. The P notation used in Table 3.3 is cumbersome because of the number of subscripts used to represent the entries. Note, however, that as an array is being constructed, only two subscripts are needed. Proceeding down the table corresponds to using consecutive points x_i with larger i, and proceeding to the right corresponds to increasing the degree of the interpolating polynomial. Since the points appear consecutively in each entry, we need to describe only a starting point and the number of additional points used in constructing the approximation.

Eric Harold Neville (1889–1961) gave this modification of the Lagrange formula in a paper published in 1932.[N]

To avoid the multiple subscripts, we let $Q_{i,j}(x)$, for $0 \le j \le i$, denote the interpolating polynomial of degree j on the $(j+1)$ numbers $x_{i-j}, x_{i-j+1}, \ldots, x_{i-1}, x_i$; that is,

$$Q_{i,j} = P_{i-j,i-j+1,\ldots,i-1,i}.$$

Using this notation provides the Q notation array in Table 3.4.

Table 3.4

x_0	$P_0 = Q_{0,0}$				
x_1	$P_1 = Q_{1,0}$	$P_{0,1} = Q_{1,1}$			
x_2	$P_2 = Q_{2,0}$	$P_{1,2} = Q_{2,1}$	$P_{0,1,2} = Q_{2,2}$		
x_3	$P_3 = Q_{3,0}$	$P_{2,3} = Q_{3,1}$	$P_{1,2,3} = Q_{3,2}$	$P_{0,1,2,3} = Q_{3,3}$	
x_4	$P_4 = Q_{4,0}$	$P_{3,4} = Q_{4,1}$	$P_{2,3,4} = Q_{4,2}$	$P_{1,2,3,4} = Q_{4,3}$	$P_{0,1,2,3,4} = Q_{4,4}$

Example 2 Values of various interpolating polynomials at $x = 1.5$ were obtained in the Illustration at the beginning of the Section using the data shown in Table 3.5. Apply Neville's method to the data by constructing a recursive table of the form shown in Table 3.4.

Table 3.5

x	$f(x)$
1.0	0.7651977
1.3	0.6200860
1.6	0.4554022
1.9	0.2818186
2.2	0.1103623

Solution Let $x_0 = 1.0$, $x_1 = 1.3$, $x_2 = 1.6$, $x_3 = 1.9$, and $x_4 = 2.2$, then $Q_{0,0} = f(1.0)$, $Q_{1,0} = f(1.3)$, $Q_{2,0} = f(1.6)$, $Q_{3,0} = f(1.9)$, and $Q_{4,0} = f(2.2)$. These are the five polynomials of degree zero (constants) that approximate $f(1.5)$, and are the same as data given in Table 3.5.

Calculating the first-degree approximation $Q_{1,1}(1.5)$ gives

$$\begin{aligned} Q_{1,1}(1.5) &= \frac{(x - x_0)Q_{1,0} - (x - x_1)Q_{0,0}}{x_1 - x_0} \\ &= \frac{(1.5 - 1.0)Q_{1,0} - (1.5 - 1.3)Q_{0,0}}{1.3 - 1.0} \\ &= \frac{0.5(0.6200860) - 0.2(0.7651977)}{0.3} = 0.5233449. \end{aligned}$$

Similarly,

$$Q_{2,1}(1.5) = \frac{(1.5 - 1.3)(0.4554022) - (1.5 - 1.6)(0.6200860)}{1.6 - 1.3} = 0.5102968,$$

$$Q_{3,1}(1.5) = 0.5132634, \quad \text{and} \quad Q_{4,1}(1.5) = 0.5104270.$$

The best linear approximation is expected to be $Q_{2,1}$ because 1.5 is between $x_1 = 1.3$ and $x_2 = 1.6$.

In a similar manner, approximations using higher-degree polynomials are given by

$$Q_{2,2}(1.5) = \frac{(1.5 - 1.0)(0.5102968) - (1.5 - 1.6)(0.5233449)}{1.6 - 1.0} = 0.5124715,$$

$$Q_{3,2}(1.5) = 0.5112857, \quad \text{and} \quad Q_{4,2}(1.5) = 0.5137361.$$

The higher-degree approximations are generated in a similar manner and are shown in Table 3.6. ■

Table 3.6

1.0	0.7651977				
1.3	0.6200860	0.5233449			
1.6	0.4554022	0.5102968	0.5124715		
1.9	0.2818186	0.5132634	0.5112857	0.5118127	
2.2	0.1103623	0.5104270	0.5137361	0.5118302	0.5118200

If the latest approximation, $Q_{4,4}$, was not sufficiently accurate, another node, x_5, could be selected, and another row added to the table:

$$x_5 \quad Q_{5,0} \quad Q_{5,1} \quad Q_{5,2} \quad Q_{5,3} \quad Q_{5,4} \quad Q_{5,5}.$$

Then $Q_{4,4}$, $Q_{5,4}$, and $Q_{5,5}$ could be compared to determine further accuracy.

The function in Example 2 is the Bessel function of the first kind of order zero, whose value at 2.5 is -0.0483838, and the next row of approximations to $f(1.5)$ is

$$2.5 \quad -0.0483838 \quad 0.4807699 \quad 0.5301984 \quad 0.5119070 \quad 0.5118430 \quad 0.5118277.$$

The final new entry, 0.5118277, is correct to all seven decimal places.

The *NumericalAnalysis* package in Maple can be used to apply Neville's method for the values of x and $f(x) = y$ in Table 3.6. After loading the package we define the data with

$xy := [[1.0, 0.7651977], [1.3, 0.6200860], [1.6, 0.4554022], [1.9, 0.2818186]]$

Neville's method using this data gives the approximation at $x = 1.5$ with the command

$p3 := PolynomialInterpolation(xy, method = neville, extrapolate = [1.5])$

The output from Maple for this command is

$$POLYINTERP([[1.0, 0.7651977], [1.3, 0.6200860], [1.6, 0.4554022], [1.9, 0.2818186]], method = neville, extrapolate = [1.5], INFO)$$

which isn't very informative. To display the information, we enter the command

$NevilleTable(p3, 1.5)$

and Maple returns an array with four rows and four columns. The nonzero entries corresponding to the top four rows of Table 3.6 (with the first column deleted), the zero entries are simply used to fill up the array.

To add the additional row to the table using the additional data (2.2, 0.1103623) we use the command

$p3a := AddPoint(p3, [2.2, 0.1103623])$

and a new array with all the approximation entries in Table 3.6 is obtained with

$NevilleTable(p3a, 1.5)$

Example 3 Table 3.7 lists the values of $f(x) = \ln x$ accurate to the places given. Use Neville's method and four-digit rounding arithmetic to approximate $f(2.1) = \ln 2.1$ by completing the Neville table.

Table 3.7

i	x_i	$\ln x_i$
0	2.0	0.6931
1	2.2	0.7885
2	2.3	0.8329

Solution Because $x - x_0 = 0.1$, $x - x_1 = -0.1$, $x - x_2 = -0.2$, and we are given $Q_{0,0} = 0.6931$, $Q_{1,0} = 0.7885$, and $Q_{2,0} = 0.8329$, we have

$$Q_{1,1} = \frac{1}{0.2}[(0.1)0.7885 - (-0.1)0.6931] = \frac{0.1482}{0.2} = 0.7410$$

and

$$Q_{2,1} = \frac{1}{0.1}[(-0.1)0.8329 - (-0.2)0.7885] = \frac{0.07441}{0.1} = 0.7441.$$

The final approximation we can obtain from this data is

$$Q_{2,1} = \frac{1}{0.3}[(0.1)0.7441 - (-0.2)0.7410] = \frac{0.2276}{0.3} = 0.7420.$$

These values are shown in Table 3.8. ■

Table 3.8

i	x_i	$x - x_i$	Q_{i0}	Q_{i1}	Q_{i2}
0	2.0	0.1	0.6931		
1	2.2	−0.1	0.7885	0.7410	
2	2.3	−0.2	0.8329	0.7441	0.7420

In the preceding example we have $f(2.1) = \ln 2.1 = 0.7419$ to four decimal places, so the absolute error is

$$|f(2.1) - P_2(2.1)| = |0.7419 - 0.7420| = 10^{-4}.$$

However, $f'(x) = 1/x$, $f''(x) = -1/x^2$, and $f'''(x) = 2/x^3$, so the Lagrange error formula (3.3) in Theorem 3.3 gives the error bound

$$|f(2.1) - P_2(2.1)| = \left|\frac{f'''(\xi(2.1))}{3!}(x - x_0)(x - x_1)(x - x_2)\right|$$
$$= \left|\frac{1}{3\,(\xi(2.1))^3}(0.1)(-0.1)(-0.2)\right| \le \frac{0.002}{3(2)^3} = 8.\overline{3} \times 10^{-5}.$$

Notice that the actual error, 10^{-4}, exceeds the error bound, $8.\overline{3} \times 10^{-5}$. This apparent contradiction is a consequence of finite-digit computations. We used four-digit rounding arithmetic, and the Lagrange error formula (3.3) assumes infinite-digit arithmetic. This caused our actual errors to exceed the theoretical error estimate.

- Remember: You cannot expect more accuracy than the arithmetic provides.

Algorithm 3.1 constructs the entries in Neville's method by rows.

ALGORITHM
3.1

Neville's Iterated Interpolation

To evaluate the interpolating polynomial P on the $n+1$ distinct numbers $x_0, \ldots, x_n$ at the number x for the function f:

INPUT numbers $x, x_0, x_1, \ldots, x_n$; values $f(x_0), f(x_1), \ldots, f(x_n)$ as the first column $Q_{0,0}, Q_{1,0}, \ldots, Q_{n,0}$ of Q.

OUTPUT the table Q with $P(x) = Q_{n,n}$.

Step 1 For $i = 1, 2, \ldots, n$
for $j = 1, 2, \ldots, i$

$$\text{set } Q_{i,j} = \frac{(x - x_{i-j})Q_{i,j-1} - (x - x_i)Q_{i-1,j-1}}{x_i - x_{i-j}}.$$

Step 2 OUTPUT (Q);
STOP. ■

The algorithm can be modified to allow for the addition of new interpolating nodes. For example, the inequality

$$|Q_{i,i} - Q_{i-1,i-1}| < \varepsilon$$

can be used as a stopping criterion, where ε is a prescribed error tolerance. If the inequality is true, $Q_{i,i}$ is a reasonable approximation to $f(x)$. If the inequality is false, a new interpolation point, x_{i+1}, is added.

EXERCISE SET 3.2

1. Use Neville's method to obtain the approximations for Lagrange interpolating polynomials of degrees one, two, and three to approximate each of the following:

a. $f(8.4)$ if $f(8.1) = 16.94410$, $f(8.3) = 17.56492$, $f(8.6) = 18.50515$, $f(8.7) = 18.82091$

b. $f\left(-\frac{1}{3}\right)$ if $f(-0.75) = -0.07181250$, $f(-0.5) = -0.02475000$, $f(-0.25) = 0.33493750$, $f(0) = 1.10100000$

c. $f(0.25)$ if $f(0.1) = 0.62049958$, $f(0.2) = -0.28398668$, $f(0.3) = 0.00660095$, $f(0.4) = 0.24842440$

d. $f(0.9)$ if $f(0.6) = -0.17694460$, $f(0.7) = 0.01375227$, $f(0.8) = 0.22363362$, $f(1.0) = 0.65809197$

2. Use Neville's method to obtain the approximations for Lagrange interpolating polynomials of degrees one, two, and three to approximate each of the following:

a. $f(0.43)$ if $f(0) = 1$, $f(0.25) = 1.64872$, $f(0.5) = 2.71828$, $f(0.75) = 4.48169$

b. $f(0)$ if $f(-0.5) = 1.93750$, $f(-0.25) = 1.33203$, $f(0.25) = 0.800781$, $f(0.5) = 0.687500$

c. $f(0.18)$ if $f(0.1) = -0.29004986$, $f(0.2) = -0.56079734$, $f(0.3) = -0.81401972$, $f(0.4) = -1.0526302$

d. $f(0.25)$ if $f(-1) = 0.86199480$, $f(-0.5) = 0.95802009$, $f(0) = 1.0986123$, $f(0.5) = 1.2943767$

3. Use Neville's method to approximate $\sqrt{3}$ with the following functions and values.

a. $f(x) = 3^x$ and the values $x_0 = -2$, $x_1 = -1$, $x_2 = 0$, $x_3 = 1$, and $x_4 = 2$.

b. $f(x) = \sqrt{x}$ and the values $x_0 = 0$, $x_1 = 1$, $x_2 = 2$, $x_3 = 4$, and $x_4 = 5$.

c. Compare the accuracy of the approximation in parts (a) and (b).

4. Let $P_3(x)$ be the interpolating polynomial for the data $(0, 0)$, $(0.5, y)$, $(1, 3)$, and $(2, 2)$. Use Neville's method to find y if $P_3(1.5) = 0$.

5. Neville's method is used to approximate $f(0.4)$, giving the following table.

$x_0 = 0$	$P_0 = 1$			
$x_1 = 0.25$	$P_1 = 2$	$P_{01} = 2.6$		
$x_2 = 0.5$	P_2	$P_{1,2}$	$P_{0,1,2}$	
$x_3 = 0.75$	$P_3 = 8$	$P_{2,3} = 2.4$	$P_{1,2,3} = 2.96$	$P_{0,1,2,3} = 3.016$

Determine $P_2 = f(0.5)$.

6. Neville's method is used to approximate $f(0.5)$, giving the following table.

$x_0 = 0$	$P_0 = 0$		
$x_1 = 0.4$	$P_1 = 2.8$	$P_{0,1} = 3.5$	
$x_2 = 0.7$	P_2	$P_{1,2}$	$P_{0,1,2} = \frac{27}{7}$

Determine $P_2 = f(0.7)$.

7. Suppose $x_j = j$, for $j = 0, 1, 2, 3$ and it is known that

$$P_{0,1}(x) = 2x + 1, \quad P_{0,2}(x) = x + 1, \quad \text{and} \quad P_{1,2,3}(2.5) = 3.$$

Find $P_{0,1,2,3}(2.5)$.

8. Suppose $x_j = j$, for $j = 0, 1, 2, 3$ and it is known that

$$P_{0,1}(x) = x + 1, \quad P_{1,2}(x) = 3x - 1, \quad \text{and} \quad P_{1,2,3}(1.5) = 4.$$

Find $P_{0,1,2,3}(1.5)$.

9. Neville's Algorithm is used to approximate $f(0)$ using $f(-2)$, $f(-1)$, $f(1)$, and $f(2)$. Suppose $f(-1)$ was understated by 2 and $f(1)$ was overstated by 3. Determine the error in the original calculation of the value of the interpolating polynomial to approximate $f(0)$.

10. Neville's Algorithm is used to approximate $f(0)$ using $f(-2)$, $f(-1)$, $f(1)$, and $f(2)$. Suppose $f(-1)$ was overstated by 2 and $f(1)$ was understated by 3. Determine the error in the original calculation of the value of the interpolating polynomial to approximate $f(0)$.

11. Construct a sequence of interpolating values y_n to $f(1 + \sqrt{10})$, where $f(x) = (1 + x^2)^{-1}$ for $-5 \le x \le 5$, as follows: For each $n = 1, 2, \ldots, 10$, let $h = 10/n$ and $y_n = P_n(1 + \sqrt{10})$, where $P_n(x)$ is the interpolating polynomial for $f(x)$ at the nodes $x_0^{(n)}, x_1^{(n)}, \ldots, x_n^{(n)}$ and $x_j^{(n)} = -5 + jh$, for each $j = 0, 1, 2, \ldots, n$. Does the sequence $\{y_n\}$ appear to converge to $f(1 + \sqrt{10})$?

Inverse Interpolation Suppose $f \in C^1[a, b]$, $f'(x) \neq 0$ on $[a, b]$ and f has one zero p in $[a, b]$. Let $x_0, \ldots, x_n$, be $n + 1$ distinct numbers in $[a, b]$ with $f(x_k) = y_k$, for each $k = 0, 1, \ldots, n$. To approximate p construct the interpolating polynomial of degree n on the nodes $y_0, \ldots, y_n$ for f^{-1}. Since $y_k = f(x_k)$ and $0 = f(p)$, it follows that $f^{-1}(y_k) = x_k$ and $p = f^{-1}(0)$. Using iterated interpolation to approximate $f^{-1}(0)$ is called *iterated inverse interpolation.*

12. Use iterated inverse interpolation to find an approximation to the solution of $x - e^{-x} = 0$, using the data

x	0.3	0.4	0.5	0.6
e^{-x}	0.740818	0.670320	0.606531	0.548812

13. Construct an algorithm that can be used for inverse interpolation.

3.3 Divided Differences

Iterated interpolation was used in the previous section to generate successively higher-degree polynomial approximations at a specific point. Divided-difference methods introduced in this section are used to successively generate the polynomials themselves.

Suppose that $P_n(x)$ is the nth Lagrange polynomial that agrees with the function f at the distinct numbers $x_0, x_1, \ldots, x_n$. Although this polynomial is unique, there are alternate algebraic representations that are useful in certain situations. The divided differences of f with respect to $x_0, x_1, \ldots, x_n$ are used to express $P_n(x)$ in the form

$$P_n(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + \cdots + a_n(x - x_0) \cdots (x - x_{n-1}), \quad (3.5)$$

for appropriate constants $a_0, a_1, \ldots, a_n$. To determine the first of these constants, a_0, note that if $P_n(x)$ is written in the form of Eq. (3.5), then evaluating $P_n(x)$ at x_0 leaves only the constant term a_0; that is,

$$a_0 = P_n(x_0) = f(x_0).$$

As in so many areas, Isaac Newton is prominent in the study of difference equations. He developed interpolation formulas as early as 1675, using his Δ notation in tables of differences. He took a very general approach to the difference formulas, so explicit examples that he produced, including Lagrange's formulas, are often known by other names.

Similarly, when $P(x)$ is evaluated at x_1, the only nonzero terms in the evaluation of $P_n(x_1)$ are the constant and linear terms,

$$f(x_0) + a_1(x_1 - x_0) = P_n(x_1) = f(x_1);$$

so

$$a_1 = \frac{f(x_1) - f(x_0)}{x_1 - x_0}. \quad (3.6)$$

We now introduce the divided-difference notation, which is related to Aitken's Δ^2 notation used in Section 2.5. The *zeroth divided difference* of the function f with respect to x_i, denoted $f[x_i]$, is simply the value of f at x_i:

$$f[x_i] = f(x_i). \quad (3.7)$$

The remaining divided differences are defined recursively; the *first divided difference* of f with respect to x_i and x_{i+1} is denoted $f[x_i, x_{i+1}]$ and defined as

$$f[x_i, x_{i+1}] = \frac{f[x_{i+1}] - f[x_i]}{x_{i+1} - x_i}. \quad (3.8)$$

The *second divided difference*, $f[x_i, x_{i+1}, x_{i+2}]$, is defined as

$$f[x_i, x_{i+1}, x_{i+2}] = \frac{f[x_{i+1}, x_{i+2}] - f[x_i, x_{i+1}]}{x_{i+2} - x_i}.$$

Similarly, after the $(k-1)$st divided differences,

$$f[x_i, x_{i+1}, x_{i+2}, \ldots, x_{i+k-1}] \quad \text{and} \quad f[x_{i+1}, x_{i+2}, \ldots, x_{i+k-1}, x_{i+k}],$$

have been determined, the ***k*th divided difference** relative to $x_i, x_{i+1}, x_{i+2}, \ldots, x_{i+k}$ is

$$f[x_i, x_{i+1}, \ldots, x_{i+k-1}, x_{i+k}] = \frac{f[x_{i+1}, x_{i+2}, \ldots, x_{i+k}] - f[x_i, x_{i+1}, \ldots, x_{i+k-1}]}{x_{i+k} - x_i}. \quad (3.9)$$

The process ends with the single *nth divided difference*,

$$f[x_0, x_1, \ldots, x_n] = \frac{f[x_1, x_2, \ldots, x_n] - f[x_0, x_1, \ldots, x_{n-1}]}{x_n - x_0}.$$

Because of Eq. (3.6) we can write $a_1 = f[x_0, x_1]$, just as a_0 can be expressed as $a_0 = f(x_0) = f[x_0]$. Hence the interpolating polynomial in Eq. (3.5) is

$$P_n(x) = f[x_0] + f[x_0, x_1](x - x_0) + a_2(x - x_0)(x - x_1) + \cdots + a_n(x - x_0)(x - x_1) \cdots (x - x_{n-1}).$$

As might be expected from the evaluation of a_0 and a_1, the required constants are

$$a_k = f[x_0, x_1, x_2, \ldots, x_k],$$

for each $k = 0, 1, \ldots, n$. So $P_n(x)$ can be rewritten in a form called Newton's Divided-Difference:

$$P_n(x) = f[x_0] + \sum_{k=1}^{n} f[x_0, x_1, \ldots, x_k](x - x_0) \cdots (x - x_{k-1}). \tag{3.10}$$

The value of $f[x_0, x_1, \ldots, x_k]$ is independent of the order of the numbers $x_0, x_1, \ldots, x_k$, as shown in Exercise 21.

The generation of the divided differences is outlined in Table 3.9. Two fourth and one fifth difference can also be determined from these data.

Table 3.9

x	$f(x)$	First divided differences	Second divided differences	Third divided differences
x_0	$f[x_0]$			
		$f[x_0, x_1] = \dfrac{f[x_1] - f[x_0]}{x_1 - x_0}$		
x_1	$f[x_1]$		$f[x_0, x_1, x_2] = \dfrac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0}$	
		$f[x_1, x_2] = \dfrac{f[x_2] - f[x_1]}{x_2 - x_1}$		$f[x_0, x_1, x_2, x_3] = \dfrac{f[x_1, x_2, x_3] - f[x_0, x_1, x_2]}{x_3 - x_0}$
x_2	$f[x_2]$		$f[x_1, x_2, x_3] = \dfrac{f[x_2, x_3] - f[x_1, x_2]}{x_3 - x_1}$	
		$f[x_2, x_3] = \dfrac{f[x_3] - f[x_2]}{x_3 - x_2}$		$f[x_1, x_2, x_3, x_4] = \dfrac{f[x_2, x_3, x_4] - f[x_1, x_2, x_3]}{x_4 - x_1}$
x_3	$f[x_3]$		$f[x_2, x_3, x_4] = \dfrac{f[x_3, x_4] - f[x_2, x_3]}{x_4 - x_2}$	
		$f[x_3, x_4] = \dfrac{f[x_4] - f[x_3]}{x_4 - x_3}$		$f[x_2, x_3, x_4, x_5] = \dfrac{f[x_3, x_4, x_5] - f[x_2, x_3, x_4]}{x_5 - x_2}$
x_4	$f[x_4]$		$f[x_3, x_4, x_5] = \dfrac{f[x_4, x_5] - f[x_3, x_4]}{x_5 - x_3}$	
		$f[x_4, x_5] = \dfrac{f[x_5] - f[x_4]}{x_5 - x_4}$		
x_5	$f[x_5]$			

Newton's Divided-Difference Formula

To obtain the divided-difference coefficients of the interpolatory polynomial P on the $(n+1)$ distinct numbers $x_0, x_1, \ldots, x_n$ for the function f:

INPUT numbers $x_0, x_1, \ldots, x_n$; values $f(x_0), f(x_1), \ldots, f(x_n)$ as $F_{0,0}, F_{1,0}, \ldots, F_{n,0}$.

OUTPUT the numbers $F_{0,0}, F_{1,1}, \ldots, F_{n,n}$ where

$$P_n(x) = F_{0,0} + \sum_{i=1}^{n} F_{i,i} \prod_{j=0}^{i-1} (x - x_j). \quad (F_{i,i} \text{ is } f[x_0, x_1, \ldots, x_i].)$$

Step 1 For $i = 1, 2, \ldots, n$
For $j = 1, 2, \ldots, i$
set $F_{i,j} = \dfrac{F_{i,j-1} - F_{i-1,j-1}}{x_i - x_{i-j}}$. $(F_{i,j} = f[x_{i-j}, \ldots, x_i].)$

Step 2 OUTPUT $(F_{0,0}, F_{1,1}, \ldots, F_{n,n})$;
STOP. ■

The form of the output in Algorithm 3.2 can be modified to produce all the divided differences, as shown in Example 1.

Example 1 Complete the divided difference table for the data used in Example 1 of Section 3.2, and reproduced in Table 3.10, and construct the interpolating polynomial that uses all this data.

Table 3.10

x	$f(x)$
1.0	0.7651977
1.3	0.6200860
1.6	0.4554022
1.9	0.2818186
2.2	0.1103623

Solution The first divided difference involving x_0 and x_1 is

$$f[x_0, x_1] = \frac{f[x_1] - f[x_0]}{x_1 - x_0} = \frac{0.6200860 - 0.7651977}{1.3 - 1.0} = -0.4837057.$$

The remaining first divided differences are found in a similar manner and are shown in the fourth column in Table 3.11.

Table 3.11

i	x_i	$f[x_i]$	$f[x_{i-1}, x_i]$	$f[x_{i-2}, x_{i-1}, x_i]$	$f[x_{i-3}, \ldots, x_i]$	$f[x_{i-4}, \ldots, x_i]$
0	1.0	0.7651977				
			−0.4837057			
1	1.3	0.6200860		−0.1087339		
			−0.5489460		0.0658784	
2	1.6	0.4554022		−0.0494433		0.0018251
			−0.5786120		0.0680685	
3	1.9	0.2818186		0.0118183		
			−0.5715210			
4	2.2	0.1103623				

The second divided difference involving x_0, x_1, and x_2 is

$$f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0} = \frac{-0.5489460 - (-0.4837057)}{1.6 - 1.0} = -0.1087339.$$

The remaining second divided differences are shown in the 5th column of Table 3.11. The third divided difference involving x_0, x_1, x_2, and x_3 and the fourth divided difference involving all the data points are, respectively,

$$f[x_0, x_1, x_2, x_3] = \frac{f[x_1, x_2, x_3] - f[x_0, x_1, x_2]}{x_3 - x_0} = \frac{-0.0494433 - (-0.1087339)}{1.9 - 1.0}$$
$$= 0.0658784,$$

and

$$f[x_0, x_1, x_2, x_3, x_4] = \frac{f[x_1, x_2, x_3, x_4] - f[x_0, x_1, x_2, x_3]}{x_4 - x_0} = \frac{0.0680685 - 0.0658784}{2.2 - 1.0}$$
$$= 0.0018251.$$

All the entries are given in Table 3.11.

The coefficients of the Newton forward divided-difference form of the interpolating polynomial are along the diagonal in the table. This polynomial is

$$P_4(x) = 0.7651977 - 0.4837057(x - 1.0) - 0.1087339(x - 1.0)(x - 1.3)$$
$$+ 0.0658784(x - 1.0)(x - 1.3)(x - 1.6)$$
$$+ 0.0018251(x - 1.0)(x - 1.3)(x - 1.6)(x - 1.9).$$

Notice that the value $P_4(1.5) = 0.5118200$ agrees with the result in Table 3.6 for Example 2 of Section 3.2, as it must because the polynomials are the same. ■

We can use Maple with the *NumericalAnalysis* package to create the Newton Divided-Difference table. First load the package and define the x and $f(x) = y$ values that will be used to generate the first four rows of Table 3.11.

$xy := [[1.0, 0.7651977], [1.3, 0.6200860], [1.6, 0.4554022], [1.9, 0.2818186]]$

The command to create the divided-difference table is

$p3 := PolynomialInterpolation(xy, independentvar = 'x', method = newton)$

A matrix containing the divided-difference table as its nonzero entries is created with the

$DividedDifferenceTable(p3)$

We can add another row to the table with the command

$p4 := AddPoint(p3, [2.2, 0.1103623])$

which produces the divided-difference table with entries corresponding to those in Table 3.11.

The Newton form of the interpolation polynomial is created with

$Interpolant(p4)$

which produces the polynomial in the form of $P_4(x)$ in Example 1, except that in place of the first two terms of $P_4(x)$:

$$0.7651977 - 0.4837057(x - 1.0)$$

Maple gives this as $1.248903367 - 0.4837056667x$.

The Mean Value Theorem 1.8 applied to Eq. (3.8) when $i = 0$,

$$f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0},$$

implies that when f' exists, $f[x_0, x_1] = f'(\xi)$ for some number ξ between x_0 and x_1. The following theorem generalizes this result.

Theorem 3.6 Suppose that $f \in C^n[a, b]$ and $x_0, x_1, \ldots, x_n$ are distinct numbers in $[a, b]$. Then a number ξ exists in (a, b) with

$$f[x_0, x_1, \ldots, x_n] = \frac{f^{(n)}(\xi)}{n!}.$$ ■

Proof Let

$$g(x) = f(x) - P_n(x).$$

Since $f(x_i) = P_n(x_i)$ for each $i = 0, 1, \ldots, n$, the function g has $n+1$ distinct zeros in $[a, b]$. Generalized Rolle's Theorem 1.10 implies that a number ξ in (a, b) exists with $g^{(n)}(\xi) = 0$, so

$$0 = f^{(n)}(\xi) - P_n^{(n)}(\xi).$$

Since $P_n(x)$ is a polynomial of degree n whose leading coefficient is $f[x_0, x_1, \ldots, x_n]$,

$$P_n^{(n)}(x) = n! f[x_0, x_1, \ldots, x_n],$$

for all values of x. As a consequence,

$$f[x_0, x_1, \ldots, x_n] = \frac{f^{(n)}(\xi)}{n!}.$$ ■ ■ ■

Newton's divided-difference formula can be expressed in a simplified form when the nodes are arranged consecutively with equal spacing. In this case, we introduce the notation $h = x_{i+1} - x_i$, for each $i = 0, 1, \ldots, n-1$ and let $x = x_0 + sh$. Then the difference $x - x_i$ is $x - x_i = (s - i)h$. So Eq. (3.10) becomes

$$\begin{aligned} P_n(x) = P_n(x_0 + sh) &= f[x_0] + shf[x_0, x_1] + s(s-1)h^2 f[x_0, x_1, x_2] \\ &\quad + \cdots + s(s-1)\cdots(s-n+1)h^n f[x_0, x_1, \ldots, x_n] \\ &= f[x_0] + \sum_{k=1}^{n} s(s-1)\cdots(s-k+1)h^k f[x_0, x_1, \ldots, x_k]. \end{aligned}$$

Using binomial-coefficient notation,

$$\binom{s}{k} = \frac{s(s-1)\cdots(s-k+1)}{k!},$$

we can express $P_n(x)$ compactly as

$$P_n(x) = P_n(x_0 + sh) = f[x_0] + \sum_{k=1}^{n} \binom{s}{k} k! h^k f[x_0, x_i, \ldots, x_k]. \tag{3.11}$$

Forward Differences

The **Newton forward-difference formula**, is constructed by making use of the forward difference notation Δ introduced in Aitken's Δ^2 method. With this notation,

$$f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{1}{h}(f(x_1) - f(x_0)) = \frac{1}{h}\Delta f(x_0)$$

$$f[x_0, x_1, x_2] = \frac{1}{2h}\left[\frac{\Delta f(x_1) - \Delta f(x_0)}{h}\right] = \frac{1}{2h^2}\Delta^2 f(x_0),$$

and, in general,

$$f[x_0, x_1, \ldots, x_k] = \frac{1}{k!h^k}\Delta^k f(x_0).$$

Since $f[x_0] = f(x_0)$, Eq. (3.11) has the following form.

Newton Forward-Difference Formula

$$P_n(x) = f(x_0) + \sum_{k=1}^{n} \binom{s}{k} \Delta^k f(x_0) \tag{3.12}$$

Backward Differences

If the interpolating nodes are reordered from last to first as $x_n, x_{n-1}, \ldots, x_0$, we can write the interpolatory formula as

$$\begin{aligned} P_n(x) = f[x_n] &+ f[x_n, x_{n-1}](x - x_n) + f[x_n, x_{n-1}, x_{n-2}](x - x_n)(x - x_{n-1}) \\ &+ \cdots + f[x_n, \ldots, x_0](x - x_n)(x - x_{n-1})\cdots(x - x_1). \end{aligned}$$

If, in addition, the nodes are equally spaced with $x = x_n + sh$ and $x = x_i + (s+n-i)h$, then

$$\begin{aligned} P_n(x) &= P_n(x_n + sh) \\ &= f[x_n] + shf[x_n, x_{n-1}] + s(s+1)h^2 f[x_n, x_{n-1}, x_{n-2}] + \cdots \\ &\quad + s(s+1)\cdots(s+n-1)h^n f[x_n, \ldots, x_0]. \end{aligned}$$

This is used to derive a commonly applied formula known as the **Newton backward-difference formula**. To discuss this formula, we need the following definition.

Definition 3.7 Given the sequence $\{p_n\}_{n=0}^{\infty}$, define the backward difference ∇p_n (read *nabla* p_n) by

$$\nabla p_n = p_n - p_{n-1}, \quad \text{for } n \geq 1.$$

Higher powers are defined recursively by

$$\nabla^k p_n = \nabla(\nabla^{k-1} p_n), \quad \text{for } k \geq 2. \qquad \blacksquare$$

Definition 3.7 implies that

$$f[x_n, x_{n-1}] = \frac{1}{h}\nabla f(x_n), \quad f[x_n, x_{n-1}, x_{n-2}] = \frac{1}{2h^2}\nabla^2 f(x_n),$$

and, in general,

$$f[x_n, x_{n-1}, \ldots, x_{n-k}] = \frac{1}{k!h^k}\nabla^k f(x_n).$$

Consequently,

$$P_n(x) = f[x_n] + s\nabla f(x_n) + \frac{s(s+1)}{2}\nabla^2 f(x_n) + \cdots + \frac{s(s+1)\cdots(s+n-1)}{n!}\nabla^n f(x_n).$$

If we extend the binomial coefficient notation to include all real values of s by letting

$$\binom{-s}{k} = \frac{-s(-s-1)\cdots(-s-k+1)}{k!} = (-1)^k \frac{s(s+1)\cdots(s+k-1)}{k!},$$

then

$$P_n(x) = f[x_n] + (-1)^1\binom{-s}{1}\nabla f(x_n) + (-1)^2\binom{-s}{2}\nabla^2 f(x_n) + \cdots + (-1)^n\binom{-s}{n}\nabla^n f(x_n).$$

This gives the following result.

Newton Backward–Difference Formula

$$P_n(x) = f[x_n] + \sum_{k=1}^{n}(-1)^k\binom{-s}{k}\nabla^k f(x_n) \tag{3.13}$$

Illustration The divided-difference Table 3.12 corresponds to the data in Example 1.

Table 3.12

		First divided differences	Second divided differences	Third divided differences	Fourth divided differences
1.0	0.7651977				
		−0.4837057			
1.3	0.6200860		−0.1087339		
		−0.5489460		0.0658784	
1.6	0.4554022		−0.0494433		0.0018251
		−0.5786120		0.0680685	
1.9	0.2818186		0.0118183		
		−0.5715210			
2.2	0.1103623				

Only one interpolating polynomial of degree at most 4 uses these five data points, but we will organize the data points to obtain the best interpolation approximations of degrees 1, 2, and 3. This will give us a sense of accuracy of the fourth-degree approximation for the given value of x.

If an approximation to $f(1.1)$ is required, the reasonable choice for the nodes would be $x_0 = 1.0$, $x_1 = 1.3$, $x_2 = 1.6$, $x_3 = 1.9$, and $x_4 = 2.2$ since this choice makes the earliest possible use of the data points closest to $x = 1.1$, and also makes use of the fourth divided difference. This implies that $h = 0.3$ and $s = \frac{1}{3}$, so the Newton forward divided-difference formula is used with the divided differences that have a *solid* underline (___) in Table 3.12:

$$\begin{aligned} P_4(1.1) &= P_4\left(1.0 + \frac{1}{3}(0.3)\right) \\ &= 0.7651977 + \frac{1}{3}(0.3)(-0.4837057) + \frac{1}{3}\left(-\frac{2}{3}\right)(0.3)^2(-0.1087339) \\ &\quad + \frac{1}{3}\left(-\frac{2}{3}\right)\left(-\frac{5}{3}\right)(0.3)^3(0.0658784) \\ &\quad + \frac{1}{3}\left(-\frac{2}{3}\right)\left(-\frac{5}{3}\right)\left(-\frac{8}{3}\right)(0.3)^4(0.0018251) \\ &= 0.7196460. \end{aligned}$$

To approximate a value when x is close to the end of the tabulated values, say, $x = 2.0$, we would again like to make the earliest use of the data points closest to x. This requires using the Newton backward divided-difference formula with $s = -\frac{2}{3}$ and the divided differences in Table 3.12 that have a *wavy* underline (~~~). Notice that the fourth divided difference is used in both formulas.

$$\begin{aligned} P_4(2.0) &= P_4\left(2.2 - \frac{2}{3}(0.3)\right) \\ &= 0.1103623 - \frac{2}{3}(0.3)(-0.5715210) - \frac{2}{3}\left(\frac{1}{3}\right)(0.3)^2(0.0118183) \\ &\quad - \frac{2}{3}\left(\frac{1}{3}\right)\left(\frac{4}{3}\right)(0.3)^3(0.0680685) - \frac{2}{3}\left(\frac{1}{3}\right)\left(\frac{4}{3}\right)\left(\frac{7}{3}\right)(0.3)^4(0.0018251) \\ &= 0.2238754. \end{aligned}$$

□

Centered Differences

The Newton forward- and backward-difference formulas are not appropriate for approximating $f(x)$ when x lies near the center of the table because neither will permit the highest-order difference to have x_0 close to x. A number of divided-difference formulas are available for this case, each of which has situations when it can be used to maximum advantage. These methods are known as **centered-difference formulas**. We will consider only one centered-difference formula, Stirling's method.

For the centered-difference formulas, we choose x_0 near the point being approximated and label the nodes directly below x_0 as $x_1, x_2, \ldots$ and those directly above as $x_{-1}, x_{-2}, \ldots$. With this convention, **Stirling's formula** is given by

$$\begin{aligned} P_n(x) = P_{2m+1}(x) = f[x_0] + \frac{sh}{2}(f[x_{-1}, x_0] + f[x_0, x_1]) + s^2h^2 f[x_{-1}, x_0, x_1] \\ + \frac{s(s^2-1)h^3}{2} f[x_{-2}, x_{-1}, x_0, x_1] + f[x_{-1}, x_0, x_1, x_2]) \\ + \cdots + s^2(s^2-1)(s^2-4)\cdots(s^2-(m-1)^2)h^{2m} f[x_{-m}, \ldots, x_m] \\ + \frac{s(s^2-1)\cdots(s^2-m^2)h^{2m+1}}{2}(f[x_{-m-1}, \ldots, x_m] + f[x_{-m}, \ldots, x_{m+1}]), \end{aligned} \tag{3.14}$$

James Stirling (1692–1770) published this and numerous other formulas in *Methodus Differentialis* in 1720. Techniques for accelerating the convergence of various series are included in this work.

if $n = 2m + 1$ is odd. If $n = 2m$ is even, we use the same formula but delete the last line. The entries used for this formula are underlined in Table 3.13.

Table 3.13

x	$f(x)$	First divided differences	Second divided differences	Third divided differences	Fourth divided differences
x_{-2}	$f[x_{-2}]$				
		$f[x_{-2}, x_{-1}]$			
x_{-1}	$f[x_{-1}]$		$f[x_{-2}, x_{-1}, x_0]$		
		$\underline{f[x_{-1}, x_0]}$		$\underline{f[x_{-2}, x_{-1}, x_0, x_1]}$	
x_0	$\underline{f[x_0]}$		$\underline{f[x_{-1}, x_0, x_1]}$		$\underline{f[x_{-2}, x_{-1}, x_0, x_1, x_2]}$
		$\underline{f[x_0, x_1]}$		$\underline{f[x_{-1}, x_0, x_1, x_2]}$	
x_1	$f[x_1]$		$f[x_0, x_1, x_2]$		
		$f[x_1, x_2]$			
x_2	$f[x_2]$				

Example 2 Consider the table of data given in the previous examples. Use Stirling's formula to approximate $f(1.5)$ with $x_0 = 1.6$.

Solution To apply Stirling's formula we use the *underlined* entries in the difference Table 3.14.

Table 3.14

x	$f(x)$	First divided differences	Second divided differences	Third divided differences	Fourth divided differences
1.0	0.7651977				
		−0.4837057			
1.3	0.6200860		−0.1087339		
		$\underline{-0.5489460}$		$\underline{0.0658784}$	
1.6	$\underline{0.4554022}$		$\underline{-0.0494433}$		$\underline{0.0018251}$
		$\underline{-0.5786120}$		$\underline{0.0680685}$	
1.9	0.2818186		0.0118183		
		−0.5715210			
2.2	0.1103623				

The formula, with $h = 0.3$, $x_0 = 1.6$, and $s = -\frac{1}{3}$, becomes

$$\begin{aligned} f(1.5) &\approx P_4\left(1.6 + \left(-\frac{1}{3}\right)(0.3)\right) \\ &= 0.4554022 + \left(-\frac{1}{3}\right)\left(\frac{0.3}{2}\right)((-0.5489460) + (-0.5786120)) \\ &\quad + \left(-\frac{1}{3}\right)^2 (0.3)^2(-0.0494433) \\ &\quad + \frac{1}{2}\left(-\frac{1}{3}\right)\left(\left(-\frac{1}{3}\right)^2 - 1\right)(0.3)^3(0.0658784 + 0.0680685) \\ &\quad + \left(-\frac{1}{3}\right)^2\left(\left(-\frac{1}{3}\right)^2 - 1\right)(0.3)^4(0.0018251) = 0.5118200. \end{aligned}$$

■

Most texts on numerical analysis written before the wide-spread use of computers have extensive treatments of divided-difference methods. If a more comprehensive treatment of this subject is needed, the book by Hildebrand [Hild] is a particularly good reference.

EXERCISE SET 3.3

1. Use Eq. (3.10) or Algorithm 3.2 to construct interpolating polynomials of degree one, two, and three for the following data. Approximate the specified value using each of the polynomials.

a. $f(8.4)$ if $f(8.1) = 16.94410$, $f(8.3) = 17.56492$, $f(8.6) = 18.50515$, $f(8.7) = 18.82091$

b. $f(0.9)$ if $f(0.6) = -0.17694460$, $f(0.7) = 0.01375227$, $f(0.8) = 0.22363362$, $f(1.0) = 0.65809197$

2. Use Eq. (3.10) or Algorithm 3.2 to construct interpolating polynomials of degree one, two, and three for the following data. Approximate the specified value using each of the polynomials.

a. $f(0.43)$ if $f(0) = 1$, $f(0.25) = 1.64872$, $f(0.5) = 2.71828$, $f(0.75) = 4.48169$

b. $f(0)$ if $f(-0.5) = 1.93750$, $f(-0.25) = 1.33203$, $f(0.25) = 0.800781$, $f(0.5) = 0.687500$

3. Use Newton the forward-difference formula to construct interpolating polynomials of degree one, two, and three for the following data. Approximate the specified value using each of the polynomials.

a. $f\left(-\frac{1}{3}\right)$ if $f(-0.75) = -0.07181250$, $f(-0.5) = -0.02475000$, $f(-0.25) = 0.33493750$, $f(0) = 1.10100000$

b. $f(0.25)$ if $f(0.1) = -0.62049958$, $f(0.2) = -0.28398668$, $f(0.3) = 0.00660095$, $f(0.4) = 0.24842440$

4. Use the Newton forward-difference formula to construct interpolating polynomials of degree one, two, and three for the following data. Approximate the specified value using each of the polynomials.

a. $f(0.43)$ if $f(0) = 1$, $f(0.25) = 1.64872$, $f(0.5) = 2.71828$, $f(0.75) = 4.48169$

b. $f(0.18)$ if $f(0.1) = -0.29004986$, $f(0.2) = -0.56079734$, $f(0.3) = -0.81401972$, $f(0.4) = -1.0526302$

5. Use the Newton backward-difference formula to construct interpolating polynomials of degree one, two, and three for the following data. Approximate the specified value using each of the polynomials.

a. $f(-1/3)$ if $f(-0.75) = -0.07181250$, $f(-0.5) = -0.02475000$, $f(-0.25) = 0.33493750$, $f(0) = 1.10100000$

b. $f(0.25)$ if $f(0.1) = -0.62049958$, $f(0.2) = -0.28398668$, $f(0.3) = 0.00660095$, $f(0.4) = 0.24842440$

6. Use the Newton backward-difference formula to construct interpolating polynomials of degree one, two, and three for the following data. Approximate the specified value using each of the polynomials.

a. $f(0.43)$ if $f(0) = 1$, $f(0.25) = 1.64872$, $f(0.5) = 2.71828$, $f(0.75) = 4.48169$

b. $f(0.25)$ if $f(-1) = 0.86199480$, $f(-0.5) = 0.95802009$, $f(0) = 1.0986123$, $f(0.5) = 1.2943767$

7. **a.** Use Algorithm 3.2 to construct the interpolating polynomial of degree three for the unequally spaced points given in the following table:

x	$f(x)$
−0.1	5.30000
0.0	2.00000
0.2	3.19000
0.3	1.00000

b. Add $f(0.35) = 0.97260$ to the table, and construct the interpolating polynomial of degree four.

8. **a.** Use Algorithm 3.2 to construct the interpolating polynomial of degree four for the unequally spaced points given in the following table:

x	$f(x)$
0.0	−6.00000
0.1	−5.89483
0.3	−5.65014
0.6	−5.17788
1.0	−4.28172

b. Add $f(1.1) = -3.99583$ to the table, and construct the interpolating polynomial of degree five.

9. **a.** Approximate $f(0.05)$ using the following data and the Newton forward-difference formula:

x	0.0	0.2	0.4	0.6	0.8
$f(x)$	1.00000	1.22140	1.49182	1.82212	2.22554

b. Use the Newton backward-difference formula to approximate $f(0.65)$.

c. Use Stirling's formula to approximate $f(0.43)$.

10. Show that the polynomial interpolating the following data has degree 3.

x	−2	−1	0	1	2	3
$f(x)$	1	4	11	16	13	−4

11. **a.** Show that the cubic polynomials

$$P(x) = 3 - 2(x+1) + 0(x+1)(x) + (x+1)(x)(x-1)$$

and

$$Q(x) = -1 + 4(x+2) - 3(x+2)(x+1) + (x+2)(x+1)(x)$$

both interpolate the data

x	−2	−1	0	1	2
$f(x)$	−1	3	1	−1	3

b. Why does part (a) not violate the uniqueness property of interpolating polynomials?

12. A fourth-degree polynomial $P(x)$ satisfies $\Delta^4 P(0) = 24$, $\Delta^3 P(0) = 6$, and $\Delta^2 P(0) = 0$, where $\Delta P(x) = P(x+1) - P(x)$. Compute $\Delta^2 P(10)$.

13. The following data are given for a polynomial $P(x)$ of unknown degree.

x	0	1	2
$P(x)$	2	-1	4

Determine the coefficient of x^2 in $P(x)$ if all third-order forward differences are 1.

14. The following data are given for a polynomial $P(x)$ of unknown degree.

x	0	1	2	3
$P(x)$	4	9	15	18

Determine the coefficient of x^3 in $P(x)$ if all fourth-order forward differences are 1.

15. The Newton forward-difference formula is used to approximate $f(0.3)$ given the following data.

x	0.0	0.2	0.4	0.6
$f(x)$	15.0	21.0	30.0	51.0

Suppose it is discovered that $f(0.4)$ was understated by 10 and $f(0.6)$ was overstated by 5. By what amount should the approximation to $f(0.3)$ be changed?

16. For a function f, the Newton divided-difference formula gives the interpolating polynomial

$$P_3(x) = 1 + 4x + 4x(x - 0.25) + \frac{16}{3}x(x - 0.25)(x - 0.5),$$

on the nodes $x_0 = 0$, $x_1 = 0.25$, $x_2 = 0.5$ and $x_3 = 0.75$. Find $f(0.75)$.

17. For a function f, the forward-divided differences are given by

$x_0 = 0.0$	$f[x_0]$		
		$f[x_0, x_1]$	
$x_1 = 0.4$	$f[x_1]$		$f[x_0, x_1, x_2] = \frac{50}{7}$
		$f[x_1, x_2] = 10$	
$x_2 = 0.7$	$f[x_2] = 6$		

Determine the missing entries in the table.

18. **a.** The introduction to this chapter included a table listing the population of the United States from 1950 to 2000. Use appropriate divided differences to approximate the population in the years 1940, 1975, and 2020.

b. The population in 1940 was approximately 132,165,000. How accurate do you think your 1975 and 2020 figures are?

19. Given

$$\begin{aligned} P_n(x) = {} & f[x_0] + f[x_0, x_1](x - x_0) + a_2(x - x_0)(x - x_1) \\ & + a_3(x - x_0)(x - x_1)(x - x_2) + \cdots \\ & + a_n(x - x_0)(x - x_1)\cdots(x - x_{n-1}), \end{aligned}$$

use $P_n(x_2)$ to show that $a_2 = f[x_0, x_1, x_2]$.

20. Show that

$$f[x_0, x_1, \ldots, x_n, x] = \frac{f^{(n+1)}(\xi(x))}{(n+1)!},$$

for some $\xi(x)$. [*Hint:* From Eq. (3.3),

$$f(x) = P_n(x) + \frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x - x_0)\cdots(x - x_n).$$

Considering the interpolation polynomial of degree $n+1$ on $x_0, x_1, \ldots, x_n, x$, we have

$$f(x) = P_{n+1}(x) = P_n(x) + f[x_0, x_1, \ldots, x_n, x](x - x_0) \cdots (x - x_n).]$$

21. Let $i_0, i_1, \ldots, i_n$ be a rearrangement of the integers $0, 1, \ldots, n$. Show that $f[x_{i_0}, x_{i_1}, \ldots, x_{i_n}] = f[x_0, x_1, \ldots, x_n]$. [*Hint:* Consider the leading coefficient of the nth Lagrange polynomial on the data $\{x_0, x_1, \ldots, x_n\} = \{x_{i_0}, x_{i_1}, \ldots, x_{i_n}\}$.]

3.4 Hermite Interpolation

The Latin word *osculum*, literally a "small mouth" or "kiss", when applied to a curve indicates that it just touches and has the same shape. Hermite interpolation has this osculating property. It matches a given curve, and its derivative forces the interpolating curve to "kiss" the given curve.

Osculating polynomials generalize both the Taylor polynomials and the Lagrange polynomials. Suppose that we are given $n+1$ distinct numbers $x_0, x_1, \ldots, x_n$ in $[a, b]$ and nonnegative integers $m_0, m_1, \ldots, m_n$, and $m = \max\{m_0, m_1, \ldots, m_n\}$. The osculating polynomial approximating a function $f \in C^m[a, b]$ at x_i, for each $i = 0, \ldots, n$, is the polynomial of least degree that has the same values as the function f and all its derivatives of order less than or equal to m_i at each x_i. The degree of this osculating polynomial is at most

$$M = \sum_{i=0}^{n} m_i + n$$

because the number of conditions to be satisfied is $\sum_{i=0}^{n} m_i + (n+1)$, and a polynomial of degree M has $M+1$ coefficients that can be used to satisfy these conditions.

Definition 3.8

Let $x_0, x_1, \ldots, x_n$ be $n+1$ distinct numbers in $[a, b]$ and for $i = 0, 1, \ldots, n$ let m_i be a nonnegative integer. Suppose that $f \in C^m[a, b]$, where $m = \max_{0 \le i \le n} m_i$.

The **osculating polynomial** approximating f is the polynomial $P(x)$ of least degree such that

$$\frac{d^k P(x_i)}{dx^k} = \frac{d^k f(x_i)}{dx^k}, \quad \text{for each } i = 0, 1, \ldots, n \quad \text{and} \quad k = 0, 1, \ldots, m_i. \qquad \blacksquare$$

Charles Hermite (1822–1901) made significant mathematical discoveries throughout his life in areas such as complex analysis and number theory, particularly involving the theory of equations. He is perhaps best known for proving in 1873 that e is transcendental, that is, it is not the solution to any algebraic equation having integer coefficients. This lead in 1882 to Lindemann's proof that π is also transcendental, which demonstrated that it is impossible to use the standard geometry tools of Euclid to construct a square that has the same area as a unit circle.

Note that when $n = 0$, the osculating polynomial approximating f is the m_0th Taylor polynomial for f at x_0. When $m_i = 0$ for each i, the osculating polynomial is the nth Lagrange polynomial interpolating f on $x_0, x_1, \ldots, x_n$.

Hermite Polynomials

The case when $m_i = 1$, for each $i = 0, 1, \ldots, n$, gives the **Hermite polynomials**. For a given function f, these polynomials agree with f at $x_0, x_1, \ldots, x_n$. In addition, since their first derivatives agree with those of f, they have the same "shape" as the function at $(x_i, f(x_i))$ in the sense that the *tangent lines* to the polynomial and the function agree. We will restrict our study of osculating polynomials to this situation and consider first a theorem that describes precisely the form of the Hermite polynomials.

Theorem 3.9

If $f \in C^1[a, b]$ and $x_0, \ldots, x_n \in [a, b]$ are distinct, the unique polynomial of least degree agreeing with f and f' at $x_0, \ldots, x_n$ is the Hermite polynomial of degree at most $2n+1$ given by

$$H_{2n+1}(x) = \sum_{j=0}^{n} f(x_j) H_{n,j}(x) + \sum_{j=0}^{n} f'(x_j) \hat{H}_{n,j}(x),$$

Hermite gave a description of a general osculatory polynomial in a letter to Carl W. Borchardt in 1878, to whom he regularly sent his new results. His demonstration is an interesting application of the use of complex integration techniques to solve a real-valued problem.

where, for $L_{n,j}(x)$ denoting the jth Lagrange coefficient polynomial of degree n, we have

$$H_{n,j}(x) = [1 - 2(x - x_j)L'_{n,j}(x_j)]L^2_{n,j}(x) \quad \text{and} \quad \hat{H}_{n,j}(x) = (x - x_j)L^2_{n,j}(x).$$

Moreover, if $f \in C^{2n+2}[a, b]$, then

$$f(x) = H_{2n+1}(x) + \frac{(x - x_0)^2 \dots (x - x_n)^2}{(2n + 2)!} f^{(2n+2)}(\xi(x)),$$

for some (generally unknown) $\xi(x)$ in the interval (a, b). ■

Proof First recall that

$$L_{n,j}(x_i) = \begin{cases} 0, & \text{if } i \neq j, \\ 1, & \text{if } i = j. \end{cases}$$

Hence when $i \neq j$,

$$H_{n,j}(x_i) = 0 \qquad \text{and} \qquad \hat{H}_{n,j}(x_i) = 0,$$

whereas, for each i,

$$H_{n,i}(x_i) = [1 - 2(x_i - x_i)L'_{n,i}(x_i)] \cdot 1 = 1 \qquad \text{and} \qquad \hat{H}_{n,i}(x_i) = (x_i - x_i) \cdot 1^2 = 0.$$

As a consequence

$$H_{2n+1}(x_i) = \sum_{\substack{j=0 \\ j \neq i}}^{n} f(x_j) \cdot 0 + f(x_i) \cdot 1 + \sum_{j=0}^{n} f'(x_j) \cdot 0 = f(x_i),$$

so H_{2n+1} agrees with f at $x_0, x_1, \dots, x_n$.

To show the agreement of H'_{2n+1} with f' at the nodes, first note that $L_{n,j}(x)$ is a factor of $H'_{n,j}(x)$, so $H'_{n,j}(x_i) = 0$ when $i \neq j$. In addition, when $i = j$ we have $L_{n,i}(x_i) = 1$, so

$$\begin{aligned} H'_{n,i}(x_i) &= -2L'_{n,i}(x_i) \cdot L^2_{n,i}(x_i) + [1 - 2(x_i - x_i)L'_{n,i}(x_i)]2L_{n,i}(x_i)L'_{n,i}(x_i) \\ &= -2L'_{n,i}(x_i) + 2L'_{n,i}(x_i) = 0. \end{aligned}$$

Hence, $H'_{n,j}(x_i) = 0$ for all i and j.

Finally,

$$\begin{aligned} \hat{H}'_{n,j}(x_i) &= L^2_{n,j}(x_i) + (x_i - x_j)2L_{n,j}(x_i)L'_{n,j}(x_i) \\ &= L_{n,j}(x_i)[L_{n,j}(x_i) + 2(x_i - x_j)L'_{n,j}(x_i)], \end{aligned}$$

so $\hat{H}'_{n,j}(x_i) = 0$ if $i \neq j$ and $\hat{H}'_{n,i}(x_i) = 1$. Combining these facts, we have

$$H'_{2n+1}(x_i) = \sum_{j=0}^{n} f(x_j) \cdot 0 + \sum_{\substack{j=0 \\ j \neq i}}^{n} f'(x_j) \cdot 0 + f'(x_i) \cdot 1 = f'(x_i).$$

Therefore, H_{2n+1} agrees with f and H'_{2n+1} with f' at $x_0, x_1, \dots, x_n$.

The uniqueness of this polynomial and the error formula are considered in Exercise 11. ■ ■ ■

Example 1 Use the Hermite polynomial that agrees with the data listed in Table 3.15 to find an approximation of $f(1.5)$.

Table 3.15

k	x_k	$f(x_k)$	$f'(x_k)$
0	1.3	0.6200860	−0.5220232
1	1.6	0.4554022	−0.5698959
2	1.9	0.2818186	−0.5811571

Solution We first compute the Lagrange polynomials and their derivatives. This gives

$$L_{2,0}(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} = \frac{50}{9}x^2 - \frac{175}{9}x + \frac{152}{9}, \qquad L'_{2,0}(x) = \frac{100}{9}x - \frac{175}{9};$$

$$L_{2,1}(x) = \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} = \frac{-100}{9}x^2 + \frac{320}{9}x - \frac{247}{9}, \qquad L'_{2,1}(x) = \frac{-200}{9}x + \frac{320}{9};$$

and

$$L_{2,2} = \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} = \frac{50}{9}x^2 - \frac{145}{9}x + \frac{104}{9}, \qquad L'_{2,2}(x) = \frac{100}{9}x - \frac{145}{9}.$$

The polynomials $H_{2,j}(x)$ and $\hat{H}_{2,j}(x)$ are then

$$\begin{aligned}
H_{2,0}(x) &= [1 - 2(x-1.3)(-5)]\left(\frac{50}{9}x^2 - \frac{175}{9}x + \frac{152}{9}\right)^2 \\
&= (10x - 12)\left(\frac{50}{9}x^2 - \frac{175}{9}x + \frac{152}{9}\right)^2, \\
H_{2,1}(x) &= 1 \cdot \left(\frac{-100}{9}x^2 + \frac{320}{9}x - \frac{247}{9}\right)^2, \\
H_{2,2}(x) &= 10(2-x)\left(\frac{50}{9}x^2 - \frac{145}{9}x + \frac{104}{9}\right)^2, \\
\hat{H}_{2,0}(x) &= (x-1.3)\left(\frac{50}{9}x^2 - \frac{175}{9}x + \frac{152}{9}\right)^2, \\
\hat{H}_{2,1}(x) &= (x-1.6)\left(\frac{-100}{9}x^2 + \frac{320}{9}x - \frac{247}{9}\right)^2,
\end{aligned}$$

and

$$\hat{H}_{2,2}(x) = (x-1.9)\left(\frac{50}{9}x^2 - \frac{145}{9}x + \frac{104}{9}\right)^2.$$

Finally

$$\begin{aligned}
H_5(x) = {} & 0.6200860H_{2,0}(x) + 0.4554022H_{2,1}(x) + 0.2818186H_{2,2}(x) \\
& - 0.5220232\hat{H}_{2,0}(x) - 0.5698959\hat{H}_{2,1}(x) - 0.5811571\hat{H}_{2,2}(x)
\end{aligned}$$

and

$$H_5(1.5) = 0.6200860\left(\frac{4}{27}\right) + 0.4554022\left(\frac{64}{81}\right) + 0.2818186\left(\frac{5}{81}\right) - 0.5220232\left(\frac{4}{405}\right) - 0.5698959\left(\frac{-32}{405}\right) - 0.5811571\left(\frac{-2}{405}\right) = 0.5118277,$$

a result that is accurate to the places listed. ■

Although Theorem 3.9 provides a complete description of the Hermite polynomials, it is clear from Example 1 that the need to determine and evaluate the Lagrange polynomials and their derivatives makes the procedure tedious even for small values of n.

Hermite Polynomials Using Divided Differences

There is an alternative method for generating Hermite approximations that has as its basis the Newton interpolatory divided-difference formula (3.10) at $x_0, x_1, \ldots, x_n$, that is,

$$P_n(x) = f[x_0] + \sum_{k=1}^{n} f[x_0, x_1, \ldots, x_k](x - x_0) \cdots (x - x_{k-1}).$$

The alternative method uses the connection between the nth divided difference and the nth derivative of f, as outlined in Theorem 3.6 in Section 3.3.

Suppose that the distinct numbers $x_0, x_1, \ldots, x_n$ are given together with the values of f and f' at these numbers. Define a new sequence $z_0, z_1, \ldots, z_{2n+1}$ by

$$z_{2i} = z_{2i+1} = x_i, \quad \text{for each } i = 0, 1, \ldots, n,$$

and construct the divided difference table in the form of Table 3.9 that uses $z_0, z_1, \ldots, z_{2n+1}$.

Since $z_{2i} = z_{2i+1} = x_i$ for each i, we cannot define $f[z_{2i}, z_{2i+1}]$ by the divided difference formula. However, if we assume, based on Theorem 3.6, that the reasonable substitution in this situation is $f[z_{2i}, z_{2i+1}] = f'(z_{2i}) = f'(x_i)$, we can use the entries

$$f'(x_0), f'(x_1), \ldots, f'(x_n)$$

in place of the undefined first divided differences

$$f[z_0, z_1], f[z_2, z_3], \ldots, f[z_{2n}, z_{2n+1}].$$

The remaining divided differences are produced as usual, and the appropriate divided differences are employed in Newton's interpolatory divided-difference formula. Table 3.16 shows the entries that are used for the first three divided-difference columns when determining the Hermite polynomial $H_5(x)$ for x_0, x_1, and x_2. The remaining entries are generated in the same manner as in Table 3.9. The Hermite polynomial is then given by

$$H_{2n+1}(x) = f[z_0] + \sum_{k=1}^{2n+1} f[z_0, \ldots, z_k](x - z_0)(x - z_1) \cdots (x - z_{k-1}).$$

A proof of this fact can be found in [Pow], p. 56.

Table 3.16

z	$f(z)$	First divided differences	Second divided differences
$z_0 = x_0$	$f[z_0] = f(x_0)$		
		$f[z_0, z_1] = f'(x_0)$	
$z_1 = x_0$	$f[z_1] = f(x_0)$		$f[z_0, z_1, z_2] = \dfrac{f[z_1, z_2] - f[z_0, z_1]}{z_2 - z_0}$
		$f[z_1, z_2] = \dfrac{f[z_2] - f[z_1]}{z_2 - z_1}$	
$z_2 = x_1$	$f[z_2] = f(x_1)$		$f[z_1, z_2, z_3] = \dfrac{f[z_2, z_3] - f[z_1, z_2]}{z_3 - z_1}$
		$f[z_2, z_3] = f'(x_1)$	
$z_3 = x_1$	$f[z_3] = f(x_1)$		$f[z_2, z_3, z_4] = \dfrac{f[z_3, z_4] - f[z_2, z_3]}{z_4 - z_2}$
		$f[z_3, z_4] = \dfrac{f[z_4] - f[z_3]}{z_4 - z_3}$	
$z_4 = x_2$	$f[z_4] = f(x_2)$		$f[z_3, z_4, z_5] = \dfrac{f[z_4, z_5] - f[z_3, z_4]}{z_5 - z_3}$
		$f[z_4, z_5] = f'(x_2)$	
$z_5 = x_2$	$f[z_5] = f(x_2)$		

Example 2 Use the data given in Example 1 and the divided difference method to determine the Hermite polynomial approximation at $x = 1.5$.

Solution The underlined entries in the first three columns of Table 3.17 are the data given in Example 1. The remaining entries in this table are generated by the standard divided-difference formula (3.9).

For example, for the second entry in the third column we use the second 1.3 entry in the second column and the first 1.6 entry in that column to obtain

$$\frac{0.4554022 - 0.6200860}{1.6 - 1.3} = -0.5489460.$$

For the first entry in the fourth column we use the first 1.3 entry in the third column and the first 1.6 entry in that column to obtain

$$\frac{-0.5489460 - (-0.5220232)}{1.6 - 1.3} = -0.0897427.$$

The value of the Hermite polynomial at 1.5 is

$$\begin{aligned} H_5(1.5) &= f[1.3] + f'(1.3)(1.5 - 1.3) + f[1.3, 1.3, 1.6](1.5 - 1.3)^2 \\ &\quad + f[1.3, 1.3, 1.6, 1.6](1.5 - 1.3)^2(1.5 - 1.6) \\ &\quad + f[1.3, 1.3, 1.6, 1.6, 1.9](1.5 - 1.3)^2(1.5 - 1.6)^2 \\ &\quad + f[1.3, 1.3, 1.6, 1.6, 1.9, 1.9](1.5 - 1.3)^2(1.5 - 1.6)^2(1.5 - 1.9) \\ &= 0.6200860 + (-0.5220232)(0.2) + (-0.0897427)(0.2)^2 \\ &\quad + 0.0663657(0.2)^2(-0.1) + 0.0026663(0.2)^2(-0.1)^2 \\ &\quad + (-0.0027738)(0.2)^2(-0.1)^2(-0.4) \\ &= 0.5118277. \end{aligned}$$

■

Table 3.17

1.3	0.6200860					
		−0.5220232				
1.3	0.6200860		−0.0897427			
		−0.5489460		0.0663657		
1.6	0.4554022		−0.0698330		0.0026663	
		−0.5698959		0.0679655		−0.0027738
1.6	0.4554022		−0.0290537		0.0010020	
		−0.5786120		0.0685667		
1.9	0.2818186		−0.0084837			
		−0.5811571				
1.9	0.2818186					

The technique used in Algorithm 3.3 can be extended for use in determining other osculating polynomials. A concise discussion of the procedures can be found in [Pow], pp. 53–57.

ALGORITHM
3.3

Hermite Interpolation

To obtain the coefficients of the Hermite interpolating polynomial $H(x)$ on the $(n+1)$ distinct numbers $x_0, \ldots, x_n$ for the function f:

INPUT numbers $x_0, x_1, \ldots, x_n$; values $f(x_0), \ldots, f(x_n)$ and $f'(x_0), \ldots, f'(x_n)$.

OUTPUT the numbers $Q_{0,0}, Q_{1,1}, \ldots, Q_{2n+1,2n+1}$ where

$$H(x) = Q_{0,0} + Q_{1,1}(x - x_0) + Q_{2,2}(x - x_0)^2 + Q_{3,3}(x - x_0)^2(x - x_1) + Q_{4,4}(x - x_0)^2(x - x_1)^2 + \cdots + Q_{2n+1,2n+1}(x - x_0)^2(x - x_1)^2 \cdots (x - x_{n-1})^2(x - x_n).$$

Step 1 For $i = 0, 1, \ldots, n$ do Steps 2 and 3.

Step 2 Set $z_{2i} = x_i$;
$z_{2i+1} = x_i$;
$Q_{2i,0} = f(x_i)$;
$Q_{2i+1,0} = f(x_i)$;
$Q_{2i+1,1} = f'(x_i)$.

Step 3 If $i \neq 0$ then set

$$Q_{2i,1} = \frac{Q_{2i,0} - Q_{2i-1,0}}{z_{2i} - z_{2i-1}}.$$

Step 4 For $i = 2, 3, \ldots, 2n+1$
for $j = 2, 3, \ldots, i$ set $Q_{i,j} = \dfrac{Q_{i,j-1} - Q_{i-1,j-1}}{z_i - z_{i-j}}$.

Step 5 OUTPUT $(Q_{0,0}, Q_{1,1}, \ldots, Q_{2n+1,2n+1})$;
STOP ■

The *NumericalAnalysis* package in Maple can be used to construct the Hermite coefficients. We first need to load the package and to define the data that is being used, in this case, x_i, $f(x_i)$, and $f'(x_i)$ for $i = 0, 1, \ldots, n$. This is done by presenting the data in the form $[x_i, f(x_i), f'(x_i)]$. For example, the data for Example 2 is entered as

$xy := [[1.3, 0.6200860, -0.5220232], [1.6, 0.4554022, -0.5698959], [1.9, 0.2818186, -0.5811571]]$

Then the command

$h5 := PolynomialInterpolation(xy, method = hermite, independentvar = 'x')$

produces an array whose nonzero entries correspond to the values in Table 3.17. The Hermite interpolating polynomial is created with the command

$Interpolant(h5))$

This gives the polynomial in (almost) Newton forward-difference form

$$1.29871616 - 0.5220232x - 0.08974266667(x - 1.3)^2 + 0.06636555557(x - 1.3)^2(x - 1.6) + 0.002666666633(x - 1.3)^2(x - 1.6)^2 - 0.002774691277(x - 1.3)^2(x - 1.6)^2(x - 1.9)$$

If a standard representation of the polynomial is needed, it is found with

$expand(Interpolant(h5))$

giving the Maple response

$$1.001944063 - 0.0082292208x - 0.2352161732x^2 - 0.01455607812x^3 + 0.02403178946x^4 - 0.002774691277x^5$$

EXERCISE SET 3.4

1. Use Theorem 3.9 or Algorithm 3.3 to construct an approximating polynomial for the following data.

a.

x	$f(x)$	$f'(x)$
8.3	17.56492	3.116256
8.6	18.50515	3.151762

b.

x	$f(x)$	$f'(x)$
0.8	0.22363362	2.1691753
1.0	0.65809197	2.0466965

c.

x	$f(x)$	$f'(x)$
−0.5	−0.0247500	0.7510000
−0.25	0.3349375	2.1890000
0	1.1010000	4.0020000

d.

x	$f(x)$	$f'(x)$
0.1	−0.62049958	3.58502082
0.2	−0.28398668	3.14033271
0.3	0.00660095	2.66668043
0.4	0.24842440	2.16529366

2. Use Theorem 3.9 or Algorithm 3.3 to construct an approximating polynomial for the following data.

a.

x	$f(x)$	$f'(x)$
0	1.00000	2.00000
0.5	2.71828	5.43656

b.

x	$f(x)$	$f'(x)$
−0.25	1.33203	0.437500
0.25	0.800781	−0.625000

c.

x	$f(x)$	$f'(x)$
0.1	−0.29004996	−2.8019975
0.2	−0.56079734	−2.6159201
0.3	−0.81401972	−2.9734038

d.

x	$f(x)$	$f'(x)$
−1	0.86199480	0.15536240
−0.5	0.95802009	0.23269654
0	1.0986123	0.33333333
0.5	1.2943767	0.45186776

3. The data in Exercise 1 were generated using the following functions. Use the polynomials constructed in Exercise 1 for the given value of x to approximate $f(x)$, and calculate the absolute error.

a. $f(x) = x \ln x$; approximate $f(8.4)$.

b. $f(x) = \sin(e^x - 2)$; approximate $f(0.9)$.

c. $f(x) = x^3 + 4.001x^2 + 4.002x + 1.101$; approximate $f(-1/3)$.

d. $f(x) = x \cos x - 2x^2 + 3x - 1$; approximate $f(0.25)$.

4. The data in Exercise 2 were generated using the following functions. Use the polynomials constructed in Exercise 2 for the given value of x to approximate $f(x)$, and calculate the absolute error.

a. $f(x) = e^{2x}$; approximate $f(0.43)$.

b. $f(x) = x^4 - x^3 + x^2 - x + 1$; approximate $f(0)$.

c. $f(x) = x^2 \cos x - 3x$; approximate $f(0.18)$.

d. $f(x) = \ln(e^x + 2)$; approximate $f(0.25)$.

5. **a.** Use the following values and five-digit rounding arithmetic to construct the Hermite interpolating polynomial to approximate $\sin 0.34$.

x	$\sin x$	$D_x \sin x = \cos x$
0.30	0.29552	0.95534
0.32	0.31457	0.94924
0.35	0.34290	0.93937

b. Determine an error bound for the approximation in part (a), and compare it to the actual error.

c. Add $\sin 0.33 = 0.32404$ and $\cos 0.33 = 0.94604$ to the data, and redo the calculations.

6. Let $f(x) = 3xe^x - e^{2x}$.

a. Approximate $f(1.03)$ by the Hermite interpolating polynomial of degree at most three using $x_0 = 1$ and $x_1 = 1.05$. Compare the actual error to the error bound.

b. Repeat (a) with the Hermite interpolating polynomial of degree at most five, using $x_0 = 1$, $x_1 = 1.05$, and $x_2 = 1.07$.

7. Use the error formula and Maple to find a bound for the errors in the approximations of $f(x)$ in parts (a) and (c) of Exercise 3.

8. Use the error formula and Maple to find a bound for the errors in the approximations of $f(x)$ in parts (a) and (c) of Exercise 4.

9. The following table lists data for the function described by $f(x) = e^{0.1x^2}$. Approximate $f(1.25)$ by using $H_5(1.25)$ and $H_3(1.25)$, where H_5 uses the nodes $x_0 = 1$, $x_1 = 2$, and $x_2 = 3$; and H_3 uses the nodes $\bar{x}_0 = 1$ and $\bar{x}_1 = 1.5$. Find error bounds for these approximations.

x	$f(x) = e^{0.1x^2}$	$f'(x) = 0.2xe^{0.1x^2}$
$x_0 = \bar{x}_0 = 1$	1.105170918	0.2210341836
$\bar{x}_1 = 1.5$	1.252322716	0.3756968148
$x_1 = 2$	1.491824698	0.5967298792
$x_2 = 3$	2.459603111	1.475761867

10. A car traveling along a straight road is clocked at a number of points. The data from the observations are given in the following table, where the time is in seconds, the distance is in feet, and the speed is in feet per second.

Time	0	3	5	8	13
Distance	0	225	383	623	993
Speed	75	77	80	74	72

a. Use a Hermite polynomial to predict the position of the car and its speed when $t = 10$ s.

b. Use the derivative of the Hermite polynomial to determine whether the car ever exceeds a 55 mi/h speed limit on the road. If so, what is the first time the car exceeds this speed?

c. What is the predicted maximum speed for the car?

11. **a.** Show that $H_{2n+1}(x)$ is the unique polynomial of least degree agreeing with f and f' at $x_0, \ldots, x_n$. [*Hint:* Assume that $P(x)$ is another such polynomial and consider $D = H_{2n+1} - P$ and D' at $x_0, x_1, \ldots, x_n$.]

b. Derive the error term in Theorem 3.9. [*Hint:* Use the same method as in the Lagrange error derivation, Theorem 3.3, defining

$$g(t) = f(t) - H_{2n+1}(t) - \frac{(t-x_0)^2 \cdots (t-x_n)^2}{(x-x_0)^2 \cdots (x-x_n)^2}[f(x) - H_{2n+1}(x)]$$

and using the fact that $g'(t)$ has $(2n+2)$ *distinct* zeros in $[a, b]$.]

12. Let $z_0 = x_0$, $z_1 = x_0$, $z_2 = x_1$, and $z_3 = x_1$. Form the following divided-difference table.

$z_0 = x_0$	$f[z_0] = f(x_0)$			
		$f[z_0, z_1] = f'(x_0)$		
$z_1 = x_0$	$f[z_1] = f(x_0)$		$f[z_0, z_1, z_2]$	
		$f[z_1, z_2]$		$f[z_0, z_1, z_2, z_3]$
$z_2 = x_1$	$f[z_2] = f(x_1)$		$f[z_1, z_2, z_3]$	
		$f[z_2, z_3] = f'(x_1)$		
$z_3 = x_1$	$f[z_3] = f(x_1)$			

Show that the cubic Hermite polynomial $H_3(x)$ can also be written as $f[z_0] + f[z_0, z_1](x - x_0) + f[z_0, z_1, z_2](x - x_0)^2 + f[z_0, z_1, z_2, z_3](x - x_0)^2(x - x_1)$.

3.5 Cubic Spline Interpolation[1]

The previous sections concerned the approximation of arbitrary functions on closed intervals using a single polynomial. However, high-degree polynomials can oscillate erratically, that is, a minor fluctuation over a small portion of the interval can induce large fluctuations over the entire range. We will see a good example of this in Figure 3.14 at the end of this section.

An alternative approach is to divide the approximation interval into a collection of subintervals and construct a (generally) different approximating polynomial on each subinterval. This is called **piecewise-polynomial approximation**.

Piecewise-Polynomial Approximation

The simplest piecewise-polynomial approximation is **piecewise-linear** interpolation, which consists of joining a set of data points

$$\{(x_0, f(x_0)), (x_1, f(x_1)), \ldots, (x_n, f(x_n))\}$$

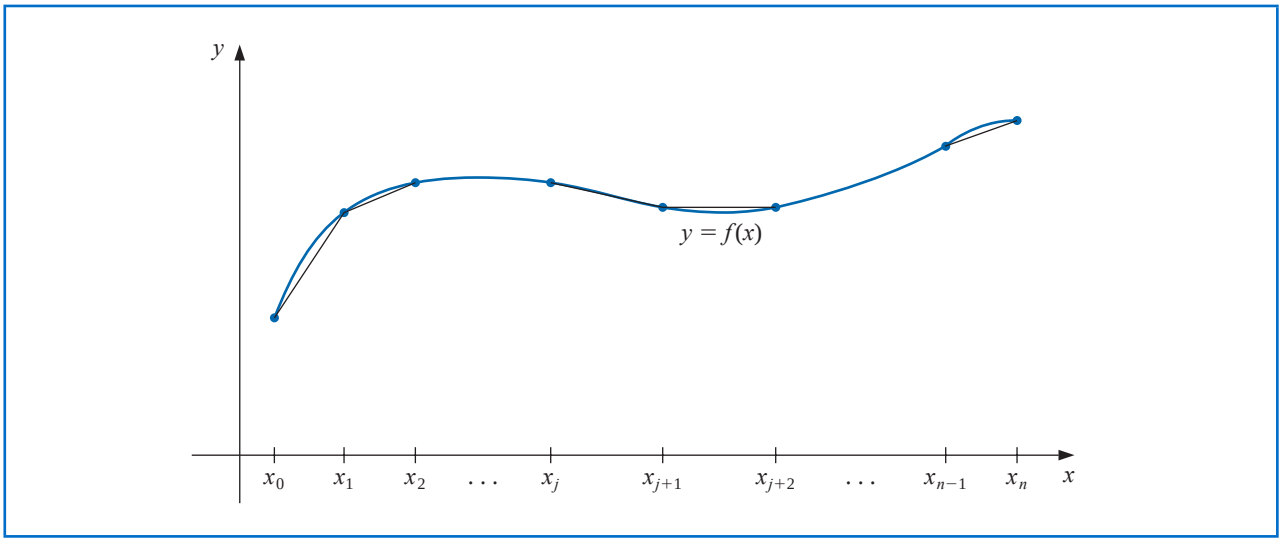
by a series of straight lines, as shown in Figure 3.7.

A disadvantage of linear function approximation is that there is likely no differentiability at the endpoints of the subintervals, which, in a geometrical context, means that the interpolating function is not "smooth." Often it is clear from physical conditions that smoothness is required, so the approximating function must be continuously differentiable.

An alternative procedure is to use a piecewise polynomial of Hermite type. For example, if the values of f and of f' are known at each of the points $x_0 < x_1 < \cdots < x_n$, a cubic Hermite polynomial can be used on each of the subintervals $[x_0, x_1], [x_1, x_2], \ldots, [x_{n-1}, x_n]$ to obtain a function that has a continuous derivative on the interval $[x_0, x_n]$.

[1]The proofs of the theorems in this section rely on results in Chapter 6.

Figure 3.7



Isaac Jacob Schoenberg (1903–1990) developed his work on splines during World War II while on leave from the University of Pennsylvania to work at the Army's Ballistic Research Laboratory in Aberdeen, Maryland. His original work involved numerical procedures for solving differential equations. The much broader application of splines to the areas of data fitting and computer-aided geometric design became evident with the widespread availability of computers in the 1960s.

To determine the appropriate Hermite cubic polynomial on a given interval is simply a matter of computing $H_3(x)$ for that interval. The Lagrange interpolating polynomials needed to determine H_3 are of first degree, so this can be accomplished without great difficulty. However, to use Hermite piecewise polynomials for general interpolation, we need to know the derivative of the function being approximated, and this is frequently unavailable.

The remainder of this section considers approximation using piecewise polynomials that require no specific derivative information, except perhaps at the endpoints of the interval on which the function is being approximated.

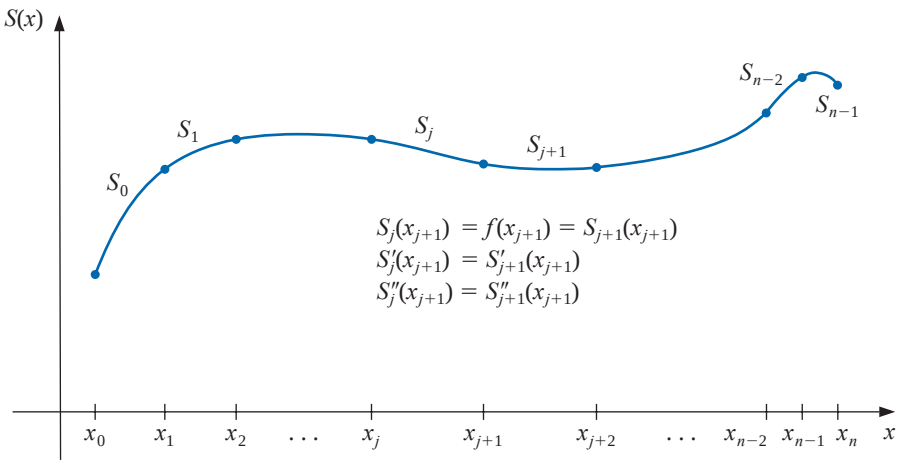
The simplest type of differentiable piecewise-polynomial function on an entire interval $[x_0, x_n]$ is the function obtained by fitting one quadratic polynomial between each successive pair of nodes. This is done by constructing a quadratic on $[x_0, x_1]$ agreeing with the function at x_0 and x_1, another quadratic on $[x_1, x_2]$ agreeing with the function at x_1 and x_2, and so on. A general quadratic polynomial has three arbitrary constants—the constant term, the coefficient of x, and the coefficient of x^2—and only two conditions are required to fit the data at the endpoints of each subinterval. So flexibility exists that permits the quadratics to be chosen so that the interpolant has a continuous derivative on $[x_0, x_n]$. The difficulty arises because we generally need to specify conditions about the derivative of the interpolant at the endpoints x_0 and x_n. There is not a sufficient number of constants to ensure that the conditions will be satisfied. (See Exercise 26.)

The root of the word "spline" is the same as that of splint. It was originally a small strip of wood that could be used to join two boards. Later the word was used to refer to a long flexible strip, generally of metal, that could be used to draw continuous smooth curves by forcing the strip to pass through specified points and tracing along the curve.

Cubic Splines

The most common piecewise-polynomial approximation uses cubic polynomials between each successive pair of nodes and is called **cubic spline interpolation**. A general cubic polynomial involves four constants, so there is sufficient flexibility in the cubic spline procedure to ensure that the interpolant is not only continuously differentiable on the interval, but also has a continuous second derivative. The construction of the cubic spline does not, however, assume that the derivatives of the interpolant agree with those of the function it is approximating, even at the nodes. (See Figure 3.8.)

Figure 3.8



Definition 3.10 Given a function f defined on $[a, b]$ and a set of nodes $a = x_0 < x_1 < \cdots < x_n = b$, a **cubic spline interpolant** S for f is a function that satisfies the following conditions:

A natural spline has no conditions imposed for the direction at its endpoints, so the curve takes the shape of a straight line after it passes through the interpolation points nearest its endpoints. The name derives from the fact that this is the natural shape a flexible strip assumes if forced to pass through specified interpolation points with no additional constraints. (See Figure 3.9.)

(a) $S(x)$ is a cubic polynomial, denoted $S_j(x)$, on the subinterval $[x_j, x_{j+1}]$ for each $j = 0, 1, \ldots, n-1$;

(b) $S_j(x_j) = f(x_j)$ and $S_j(x_{j+1}) = f(x_{j+1})$ for each $j = 0, 1, \ldots, n-1$;

(c) $S_{j+1}(x_{j+1}) = S_j(x_{j+1})$ for each $j = 0, 1, \ldots, n-2$; (*Implied by* **(b)**.)

(d) $S'_{j+1}(x_{j+1}) = S'_j(x_{j+1})$ for each $j = 0, 1, \ldots, n-2$;

(e) $S''_{j+1}(x_{j+1}) = S''_j(x_{j+1})$ for each $j = 0, 1, \ldots, n-2$;

(f) One of the following sets of boundary conditions is satisfied:

(i) $S''(x_0) = S''(x_n) = 0$ (**natural** *(or free)* **boundary**);

(ii) $S'(x_0) = f'(x_0)$ and $S'(x_n) = f'(x_n)$ (**clamped boundary**). ■

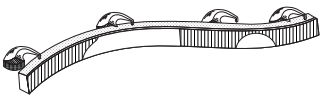


Figure 3.9

Although cubic splines are defined with other boundary conditions, the conditions given in **(f)** are sufficient for our purposes. When the free boundary conditions occur, the spline is called a **natural spline**, and its graph approximates the shape that a long flexible rod would assume if forced to go through the data points $\{(x_0, f(x_0)), (x_1, f(x_1)), \ldots, (x_n, f(x_n))\}$.

In general, clamped boundary conditions lead to more accurate approximations because they include more information about the function. However, for this type of boundary condition to hold, it is necessary to have either the values of the derivative at the endpoints or an accurate approximation to those values.

Example 1 Construct a natural cubic spline that passes through the points $(1, 2)$, $(2, 3)$, and $(3, 5)$.

Solution This spline consists of two cubics. The first for the interval $[1, 2]$, denoted

$$S_0(x) = a_0 + b_0(x - 1) + c_0(x - 1)^2 + d_0(x - 1)^3,$$

and the other for $[2, 3]$, denoted

$$S_1(x) = a_1 + b_1(x-2) + c_1(x-2)^2 + d_1(x-2)^3.$$

There are 8 constants to be determined, which requires 8 conditions. Four conditions come from the fact that the splines must agree with the data at the nodes. Hence

$$2 = f(1) = a_0, \quad 3 = f(2) = a_0 + b_0 + c_0 + d_0, \quad 3 = f(2) = a_1, \quad \text{and}$$
$$5 = f(3) = a_1 + b_1 + c_1 + d_1.$$

Two more come from the fact that $S_0'(2) = S_1'(2)$ and $S_0''(2) = S_1''(2)$. These are

$$S_0'(2) = S_1'(2): \quad b_0 + 2c_0 + 3d_0 = b_1 \qquad \text{and} \qquad S_0''(2) = S_1''(2): \quad 2c_0 + 6d_0 = 2c_1$$

The final two come from the natural boundary conditions:

$$S_0''(1) = 0: \quad 2c_0 = 0 \qquad \text{and} \qquad S_1''(3) = 0: \quad 2c_1 + 6d_1 = 0.$$

Solving this system of equations gives the spline

$$S(x) = \begin{cases} 2 + \frac{3}{4}(x-1) + \frac{1}{4}(x-1)^3, & \text{for } x \in [1,2] \\ 3 + \frac{3}{2}(x-2) + \frac{3}{4}(x-2)^2 - \frac{1}{4}(x-2)^3, & \text{for } x \in [2,3] \end{cases}$$

■

Construction of a Cubic Spline

As the preceding example demonstrates, a spline defined on an interval that is divided into n subintervals will require determining $4n$ constants. To construct the cubic spline interpolant for a given function f, the conditions in the definition are applied to the cubic polynomials

$$S_j(x) = a_j + b_j(x - x_j) + c_j(x - x_j)^2 + d_j(x - x_j)^3,$$

Clamping a spline indicates that the ends of the flexible strip are fixed so that it is forced to take a specific direction at each of its endpoints. This is important, for example, when two spline functions should match at their endpoints. This is done mathematically by specifying the values of the derivative of the curve at the endpoints of the spline.

for each $j = 0, 1, \ldots, n-1$. Since $S_j(x_j) = a_j = f(x_j)$, condition **(c)** can be applied to obtain

$$a_{j+1} = S_{j+1}(x_{j+1}) = S_j(x_{j+1}) = a_j + b_j(x_{j+1} - x_j) + c_j(x_{j+1} - x_j)^2 + d_j(x_{j+1} - x_j)^3,$$

for each $j = 0, 1, \ldots, n-2$.

The terms $x_{j+1} - x_j$ are used repeatedly in this development, so it is convenient to introduce the simpler notation

$$h_j = x_{j+1} - x_j,$$

for each $j = 0, 1, \ldots, n-1$. If we also define $a_n = f(x_n)$, then the equation

$$a_{j+1} = a_j + b_j h_j + c_j h_j^2 + d_j h_j^3 \tag{3.15}$$

holds for each $j = 0, 1, \ldots, n-1$.

In a similar manner, define $b_n = S'(x_n)$ and observe that

$$S'_j(x) = b_j + 2c_j(x - x_j) + 3d_j(x - x_j)^2$$

implies $S'_j(x_j) = b_j$, for each $j = 0, 1, \ldots, n-1$. Applying condition **(d)** gives

$$b_{j+1} = b_j + 2c_j h_j + 3d_j h_j^2, \tag{3.16}$$

for each $j = 0, 1, \ldots, n-1$.

Another relationship between the coefficients of S_j is obtained by defining $c_n = S''(x_n)/2$ and applying condition **(e)**. Then, for each $j = 0, 1, \ldots, n-1$,

$$c_{j+1} = c_j + 3d_j h_j. \tag{3.17}$$

Solving for d_j in Eq. (3.17) and substituting this value into Eqs. (3.15) and (3.16) gives, for each $j = 0, 1, \ldots, n-1$, the new equations

$$a_{j+1} = a_j + b_j h_j + \frac{h_j^2}{3}(2c_j + c_{j+1}) \tag{3.18}$$

and

$$b_{j+1} = b_j + h_j(c_j + c_{j+1}). \tag{3.19}$$

The final relationship involving the coefficients is obtained by solving the appropriate equation in the form of equation (3.18), first for b_j,

$$b_j = \frac{1}{h_j}(a_{j+1} - a_j) - \frac{h_j}{3}(2c_j + c_{j+1}), \tag{3.20}$$

and then, with a reduction of the index, for b_{j-1}. This gives

$$b_{j-1} = \frac{1}{h_{j-1}}(a_j - a_{j-1}) - \frac{h_{j-1}}{3}(2c_{j-1} + c_j).$$

Substituting these values into the equation derived from Eq. (3.19), with the index reduced by one, gives the linear system of equations

$$h_{j-1}c_{j-1} + 2(h_{j-1} + h_j)c_j + h_j c_{j+1} = \frac{3}{h_j}(a_{j+1} - a_j) - \frac{3}{h_{j-1}}(a_j - a_{j-1}), \tag{3.21}$$

for each $j = 1, 2, \ldots, n-1$. This system involves only the $\{c_j\}_{j=0}^n$ as unknowns. The values of $\{h_j\}_{j=0}^{n-1}$ and $\{a_j\}_{j=0}^n$ are given, respectively, by the spacing of the nodes $\{x_j\}_{j=0}^n$ and the values of f at the nodes. So once the values of $\{c_j\}_{j=0}^n$ are determined, it is a simple matter to find the remainder of the constants $\{b_j\}_{j=0}^{n-1}$ from Eq. (3.20) and $\{d_j\}_{j=0}^{n-1}$ from Eq. (3.17). Then we can construct the cubic polynomials $\{S_j(x)\}_{j=0}^{n-1}$.

The major question that arises in connection with this construction is whether the values of $\{c_j\}_{j=0}^n$ can be found using the system of equations given in (3.21) and, if so, whether these values are unique. The following theorems indicate that this is the case when either of the boundary conditions given in part **(f)** of the definition are imposed. The proofs of these theorems require material from linear algebra, which is discussed in Chapter 6.

Natural Splines

Theorem 3.11 If f is defined at $a = x_0 < x_1 < \cdots < x_n = b$, then f has a unique natural spline interpolant S on the nodes $x_0, x_1, \ldots, x_n$; that is, a spline interpolant that satisfies the natural boundary conditions $S''(a) = 0$ and $S''(b) = 0$. ■

Proof The boundary conditions in this case imply that $c_n = S''(x_n)/2 = 0$ and that

$$0 = S''(x_0) = 2c_0 + 6d_0(x_0 - x_0),$$

so $c_0 = 0$. The two equations $c_0 = 0$ and $c_n = 0$ together with the equations in (3.21) produce a linear system described by the vector equation $A\mathbf{x} = \mathbf{b}$, where A is the $(n+1) \times (n+1)$ matrix

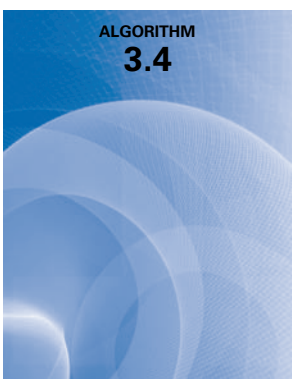
$$A = \begin{bmatrix} 1 & 0 & 0 & \cdots & & & 0 \\ h_0 & 2(h_0 + h_1) & h_1 & & & & \\ 0 & h_1 & 2(h_1 + h_2) & h_2 & & & \\ \vdots & \ddots & \ddots & \ddots & \ddots & & \\ & & & h_{n-2} & 2(h_{n-2} + h_{n-1}) & h_{n-1} & \\ 0 & \cdots & & & 0 & 0 & 1 \end{bmatrix},$$

and $\mathbf{b}$ and $\mathbf{x}$ are the vectors

$$\mathbf{b} = \begin{bmatrix} 0 \\ \frac{3}{h_1}(a_2 - a_1) - \frac{3}{h_0}(a_1 - a_0) \\ \vdots \\ \frac{3}{h_{n-1}}(a_n - a_{n-1}) - \frac{3}{h_{n-2}}(a_{n-1} - a_{n-2}) \\ 0 \end{bmatrix} \quad \text{and} \quad \mathbf{x} = \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_n \end{bmatrix}.$$

The matrix A is strictly diagonally dominant, that is, in each row the magnitude of the diagonal entry exceeds the sum of the magnitudes of all the other entries in the row. A linear system with a matrix of this form will be shown by Theorem 6.21 in Section 6.6 to have a unique solution for $c_0, c_1, \ldots, c_n$. ■ ■ ■

The solution to the cubic spline problem with the boundary conditions $S''(x_0) = S''(x_n) = 0$ can be obtained by applying Algorithm 3.4.



Natural Cubic Spline

To construct the cubic spline interpolant S for the function f, defined at the numbers $x_0 < x_1 < \cdots < x_n$, satisfying $S''(x_0) = S''(x_n) = 0$:

INPUT $n; x_0, x_1, \ldots, x_n; a_0 = f(x_0), a_1 = f(x_1), \ldots, a_n = f(x_n)$.

OUTPUT a_j, b_j, c_j, d_j for $j = 0, 1, \ldots, n-1$.

(*Note:* $S(x) = S_j(x) = a_j + b_j(x - x_j) + c_j(x - x_j)^2 + d_j(x - x_j)^3$ *for* $x_j \le x \le x_{j+1}$.)

Step 1 For $i = 0, 1, \ldots, n-1$ set $h_i = x_{i+1} - x_i$.

Step 2 For $i = 1, 2, \ldots, n-1$ set

$$\alpha_i = \frac{3}{h_i}(a_{i+1} - a_i) - \frac{3}{h_{i-1}}(a_i - a_{i-1}).$$

Step 3 Set $l_0 = 1$; (*Steps 3, 4, 5, and part of Step 6 solve a tridiagonal linear system using a method described in Algorithm 6.7.*)
$\mu_0 = 0$;
$z_0 = 0$.

Step 4 For $i = 1, 2, \ldots, n-1$
set $l_i = 2(x_{i+1} - x_{i-1}) - h_{i-1}\mu_{i-1}$;
$\mu_i = h_i/l_i$;
$z_i = (\alpha_i - h_{i-1}z_{i-1})/l_i$.

Step 5 Set $l_n = 1$;
$z_n = 0$;
$c_n = 0$.

Step 6 For $j = n-1, n-2, \ldots, 0$
set $c_j = z_j - \mu_j c_{j+1}$;
$b_j = (a_{j+1} - a_j)/h_j - h_j(c_{j+1} + 2c_j)/3$;
$d_j = (c_{j+1} - c_j)/(3h_j)$.

Step 7 OUTPUT $(a_j, b_j, c_j, d_j$ for $j = 0, 1, \ldots, n-1)$;
STOP. ■

Example 2 At the beginning of Chapter 3 we gave some Taylor polynomials to approximate the exponential $f(x) = e^x$. Use the data points $(0, 1)$, $(1, e)$, $(2, e^2)$, and $(3, e^3)$ to form a natural spline $S(x)$ that approximates $f(x) = e^x$.

Solution We have $n = 3$, $h_0 = h_1 = h_2 = 1$, $a_0 = 1$, $a_1 = e$, $a_2 = e^2$, and $a_3 = e^3$. So the matrix A and the vectors $\mathbf{b}$ and $\mathbf{x}$ given in Theorem 3.11 have the forms

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 0 \\ 3(e^2 - 2e + 1) \\ 3(e^3 - 2e^2 + e) \\ 0 \end{bmatrix}, \quad \text{and} \quad \mathbf{x} = \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{bmatrix}.$$

The vector-matrix equation $A\mathbf{x} = \mathbf{b}$ is equivalent to the system of equations

$$\begin{aligned} c_0 &= 0, \\ c_0 + 4c_1 + c_2 &= 3(e^2 - 2e + 1), \\ c_1 + 4c_2 + c_3 &= 3(e^3 - 2e^2 + e), \\ c_3 &= 0. \end{aligned}$$

This system has the solution $c_0 = c_3 = 0$, and to 5 decimal places,

$$c_1 = \frac{1}{5}(-e^3 + 6e^2 - 9e + 4) \approx 0.75685, \quad \text{and} \quad c_2 = \frac{1}{5}(4e^3 - 9e^2 + 6e - 1) \approx 5.83007.$$

Solving for the remaining constants gives

$$\begin{aligned}
b_0 &= \frac{1}{h_0}(a_1 - a_0) - \frac{h_0}{3}(c_1 + 2c_0) \\
&= (e-1) - \frac{1}{15}(-e^3 + 6e^2 - 9e + 4) \approx 1.46600, \\
b_1 &= \frac{1}{h_1}(a_2 - a_1) - \frac{h_1}{3}(c_2 + 2c_1) \\
&= (e^2 - e) - \frac{1}{15}(2e^3 + 3e^2 - 12e + 7) \approx 2.22285, \\
b_2 &= \frac{1}{h_2}(a_3 - a_2) - \frac{h_2}{3}(c_3 + 2c_2) \\
&= (e^3 - e^2) - \frac{1}{15}(8e^3 - 18e^2 + 12e - 2) \approx 8.80977, \\
d_0 &= \frac{1}{3h_0}(c_1 - c_0) = \frac{1}{15}(-e^3 + 6e^2 - 9e + 4) \approx 0.25228, \\
d_1 &= \frac{1}{3h_1}(c_2 - c_1) = \frac{1}{3}(e^3 - 3e^2 + 3e - 1) \approx 1.69107,
\end{aligned}$$

and

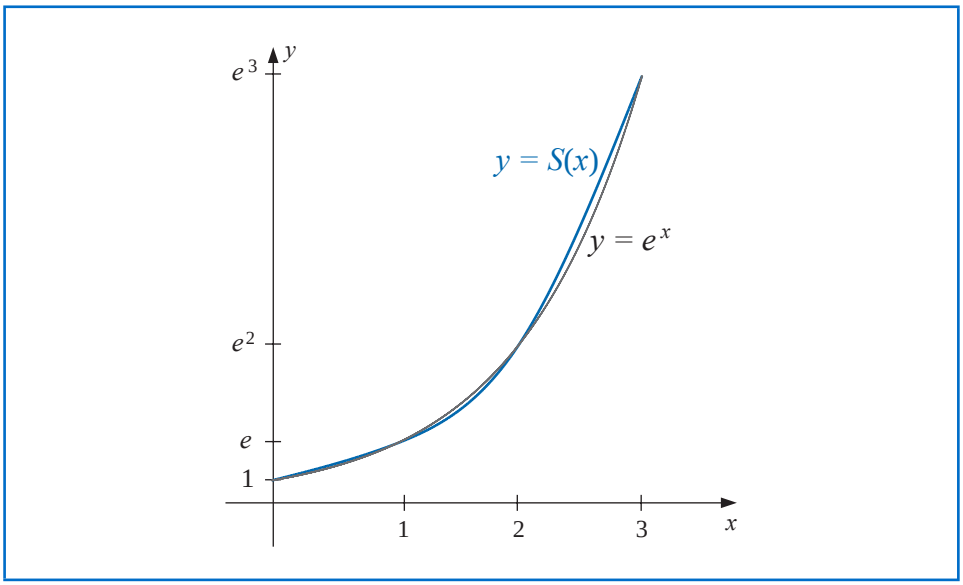
$$d_2 = \frac{1}{3h_2}(c_3 - c_1) = \frac{1}{15}(-4e^3 + 9e^2 - 6e + 1) \approx -1.94336.$$

The natural cubic spine is described piecewise by

$$S(x) = \begin{cases} 1 + 1.46600x + 0.25228x^3, & \text{for } x \in [0,1], \\ 2.71828 + 2.22285(x-1) + 0.75685(x-1)^2 + 1.69107(x-1)^3, & \text{for } x \in [1,2], \\ 7.38906 + 8.80977(x-2) + 5.83007(x-2)^2 - 1.94336(x-2)^3, & \text{for } x \in [2,3]. \end{cases}$$

The spline and its agreement with $f(x) = e^x$ are shown in Figure 3.10. ■

Figure 3.10



The *NumericalAnalysis* package can be used to create a cubic spline in a manner similar to other constructions in this chapter. However, the *CurveFitting* Package in Maple can also be used, and since this has not been discussed previously we will use it to create the natural spline in Example 2. First we load the package with the command

with(*CurveFitting*)

and define the function being approximated with

$f := x \to e^x$

To create a spline we need to specify the nodes, variable, the degree, and the natural endpoints. This is done with

$sn := t \to Spline([[0., 1.0], [1.0, f(1.0)], [2.0, f(2.0)], [3.0, f(3.0)]], t, degree = 3, endpoints = \text{'natural'})$

Maple returns

$$t \to CurveFitting\text{:-}Spline([[0., 1.0], [1.0, f(1.0)], [2.0, f(2.0)], [3.0, f(3.0)]], t, degree = 3, endpoints = \text{'natural'})$$

The form of the natural spline is seen with the command

$sn(t)$

which produces

$$\begin{cases} 1. + 1.465998t^2 + 0.2522848t^3 & t < 1.0 \\ 0.495432 + 2.22285t + 0.756853(t-1.0)^2 + 1.691071(t-1.0)^3 & t < 2.0 \\ -10.230483 + 8.809770t + 5.830067(t-2.0)^2 - 1.943356(t-2.0)^3 & otherwise \end{cases}$$

Once we have determined a spline approximation for a function we can use it to approximate other properties of the function. The next illustration involves the integral of the spline we found in the previous example.

Illustration To approximate the integral of $f(x) = e^x$ on $[0, 3]$, which has the value

$$\int_0^3 e^x\,dx = e^3 - 1 \approx 20.08553692 - 1 = 19.08553692,$$

we can piecewise integrate the spline that approximates f on this integral. This gives

$$\begin{aligned} \int_0^3 S(x) = &\int_0^1 1 + 1.46600x + 0.25228x^3\,dx \\ &+ \int_1^2 2.71828 + 2.22285(x-1) + 0.75685(x-1)^2 + 1.69107(x-1)^3\,dx \\ &+ \int_2^3 7.38906 + 8.80977(x-2) + 5.83007(x-2)^2 - 1.94336(x-2)^3\,dx. \end{aligned}$$

Integrating and collecting values from like powers gives

$$\begin{aligned}\int_0^3 S(x) &= \left[x + 1.46600\frac{x^2}{2} + 0.25228\frac{x^4}{4}\right]_0^1 \\ &\quad + \left[2.71828(x-1) + 2.22285\frac{(x-1)^2}{2} + 0.75685\frac{(x-1)^3}{3} + 1.69107\frac{(x-1)^4}{4}\right]_1^2 \\ &\quad + \left[7.38906(x-2) + 8.80977\frac{(x-2)^2}{2} + 5.83007\frac{(x-2)^3}{3} - 1.94336\frac{(x-2)^4}{4}\right]_2^3 \\ &= (1 + 2.71828 + 7.38906) + \frac{1}{2}(1.46600 + 2.22285 + 8.80977) \\ &\quad + \frac{1}{3}(0.75685 + 5.83007) + \frac{1}{4}(0.25228 + 1.69107 - 1.94336) \\ &= 19.55229.\end{aligned}$$

Because the nodes are equally spaced in this example the integral approximation is simply

$$\int_0^3 S(x)\,dx = (a_0+a_1+a_2)+\frac{1}{2}(b_0+b_1+b_2)+\frac{1}{3}(c_0+c_1+c_2)+\frac{1}{4}(d_0+d_1+d_2). \quad (3.22)$$

□

If we create the natural spline using Maple as described after Example 2, we can then use Maple's integration command to find the value in the Illustration. Simply enter

int$(sn(t), t = 0 .. 3)$

19.55228648

Clamped Splines

Example 3 In Example 1 we found a natural spline S that passes through the points $(1, 2)$, $(2, 3)$, and $(3, 5)$. Construct a clamped spline s through these points that has $s'(1) = 2$ and $s'(3) = 1$.

Solution Let

$$s_0(x) = a_0 + b_0(x-1) + c_0(x-1)^2 + d_0(x-1)^3,$$

be the cubic on $[1, 2]$ and the cubic on $[2, 3]$ be

$$s_1(x) = a_1 + b_1(x-2) + c_1(x-2)^2 + d_1(x-2)^3.$$

Then most of the conditions to determine the 8 constants are the same as those in Example 1. That is,

$$2 = f(1) = a_0, \quad 3 = f(2) = a_0 + b_0 + c_0 + d_0, \quad 3 = f(2) = a_1, \qquad \text{and}$$

$$5 = f(3) = a_1 + b_1 + c_1 + d_1.$$

$$s_0'(2) = s_1'(2): \quad b_0 + 2c_0 + 3d_0 = b_1 \qquad \text{and} \qquad s_0''(2) = s_1''(2): \quad 2c_0 + 6d_0 = 2c_1$$

However, the boundary conditions are now

$$s_0'(1) = 2: \quad b_0 = 2 \quad \text{and} \quad s_1'(3) = 1: \quad b_1 + 2c_1 + 3d_1 = 1.$$

Solving this system of equations gives the spline as

$$s(x) = \begin{cases} 2 + 2(x-1) - \frac{5}{2}(x-1)^2 + \frac{3}{2}(x-1)^3, & \text{for } x \in [1,2] \\ 3 + \frac{3}{2}(x-2) + 2(x-2)^2 - \frac{3}{2}(x-2)^3, & \text{for } x \in [2,3] \end{cases}$$ ■

In the case of general clamped boundary conditions we have a result that is similar to the theorem for natural boundary conditions described in Theorem 3.11.

Theorem 3.12 If f is defined at $a = x_0 < x_1 < \cdots < x_n = b$ and differentiable at a and b, then f has a unique clamped spline interpolant S on the nodes $x_0, x_1, \ldots, x_n$; that is, a spline interpolant that satisfies the clamped boundary conditions $S'(a) = f'(a)$ and $S'(b) = f'(b)$. ■

Proof Since $f'(a) = S'(a) = S'(x_0) = b_0$, Eq. (3.20) with $j = 0$ implies

$$f'(a) = \frac{1}{h_0}(a_1 - a_0) - \frac{h_0}{3}(2c_0 + c_1).$$

Consequently,

$$2h_0c_0 + h_0c_1 = \frac{3}{h_0}(a_1 - a_0) - 3f'(a).$$

Similarly,

$$f'(b) = b_n = b_{n-1} + h_{n-1}(c_{n-1} + c_n),$$

so Eq. (3.20) with $j = n - 1$ implies that

$$\begin{aligned} f'(b) &= \frac{a_n - a_{n-1}}{h_{n-1}} - \frac{h_{n-1}}{3}(2c_{n-1} + c_n) + h_{n-1}(c_{n-1} + c_n) \\ &= \frac{a_n - a_{n-1}}{h_{n-1}} + \frac{h_{n-1}}{3}(c_{n-1} + 2c_n), \end{aligned}$$

and

$$h_{n-1}c_{n-1} + 2h_{n-1}c_n = 3f'(b) - \frac{3}{h_{n-1}}(a_n - a_{n-1}).$$

Equations (3.21) together with the equations

$$2h_0c_0 + h_0c_1 = \frac{3}{h_0}(a_1 - a_0) - 3f'(a)$$

and

$$h_{n-1}c_{n-1} + 2h_{n-1}c_n = 3f'(b) - \frac{3}{h_{n-1}}(a_n - a_{n-1})$$

determine the linear system $A\mathbf{x} = \mathbf{b}$, where

$$A = \begin{bmatrix} 2h_0 & h_0 & 0 & \cdots & 0 \\ h_0 & 2(h_0+h_1) & h_1 & & \\ 0 & h_1 & 2(h_1+h_2) & h_2 & \\ \vdots & & \ddots & \ddots & \ddots \\ & & h_{n-2} & 2(h_{n-2}+h_{n-1}) & h_{n-1} \\ 0 & \cdots & 0 & h_{n-1} & 2h_{n-1} \end{bmatrix},$$

$$\mathbf{b} = \begin{bmatrix} \frac{3}{h_0}(a_1 - a_0) - 3f'(a) \\ \frac{3}{h_1}(a_2 - a_1) - \frac{3}{h_0}(a_1 - a_0) \\ \vdots \\ \frac{3}{h_{n-1}}(a_n - a_{n-1}) - \frac{3}{h_{n-2}}(a_{n-1} - a_{n-2}) \\ 3f'(b) - \frac{3}{h_{n-1}}(a_n - a_{n-1}) \end{bmatrix}, \quad \text{and} \quad \mathbf{x} = \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_n \end{bmatrix}.$$

This matrix A is also strictly diagonally dominant, so it satisfies the conditions of Theorem 6.21 in Section 6.6. Therefore, the linear system has a unique solution for $c_0, c_1, \ldots, c_n$. ■ ■ ■

The solution to the cubic spline problem with the boundary conditions $S'(x_0) = f'(x_0)$ and $S'(x_n) = f'(x_n)$ can be obtained by applying Algorithm 3.5.

ALGORITHM
3.5

Clamped Cubic Spline

To construct the cubic spline interpolant S for the function f defined at the numbers $x_0 < x_1 < \cdots < x_n$, satisfying $S'(x_0) = f'(x_0)$ and $S'(x_n) = f'(x_n)$:

INPUT n; $x_0, x_1, \ldots, x_n$; $a_0 = f(x_0)$, $a_1 = f(x_1), \ldots, a_n = f(x_n)$; $FPO = f'(x_0)$; $FPN = f'(x_n)$.

OUTPUT a_j, b_j, c_j, d_j for $j = 0, 1, \ldots, n-1$.

(*Note:* $S(x) = S_j(x) = a_j + b_j(x - x_j) + c_j(x - x_j)^2 + d_j(x - x_j)^3$ *for* $x_j \le x \le x_{j+1}$.)

Step 1 For $i = 0, 1, \ldots, n-1$ set $h_i = x_{i+1} - x_i$.

Step 2 Set $\alpha_0 = 3(a_1 - a_0)/h_0 - 3FPO$;
$\alpha_n = 3FPN - 3(a_n - a_{n-1})/h_{n-1}$.

Step 3 For $i = 1, 2, \ldots, n-1$

set $\alpha_i = \dfrac{3}{h_i}(a_{i+1} - a_i) - \dfrac{3}{h_{i-1}}(a_i - a_{i-1})$.

Step 4 Set $l_0 = 2h_0$; (*Steps 4,5,6, and part of Step 7 solve a tridiagonal linear system using a method described in Algorithm 6.7.*)
$\mu_0 = 0.5$;
$z_0 = \alpha_0/l_0$.

Step 5 For $i = 1, 2, \ldots, n-1$
set $l_i = 2(x_{i+1} - x_{i-1}) - h_{i-1}\mu_{i-1}$;
$\mu_i = h_i/l_i$;
$z_i = (\alpha_i - h_{i-1}z_{i-1})/l_i$.

Step 6 Set $l_n = h_{n-1}(2 - \mu_{n-1})$;
$z_n = (\alpha_n - h_{n-1}z_{n-1})/l_n$;
$c_n = z_n$.

Step 7 For $j = n-1, n-2, \ldots, 0$
set $c_j = z_j - \mu_j c_{j+1}$;
$b_j = (a_{j+1} - a_j)/h_j - h_j(c_{j+1} + 2c_j)/3$;
$d_j = (c_{j+1} - c_j)/(3h_j)$.

Step 8 OUTPUT $(a_j, b_j, c_j, d_j$ for $j = 0, 1, \ldots, n-1)$;
STOP.

■

Example 4 Example 2 used a natural spline and the data points $(0, 1)$, $(1, e)$, $(2, e^2)$, and $(3, e^3)$ to form a new approximating function $S(x)$. Determine the clamped spline $s(x)$ that uses this data and the additional information that, since $f'(x) = e^x$, so $f'(0) = 1$ and $f'(3) = e^3$.

Solution As in Example 2, we have $n = 3$, $h_0 = h_1 = h_2 = 1$, $a_0 = 0$, $a_1 = e$, $a_2 = e^2$, and $a_3 = e^3$. This together with the information that $f'(0) = 1$ and $f'(3) = e^3$ gives the the matrix A and the vectors $\mathbf{b}$ and $\mathbf{x}$ with the forms

$$A = \begin{bmatrix} 2 & 1 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 1 & 2 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 3(e-2) \\ 3(e^2 - 2e + 1) \\ 3(e^3 - 2e^2 + e) \\ 3e^2 \end{bmatrix}, \quad \text{and} \quad \mathbf{x} = \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{bmatrix}.$$

The vector-matrix equation $A\mathbf{x} = \mathbf{b}$ is equivalent to the system of equations

$$\begin{aligned} 2c_0 + c_1 &= 3(e-2), \\ c_0 + 4c_1 + c_2 &= 3(e^2 - 2e + 1), \\ c_1 + 4c_2 + c_3 &= 3(e^3 - 2e^2 + e), \\ c_2 + 2c_3 &= 3e^2. \end{aligned}$$

Solving this system simultaneously for c_0, c_1, c_2 and c_3 gives, to 5 decimal places,

$$\begin{aligned} c_0 &= \frac{1}{15}(2e^3 - 12e^2 + 42e - 59) = 0.44468, \\ c_1 &= \frac{1}{15}(-4e^3 + 24e^2 - 39e + 28) = 1.26548, \\ c_2 &= \frac{1}{15}(14e^3 - 39e^2 + 24e - 8) = 3.35087, \\ c_3 &= \frac{1}{15}(-7e^3 + 42e^2 - 12e + 4) = 9.40815. \end{aligned}$$

Solving for the remaining constants in the same manner as Example 2 gives

$$b_0 = 1.00000, \quad b_1 = 2.71016, \quad b_2 = 7.32652,$$

and

$$d_0 = 0.27360, \quad d_1 = 0.69513, \quad d_2 = 2.01909.$$

This gives the clamped cubic spine

$$s(x) = \begin{cases} 1 + x + 0.44468x^2 + 0.27360x^3, & \text{if } 0 \le x < 1, \\ 2.71828 + 2.71016(x-1) + 1.26548(x-1)^2 + 0.69513(x-1)^3, & \text{if } 1 \le x < 2, \\ 7.38906 + 7.32652(x-2) + 3.35087(x-2)^2 + 2.01909(x-2)^3, & \text{if } 2 \le x \le 3. \end{cases}$$

The graph of the clamped spline and $f(x) = e^x$ are so similar that no difference can be seen. ■

We can create the clamped cubic spline in Example 4 with the same commands we used for the natural spline, the only change that is needed is to specify the derivative at the endpoints. In this case we use

$sn := t \to Spline([[0., 1.0], [1.0, f(1.0)], [2.0, f(2.0)], [3.0, f(3.0)]], t, degree = 3,$
$endpoints = [1.0, e^{3.0}])$

giving essentially the same results as in the example.

We can also approximate the integral of f on $[0, 3]$, by integrating the clamped spline. The exact value of the integral is

$$\int_0^3 e^x \, dx = e^3 - 1 \approx 20.08554 - 1 = 19.08554.$$

Because the data is equally spaced, piecewise integrating the clamped spline results in the same formula as in (3.22), that is,

$$\int_0^3 s(x) \, dx = (a_0 + a_1 + a_2) + \frac{1}{2}(b_0 + b_1 + b_2) + \frac{1}{3}(c_0 + c_1 + c_2) + \frac{1}{4}(d_0 + d_1 + d_2).$$

Hence the integral approximation is

$$\begin{aligned} \int_0^3 s(x) \, dx &= (1 + 2.71828 + 7.38906) + \frac{1}{2}(1 + 2.71016 + 7.32652) \\ &\quad + \frac{1}{3}(0.44468 + 1.26548 + 3.35087) + \frac{1}{4}(0.27360 + 0.69513 + 2.01909) \\ &= 19.05965. \end{aligned}$$

The absolute error in the integral approximation using the clamped and natural splines are

$$\text{Natural} : |19.08554 - 19.55229| = 0.46675$$

and

$$\text{Clamped} : |19.08554 - 19.05965| = 0.02589.$$

For integration purposes the clamped spline is vastly superior. This should be no surprise since the boundary conditions for the clamped spline are exact, whereas for the natural spline we are essentially assuming that, since $f''(x) = e^x$,

$$0 = S''(0) \approx f''(0) = e^1 = 1 \quad \text{and} \quad 0 = S''(3) \approx f''(3) = e^3 \approx 20.$$

The next illustration uses a spine to approximate a curve that has no given functional representation.

Illustration Figure 3.11 shows a ruddy duck in flight. To approximate the top profile of the duck, we have chosen points along the curve through which we want the approximating curve to pass. Table 3.18 lists the coordinates of 21 data points relative to the superimposed coordinate system shown in Figure 3.12. Notice that more points are used when the curve is changing rapidly than when it is changing more slowly.

Figure 3.11

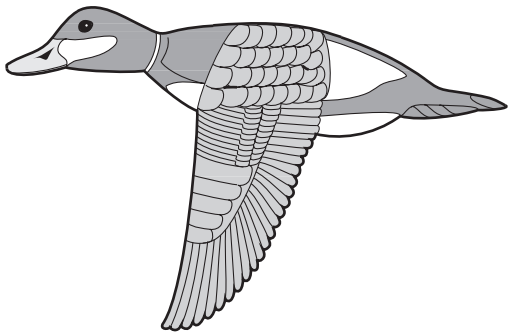
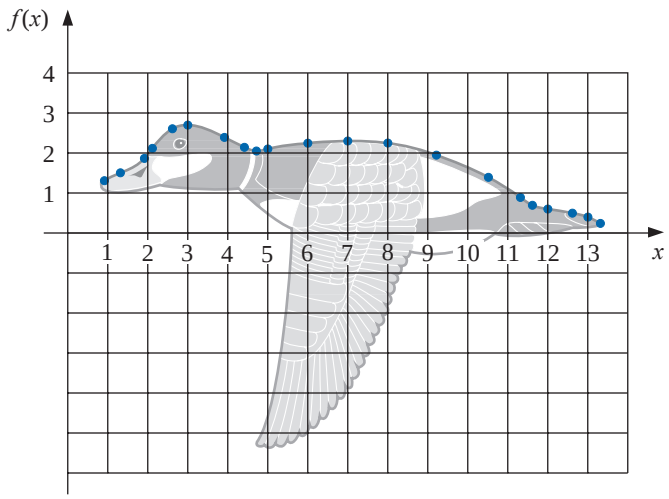


Table 3.18

x	0.9	1.3	1.9	2.1	2.6	3.0	3.9	4.4	4.7	5.0	6.0	7.0	8.0	9.2	10.5	11.3	11.6	12.0	12.6	13.0	13.3
$f(x)$	1.3	1.5	1.85	2.1	2.6	2.7	2.4	2.15	2.05	2.1	2.25	2.3	2.25	1.95	1.4	0.9	0.7	0.6	0.5	0.4	0.25

Figure 3.12

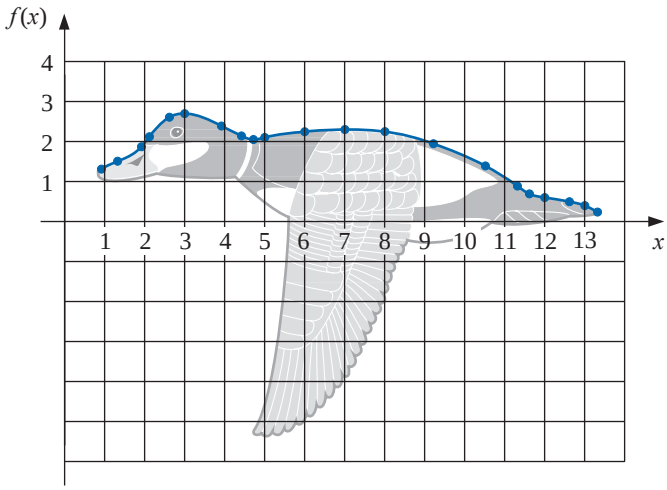


Using Algorithm 3.4 to generate the natural cubic spline for this data produces the coefficients shown in Table 3.19. This spline curve is nearly identical to the profile, as shown in Figure 3.13.

Table 3.19

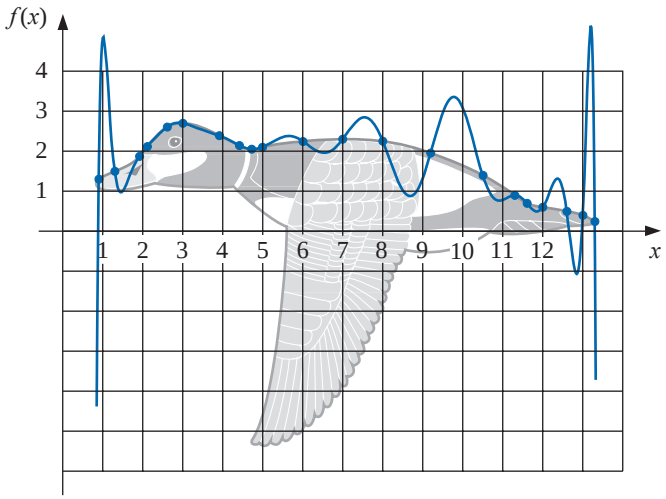
j	x_j	a_j	b_j	c_j	d_j
0	0.9	1.3	5.40	0.00	−0.25
1	1.3	1.5	0.42	−0.30	0.95
2	1.9	1.85	1.09	1.41	−2.96
3	2.1	2.1	1.29	−0.37	−0.45
4	2.6	2.6	0.59	−1.04	0.45
5	3.0	2.7	−0.02	−0.50	0.17
6	3.9	2.4	−0.50	−0.03	0.08
7	4.4	2.15	−0.48	0.08	1.31
8	4.7	2.05	−0.07	1.27	−1.58
9	5.0	2.1	0.26	−0.16	0.04
10	6.0	2.25	0.08	−0.03	0.00
11	7.0	2.3	0.01	−0.04	−0.02
12	8.0	2.25	−0.14	−0.11	0.02
13	9.2	1.95	−0.34	−0.05	−0.01
14	10.5	1.4	−0.53	−0.10	−0.02
15	11.3	0.9	−0.73	−0.15	1.21
16	11.6	0.7	−0.49	0.94	−0.84
17	12.0	0.6	−0.14	−0.06	0.04
18	12.6	0.5	−0.18	0.00	−0.45
19	13.0	0.4	−0.39	−0.54	0.60
20	13.3	0.25			

Figure 3.13



For comparison purposes, Figure 3.14 gives an illustration of the curve that is generated using a Lagrange interpolating polynomial to fit the data given in Table 3.18. The interpolating polynomial in this case is of degree 20 and oscillates wildly. It produces a very strange illustration of the back of a duck, in flight or otherwise.

Figure 3.14



To use a clamped spline to approximate this curve we would need derivative approximations for the endpoints. Even if these approximations were available, we could expect little improvement because of the close agreement of the natural cubic spline to the curve of the top profile. □

Constructing a cubic spline to approximate the lower profile of the ruddy duck would be more difficult since the curve for this portion cannot be expressed as a function of x, and at certain points the curve does not appear to be smooth. These problems can be resolved by using separate splines to represent various portions of the curve, but a more effective approach to approximating curves of this type is considered in the next section.

The clamped boundary conditions are generally preferred when approximating functions by cubic splines, so the derivative of the function must be known or approximated at the endpoints of the interval. When the nodes are equally spaced near both endpoints, approximations can be obtained by any of the appropriate formulas given in Sections 4.1 and 4.2. When the nodes are unequally spaced, the problem is considerably more difficult.

To conclude this section, we list an error-bound formula for the cubic spline with clamped boundary conditions. The proof of this result can be found in [Schul], pp. 57–58.

Theorem 3.13 Let $f \in C^4[a,b]$ with $\max_{a\le x\le b} |f^{(4)}(x)| = M$. If S is the unique clamped cubic spline interpolant to f with respect to the nodes $a = x_0 < x_1 < \cdots < x_n = b$, then for all x in $[a,b]$,

$$|f(x) - S(x)| \le \frac{5M}{384} \max_{0\le j\le n-1} (x_{j+1} - x_j)^4.$$ ■

A fourth-order error-bound result also holds in the case of natural boundary conditions, but it is more difficult to express. (See [BD], pp. 827–835.)

The natural boundary conditions will generally give less accurate results than the clamped conditions near the ends of the interval $[x_0, x_n]$ unless the function f happens

to nearly satisfy $f''(x_0) = f''(x_n) = 0$. An alternative to the natural boundary condition that does not require knowledge of the derivative of f is the *not-a-knot* condition, (see [Deb2], pp. 55–56). This condition requires that $S'''(x)$ be continuous at x_1 and at x_{n-1}.

EXERCISE SET 3.5

1. Determine the natural cubic spline S that interpolates the data $f(0) = 0$, $f(1) = 1$, and $f(2) = 2$.

2. Determine the clamped cubic spline s that interpolates the data $f(0) = 0$, $f(1) = 1$, $f(2) = 2$ and satisfies $s'(0) = s'(2) = 1$.

3. Construct the natural cubic spline for the following data.

a.

x	$f(x)$
8.3	17.56492
8.6	18.50515

b.

x	$f(x)$
0.8	0.22363362
1.0	0.65809197

c.

x	$f(x)$
−0.5	−0.0247500
−0.25	0.3349375
0	1.1010000

d.

x	$f(x)$
0.1	−0.62049958
0.2	−0.28398668
0.3	0.00660095
0.4	0.24842440

4. Construct the natural cubic spline for the following data.

a.

x	$f(x)$
0	1.00000
0.5	2.71828

b.

x	$f(x)$
−0.25	1.33203
0.25	0.800781

c.

x	$f(x)$
0.1	−0.29004996
0.2	−0.56079734
0.3	−0.81401972

d.

x	$f(x)$
−1	0.86199480
−0.5	0.95802009
0	1.0986123
0.5	1.2943767

5. The data in Exercise 3 were generated using the following functions. Use the cubic splines constructed in Exercise 3 for the given value of x to approximate $f(x)$ and $f'(x)$, and calculate the actual error.

a. $f(x) = x \ln x$; approximate $f(8.4)$ and $f'(8.4)$.

b. $f(x) = \sin(e^x - 2)$; approximate $f(0.9)$ and $f'(0.9)$.

c. $f(x) = x^3 + 4.001x^2 + 4.002x + 1.101$; approximate $f(-\frac{1}{3})$ and $f'(-\frac{1}{3})$.

d. $f(x) = x \cos x - 2x^2 + 3x - 1$; approximate $f(0.25)$ and $f'(0.25)$.

6. The data in Exercise 4 were generated using the following functions. Use the cubic splines constructed in Exercise 4 for the given value of x to approximate $f(x)$ and $f'(x)$, and calculate the actual error.

a. $f(x) = e^{2x}$; approximate $f(0.43)$ and $f'(0.43)$.

b. $f(x) = x^4 - x^3 + x^2 - x + 1$; approximate $f(0)$ and $f'(0)$.

c. $f(x) = x^2 \cos x - 3x$; approximate $f(0.18)$ and $f'(0.18)$.

d. $f(x) = \ln(e^x + 2)$; approximate $f(0.25)$ and $f'(0.25)$.

7. Construct the clamped cubic spline using the data of Exercise 3 and the fact that

a. $f'(8.3) = 3.116256$ and $f'(8.6) = 3.151762$

b. $f'(0.8) = 2.1691753$ and $f'(1.0) = 2.0466965$

c. $f'(-0.5) = 0.7510000$ and $f'(0) = 4.0020000$

d. $f'(0.1) = 3.58502082$ and $f'(0.4) = 2.16529366$

8. Construct the clamped cubic spline using the data of Exercise 4 and the fact that

a. $f'(0) = 2$ and $f'(0.5) = 5.43656$

b. $f'(-0.25) = 0.437500$ and $f'(0.25) = -0.625000$

c. $f'(0.1) = -2.8004996$ and $f'(0) = -2.9734038$

d. $f'(-1) = 0.15536240$ and $f'(0.5) = 0.45186276$

9. Repeat Exercise 5 using the clamped cubic splines constructed in Exercise 7.

10. Repeat Exercise 6 using the clamped cubic splines constructed in Exercise 8.

11. A natural cubic spline S on $[0, 2]$ is defined by

$$S(x) = \begin{cases} S_0(x) = 1 + 2x - x^3, & \text{if } 0 \le x < 1, \\ S_1(x) = 2 + b(x-1) + c(x-1)^2 + d(x-1)^3, & \text{if } 1 \le x \le 2. \end{cases}$$

Find b, c, and d.

12. A clamped cubic spline s for a function f is defined on $[1, 3]$ by

$$s(x) = \begin{cases} s_0(x) = 3(x-1) + 2(x-1)^2 - (x-1)^3, & \text{if } 1 \le x < 2, \\ s_1(x) = a + b(x-2) + c(x-2)^2 + d(x-2)^3, & \text{if } 2 \le x \le 3. \end{cases}$$

Given $f'(1) = f'(3)$, find a, b, c, and d.

13. A natural cubic spline S is defined by

$$S(x) = \begin{cases} S_0(x) = 1 + B(x-1) - D(x-1)^3, & \text{if } 1 \le x < 2, \\ S_1(x) = 1 + b(x-2) - \frac{3}{4}(x-2)^2 + d(x-2)^3, & \text{if } 2 \le x \le 3. \end{cases}$$

If S interpolates the data $(1, 1)$, $(2, 1)$, and $(3, 0)$, find B, D, b, and d.

14. A clamped cubic spline s for a function f is defined by

$$s(x) = \begin{cases} s_0(x) = 1 + Bx + 2x^2 - 2x^3, & \text{if } 0 \le x < 1, \\ s_1(x) = 1 + b(x-1) - 4(x-1)^2 + 7(x-1)^3, & \text{if } 1 \le x \le 2. \end{cases}$$

Find $f'(0)$ and $f'(2)$.

15. Construct a natural cubic spline to approximate $f(x) = \cos \pi x$ by using the values given by $f(x)$ at $x = 0, 0.25, 0.5, 0.75$, and 1.0. Integrate the spline over $[0, 1]$, and compare the result to $\int_0^1 \cos \pi x \, dx = 0$. Use the derivatives of the spline to approximate $f'(0.5)$ and $f''(0.5)$. Compare these approximations to the actual values.

16. Construct a natural cubic spline to approximate $f(x) = e^{-x}$ by using the values given by $f(x)$ at $x = 0$, 0.25, 0.75, and 1.0. Integrate the spline over $[0, 1]$, and compare the result to $\int_0^1 e^{-x} \, dx = 1 - 1/e$. Use the derivatives of the spline to approximate $f'(0.5)$ and $f''(0.5)$. Compare the approximations to the actual values.

17. Repeat Exercise 15, constructing instead the clamped cubic spline with $f'(0) = f'(1) = 0$.

18. Repeat Exercise 16, constructing instead the clamped cubic spline with $f'(0) = -1$, $f'(1) = -e^{-1}$.

19. Suppose that $f(x)$ is a polynomial of degree 3. Show that $f(x)$ is its own clamped cubic spline, but that it cannot be its own natural cubic spline.

20. Suppose the data $\{x_i, f(x_i))\}_{i=1}^n$ lie on a straight line. What can be said about the natural and clamped cubic splines for the function f? [*Hint:* Take a cue from the results of Exercises 1 and 2.]

21. Given the partition $x_0 = 0$, $x_1 = 0.05$, and $x_2 = 0.1$ of $[0, 0.1]$, find the piecewise linear interpolating function F for $f(x) = e^{2x}$. Approximate $\int_0^{0.1} e^{2x} \, dx$ with $\int_0^{0.1} F(x) \, dx$, and compare the results to the actual value.

22. Let $f \in C^2[a, b]$, and let the nodes $a = x_0 < x_1 < \cdots < x_n = b$ be given. Derive an error estimate similar to that in Theorem 3.13 for the piecewise linear interpolating function F. Use this estimate to derive error bounds for Exercise 21.

23. Extend Algorithms 3.4 and 3.5 to include as output the first and second derivatives of the spline at the nodes.

24. Extend Algorithms 3.4 and 3.5 to include as output the integral of the spline over the interval $[x_0, x_n]$.

25. Given the partition $x_0 = 0$, $x_1 = 0.05$, $x_2 = 0.1$ of $[0, 0.1]$ and $f(x) = e^{2x}$:

a. Find the cubic spline s with clamped boundary conditions that interpolates f.

b. Find an approximation for $\int_0^{0.1} e^{2x} \, dx$ by evaluating $\int_0^{0.1} s(x) \, dx$.

c. Use Theorem 3.13 to estimate $\max_{0\le x\le 0.1} |f(x) - s(x)|$ and

$$\left| \int_0^{0.1} f(x)\,dx - \int_0^{0.1} s(x)\,dx \right|.$$

d. Determine the cubic spline S with natural boundary conditions, and compare $S(0.02)$, $s(0.02)$, and $e^{0.04} = 1.04081077$.

26. Let f be defined on $[a, b]$, and let the nodes $a = x_0 < x_1 < x_2 = b$ be given. A quadratic spline interpolating function S consists of the quadratic polynomial

$$S_0(x) = a_0 + b_0(x - x_0) + c_0(x - x_0)^2 \quad \text{on } [x_0, x_1]$$

and the quadratic polynomial

$$S_1(x) = a_1 + b_1(x - x_1) + c_1(x - x_1)^2 \quad \text{on } [x_1, x_2],$$

such that

i. $S(x_0) = f(x_0), S(x_1) = f(x_1)$, and $S(x_2) = f(x_2)$,

ii. $S \in C^1[x_0, x_2]$.

Show that conditions (i) and (ii) lead to five equations in the six unknowns a_0, b_0, c_0, a_1, b_1, and c_1. The problem is to decide what additional condition to impose to make the solution unique. Does the condition $S \in C^2[x_0, x_2]$ lead to a meaningful solution?

27. Determine a quadratic spline s that interpolates the data $f(0) = 0$, $f(1) = 1$, $f(2) = 2$ and satisfies $s'(0) = 2$.

28. **a.** The introduction to this chapter included a table listing the population of the United States from 1950 to 2000. Use natural cubic spline interpolation to approximate the population in the years 1940, 1975, and 2020.

b. The population in 1940 was approximately 132,165,000. How accurate do you think your 1975 and 2020 figures are?

29. A car traveling along a straight road is clocked at a number of points. The data from the observations are given in the following table, where the time is in seconds, the distance is in feet, and the speed is in feet per second.

Time	0	3	5	8	13
Distance	0	225	383	623	993
Speed	75	77	80	74	72

a. Use a clamped cubic spline to predict the position of the car and its speed when $t = 10$ s.

b. Use the derivative of the spline to determine whether the car ever exceeds a 55-mi/h speed limit on the road; if so, what is the first time the car exceeds this speed?

c. What is the predicted maximum speed for the car?

30. The 2009 Kentucky Derby was won by a horse named Mine That Bird (at more than 50:1 odds) in a time of 2:02.66 (2 minutes and 2.66 seconds) for the $1\frac{1}{4}$-mile race. Times at the quarter-mile, half-mile, and mile poles were 0:22.98, 0:47.23, and 1:37.49.

a. Use these values together with the starting time to construct a natural cubic spline for Mine That Bird's race.

b. Use the spline to predict the time at the three-quarter-mile pole, and compare this to the actual time of 1:12.09.

c. Use the spline to approximate Mine That Bird's starting speed and speed at the finish line.

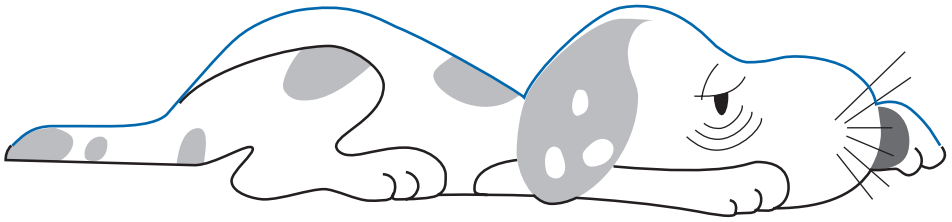
31. It is suspected that the high amounts of tannin in mature oak leaves inhibit the growth of the winter moth (*Operophtera bromata L., Geometridae*) larvae that extensively damage these trees in certain years. The following table lists the average weight of two samples of larvae at times in the first 28 days after birth. The first sample was reared on young oak leaves, whereas the second sample was reared on mature leaves from the same tree.

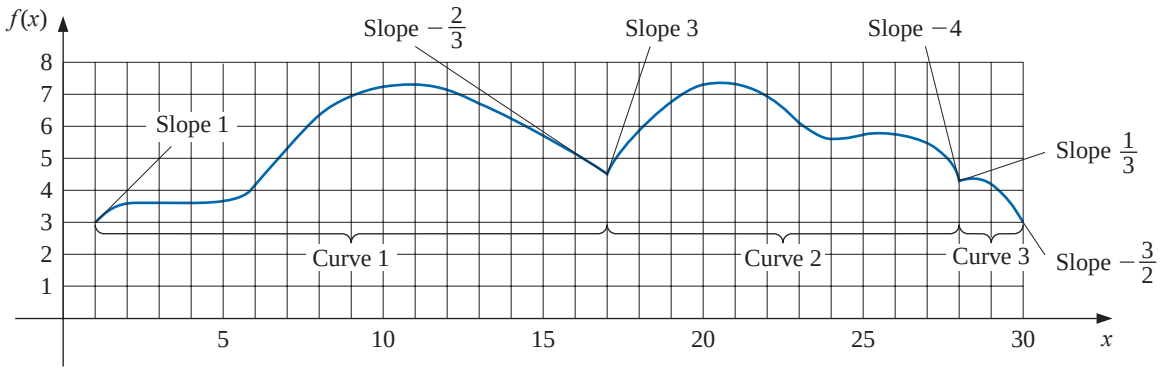
a. Use a natural cubic spline to approximate the average weight curve for each sample.

b. Find an approximate maximum average weight for each sample by determining the maximum of the spline.

Day	0	6	10	13	17	20	28
Sample 1 average weight (mg)	6.67	17.33	42.67	37.33	30.10	29.31	28.74
Sample 2 average weight (mg)	6.67	16.11	18.89	15.00	10.56	9.44	8.89

32. The upper portion of this noble beast is to be approximated using clamped cubic spline interpolants. The curve is drawn on a grid from which the table is constructed. Use Algorithm 3.5 to construct the three clamped cubic splines.





	Curve 1				Curve 2				Curve 3		
i	x_i	$f(x_i)$	$f'(x_i)$	i	x_i	$f(x_i)$	$f'(x_i)$	i	x_i	$f(x_i)$	$f'(x_i)$
0	1	3.0	1.0	0	17	4.5	3.0	0	27.7	4.1	0.33
1	2	3.7		1	20	7.0		1	28	4.3	
2	5	3.9		2	23	6.1		2	29	4.1	
3	6	4.2		3	24	5.6		3	30	3.0	−1.5
4	7	5.7		4	25	5.8					
5	8	6.6		5	27	5.2					
6	10	7.1		6	27.7	4.1	−4.0				
7	13	6.7									
8	17	4.5	−0.67								

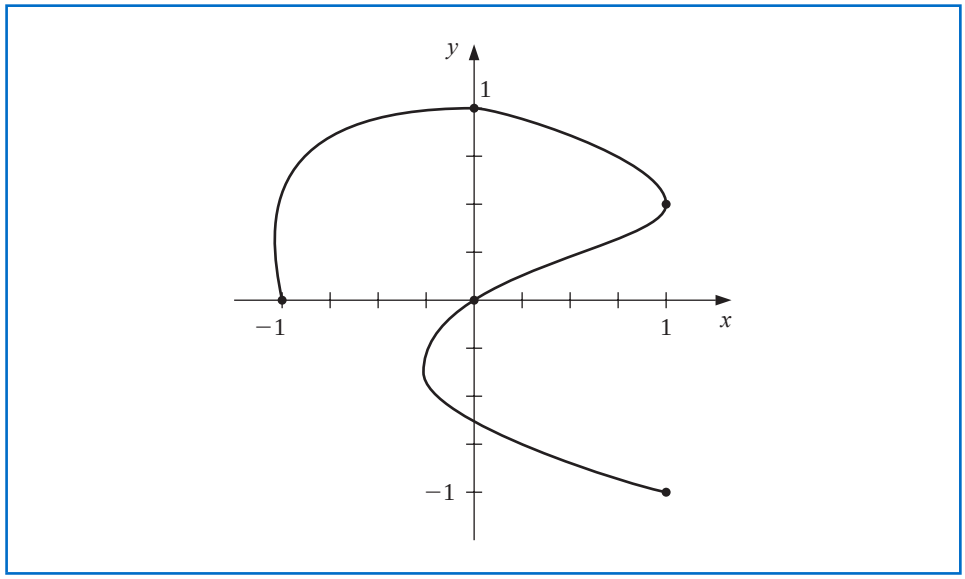
33. Repeat Exercise 32, constructing three natural splines using Algorithm 3.4.

3.6 Parametric Curves

None of the techniques developed in this chapter can be used to generate curves of the form shown in Figure 3.15 because this curve cannot be expressed as a function of one coordinate variable in terms of the other. In this section we will see how to represent general curves by using a parameter to express both the x- and y-coordinate variables. Any good book

on computer graphics will show how this technique can be extended to represent general curves and surfaces in space. (See, for example, [FVFH].)

Figure 3.15



A straightforward parametric technique for determining a polynomial or piecewise polynomial to connect the points (x_0, y_0), (x_1, y_1), ..., (x_n, y_n) in the order given is to use a parameter t on an interval $[t_0, t_n]$, with $t_0 < t_1 < \cdots < t_n$, and construct approximation functions with

$$x_i = x(t_i) \quad \text{and} \quad y_i = y(t_i), \quad \text{for each } i = 0, 1, \ldots, n.$$

The following example demonstrates the technique in the case where both approximating functions are Lagrange interpolating polynomials.

Example 1 Construct a pair of Lagrange polynomials to approximate the curve shown in Figure 3.15, using the data points shown on the curve.

Solution There is flexibility in choosing the parameter, and we will choose the points $\{t_i\}_{i=0}^{4}$ equally spaced in [0,1], which gives the data in Table 3.20.

Table 3.20

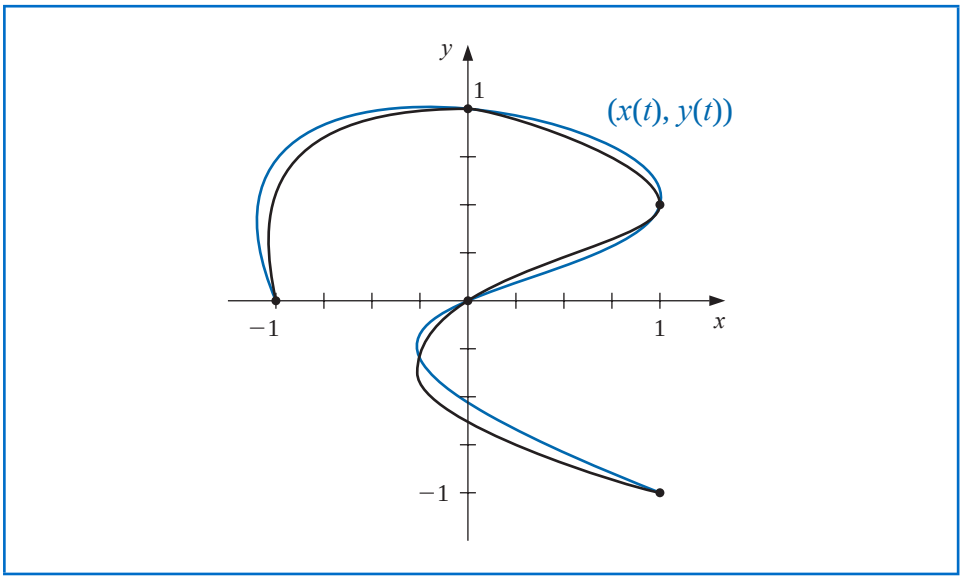
i	0	1	2	3	4
t_i	0	0.25	0.5	0.75	1
x_i	−1	0	1	0	1
y_i	0	1	0.5	0	−1

This produces the interpolating polynomials

$$x(t) = \left(\left(64t - \tfrac{352}{3}\right)t + 60\right)t - \tfrac{14}{3}\,t - 1 \quad \text{and} \quad y(t) = \left(\left(-\tfrac{64}{3}t + 48\right)t - \tfrac{116}{3}\right)t + 11\,t.$$

Plotting this parametric system produces the graph shown in blue in Figure 3.16. Although it passes through the required points and has the same basic shape, it is quite a crude approximation to the original curve. A more accurate approximation would require additional nodes, with the accompanying increase in computation. ■

Figure 3.16



Parametric Hermite and spline curves can be generated in a similar manner, but these also require extensive computational effort.

Applications in computer graphics require the rapid generation of smooth curves that can be easily and quickly modified. For both aesthetic and computational reasons, changing one portion of these curves should have little or no effect on other portions of the curves. This eliminates the use of interpolating polynomials and splines since changing one portion of these curves affects the whole curve.

A successful computer design system needs to be based on a formal mathematical theory so that the results are predictable, but this theory should be performed in the background so that the artist can base the design on aesthetics.

The choice of curve for use in computer graphics is generally a form of the piecewise cubic Hermite polynomial. Each portion of a cubic Hermite polynomial is completely determined by specifying its endpoints and the derivatives at these endpoints. As a consequence, one portion of the curve can be changed while leaving most of the curve the same. Only the adjacent portions need to be modified to ensure smoothness at the endpoints. The computations can be performed quickly, and the curve can be modified a section at a time.

The problem with Hermite interpolation is the need to specify the derivatives at the endpoints of each section of the curve. Suppose the curve has $n + 1$ data points $(x(t_0), y(t_0)), \ldots, (x(t_n), y(t_n))$, and we wish to parameterize the cubic to allow complex features. Then we must specify $x'(t_i)$ and $y'(t_i)$, for each $i = 0, 1, \ldots, n$. This is not as difficult as it would first appear, since each portion is generated independently. We must ensure only that the derivatives at the endpoints of each portion match those in the adjacent portion. Essentially, then, we can simplify the process to one of determining a pair of cubic Hermite polynomials in the parameter t, where $t_0 = 0$ and $t_1 = 1$, given the endpoint data $(x(0), y(0))$ and $(x(1), y(1))$ and the derivatives dy/dx (at $t = 0$) and dy/dx (at $t = 1$).

Notice, however, that we are specifying only six conditions, and the cubic polynomials in $x(t)$ and $y(t)$ each have four parameters, for a total of eight. This provides flexibility in choosing the pair of cubic Hermite polynomials to satisfy the conditions, because the natural form for determining $x(t)$ and $y(t)$ requires that we specify $x'(0)$, $x'(1)$, $y'(0)$, and $y'(1)$. The explicit Hermite curve in x and y requires specifying only the quotients

$$\frac{dy}{dx}(t = 0) = \frac{y'(0)}{x'(0)} \quad \text{and} \quad \frac{dy}{dx}(t = 1) = \frac{y'(1)}{x'(1)}.$$

By multiplying $x'(0)$ and $y'(0)$ by a common scaling factor, the tangent line to the curve at $(x(0), y(0))$ remains the same, but the shape of the curve varies. The larger the scaling

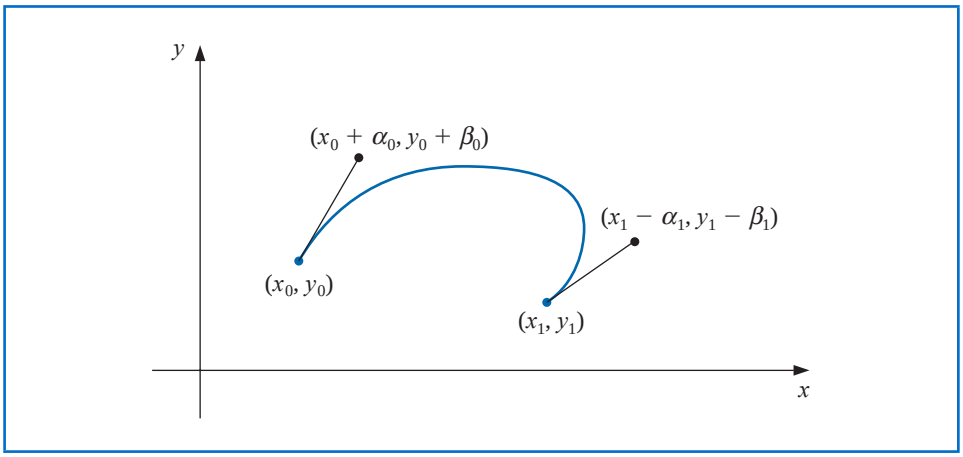
factor, the closer the curve comes to approximating the tangent line near $(x(0), y(0))$. A similar situation exists at the other endpoint $(x(1), y(1))$.

To further simplify the process in interactive computer graphics, the derivative at an endpoint is specified by using a second point, called a *guidepoint*, on the desired tangent line. The farther the guidepoint is from the node, the more closely the curve approximates the tangent line near the node.

In Figure 3.17, the nodes occur at (x_0, y_0) and (x_1, y_1), the guidepoint for (x_0, y_0) is $(x_0 + \alpha_0, y_0 + \beta_0)$, and the guidepoint for (x_1, y_1) is $(x_1 - \alpha_1, y_1 - \beta_1)$. The cubic Hermite polynomial $x(t)$ on $[0, 1]$ satisfies

$$x(0) = x_0, \quad x(1) = x_1, \quad x'(0) = \alpha_0, \quad \text{and} \quad x'(1) = \alpha_1.$$

Figure 3.17



The unique cubic polynomial satisfying these conditions is

$$x(t) = [2(x_0 - x_1) + (\alpha_0 + \alpha_1)]t^3 + [3(x_1 - x_0) - (\alpha_1 + 2\alpha_0)]t^2 + \alpha_0 t + x_0. \quad (3.23)$$

In a similar manner, the unique cubic polynomial satisfying

$$y(0) = y_0, \quad y(1) = y_1, \quad y'(0) = \beta_0, \quad \text{and} \quad y'(1) = \beta_1$$

is

$$y(t) = [2(y_0 - y_1) + (\beta_0 + \beta_1)]t^3 + [3(y_1 - y_0) - (\beta_1 + 2\beta_0)]t^2 + \beta_0 t + y_0. \quad (3.24)$$

Example 2 Determine the graph of the parametric curve generated Eq. (3.23) and (3.24) when the end points are $(x_0, y_0) = (0, 0)$ and $(x_1, y_1) = (1, 0)$, and respective guide points, as shown in Figure 3.18 are $(1, 1)$ and $(0, 1)$.

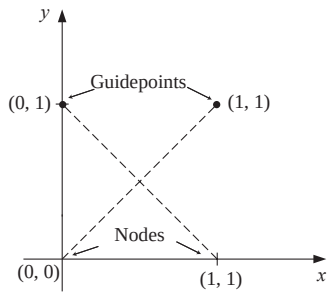


Figure 3.18

Solution The endpoint information implies that $x_0 = 0$, $x_1 = 1$, $y_0 = 0$, and $y_1 = 0$, and the guide points at $(1, 1)$ and $(0, 1)$ imply that $\alpha_0 = 1$, $\alpha_1 = 1$, $\beta_0 = 1$, and $\beta_1 = -1$. Note that the slopes of the guide lines at $(0, 0)$ and $(1, 0)$ are, respectively

$$\frac{\beta_0}{\alpha_0} = \frac{1}{1} = 1 \quad \text{and} \quad \frac{\beta_1}{\alpha_1} = \frac{-1}{1} = -1.$$

Equations (3.23) and (3.24) imply that for $t \in [0, 1]$ we have

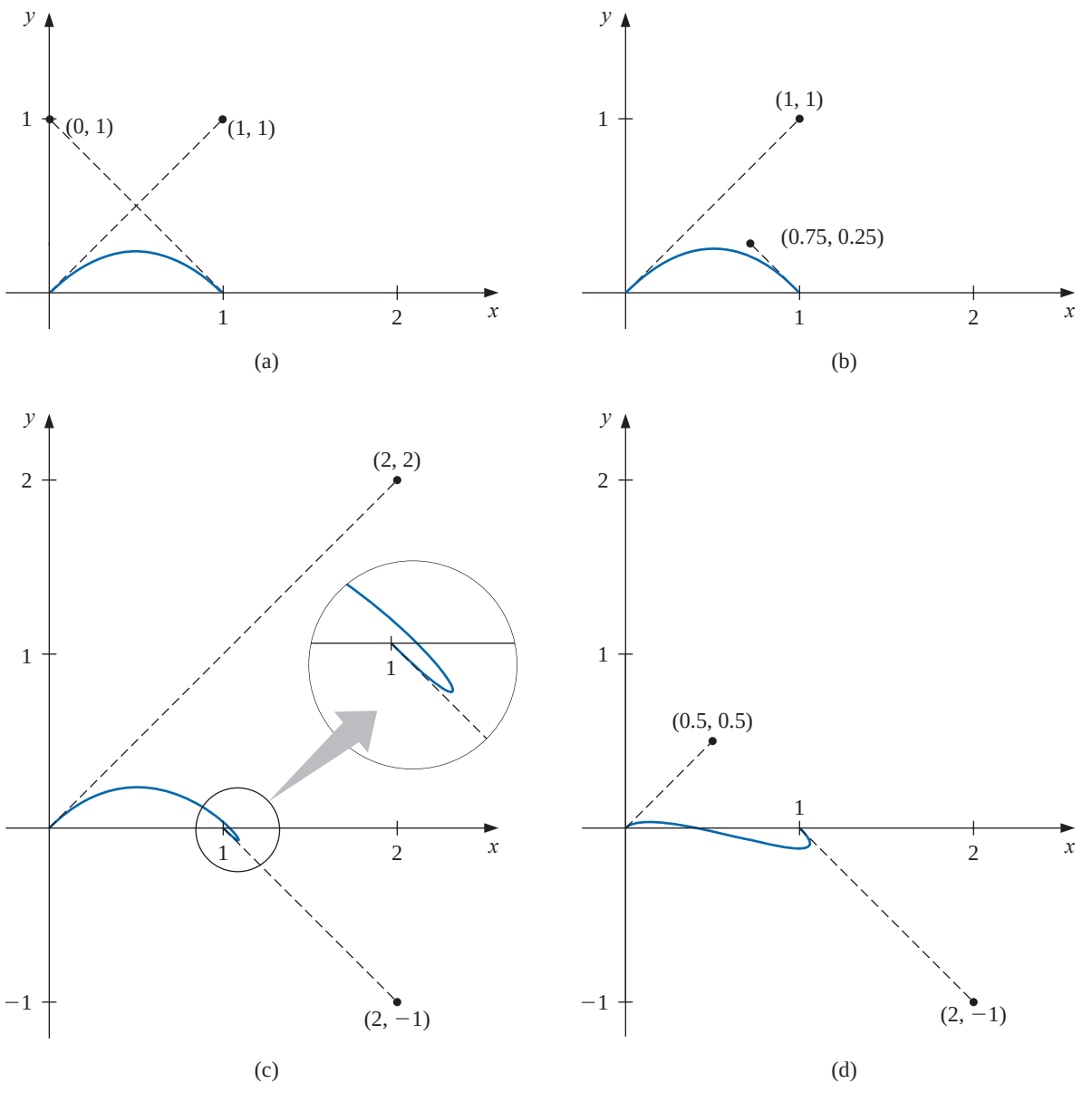
$$x(t) = [2(0 - 1) + (1 + 1)]t^3 + [3(0 - 0) - (1 + 2 \cdot 1)]t^2 + 1 \cdot t + 0 = t$$

and

$$y(t) = [2(0-0) + (1 + (-1))]t^3 + [3(0-0) - (-1 + 2 \cdot 1)]t^2 + 1 \cdot t + 0 = -t^2 + t.$$

This graph is shown as (a) in Figure 3.19, together with some other possibilities of curves produced by Eqs. (3.23) and (3.24) when the nodes are $(0, 0)$ and $(1, 0)$ and the slopes at these nodes are 1 and -1, respectively. ■

Figure 3.19



Pierre Etienne Bézier (1910–1999) was head of design and production for Renault motorcars for most of his professional life. He began his research into computer-aided design and manufacturing in 1960, developing interactive tools for curve and surface design, and initiated computer-generated milling for automobile modeling.

The Bézier curves that bear his name have the advantage of being based on a rigorous mathematical theory that does not need to be explicitly recognized by the practitioner who simply wants to make an aesthetically pleasing curve or surface. These are the curves that are the basis of the powerful Adobe Postscript system, and produce the freehand curves that are generated in most sufficiently powerful computer graphics packages.

The standard procedure for determining curves in an interactive graphics mode is to first use a mouse or touchpad to set the nodes and guidepoints to generate a first approximation to the curve. These can be set manually, but most graphics systems permit you to use your input device to draw the curve on the screen freehand and will select appropriate nodes and guidepoints for your freehand curve.

The nodes and guidepoints can then be manipulated into a position that produces an aesthetically pleasing curve. Since the computation is minimal, the curve can be determined so quickly that the resulting change is seen immediately. Moreover, all the data needed to compute the curves are imbedded in the coordinates of the nodes and guidepoints, so no analytical knowledge is required of the user.

Popular graphics programs use this type of system for their freehand graphic representations in a slightly modified form. The Hermite cubics are described as **Bézier polynomials**, which incorporate a scaling factor of 3 when computing the derivatives at the endpoints. This modifies the parametric equations to

$$x(t) = [2(x_0 - x_1) + 3(\alpha_0 + \alpha_1)]t^3 + [3(x_1 - x_0) - 3(\alpha_1 + 2\alpha_0)]t^2 + 3\alpha_0 t + x_0, \quad (3.25)$$

and

$$y(t) = [2(y_0 - y_1) + 3(\beta_0 + \beta_1)]t^3 + [3(y_1 - y_0) - 3(\beta_1 + 2\beta_0)]t^2 + 3\beta_0 t + y_0, \quad (3.26)$$

for $0 \le t \le 1$, but this change is transparent to the user of the system.

Algorithm 3.6 constructs a set of Bézier curves based on the parametric equations in Eqs. (3.25) and (3.26).

ALGORITHM
3.6

Bézier Curve

To construct the cubic Bézier curves $C_0, \ldots, C_{n-1}$ in parametric form, where C_i is represented by

$$(x_i(t), y_i(t)) = (a_0^{(i)} + a_1^{(i)}t + a_2^{(i)}t^2 + a_3^{(i)}t^3, b_0^{(i)} + b_1^{(i)}t + b_2^{(i)}t^2 + b_3^{(i)}t^3),$$

for $0 \le t \le 1$, as determined by the left endpoint (x_i, y_i), left guidepoint (x_i^+, y_i^+), right endpoint (x_{i+1}, y_{i+1}), and right guidepoint (x_{i+1}^-, y_{i+1}^-) for each $i = 0, 1, \ldots, n-1$:

INPUT n; $(x_0, y_0), \ldots, (x_n, y_n)$; $(x_0^+, y_0^+), \ldots, (x_{n-1}^+, y_{n-1}^+)$; $(x_1^-, y_1^-), \ldots, (x_n^-, y_n^-)$.

OUTPUT coefficients $\{a_0^{(i)}, a_1^{(i)}, a_2^{(i)}, a_3^{(i)}, b_0^{(i)}, b_1^{(i)}, b_2^{(i)}, b_3^{(i)}$, for $0 \le i \le n-1\}$.

Step 1 For each $i = 0, 1, \ldots, n-1$ do Steps 2 and 3.

Step 2 Set
$$a_0^{(i)} = x_i;$$
$$b_0^{(i)} = y_i;$$
$$a_1^{(i)} = 3(x_i^+ - x_i);$$
$$b_1^{(i)} = 3(y_i^+ - y_i);$$
$$a_2^{(i)} = 3(x_i + x_{i+1}^- - 2x_i^+);$$
$$b_2^{(i)} = 3(y_i + y_{i+1}^- - 2y_i^+);$$
$$a_3^{(i)} = x_{i+1} - x_i + 3x_i^+ - 3x_{i+1}^-;$$
$$b_3^{(i)} = y_{i+1} - y_i + 3y_i^+ - 3y_{i+1}^-;$$

Step 3 OUTPUT $(a_0^{(i)}, a_1^{(i)}, a_2^{(i)}, a_3^{(i)}, b_0^{(i)}, b_1^{(i)}, b_2^{(i)}, b_3^{(i)})$.

Step 4 STOP. ■

Three-dimensional curves are generated in a similar manner by additionally specifying third components z_0 and z_1 for the nodes and $z_0+\gamma_0$ and $z_1-\gamma_1$ for the guidepoints. The more difficult problem involving the representation of three-dimensional curves concerns the loss of the third dimension when the curve is projected onto a two-dimensional computer screen. Various projection techniques are used, but this topic lies within the realm of computer graphics. For an introduction to this topic and ways that the technique can be modified for surface representations, see one of the many books on computer graphics methods, such as [FVFH].

EXERCISE SET 3.6

1. Let $(x_0, y_0) = (0, 0)$ and $(x_1, y_1) = (5, 2)$ be the endpoints of a curve. Use the given guidepoints to construct parametric cubic Hermite approximations $(x(t), y(t))$ to the curve, and graph the approximations.

a. $(1, 1)$ and $(6, 1)$

b. $(0.5, 0.5)$ and $(5.5, 1.5)$

c. $(1, 1)$ and $(6, 3)$

d. $(2, 2)$ and $(7, 0)$

2. Repeat Exercise 1 using cubic Bézier polynomials.

3. Construct and graph the cubic Bézier polynomials given the following points and guidepoints.

a. Point $(1, 1)$ with guidepoint $(1.5, 1.25)$ to point $(6, 2)$ with guidepoint $(7, 3)$

b. Point $(1, 1)$ with guidepoint $(1.25, 1.5)$ to point $(6, 2)$ with guidepoint $(5, 3)$

c. Point $(0, 0)$ with guidepoint $(0.5, 0.5)$ to point $(4, 6)$ with entering guidepoint $(3.5, 7)$ and exiting guidepoint $(4.5, 5)$ to point $(6, 1)$ with guidepoint $(7, 2)$

d. Point $(0, 0)$ with guidepoint $(0.5, 0.5)$ to point $(2, 1)$ with entering guidepoint $(3, 1)$ and exiting guidepoint $(3, 1)$ to point $(4, 0)$ with entering guidepoint $(5, 1)$ and exiting guidepoint $(3, -1)$ to point $(6, -1)$ with guidepoint $(6.5, -0.25)$

4. Use the data in the following table and Algorithm 3.6 to approximate the shape of the letter $\mathcal{N}$.

i	x_i	y_i	α_i	β_i	α'_i	β'_i
0	3	6	3.3	6.5		
1	2	2	2.8	3.0	2.5	2.5
2	6	6	5.8	5.0	5.0	5.8
3	5	2	5.5	2.2	4.5	2.5
4	6.5	3			6.4	2.8

5. Suppose a cubic Bézier polynomial is placed through (u_0, v_0) and (u_3, v_3) with guidepoints (u_1, v_1) and (u_2, v_2), respectively.

a. Derive the parametric equations for $u(t)$ and $v(t)$ assuming that

$$u(0) = u_0, \quad u(1) = u_3, \quad u'(0) = u_1 - u_0, \quad u'(1) = u_3 - u_2$$

and

$$v(0) = v_0, \quad v(1) = v_3, \quad v'(0) = v_1 - v_0, \quad v'(1) = v_3 - v_2.$$

b. Let $f(i/3) = u_i$, for $i = 0, 1, 2, 3$ and $g(i/3) = v_i$, for $i = 0, 1, 2, 3$. Show that the Bernstein polynomial of degree 3 in t for f is $u(t)$ and the Bernstein polynomial of degree three in t for g is $v(t)$. (See Exercise 23 of Section 3.1.)

3.7 Survey of Methods and Software

In this chapter we have considered approximating a function using polynomials and piecewise polynomials. The function can be specified by a given defining equation or by providing points in the plane through which the graph of the function passes. A set of nodes $x_0, x_1, \ldots, x_n$ is given in each case, and more information, such as the value of various derivatives, may also be required. We need to find an approximating function that satisfies the conditions specified by these data.

The interpolating polynomial $P(x)$ is the polynomial of least degree that satisfies, for a function f,

$$P(x_i) = f(x_i), \quad \text{for each } i = 0, 1, \ldots, n.$$

Although this interpolating polynomial is unique, it can take many different forms. The Lagrange form is most often used for interpolating tables when n is small and for deriving formulas for approximating derivatives and integrals. Neville's method is used for evaluating several interpolating polynomials at the same value of x. Newton's forms of the polynomial are more appropriate for computation and are also used extensively for deriving formulas for solving differential equations. However, polynomial interpolation has the inherent weaknesses of oscillation, particularly if the number of nodes is large. In this case there are other methods that can be better applied.

The Hermite polynomials interpolate a function and its derivative at the nodes. They can be very accurate but require more information about the function being approximated. When there are a large number of nodes, the Hermite polynomials also exhibit oscillation weaknesses.

The most commonly used form of interpolation is piecewise-polynomial interpolation. If function and derivative values are available, piecewise cubic Hermite interpolation is recommended. This is the preferred method for interpolating values of a function that is the solution to a differential equation. When only the function values are available, natural cubic spline interpolation can be used. This spline forces the second derivative of the spline to be zero at the endpoints. Other cubic splines require additional data. For example, the clamped cubic spline needs values of the derivative of the function at the endpoints of the interval.

Other methods of interpolation are commonly used. Trigonometric interpolation, in particular the Fast Fourier Transform discussed in Chapter 8, is used with large amounts of data when the function is assumed to have a periodic nature. Interpolation by rational functions is also used.

If the data are suspected to be inaccurate, smoothing techniques can be applied, and some form of least squares fit of data is recommended. Polynomials, trigonometric functions, rational functions, and splines can be used in least squares fitting of data. We consider these topics in Chapter 8.

Interpolation routines included in the IMSL Library are based on the book *A Practical Guide to Splines* by Carl de Boor [Deb] and use interpolation by cubic splines. There are cubic splines to minimize oscillations and to preserve concavity. Methods for two-dimensional interpolation by bicubic splines are also included.

The NAG library contains subroutines for polynomial and Hermite interpolation, for cubic spline interpolation, and for piecewise cubic Hermite interpolation. NAG also contains subroutines for interpolating functions of two variables.

The netlib library contains the subroutines to compute the cubic spline with various endpoint conditions. One package produces the Newton's divided difference coefficients for

a discrete set of data points, and there are various routines for evaluating Hermite piecewise polynomials.

MATLAB can be used to interpolate a discrete set of data points, using either nearest neighbor interpolation, linear interpolation, cubic spline interpolation, or cubic interpolation. Cubic splines can also be produced.

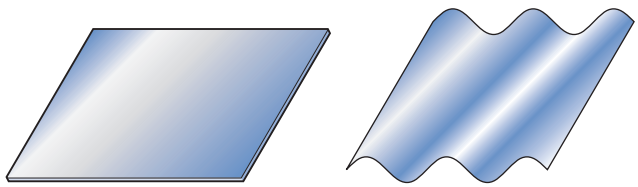
General references to the methods in this chapter are the books by Powell [Pow] and by Davis [Da]. The seminal paper on splines is due to Schoenberg [Scho]. Important books on splines are by Schultz [Schul], De Boor [Deb2], Dierckx [Di], and Schumaker [Schum].

CHAPTER

4 Numerical Differentiation and Integration

Introduction

A sheet of corrugated roofing is constructed by pressing a flat sheet of aluminum into one whose cross section has the form of a sine wave.



A corrugated sheet 4 ft long is needed, the height of each wave is 1 in. from the center line, and each wave has a period of approximately 2π in. The problem of finding the length of the initial flat sheet is one of determining the length of the curve given by $f(x) = \sin x$ from $x = 0$ in. to $x = 48$ in. From calculus we know that this length is

$$L = \int_0^{48} \sqrt{1 + (f'(x))^2}\, dx = \int_0^{48} \sqrt{1 + (\cos x)^2}\, dx,$$

so the problem reduces to evaluating this integral. Although the sine function is one of the most common mathematical functions, the calculation of its length involves an elliptic integral of the second kind, which cannot be evaluated explicitly. Methods are developed in this chapter to approximate the solution to problems of this type. This particular problem is considered in Exercise 25 of Section 4.4 and Exercise 12 of Section 4.5.

We mentioned in the introduction to Chapter 3 that one reason for using algebraic polynomials to approximate an arbitrary set of data is that, given any continuous function defined on a closed interval, there exists a polynomial that is arbitrarily close to the function at every point in the interval. Also, the derivatives and integrals of polynomials are easily obtained and evaluated. It should not be surprising, then, that many procedures for approximating derivatives and integrals use the polynomials that approximate the function.

4.1 Numerical Differentiation

The derivative of the function f at x_0 is

$$f'(x_0) = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h}.$$

This formula gives an obvious way to generate an approximation to $f'(x_0)$; simply compute

$$\frac{f(x_0 + h) - f(x_0)}{h}$$

for small values of h. Although this may be obvious, it is not very successful, due to our old nemesis round-off error. But it is certainly a place to start.

To approximate $f'(x_0)$, suppose first that $x_0 \in (a, b)$, where $f \in C^2[a, b]$, and that $x_1 = x_0 + h$ for some $h \neq 0$ that is sufficiently small to ensure that $x_1 \in [a, b]$. We construct the first Lagrange polynomial $P_{0,1}(x)$ for f determined by x_0 and x_1, with its error term:

$$\begin{aligned} f(x) &= P_{0,1}(x) + \frac{(x - x_0)(x - x_1)}{2!} f''(\xi(x)) \\ &= \frac{f(x_0)(x - x_0 - h)}{-h} + \frac{f(x_0 + h)(x - x_0)}{h} + \frac{(x - x_0)(x - x_0 - h)}{2} f''(\xi(x)), \end{aligned}$$

for some $\xi(x)$ between x_0 and x_1. Differentiating gives

$$\begin{aligned} f'(x) &= \frac{f(x_0 + h) - f(x_0)}{h} + D_x \left[\frac{(x - x_0)(x - x_0 - h)}{2} f''(\xi(x)) \right] \\ &= \frac{f(x_0 + h) - f(x_0)}{h} + \frac{2(x - x_0) - h}{2} f''(\xi(x)) \\ &\quad + \frac{(x - x_0)(x - x_0 - h)}{2} D_x(f''(\xi(x))). \end{aligned}$$

Deleting the terms involving $\xi(x)$ gives

$$f'(x) \approx \frac{f(x_0 + h) - f(x_0)}{h}.$$

Difference equations were used and popularized by Isaac Newton in the last quarter of the 17th century, but many of these techniques had previously been developed by Thomas Harriot (1561–1621) and Henry Briggs (1561–1630). Harriot made significant advances in navigation techniques, and Briggs was the person most responsible for the acceptance of logarithms as an aid to computation.

One difficulty with this formula is that we have no information about $D_x f''(\xi(x))$, so the truncation error cannot be estimated. When x is x_0, however, the coefficient of $D_x f''(\xi(x))$ is 0, and the formula simplifies to

$$f'(x_0) = \frac{f(x_0 + h) - f(x_0)}{h} - \frac{h}{2} f''(\xi). \tag{4.1}$$

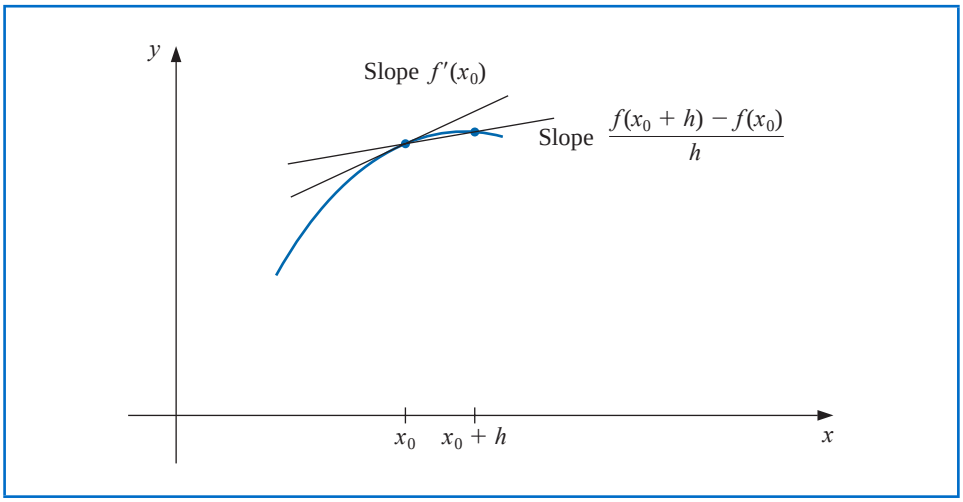
For small values of h, the difference quotient $[f(x_0 + h) - f(x_0)]/h$ can be used to approximate $f'(x_0)$ with an error bounded by $M|h|/2$, where M is a bound on $|f''(x)|$ for x between x_0 and $x_0 + h$. This formula is known as the **forward-difference formula** if $h > 0$ (see Figure 4.1) and the **backward-difference formula** if $h < 0$.

Example 1 Use the forward-difference formula to approximate the derivative of $f(x) = \ln x$ at $x_0 = 1.8$ using $h = 0.1$, $h = 0.05$, and $h = 0.01$, and determine bounds for the approximation errors.

Solution The forward-difference formula

$$\frac{f(1.8 + h) - f(1.8)}{h}$$

Figure 4.1



with $h = 0.1$ gives

$$\frac{\ln 1.9 - \ln 1.8}{0.1} = \frac{0.64185389 - 0.58778667}{0.1} = 0.5406722.$$

Because $f''(x) = -1/x^2$ and $1.8 < \xi < 1.9$, a bound for this approximation error is

$$\frac{|hf''(\xi)|}{2} = \frac{|h|}{2\xi^2} < \frac{0.1}{2(1.8)^2} = 0.0154321.$$

The approximation and error bounds when $h = 0.05$ and $h = 0.01$ are found in a similar manner and the results are shown in Table 4.1.

Table 4.1

h	$f(1.8 + h)$	$\frac{f(1.8 + h) - f(1.8)}{h}$	$\frac{\lvert h\rvert}{2(1.8)^2}$
0.1	0.64185389	0.5406722	0.0154321
0.05	0.61518564	0.5479795	0.0077160
0.01	0.59332685	0.5540180	0.0015432

Since $f'(x) = 1/x$, the exact value of $f'(1.8)$ is $0.55\overline{5}$, and in this case the error bounds are quite close to the true approximation error. ■

To obtain general derivative approximation formulas, suppose that $\{x_0, x_1, \ldots, x_n\}$ are $(n + 1)$ distinct numbers in some interval I and that $f \in C^{n+1}(I)$. From Theorem 3.3 on page 112,

$$f(x) = \sum_{k=0}^{n} f(x_k)L_k(x) + \frac{(x - x_0)\cdots(x - x_n)}{(n + 1)!} f^{(n+1)}(\xi(x)),$$

for some $\xi(x)$ in I, where $L_k(x)$ denotes the kth Lagrange coefficient polynomial for f at $x_0, x_1, \ldots, x_n$. Differentiating this expression gives

$$f'(x) = \sum_{k=0}^{n} f(x_k)L_k'(x) + D_x\left[\frac{(x-x_0)\cdots(x-x_n)}{(n+1!)}\right] f^{(n+1)}(\xi(x)) + \frac{(x-x_0)\cdots(x-x_n)}{(n+1)!} D_x[f^{(n+1)}(\xi(x))].$$

We again have a problem estimating the truncation error unless x is one of the numbers x_j. In this case, the term multiplying $D_x[f^{(n+1)}(\xi(x))]$ is 0, and the formula becomes

$$f'(x_j) = \sum_{k=0}^{n} f(x_k)L_k'(x_j) + \frac{f^{(n+1)}(\xi(x_j))}{(n+1)!} \prod_{\substack{k=0 \\ k\neq j}}^{n} (x_j - x_k), \tag{4.2}$$

which is called an **$(n+1)$-point formula** to approximate $f'(x_j)$.

In general, using more evaluation points in Eq. (4.2) produces greater accuracy, although the number of functional evaluations and growth of round-off error discourages this somewhat. The most common formulas are those involving three and five evaluation points.

We first derive some useful three-point formulas and consider aspects of their errors. Because

$$L_0(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}, \quad \text{we have} \quad L_0'(x) = \frac{2x-x_1-x_2}{(x_0-x_1)(x_0-x_2)}.$$

Similarly,

$$L_1'(x) = \frac{2x-x_0-x_2}{(x_1-x_0)(x_1-x_2)} \quad \text{and} \quad L_2'(x) = \frac{2x-x_0-x_1}{(x_2-x_0)(x_2-x_1)}.$$

Hence, from Eq. (4.2),

$$f'(x_j) = f(x_0)\left[\frac{2x_j-x_1-x_2}{(x_0-x_1)(x_0-x_2)}\right] + f(x_1)\left[\frac{2x_j-x_0-x_2}{(x_1-x_0)(x_1-x_2)}\right] + f(x_2)\left[\frac{2x_j-x_0-x_1}{(x_2-x_0)(x_2-x_1)}\right] + \frac{1}{6}f^{(3)}(\xi_j)\prod_{\substack{k=0 \\ k\neq j}}^{2}(x_j-x_k), \tag{4.3}$$

for each $j = 0, 1, 2$, where the notation ξ_j indicates that this point depends on x_j.

Three-Point Formulas

The formulas from Eq. (4.3) become especially useful if the nodes are equally spaced, that is, when

$$x_1 = x_0 + h \quad \text{and} \quad x_2 = x_0 + 2h, \quad \text{for some } h \neq 0.$$

We will assume equally-spaced nodes throughout the remainder of this section.

Using Eq. (4.3) with $x_j = x_0, x_1 = x_0 + h$, and $x_2 = x_0 + 2h$ gives

$$f'(x_0) = \frac{1}{h}\left[-\frac{3}{2}f(x_0) + 2f(x_1) - \frac{1}{2}f(x_2)\right] + \frac{h^2}{3}f^{(3)}(\xi_0).$$

Doing the same for $x_j = x_1$ gives

$$f'(x_1) = \frac{1}{h}\left[-\frac{1}{2}f(x_0) + \frac{1}{2}f(x_2)\right] - \frac{h^2}{6}f^{(3)}(\xi_1),$$

and for $x_j = x_2$,

$$f'(x_2) = \frac{1}{h}\left[\frac{1}{2}f(x_0) - 2f(x_1) + \frac{3}{2}f(x_2)\right] + \frac{h^2}{3}f^{(3)}(\xi_2).$$

Since $x_1 = x_0 + h$ and $x_2 = x_0 + 2h$, these formulas can also be expressed as

$$f'(x_0) = \frac{1}{h}\left[-\frac{3}{2}f(x_0) + 2f(x_0 + h) - \frac{1}{2}f(x_0 + 2h)\right] + \frac{h^2}{3}f^{(3)}(\xi_0),$$

$$f'(x_0 + h) = \frac{1}{h}\left[-\frac{1}{2}f(x_0) + \frac{1}{2}f(x_0 + 2h)\right] - \frac{h^2}{6}f^{(3)}(\xi_1),$$

and

$$f'(x_0 + 2h) = \frac{1}{h}\left[\frac{1}{2}f(x_0) - 2f(x_0 + h) + \frac{3}{2}f(x_0 + 2h)\right] + \frac{h^2}{3}f^{(3)}(\xi_2).$$

As a matter of convenience, the variable substitution x_0 for $x_0 + h$ is used in the middle equation to change this formula to an approximation for $f'(x_0)$. A similar change, x_0 for $x_0 + 2h$, is used in the last equation. This gives three formulas for approximating $f'(x_0)$:

$$f'(x_0) = \frac{1}{2h}[-3f(x_0) + 4f(x_0 + h) - f(x_0 + 2h)] + \frac{h^2}{3}f^{(3)}(\xi_0),$$

$$f'(x_0) = \frac{1}{2h}[-f(x_0 - h) + f(x_0 + h)] - \frac{h^2}{6}f^{(3)}(\xi_1),$$

and

$$f'(x_0) = \frac{1}{2h}[f(x_0 - 2h) - 4f(x_0 - h) + 3f(x_0)] + \frac{h^2}{3}f^{(3)}(\xi_2).$$

Finally, note that the last of these equations can be obtained from the first by simply replacing h with $-h$, so there are actually only two formulas:

Three-Point Endpoint Formula

- $$f'(x_0) = \frac{1}{2h}[-3f(x_0) + 4f(x_0 + h) - f(x_0 + 2h)] + \frac{h^2}{3}f^{(3)}(\xi_0), \tag{4.4}$$

where ξ_0 lies between x_0 and $x_0 + 2h$.

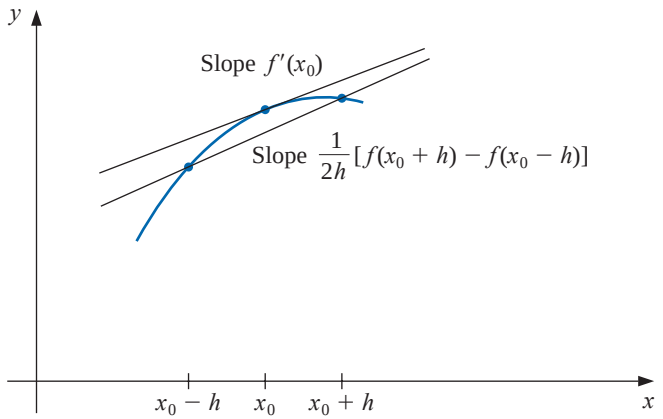
Three-Point Midpoint Formula

- $$f'(x_0) = \frac{1}{2h}[f(x_0 + h) - f(x_0 - h)] - \frac{h^2}{6}f^{(3)}(\xi_1), \tag{4.5}$$

where ξ_1 lies between $x_0 - h$ and $x_0 + h$.

Although the errors in both Eq. (4.4) and Eq. (4.5) are $O(h^2)$, the error in Eq. (4.5) is approximately half the error in Eq. (4.4). This is because Eq. (4.5) uses data on both sides of x_0 and Eq. (4.4) uses data on only one side. Note also that f needs to be evaluated at only two points in Eq. (4.5), whereas in Eq. (4.4) three evaluations are needed. Figure 4.2 on page 178 gives an illustration of the approximation produced from Eq. (4.5). The approximation in Eq. (4.4) is useful near the ends of an interval, because information about f outside the interval may not be available.

Figure 4.2



Five-Point Formulas

The methods presented in Eqs. (4.4) and (4.5) are called **three-point formulas** (even though the third point $f(x_0)$ does not appear in Eq. (4.5)). Similarly, there are **five-point formulas** that involve evaluating the function at two additional points. The error term for these formulas is $O(h^4)$. One common five-point formula is used to determine approximations for the derivative at the midpoint.

Five-Point Midpoint Formula

- $$f'(x_0) = \frac{1}{12h}[f(x_0 - 2h) - 8f(x_0 - h) + 8f(x_0 + h) - f(x_0 + 2h)] + \frac{h^4}{30}f^{(5)}(\xi), \tag{4.6}$$

  where ξ lies between $x_0 - 2h$ and $x_0 + 2h$.

The derivation of this formula is considered in Section 4.2. The other five-point formula is used for approximations at the endpoints.

Five-Point Endpoint Formula

- $$f'(x_0) = \frac{1}{12h}[-25f(x_0) + 48f(x_0 + h) - 36f(x_0 + 2h) + 16f(x_0 + 3h) - 3f(x_0 + 4h)] + \frac{h^4}{5}f^{(5)}(\xi), \tag{4.7}$$

  where ξ lies between x_0 and $x_0 + 4h$.

Left-endpoint approximations are found using this formula with $h > 0$ and right-endpoint approximations with $h < 0$. The five-point endpoint formula is particularly useful for the clamped cubic spline interpolation of Section 3.5.

Example 2 Values for $f(x) = xe^x$ are given in Table 4.2. Use all the applicable three-point and five-point formulas to approximate $f'(2.0)$.

Table 4.2

x	$f(x)$
1.8	10.889365
1.9	12.703199
2.0	14.778112
2.1	17.148957
2.2	19.855030

Solution The data in the table permit us to find four different three-point approximations. We can use the endpoint formula (4.4) with $h = 0.1$ or with $h = -0.1$, and we can use the midpoint formula (4.5) with $h = 0.1$ or with $h = 0.2$.

Using the endpoint formula (4.4) with $h = 0.1$ gives

$$\frac{1}{0.2}[-3f(2.0) + 4f(2.1) - f(2.2] = 5[-3(14.778112) + 4(17.148957) - 19.855030)] = 22.032310,$$

and with $h = -0.1$ gives 22.054525.

Using the midpoint formula (4.5) with $h = 0.1$ gives

$$\frac{1}{0.2}[f(2.1) - f(1.9)] = 5(17.148957 - 12.7703199) = 22.228790,$$

and with $h = 0.2$ gives 22.414163.

The only five-point formula for which the table gives sufficient data is the midpoint formula (4.6) with $h = 0.1$. This gives

$$\frac{1}{1.2}[f(1.8) - 8f(1.9) + 8f(2.1) - f(2.2)] = \frac{1}{1.2}[10.889365 - 8(12.703199) + 8(17.148957) - 19.855030] = 22.166999$$

If we had no other information we would accept the five-point midpoint approximation using $h = 0.1$ as the most accurate, and expect the true value to be between that approximation and the three-point mid-point approximation that is in the interval [22.166, 22.229].

The true value in this case is $f'(2.0) = (2 + 1)e^2 = 22.167168$, so the approximation errors are actually:

Three-point endpoint with $h = 0.1$: 1.35×10^{-1};

Three-point endpoint with $h = -0.1$: 1.13×10^{-1};

Three-point midpoint with $h = 0.1$: -6.16×10^{-2};

Three-point midpoint with $h = 0.2$: -2.47×10^{-1};

Five-point midpoint with $h = 0.1$: 1.69×10^{-4}. ■

Methods can also be derived to find approximations to higher derivatives of a function using only tabulated values of the function at various points. The derivation is algebraically tedious, however, so only a representative procedure will be presented.

Expand a function f in a third Taylor polynomial about a point x_0 and evaluate at $x_0 + h$ and $x_0 - h$. Then

$$f(x_0 + h) = f(x_0) + f'(x_0)h + \frac{1}{2}f''(x_0)h^2 + \frac{1}{6}f'''(x_0)h^3 + \frac{1}{24}f^{(4)}(\xi_1)h^4$$

and

$$f(x_0 - h) = f(x_0) - f'(x_0)h + \frac{1}{2}f''(x_0)h^2 - \frac{1}{6}f'''(x_0)h^3 + \frac{1}{24}f^{(4)}(\xi_{-1})h^4,$$

where $x_0 - h < \xi_{-1} < x_0 < \xi_1 < x_0 + h$.

If we add these equations, the terms involving $f'(x_0)$ and $-f'(x_0)$ cancel, so

$$f(x_0 + h) + f(x_0 - h) = 2f(x_0) + f''(x_0)h^2 + \frac{1}{24}[f^{(4)}(\xi_1) + f^{(4)}(\xi_{-1})]h^4.$$

Solving this equation for $f''(x_0)$ gives

$$f''(x_0) = \frac{1}{h^2}[f(x_0 - h) - 2f(x_0) + f(x_0 + h)] - \frac{h^2}{24}[f^{(4)}(\xi_1) + f^{(4)}(\xi_{-1})]. \qquad (4.8)$$

Suppose $f^{(4)}$ is continuous on $[x_0 - h, x_0 + h]$. Since $\frac{1}{2}[f^{(4)}(\xi_1) + f^{(4)}(\xi_{-1})]$ is between $f^{(4)}(\xi_1)$ and $f^{(4)}(\xi_{-1})$, the Intermediate Value Theorem implies that a number ξ exists between ξ_1 and ξ_{-1}, and hence in $(x_0 - h, x_0 + h)$, with

$$f^{(4)}(\xi) = \frac{1}{2}\left[f^{(4)}(\xi_1) + f^{(4)}(\xi_{-1})\right].$$

This permits us to rewrite Eq. (4.8) in its final form.

Second Derivative Midpoint Formula

- $$f''(x_0) = \frac{1}{h^2}[f(x_0 - h) - 2f(x_0) + f(x_0 + h)] - \frac{h^2}{12}f^{(4)}(\xi), \qquad (4.9)$$

  for some ξ, where $x_0 - h < \xi < x_0 + h$.

If $f^{(4)}$ is continuous on $[x_0 - h, x_0 + h]$ it is also bounded, and the approximation is $O(h^2)$.

Example 3 In Example 2 we used the data shown in Table 4.3 to approximate the first derivative of $f(x) = xe^x$ at $x = 2.0$. Use the second derivative formula (4.9) to approximate $f''(2.0)$.

Table 4.3

x	$f(x)$
1.8	10.889365
1.9	12.703199
2.0	14.778112
2.1	17.148957
2.2	19.855030

Solution The data permits us to determine two approximations for $f''(2.0)$. Using (4.9) with $h = 0.1$ gives

$$\frac{1}{0.01}[f(1.9) - 2f(2.0) + f(2.1)] = 100[12.703199 - 2(14.778112) + 17.148957]$$
$$= 29.593200,$$

and using (4.9) with $h = 0.2$ gives

$$\frac{1}{0.04}[f(1.8) - 2f(2.0) + f(2.2)] = 25[10.889365 - 2(14.778112) + 19.855030]$$
$$= 29.704275.$$

Because $f''(x) = (x + 2)e^x$, the exact value is $f''(2.0) = 29.556224$. Hence the actual errors are -3.70×10^{-2} and -1.48×10^{-1}, respectively. ■

Round-Off Error Instability

It is particularly important to pay attention to round-off error when approximating derivatives. To illustrate the situation, let us examine the three-point midpoint formula Eq. (4.5),

$$f'(x_0) = \frac{1}{2h}[f(x_0 + h) - f(x_0 - h)] - \frac{h^2}{6}f^{(3)}(\xi_1),$$

more closely. Suppose that in evaluating $f(x_0 + h)$ and $f(x_0 - h)$ we encounter round-off errors $e(x_0 + h)$ and $e(x_0 - h)$. Then our computations actually use the values $\tilde{f}(x_0 + h)$ and $\tilde{f}(x_0 - h)$, which are related to the true values $f(x_0 + h)$ and $f(x_0 - h)$ by

$$f(x_0 + h) = \tilde{f}(x_0 + h) + e(x_0 + h) \quad \text{and} \quad f(x_0 - h) = \tilde{f}(x_0 - h) + e(x_0 - h).$$

The total error in the approximation,

$$f'(x_0) - \frac{\tilde{f}(x_0 + h) - \tilde{f}(x_0 - h)}{2h} = \frac{e(x_0 + h) - e(x_0 - h)}{2h} - \frac{h^2}{6} f^{(3)}(\xi_1),$$

is due both to round-off error, the first part, and to truncation error. If we assume that the round-off errors $e(x_0 \pm h)$ are bounded by some number $\varepsilon > 0$ and that the third derivative of f is bounded by a number $M > 0$, then

$$\left| f'(x_0) - \frac{\tilde{f}(x_0 + h) - \tilde{f}(x_0 - h)}{2h} \right| \le \frac{\varepsilon}{h} + \frac{h^2}{6} M.$$

To reduce the truncation error, $h^2M/6$, we need to reduce h. But as h is reduced, the round-off error ε/h grows. In practice, then, it is seldom advantageous to let h be too small, because in that case the round-off error will dominate the calculations.

Illustration Consider using the values in Table 4.4 to approximate $f'(0.900)$, where $f(x) = \sin x$. The true value is $\cos 0.900 = 0.62161$. The formula

$$f'(0.900) \approx \frac{f(0.900 + h) - f(0.900 - h)}{2h},$$

with different values of h, gives the approximations in Table 4.5.

Table 4.4

x	$\sin x$	x	$\sin x$
0.800	0.71736	0.901	0.78395
0.850	0.75128	0.902	0.78457
0.880	0.77074	0.905	0.78643
0.890	0.77707	0.910	0.78950
0.895	0.78021	0.920	0.79560
0.898	0.78208	0.950	0.81342
0.899	0.78270	1.000	0.84147

Table 4.5

h	Approximation to $f'(0.900)$	Error
0.001	0.62500	0.00339
0.002	0.62250	0.00089
0.005	0.62200	0.00039
0.010	0.62150	−0.00011
0.020	0.62150	−0.00011
0.050	0.62140	−0.00021
0.100	0.62055	−0.00106

The optimal choice for h appears to lie between 0.005 and 0.05. We can use calculus to verify (see Exercise 29) that a minimum for

$$e(h) = \frac{\varepsilon}{h} + \frac{h^2}{6} M,$$

occurs at $h = \sqrt[3]{3\varepsilon/M}$, where

$$M = \max_{x \in [0.800, 1.00]} |f'''(x)| = \max_{x \in [0.800, 1.00]} |\cos x| = \cos 0.8 \approx 0.69671.$$

Because values of f are given to five decimal places, we will assume that the round-off error is bounded by $\varepsilon = 5 \times 10^{-6}$. Therefore, the optimal choice of h is approximately

$$h = \sqrt[3]{\frac{3(0.000005)}{0.69671}} \approx 0.028,$$

which is consistent with the results in Table 4.6. □

In practice, we cannot compute an optimal h to use in approximating the derivative, since we have no knowledge of the third derivative of the function. But we must remain aware that reducing the step size will not always improve the approximation. □

We have considered only the round-off error problems that are presented by the three-point formula Eq. (4.5), but similar difficulties occur with all the differentiation formulas. The reason can be traced to the need to divide by a power of h. As we found in Section 1.2 (see, in particular, Example 3), division by small numbers tends to exaggerate round-off error, and this operation should be avoided if possible. In the case of numerical differentiation, we cannot avoid the problem entirely, although the higher-order methods reduce the difficulty.

Keep in mind that difference method approximations might be unstable.

As approximation methods, numerical differentiation is *unstable*, since the small values of h needed to reduce truncation error also cause the round-off error to grow. This is the first class of unstable methods we have encountered, and these techniques would be avoided if it were possible. However, in addition to being used for computational purposes, the formulas are needed for approximating the solutions of ordinary and partial-differential equations.

EXERCISE SET 4.1

1. Use the forward-difference formulas and backward-difference formulas to determine each missing entry in the following tables.

a.

x	$f(x)$	$f'(x)$
0.5	0.4794	
0.6	0.5646	
0.7	0.6442	

b.

x	$f(x)$	$f'(x)$
0.0	0.00000	
0.2	0.74140	
0.4	1.3718	

2. Use the forward-difference formulas and backward-difference formulas to determine each missing entry in the following tables.

a.

x	$f(x)$	$f'(x)$
−0.3	1.9507	
−0.2	2.0421	
−0.1	2.0601	

b.

x	$f(x)$	$f'(x)$
1.0	1.0000	
1.2	1.2625	
1.4	1.6595	

3. The data in Exercise 1 were taken from the following functions. Compute the actual errors in Exercise 1, and find error bounds using the error formulas.

a. $f(x) = \sin x$ **b.** $f(x) = e^x - 2x^2 + 3x - 1$

4. The data in Exercise 2 were taken from the following functions. Compute the actual errors in Exercise 2, and find error bounds using the error formulas.

a. $f(x) = 2\cos 2x - x$ **b.** $f(x) = x^2 \ln x + 1$

5. Use the most accurate three-point formula to determine each missing entry in the following tables.

a.

x	$f(x)$	$f'(x)$
1.1	9.025013	
1.2	11.02318	
1.3	13.46374	
1.4	16.44465	

b.

x	$f(x)$	$f'(x)$
8.1	16.94410	
8.3	17.56492	
8.5	18.19056	
8.7	18.82091	

c.

x	$f(x)$	$f'(x)$
2.9	−4.827866	
3.0	−4.240058	
3.1	−3.496909	
3.2	−2.596792	

d.

x	$f(x)$	$f'(x)$
2.0	3.6887983	
2.1	3.6905701	
2.2	3.6688192	
2.3	3.6245909	

6. Use the most accurate three-point formula to determine each missing entry in the following tables.

a.

x	$f(x)$	$f'(x)$
−0.3	−0.27652	
−0.2	−0.25074	
−0.1	−0.16134	
0	0	

b.

x	$f(x)$	$f'(x)$
7.4	−68.3193	
7.6	−71.6982	
7.8	−75.1576	
8.0	−78.6974	

c.

x	$f(x)$	$f'(x)$
1.1	1.52918	
1.2	1.64024	
1.3	1.70470	
1.4	1.71277	

d.

x	$f(x)$	$f'(x)$
−2.7	0.054797	
−2.5	0.11342	
−2.3	0.65536	
−2.1	0.98472	

7. The data in Exercise 5 were taken from the following functions. Compute the actual errors in Exercise 5, and find error bounds using the error formulas.

a. $f(x) = e^{2x}$ **b.** $f(x) = x \ln x$

c. $f(x) = x \cos x - x^2 \sin x$ **d.** $f(x) = 2(\ln x)^2 + 3 \sin x$

8. The data in Exercise 6 were taken from the following functions. Compute the actual errors in Exercise 6, and find error bounds using the error formulas.

a. $f(x) = e^{2x} - \cos 2x$ **b.** $f(x) = \ln(x+2) - (x+1)^2$

c. $f(x) = x \sin x + x^2 \cos x$ **d.** $f(x) = (\cos 3x)^2 - e^{2x}$

9. Use the formulas given in this section to determine, as accurately as possible, approximations for each missing entry in the following tables.

a.

x	$f(x)$	$f'(x)$
2.1	−1.709847	
2.2	−1.373823	
2.3	−1.119214	
2.4	−0.9160143	
2.5	−0.7470223	
2.6	−0.6015966	

b.

x	$f(x)$	$f'(x)$
−3.0	9.367879	
−2.8	8.233241	
−2.6	7.180350	
−2.4	6.209329	
−2.2	5.320305	
−2.0	4.513417	

10. Use the formulas given in this section to determine, as accurately as possible, approximations for each missing entry in the following tables.

a.

x	$f(x)$	$f'(x)$
1.05	−1.709847	
1.10	−1.373823	
1.15	−1.119214	
1.20	−0.9160143	
1.25	−0.7470223	
1.30	−0.6015966	

b.

x	$f(x)$	$f'(x)$
−3.0	16.08554	
−2.8	12.64465	
−2.6	9.863738	
−2.4	7.623176	
−2.2	5.825013	
−2.0	4.389056	

11. The data in Exercise 9 were taken from the following functions. Compute the actual errors in Exercise 9, and find error bounds using the error formulas and Maple.

a. $f(x) = \tan x$ **b.** $f(x) = e^{x/3} + x^2$

12. The data in Exercise 10 were taken from the following functions. Compute the actual errors in Exercise 10, and find error bounds using the error formulas and Maple.

a. $f(x) = \tan 2x$ **b.** $f(x) = e^{-x} - 1 + x$

13. Use the following data and the knowledge that the first five derivatives of f are bounded on $[1, 5]$ by 2, 3, 6, 12 and 23, respectively, to approximate $f'(3)$ as accurately as possible. Find a bound for the error.

x	1	2	3	4	5
$f(x)$	2.4142	2.6734	2.8974	3.0976	3.2804

14. Repeat Exercise 13, assuming instead that the third derivative of f is bounded on $[1, 5]$ by 4.

15. Repeat Exercise 1 using four-digit rounding arithmetic, and compare the errors to those in Exercise 3.

16. Repeat Exercise 5 using four-digit chopping arithmetic, and compare the errors to those in Exercise 7.

17. Repeat Exercise 9 using four-digit rounding arithmetic, and compare the errors to those in Exercise 11.

18. Consider the following table of data:

x	0.2	0.4	0.6	0.8	1.0
$f(x)$	0.9798652	0.9177710	0.808038	0.6386093	0.3843735

a. Use all the appropriate formulas given in this section to approximate $f'(0.4)$ and $f''(0.4)$.

b. Use all the appropriate formulas given in this section to approximate $f'(0.6)$ and $f''(0.6)$.

19. Let $f(x) = \cos \pi x$. Use Eq. (4.9) and the values of $f(x)$ at $x = 0.25$, 0.5, and 0.75 to approximate $f''(0.5)$. Compare this result to the exact value and to the approximation found in Exercise 15 of Section 3.5. Explain why this method is particularly accurate for this problem, and find a bound for the error.

20. Let $f(x) = 3xe^x - \cos x$. Use the following data and Eq. (4.9) to approximate $f''(1.3)$ with $h = 0.1$ and with $h = 0.01$.

x	1.20	1.29	1.30	1.31	1.40
$f(x)$	11.59006	13.78176	14.04276	14.30741	16.86187

Compare your results to $f''(1.3)$.

21. Consider the following table of data:

x	0.2	0.4	0.6	0.8	1.0
$f(x)$	0.9798652	0.9177710	0.8080348	0.6386093	0.3843735

a. Use Eq. (4.7) to approximate $f'(0.2)$.

b. Use Eq. (4.7) to approximate $f'(1.0)$.

c. Use Eq. (4.6) to approximate $f'(0.6)$.

22. Derive an $O(h^4)$ five-point formula to approximate $f'(x_0)$ that uses $f(x_0 - h)$, $f(x_0)$, $f(x_0 + h)$, $f(x_0 + 2h)$, and $f(x_0 + 3h)$. [*Hint:* Consider the expression $Af(x_0 - h) + Bf(x_0 + h) + Cf(x_0 + 2h) + Df(x_0 + 3h)$. Expand in fourth Taylor polynomials, and choose A, B, C, and D appropriately.]

23. Use the formula derived in Exercise 22 and the data of Exercise 21 to approximate $f'(0.4)$ and $f'(0.8)$.

24. **a.** Analyze the round-off errors, as in Example 4, for the formula

$$f'(x_0) = \frac{f(x_0 + h) - f(x_0)}{h} - \frac{h}{2} f''(\xi_0).$$

b. Find an optimal $h > 0$ for the function given in Example 2.

25. In Exercise 10 of Section 3.4 data were given describing a car traveling on a straight road. That problem asked to predict the position and speed of the car when $t = 10$ s. Use the following times and positions to predict the speed at each time listed.

Time	0	3	5	8	10	13
Distance	0	225	383	623	742	993

26. In a circuit with impressed voltage $\mathcal{E}(t)$ and inductance L, Kirchhoff's first law gives the relationship

$$\mathcal{E}(t) = L\frac{di}{dt} + Ri,$$

where R is the resistance in the circuit and i is the current. Suppose we measure the current for several values of t and obtain:

t	1.00	1.01	1.02	1.03	1.0
i	3.10	3.12	3.14	3.18	3.24

where t is measured in seconds, i is in amperes, the inductance L is a constant 0.98 henries, and the resistance is 0.142 ohms. Approximate the voltage $\mathcal{E}(t)$ when $t = 1.00, 1.01, 1.02, 1.03$, and 1.04.

27. All calculus students know that the derivative of a function f at x can be defined as

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}.$$

Choose your favorite function f, nonzero number x, and computer or calculator. Generate approximations $f_n'(x)$ to $f'(x)$ by

$$f_n'(x) = \frac{f(x + 10^{-n}) - f(x)}{10^{-n}},$$

for $n = 1, 2, \ldots, 20$, and describe what happens.

28. Derive a method for approximating $f''''(x_0)$ whose error term is of order h^2 by expanding the function f in a fourth Taylor polynomial about x_0 and evaluating at $x_0 \pm h$ and $x_0 \pm 2h$.

29. Consider the function

$$e(h) = \frac{\varepsilon}{h} + \frac{h^2}{6}M,$$

where M is a bound for the third derivative of a function. Show that $e(h)$ has a minimum at $\sqrt[3]{3\varepsilon/M}$.

4.2 Richardson's Extrapolation

Lewis Fry Richardson (1881–1953) was the first person to systematically apply mathematics to weather prediction while working in England for the Meteorological Office. As a conscientious objector during World War I, he wrote extensively about the economic futility of warfare, using systems of differential equations to model rational interactions between countries. The extrapolation technique that bears his name was the rediscovery of a technique with roots that are at least as old as Christiaan Hugyens (1629–1695), and possibly Archimedes (287–212 B.C.E.).

Richardson's extrapolation is used to generate high-accuracy results while using low-order formulas. Although the name attached to the method refers to a paper written by L. F. Richardson and J. A. Gaunt [RG] in 1927, the idea behind the technique is much older. An interesting article regarding the history and application of extrapolation can be found in [Joy].

Extrapolation can be applied whenever it is known that an approximation technique has an error term with a predictable form, one that depends on a parameter, usually the step size h. Suppose that for each number $h \neq 0$ we have a formula $N_1(h)$ that approximates an unknown constant M, and that the truncation error involved with the approximation has the form

$$M - N_1(h) = K_1 h + K_2 h^2 + K_3 h^3 + \cdots,$$

for some collection of (unknown) constants $K_1, K_2, K_3, \ldots$.

The truncation error is $O(h)$, so unless there was a large variation in magnitude among the constants $K_1, K_2, K_3, \ldots$,

$$M - N_1(0.1) \approx 0.1K_1, \quad M - N_1(0.01) \approx 0.01K_1,$$

and, in general, $M - N_1(h) \approx K_1 h$.

The object of extrapolation is to find an easy way to combine these rather inaccurate $O(h)$ approximations in an appropriate way to produce formulas with a higher-order truncation error.

Suppose, for example, we can combine the $N_1(h)$ formulas to produce an $O(h^2)$ approximation formula, $N_2(h)$, for M with

$$M - N_2(h) = \hat{K}_2 h^2 + \hat{K}_3 h^3 + \cdots,$$

for some, again unknown, collection of constants $\hat{K}_2, \hat{K}_3, \ldots$. Then we would have

$$M - N_2(0.1) \approx 0.01\hat{K}_2, \quad M - N_2(0.01) \approx 0.0001\hat{K}_2,$$

and so on. If the constants K_1 and $\hat{K}_2$ are roughly of the same magnitude, then the $N_2(h)$ approximations would be much better than the corresponding $N_1(h)$ approximations. The extrapolation continues by combining the $N_2(h)$ approximations in a manner that produces formulas with $O(h^3)$ truncation error, and so on.

To see specifically how we can generate the extrapolation formulas, consider the $O(h)$ formula for approximating M

$$M = N_1(h) + K_1 h + K_2 h^2 + K_3 h^3 + \cdots. \tag{4.10}$$

The formula is assumed to hold for all positive h, so we replace the parameter h by half its value. Then we have a second $O(h)$ approximation formula

$$M = N_1\left(\frac{h}{2}\right) + K_1\frac{h}{2} + K_2\frac{h^2}{4} + K_3\frac{h^3}{8} + \cdots. \tag{4.11}$$

Subtracting Eq. (4.10) from twice Eq. (4.11) eliminates the term involving K_1 and gives

$$M = N_1\left(\frac{h}{2}\right) + \left[N_1\left(\frac{h}{2}\right) - N_1(h)\right] + K_2\left(\frac{h^2}{2} - h^2\right) + K_3\left(\frac{h^3}{4} - h^3\right) + \cdots. \tag{4.12}$$

Define

$$N_2(h) = N_1\left(\frac{h}{2}\right) + \left[N_1\left(\frac{h}{2}\right) - N_1(h)\right].$$

Then Eq. (4.12) is an $O(h^2)$ approximation formula for M:

$$M = N_2(h) - \frac{K_2}{2}h^2 - \frac{3K_3}{4}h^3 - \cdots. \tag{4.13}$$

Example 1 In Example 1 of Section 4.1 we use the forward-difference method with $h = 0.1$ and $h = 0.05$ to find approximations to $f'(1.8)$ for $f(x) = \ln(x)$. Assume that this formula has truncation error $O(h)$ and use extrapolation on these values to see if this results in a better approximation.

Solution In Example 1 of Section 4.1 we found that

$$\text{with } h = 0.1\text{: } f'(1.8) \approx 0.5406722, \quad \text{and} \quad \text{with } h = 0.05\text{: } f'(1.8) \approx 0.5479795.$$

This implies that

$$N_1(0.1) = 0.5406722 \quad \text{and} \quad N_1(0.05) = 0.5479795.$$

Extrapolating these results gives the new approximation

$$N_2(0.1) = N_1(0.05) + (N_1(0.05) - N_1(0.1)) = 0.5479795 + (0.5479795 - 0.5406722)$$
$$= 0.555287.$$

The $h = 0.1$ and $h = 0.05$ results were found to be accurate to within 1.5×10^{-2} and 7.7×10^{-3}, respectively. Because $f'(1.8) = 1/1.8 = 0.\overline{5}$, the extrapolated value is accurate to within 2.7×10^{-4}. ■

Extrapolation can be applied whenever the truncation error for a formula has the form

$$\sum_{j=1}^{m-1} K_j h^{\alpha_j} + O(h^{\alpha_m}),$$

for a collection of constants K_j and when $\alpha_1 < \alpha_2 < \alpha_3 < \cdots < \alpha_m$. Many formulas used for extrapolation have truncation errors that contain only even powers of h, that is, have the form

$$M = N_1(h) + K_1 h^2 + K_2 h^4 + K_3 h^6 + \cdots . \tag{4.14}$$

The extrapolation is much more effective than when all powers of h are present because the averaging process produces results with errors $O(h^2)$, $O(h^4)$, $O(h^6)$, ..., with essentially no increase in computation, over the results with errors, $O(h)$, $O(h^2)$, $O(h^3)$, ...

Assume that approximation has the form of Eq. (4.14). Replacing h with $h/2$ gives the $O(h^2)$ approximation formula

$$M = N_1\left(\frac{h}{2}\right) + K_1\frac{h^2}{4} + K_2\frac{h^4}{16} + K_3\frac{h^6}{64} + \cdots .$$

Subtracting Eq. (4.14) from 4 times this equation eliminates the h^2 term,

$$3M = \left[4N_1\left(\frac{h}{2}\right) - N_1(h)\right] + K_2\left(\frac{h^4}{4} - h^4\right) + K_3\left(\frac{h^6}{16} - h^6\right) + \cdots .$$

Dividing this equation by 3 produces an $O(h^4)$ formula

$$M = \frac{1}{3}\left[4N_1\left(\frac{h}{2}\right) - N_1(h)\right] + \frac{K_2}{3}\left(\frac{h^4}{4} - h^4\right) + \frac{K_3}{3}\left(\frac{h^6}{16} - h^6\right) + \cdots .$$

Defining

$$N_2(h) = \frac{1}{3}\left[4N_1\left(\frac{h}{2}\right) - N_1(h)\right] = N_1\left(\frac{h}{2}\right) + \frac{1}{3}\left[N_1\left(\frac{h}{2}\right) - N_1(h)\right],$$

produces the approximation formula with truncation error $O(h^4)$:

$$M = N_2(h) - K_2\frac{h^4}{4} - K_3\frac{5h^6}{16} + \cdots . \tag{4.15}$$

Now replace h in Eq. (4.15) with $h/2$ to produce a second $O(h^4)$ formula

$$M = N_2\left(\frac{h}{2}\right) - K_2\frac{h^4}{64} - K_3\frac{5h^6}{1024} - \cdots .$$

Subtracting Eq. (4.15) from 16 times this equation eliminates the h^4 term and gives

$$15M = \left[16N_2\left(\frac{h}{2}\right) - N_2(h)\right] + K_3\frac{15h^6}{64} + \cdots .$$

Dividing this equation by 15 produces the new $O(h^6)$ formula

$$M = \frac{1}{15}\left[16N_2\left(\frac{h}{2}\right) - N_2(h)\right] + K_3\frac{h^6}{64} + \cdots .$$

We now have the $O(h^6)$ approximation formula

$$N_3(h) = \frac{1}{15}\left[16N_2\left(\frac{h}{2}\right) - N_2(h)\right] = N_2\left(\frac{h}{2}\right) + \frac{1}{15}\left[N_2\left(\frac{h}{2}\right) - N_2(h)\right].$$

Continuing this procedure gives, for each $j = 2, 3, \ldots$, the $O(h^{2j})$ approximation

$$N_j(h) = N_{j-1}\left(\frac{h}{2}\right) + \frac{N_{j-1}(h/2) - N_{j-1}(h)}{4^{j-1} - 1}.$$

Table 4.6 shows the order in which the approximations are generated when

$$M = N_1(h) + K_1h^2 + K_2h^4 + K_3h^6 + \cdots. \tag{4.16}$$

It is conservatively assumed that the true result is accurate at least to within the agreement of the bottom two results in the diagonal, in this case, to within $|N_3(h) - N_4(h)|$.

Table 4.6

$O(h^2)$	$O(h^4)$	$O(h^6)$	$O(h^8)$
1: $N_1(h)$			
2: $N_1(\frac{h}{2})$	**3:** $N_2(h)$		
4: $N_1(\frac{h}{4})$	**5:** $N_2(\frac{h}{2})$	**6:** $N_3(h)$	
7: $N_1(\frac{h}{8})$	**8:** $N_2(\frac{h}{4})$	**9:** $N_3(\frac{h}{2})$	**10:** $N_4(h)$

Example 2 Taylor's theorem can be used to show that centered-difference formula in Eq. (4.5) to approximate $f'(x_0)$ can be expressed with an error formula:

$$f'(x_0) = \frac{1}{2h}[f(x_0 + h) - f(x_0 - h)] - \frac{h^2}{6}f'''(x_0) - \frac{h^4}{120}f^{(5)}(x_0) - \cdots.$$

Find approximations of order $O(h^2)$, $O(h^4)$, and $O(h^6)$ for $f'(2.0)$ when $f(x) = xe^x$ and $h = 0.2$.

Solution The constants $K_1 = -f'''(x_0)/6$, $K_2 = -f^{(5)}(x_0)/120, \cdots$, are not likely to be known, but this is not important. We only need to know that these constants exist in order to apply extrapolation.

We have the $O(h^2)$ approximation

$$f'(x_0) = N_1(h) - \frac{h^2}{6}f'''(x_0) - \frac{h^4}{120}f^{(5)}(x_0) - \cdots, \tag{4.17}$$

where

$$N_1(h) = \frac{1}{2h}[f(x_0 + h) - f(x_0 - h)].$$

This gives us the first $O(h^2)$ approximations

$$N_1(0.2) = \frac{1}{0.4}[f(2.2) - f(1.8)] = 2.5(19.855030 - 10.889365) = 22.414160,$$

and

$$N_1(0.1) = \frac{1}{0.2}[f(2.1) - f(1.9)] = 5(17.148957 - 12.703199) = 22.228786.$$

Combining these to produce the first $O(h^4)$ approximation gives

$$\begin{aligned} N_2(0.2) &= N_1(0.1) + \frac{1}{3}(N_1(0.1) - N_1(0.2)) \\ &= 22.228786 + \frac{1}{3}(22.228786 - 22.414160) = 22.166995. \end{aligned}$$

To determine an $O(h^6)$ formula we need another $O(h^4)$ result, which requires us to find the third $O(h^2)$ approximation

$$N_1(0.05) = \frac{1}{0.1}[f(2.05) - f(1.95)] = 10(15.924197 - 13.705941) = 22.182564.$$

We can now find the $O(h^4)$ approximation

$$\begin{aligned} N_2(0.1) &= N_1(0.05) + \frac{1}{3}(N_1(0.05) - N_1(0.1)) \\ &= 22.182564 + \frac{1}{3}(22.182564 - 22.228786) = 22.167157. \end{aligned}$$

and finally the $O(h^6)$ approximation

$$\begin{aligned} N_3(0.2) &= N_2(0.1) + \frac{1}{15}(N_2(0.1) - N_1(0.2)) \\ &= 22.167157 + \frac{1}{15}(22.167157 - 22.166995) = 22.167168. \end{aligned}$$

We would expect the final approximation to be accurate to at least the value 22.167 because the $N_2(0.2)$ and $N_3(0.2)$ give this same value. In fact, $N_3(0.2)$ is accurate to all the listed digits. ■

Each column beyond the first in the extrapolation table is obtained by a simple averaging process, so the technique can produce high-order approximations with minimal computational cost. However, as k increases, the round-off error in $N_1(h/2^k)$ will generally increase because the instability of numerical differentiation is related to the step size $h/2^k$. Also, the higher-order formulas depend increasingly on the entry to their immediate left in the table, which is the reason we recommend comparing the final diagonal entries to ensure accuracy.

In Section 4.1, we discussed both three- and five-point methods for approximating $f'(x_0)$ given various functional values of f. The three-point methods were derived by differentiating a Lagrange interpolating polynomial for f. The five-point methods can be obtained in a similar manner, but the derivation is tedious. Extrapolation can be used to more easily derive these formulas, as illustrated below.

Illustration Suppose we expand the function f in a fourth Taylor polynomial about x_0. Then

$$\begin{aligned} f(x) = &f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}f''(x_0)(x - x_0)^2 + \frac{1}{6}f'''(x_0)(x - x_0)^3 \\ &+ \frac{1}{24}f^{(4)}(x_0)(x - x_0)^4 + \frac{1}{120}f^{(5)}(\xi)(x - x_0)^5, \end{aligned}$$

for some number ξ between x and x_0. Evaluating f at $x_0 + h$ and $x_0 - h$ gives

$$\begin{aligned} f(x_0 + h) = &f(x_0) + f'(x_0)h + \frac{1}{2}f''(x_0)h^2 + \frac{1}{6}f'''(x_0)h^3 \\ &+ \frac{1}{24}f^{(4)}(x_0)h^4 + \frac{1}{120}f^{(5)}(\xi_1)h^5 \end{aligned} \tag{4.18}$$

and

$$\begin{aligned} f(x_0 - h) = &f(x_0) - f'(x_0)h + \frac{1}{2}f''(x_0)h^2 - \frac{1}{6}f'''(x_0)h^3 \\ &+ \frac{1}{24}f^{(4)}(x_0)h^4 - \frac{1}{120}f^{(5)}(\xi_2)h^5, \end{aligned} \tag{4.19}$$

where $x_0 - h < \xi_2 < x_0 < \xi_1 < x_0 + h$.

Subtracting Eq. (4.19) from Eq. (4.18) gives a new approximation for $f'(x)$.

$$f(x_0 + h) - f(x_0 - h) = 2hf'(x_0) + \frac{h^3}{3}f'''(x_0) + \frac{h^5}{120}[f^{(5)}(\xi_1) + f^{(5)}(\xi_2)], \quad (4.20)$$

which implies that

$$f'(x_0) = \frac{1}{2h}[f(x_0 + h) - f(x_0 - h)] - \frac{h^2}{6}f'''(x_0) - \frac{h^4}{240}[f^{(5)}(\xi_1) + f^{(5)}(\xi_2)].$$

If $f^{(5)}$ is continuous on $[x_0 - h, x_0 + h]$, the Intermediate Value Theorem 1.11 implies that a number $\tilde{\xi}$ in $(x_0 - h, x_0 + h)$ exists with

$$f^{(5)}(\tilde{\xi}) = \frac{1}{2}\left[f^{(5)}(\xi_1) + f^{(5)}(\xi_2)\right].$$

As a consequence,we have the $O(h^2)$ approximation

$$f'(x_0) = \frac{1}{2h}[f(x_0 + h) - f(x_0 - h)] - \frac{h^2}{6}f'''(x_0) - \frac{h^4}{120}f^{(5)}(\tilde{\xi}). \quad (4.21)$$

Although the approximation in Eq. (4.21) is the same as that given in the three-point formula in Eq. (4.5), the unknown evaluation point occurs now in $f^{(5)}$, rather than in f'''. Extrapolation takes advantage of this by first replacing h in Eq. (4.21) with $2h$ to give the new formula

$$f'(x_0) = \frac{1}{4h}[f(x_0 + 2h) - f(x_0 - 2h)] - \frac{4h^2}{6}f'''(x_0) - \frac{16h^4}{120}f^{(5)}(\hat{\xi}), \quad (4.22)$$

where $\hat{\xi}$ is between $x_0 - 2h$ and $x_0 + 2h$.

Multiplying Eq. (4.21) by 4 and subtracting Eq. (4.22) produces

$$3f'(x_0) = \frac{2}{h}[f(x_0 + h) - f(x_0 - h)] - \frac{1}{4h}[f(x_0 + 2h) - f(x_0 - 2h)] - \frac{h^4}{30}f^{(5)}(\tilde{\xi}) + \frac{2h^4}{15}f^{(5)}(\hat{\xi}).$$

Even if $f^{(5)}$ is continuous on $[x_0 - 2h, x_0 + 2h]$, the Intermediate Value Theorem 1.11 cannot be applied as we did to derive Eq. (4.21) because here we have the *difference* of terms involving $f^{(5)}$. However, an alternative method can be used to show that $f^{(5)}(\tilde{\xi})$ and $f^{(5)}(\hat{\xi})$ can still be replaced by a common value $f^{(5)}(\xi)$. Assuming this and dividing by 3 produces the five-point midpoint formula Eq. (4.6) that we saw in Section 4.1

$$f'(x_0) = \frac{1}{12h}[f(x_0 - 2h) - 8f(x_0 - h) + 8f(x_0 + h) - f(x_0 + 2h)] + \frac{h^4}{30}f^{(5)}(\xi). \quad \square$$

Other formulas for first and higher derivatives can be derived in a similar manner. See, for example, Exercise 8.

The technique of extrapolation is used throughout the text. The most prominent applications occur in approximating integrals in Section 4.5 and for determining approximate solutions to differential equations in Section 5.8.

EXERCISE SET 4.2

1. Apply the extrapolation process described in Example 1 to determine $N_3(h)$, an approximation to $f'(x_0)$, for the following functions and stepsizes.

a. $f(x) = \ln x, x_0 = 1.0, h = 0.4$

b. $f(x) = x + e^x, x_0 = 0.0, h = 0.4$

c. $f(x) = 2^x \sin x, x_0 = 1.05, h = 0.4$

d. $f(x) = x^3 \cos x, x_0 = 2.3, h = 0.4$

2. Add another line to the extrapolation table in Exercise 1 to obtain the approximation $N_4(h)$.

3. Repeat Exercise 1 using four-digit rounding arithmetic.

4. Repeat Exercise 2 using four-digit rounding arithmetic.

5. The following data give approximations to the integral

$$M = \int_0^{\pi} \sin x \, dx.$$

$$N_1(h) = 1.570796, \quad N_1\left(\frac{h}{2}\right) = 1.896119, \quad N_1\left(\frac{h}{4}\right) = 1.974232, \quad N_1\left(\frac{h}{8}\right) = 1.993570.$$

Assuming $M = N_1(h) + K_1h^2 + K_2h^4 + K_3h^6 + K_4h^8 + O(h^{10})$, construct an extrapolation table to determine $N_4(h)$.

6. The following data can be used to approximate the integral

$$M = \int_0^{3\pi/2} \cos x \, dx.$$

$$N_1(h) = 2.356194, \qquad N_1\left(\frac{h}{2}\right) = -0.4879837,$$

$$N_1\left(\frac{h}{4}\right) = -0.8815732, \quad N_1\left(\frac{h}{8}\right) = -0.9709157.$$

Assume a formula exists of the type given in Exercise 5 and determine $N_4(h)$.

7. Show that the five-point formula in Eq. (4.6) applied to $f(x) = xe^x$ at $x_0 = 2.0$ gives $N_2(0.2)$ in Table 4.6 when $h = 0.1$ and $N_2(0.1)$ when $h = 0.05$.

8. The forward-difference formula can be expressed as

$$f'(x_0) = \frac{1}{h}[f(x_0 + h) - f(x_0)] - \frac{h}{2}f''(x_0) - \frac{h^2}{6}f'''(x_0) + O(h^3).$$

Use extrapolation to derive an $O(h^3)$ formula for $f'(x_0)$.

9. Suppose that $N(h)$ is an approximation to M for every $h > 0$ and that

$$M = N(h) + K_1h + K_2h^2 + K_3h^3 + \cdots,$$

for some constants K_1, K_2, K_3, ... Use the values $N(h)$, $N\left(\frac{h}{3}\right)$, and $N\left(\frac{h}{9}\right)$ to produce an $O(h^3)$ approximation to M.

10. Suppose that $N(h)$ is an approximation to M for every $h > 0$ and that

$$M = N(h) + K_1h^2 + K_2h^4 + K_3h^6 + \cdots,$$

for some constants K_1, K_2, K_3, ... Use the values $N(h)$, $N\left(\frac{h}{3}\right)$, and $N\left(\frac{h}{9}\right)$ to produce an $O(h^6)$ approximation to M.

11. In calculus, we learn that $e = \lim_{h\to 0}(1 + h)^{1/h}$.

a. Determine approximations to e corresponding to $h = 0.04$, 0.02, and 0.01.

b. Use extrapolation on the approximations, assuming that constants K_1, K_2, ... exist with $e = (1 + h)^{1/h} + K_1h + K_2h^2 + K_3h^3 + \cdots$, to produce an $O(h^3)$ approximation to e, where $h = 0.04$.

c. Do you think that the assumption in part (b) is correct?

12. **a.** Show that

$$\lim_{h\to 0}\left(\frac{2+h}{2-h}\right)^{1/h} = e.$$

b. Compute approximations to e using the formula $N(h) = \left(\frac{2+h}{2-h}\right)^{1/h}$, for $h = 0.04$, 0.02, and 0.01.

c. Assume that $e = N(h) + K_1h + K_2h^2 + K_3h^3 + \cdots$. Use extrapolation, with at least 16 digits of precision, to compute an $O(h^3)$ approximation to e with $h = 0.04$. Do you think the assumption is correct?

d. Show that $N(-h) = N(h)$.

e. Use part (d) to show that $K_1 = K_3 = K_5 = \cdots = 0$ in the formula

$$e = N(h) + K_1h + K_2h^2 + K_3h^3K_4h^4 + K_5h^5 + \cdots,$$

so that the formula reduces to

$$e = N(h) + K_2h^2 + K_4h^4 + K_6h^6 + \cdots.$$

f. Use the results of part (e) and extrapolation to compute an $O(h^6)$ approximation to e with $h = 0.04$.

13. Suppose the following extrapolation table has been constructed to approximate the number M with $M = N_1(h) + K_1h^2 + K_2h^4 + K_3h^6$:

$N_1(h)$		
$N_1\left(\frac{h}{2}\right)$	$N_2(h)$	
$N_1\left(\frac{h}{4}\right)$	$N_2\left(\frac{h}{2}\right)$	$N_3(h)$

a. Show that the linear interpolating polynomial $P_{0,1}(h)$ through $(h^2, N_1(h))$ and $(h^2/4, N_1(h/2))$ satisfies $P_{0,1}(0) = N_2(h)$. Similarly, show that $P_{1,2}(0) = N_2(h/2)$.

b. Show that the linear interpolating polynomial $P_{0,2}(h)$ through $(h^4, N_2(h))$ and $(h^4/16, N_2(h/2))$ satisfies $P_{0,2}(0) = N_3(h)$.

14. Suppose that $N_1(h)$ is a formula that produces $O(h)$ approximations to a number M and that

$$M = N_1(h) + K_1h + K_2h^2 + \cdots,$$

for a collection of positive constants $K_1, K_2, \ldots$. Then $N_1(h)$, $N_1(h/2)$, $N_1(h/4)$, ... are all lower bounds for M. What can be said about the extrapolated approximations $N_2(h), N_3(h), \ldots$?

15. The semiperimeters of regular polygons with k sides that inscribe and circumscribe the unit circle were used by Archimedes before 200 B.C.E. to approximate π, the circumference of a semicircle. Geometry can be used to show that the sequence of inscribed and circumscribed semiperimeters $\{p_k\}$ and $\{P_k\}$, respectively, satisfy

$$p_k = k\sin\left(\frac{\pi}{k}\right) \quad \text{and} \quad P_k = k\tan\left(\frac{\pi}{k}\right),$$

with $p_k < \pi < P_k$, whenever $k \geq 4$.

a. Show that $p_4 = 2\sqrt{2}$ and $P_4 = 4$.

b. Show that for $k \geq 4$, the sequences satisfy the recurrence relations

$$P_{2k} = \frac{2p_kP_k}{p_k + P_k} \quad \text{and} \quad p_{2k} = \sqrt{p_kP_{2k}}.$$

c. Approximate π to within 10^{-4} by computing p_k and P_k until $P_k - p_k < 10^{-4}$.

d. Use Taylor Series to show that

$$\pi = p_k + \frac{\pi^3}{3!}\left(\frac{1}{k}\right)^2 - \frac{\pi^5}{5!}\left(\frac{1}{k}\right)^4 + \cdots$$

and

$$\pi = P_k - \frac{\pi^3}{3}\left(\frac{1}{k}\right)^2 + \frac{2\pi^5}{15}\left(\frac{1}{k}\right)^4 - \cdots.$$

e. Use extrapolation with $h = 1/k$ to better approximate π.

4.3 Elements of Numerical Integration

The need often arises for evaluating the definite integral of a function that has no explicit antiderivative or whose antiderivative is not easy to obtain. The basic method involved in approximating $\int_a^b f(x)\,dx$ is called **numerical quadrature**. It uses a sum $\sum_{i=0}^n a_i f(x_i)$ to approximate $\int_a^b f(x)\,dx$.

The methods of quadrature in this section are based on the interpolation polynomials given in Chapter 3. The basic idea is to select a set of distinct nodes $\{x_0, \ldots, x_n\}$ from the interval $[a, b]$. Then integrate the Lagrange interpolating polynomial

$$P_n(x) = \sum_{i=0}^{n} f(x_i)L_i(x)$$

and its truncation error term over $[a, b]$ to obtain

$$\begin{aligned}\int_a^b f(x)\,dx &= \int_a^b \sum_{i=0}^{n} f(x_i)L_i(x)\,dx + \int_a^b \prod_{i=0}^{n}(x - x_i)\frac{f^{(n+1)}(\xi(x))}{(n+1)!}\,dx \\ &= \sum_{i=0}^{n} a_i f(x_i) + \frac{1}{(n+1)!}\int_a^b \prod_{i=0}^{n}(x - x_i) f^{(n+1)}(\xi(x))\,dx,\end{aligned}$$

where $\xi(x)$ is in $[a, b]$ for each x and

$$a_i = \int_a^b L_i(x)\,dx, \quad \text{for each } i = 0, 1, \ldots, n.$$

The quadrature formula is, therefore,

$$\int_a^b f(x)\,dx \approx \sum_{i=0}^{n} a_i f(x_i),$$

with error given by

$$E(f) = \frac{1}{(n+1)!}\int_a^b \prod_{i=0}^{n}(x - x_i) f^{(n+1)}(\xi(x))\,dx.$$

Before discussing the general situation of quadrature formulas, let us consider formulas produced by using first and second Lagrange polynomials with equally-spaced nodes. This gives the **Trapezoidal rule** and **Simpson's rule**, which are commonly introduced in calculus courses.

The Trapezoidal Rule

To derive the Trapezoidal rule for approximating $\int_a^b f(x)\,dx$, let $x_0 = a$, $x_1 = b$, $h = b - a$ and use the linear Lagrange polynomial:

$$P_1(x) = \frac{(x - x_1)}{(x_0 - x_1)} f(x_0) + \frac{(x - x_0)}{(x_1 - x_0)} f(x_1).$$

Then

$$\int_a^b f(x)\,dx = \int_{x_0}^{x_1} \left[\frac{(x - x_1)}{(x_0 - x_1)} f(x_0) + \frac{(x - x_0)}{(x_1 - x_0)} f(x_1) \right] dx + \frac{1}{2} \int_{x_0}^{x_1} f''(\xi(x))(x - x_0)(x - x_1)\,dx. \tag{4.23}$$

The product $(x - x_0)(x - x_1)$ does not change sign on $[x_0, x_1]$, so the Weighted Mean Value Theorem for Integrals 1.13 can be applied to the error term to give, for some ξ in (x_0, x_1),

$$\begin{aligned} \int_{x_0}^{x_1} f''(\xi(x))(x - x_0)(x - x_1)\,dx &= f''(\xi) \int_{x_0}^{x_1} (x - x_0)(x - x_1)\,dx \\ &= f''(\xi) \left[\frac{x^3}{3} - \frac{(x_1 + x_0)}{2} x^2 + x_0 x_1 x \right]_{x_0}^{x_1} \\ &= -\frac{h^3}{6} f''(\xi). \end{aligned}$$

When we use the term *trapezoid* we mean a four-sided figure that has at least two of its sides parallel. The European term for this figure is *trapezium*. To further confuse the issue, the European word trapezoidal refers to a four-sided figure with no sides equal, and the American word for this type of figure is trapezium.

Consequently, Eq. (4.23) implies that

$$\begin{aligned} \int_a^b f(x)\,dx &= \left[\frac{(x - x_1)^2}{2(x_0 - x_1)} f(x_0) + \frac{(x - x_0)^2}{2(x_1 - x_0)} f(x_1) \right]_{x_0}^{x_1} - \frac{h^3}{12} f''(\xi) \\ &= \frac{(x_1 - x_0)}{2} [f(x_0) + f(x_1)] - \frac{h^3}{12} f''(\xi). \end{aligned}$$

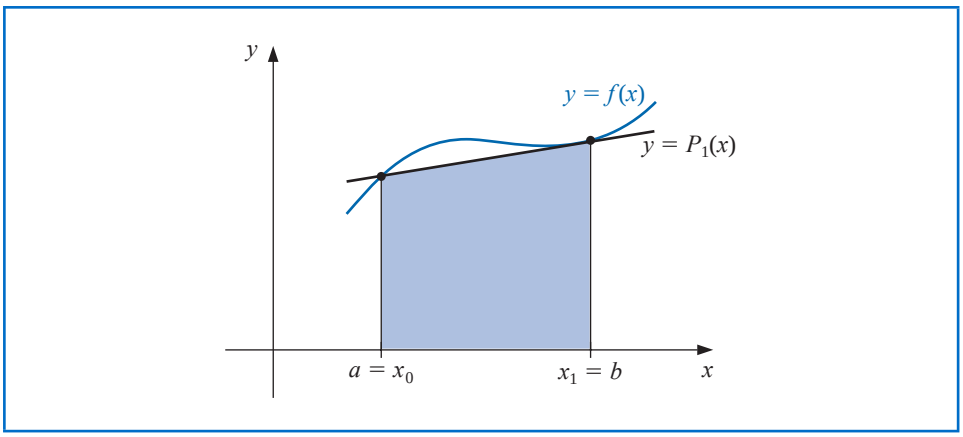
Using the notation $h = x_1 - x_0$ gives the following rule:

Trapezoidal Rule:

$$\int_a^b f(x)\,dx = \frac{h}{2} [f(x_0) + f(x_1)] - \frac{h^3}{12} f''(\xi).$$

This is called the Trapezoidal rule because when f is a function with positive values, $\int_a^b f(x)\,dx$ is approximated by the area in a trapezoid, as shown in Figure 4.3.

Figure 4.3

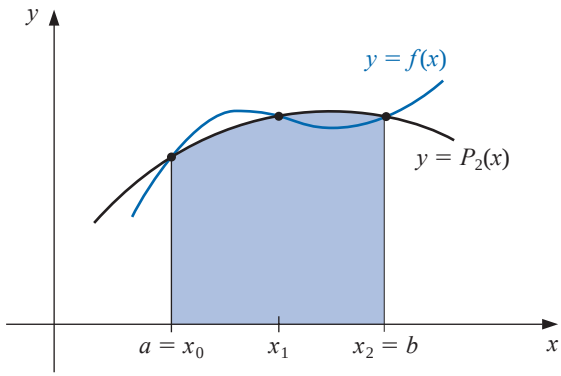


The error term for the Trapezoidal rule involves f'', so the rule gives the exact result when applied to any function whose second derivative is identically zero, that is, any polynomial of degree one or less.

Simpson's Rule

Simpson's rule results from integrating over $[a, b]$ the second Lagrange polynomial with equally-spaced nodes $x_0 = a$, $x_2 = b$, and $x_1 = a + h$, where $h = (b - a)/2$. (See Figure 4.4.)

Figure 4.4



Therefore

$$\int_a^b f(x)\,dx = \int_{x_0}^{x_2} \left[\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} f(x_0) + \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} f(x_1) + \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} f(x_2) \right] dx + \int_{x_0}^{x_2} \frac{(x-x_0)(x-x_1)(x-x_2)}{6} f^{(3)}(\xi(x))\,dx.$$

Deriving Simpson's rule in this manner, however, provides only an $O(h^4)$ error term involving $f^{(3)}$. By approaching the problem in another way, a higher-order term involving $f^{(4)}$ can be derived.

To illustrate this alternative method, suppose that f is expanded in the third Taylor polynomial about x_1. Then for each x in $[x_0, x_2]$, a number $\xi(x)$ in (x_0, x_2) exists with

$$f(x) = f(x_1) + f'(x_1)(x-x_1) + \frac{f''(x_1)}{2}(x-x_1)^2 + \frac{f'''(x_1)}{6}(x-x_1)^3 + \frac{f^{(4)}(\xi(x))}{24}(x-x_1)^4$$

and

$$\int_{x_0}^{x_2} f(x)\,dx = \left[f(x_1)(x-x_1) + \frac{f'(x_1)}{2}(x-x_1)^2 + \frac{f''(x_1)}{6}(x-x_1)^3 + \frac{f'''(x_1)}{24}(x-x_1)^4 \right]_{x_0}^{x_2} + \frac{1}{24}\int_{x_0}^{x_2} f^{(4)}(\xi(x))(x-x_1)^4\,dx. \quad (4.24)$$

Because $(x - x_1)^4$ is never negative on $[x_0, x_2]$, the Weighted Mean Value Theorem for Integrals 1.13 implies that

$$\frac{1}{24}\int_{x_0}^{x_2} f^{(4)}(\xi(x))(x - x_1)^4\,dx = \frac{f^{(4)}(\xi_1)}{24}\int_{x_0}^{x_2}(x - x_1)^4\,dx = \frac{f^{(4)}(\xi_1)}{120}(x - x_1)^5\Big]_{x_0}^{x_2},$$

for some number ξ_1 in (x_0, x_2).

However, $h = x_2 - x_1 = x_1 - x_0$, so

$$(x_2 - x_1)^2 - (x_0 - x_1)^2 = (x_2 - x_1)^4 - (x_0 - x_1)^4 = 0,$$

whereas

$$(x_2 - x_1)^3 - (x_0 - x_1)^3 = 2h^3 \quad \text{and} \quad (x_2 - x_1)^5 - (x_0 - x_1)^5 = 2h^5.$$

Consequently, Eq. (4.24) can be rewritten as

$$\int_{x_0}^{x_2} f(x)\,dx = 2hf(x_1) + \frac{h^3}{3}f''(x_1) + \frac{f^{(4)}(\xi_1)}{60}h^5.$$

If we now replace $f''(x_1)$ by the approximation given in Eq. (4.9) of Section 4.1, we have

$$\begin{aligned}\int_{x_0}^{x_2} f(x)\,dx &= 2hf(x_1) + \frac{h^3}{3}\left\{\frac{1}{h^2}[f(x_0) - 2f(x_1) + f(x_2)] - \frac{h^2}{12}f^{(4)}(\xi_2)\right\} + \frac{f^{(4)}(\xi_1)}{60}h^5\\ &= \frac{h}{3}[f(x_0) + 4f(x_1) + f(x_2)] - \frac{h^5}{12}\left[\frac{1}{3}f^{(4)}(\xi_2) - \frac{1}{5}f^{(4)}(\xi_1)\right].\end{aligned}$$

It can be shown by alternative methods (see Exercise 24) that the values ξ_1 and ξ_2 in this expression can be replaced by a common value ξ in (x_0, x_2). This gives Simpson's rule.

Thomas Simpson (1710–1761) was a self-taught mathematician who supported himself during his early years as a weaver. His primary interest was probability theory, although in 1750 he published a two-volume calculus book entitled *The Doctrine and Application of Fluxions*.

Simpson's Rule:

$$\int_{x_0}^{x_2} f(x)\,dx = \frac{h}{3}[f(x_0) + 4f(x_1) + f(x_2)] - \frac{h^5}{90}f^{(4)}(\xi).$$

The error term in Simpson's rule involves the fourth derivative of f, so it gives exact results when applied to any polynomial of degree three or less.

Example 1 Compare the Trapezoidal rule and Simpson's rule approximations to $\int_0^2 f(x)\,dx$ when $f(x)$ is

(a) x^2 **(b)** x^4 **(c)** $(x+1)^{-1}$
(d) $\sqrt{1+x^2}$ **(e)** $\sin x$ **(f)** e^x

Solution On $[0, 2]$ the Trapezoidal and Simpson's rule have the forms

$$\text{Trapezoid:} \quad \int_0^2 f(x)\,dx \approx f(0) + f(2) \quad \text{and}$$

$$\text{Simpson's:} \quad \int_0^2 f(x)\,dx \approx \frac{1}{3}[f(0) + 4f(1) + f(2)].$$

When $f(x) = x^2$ they give

$$\text{Trapezoid: } \int_0^2 f(x)\,dx \approx 0^2 + 2^2 = 4 \quad \text{and}$$

$$\text{Simpson's: } \int_0^2 f(x)\,dx \approx \frac{1}{3}[(0^2) + 4 \cdot 1^2 + 2^2] = \frac{8}{3}.$$

The approximation from Simpson's rule is exact because its truncation error involves $f^{(4)}$, which is identically 0 when $f(x) = x^2$.

The results to three places for the functions are summarized in Table 4.7. Notice that in each instance Simpson's Rule is significantly superior. ■

Table 4.7

	(a)	(b)	(c)	(d)	(e)	(f)
$f(x)$	x^2	x^4	$(x+1)^{-1}$	$\sqrt{1+x^2}$	$\sin x$	e^x
Exact value	2.667	6.400	1.099	2.958	1.416	6.389
Trapezoidal	4.000	16.000	1.333	3.326	0.909	8.389
Simpson's	2.667	6.667	1.111	2.964	1.425	6.421

Measuring Precision

The standard derivation of quadrature error formulas is based on determining the class of polynomials for which these formulas produce exact results. The next definition is used to facilitate the discussion of this derivation.

Definition 4.1 The **degree of accuracy**, or **precision**, of a quadrature formula is the largest positive integer n such that the formula is exact for x^k, for each $k = 0, 1, \ldots, n$. ■

The improved accuracy of Simpson's rule over the Trapezoidal rule is intuitively explained by the fact that Simpson's rule includes a midpoint evaluation that provides better balance to the approximation.

Definition 4.1 implies that the Trapezoidal and Simpson's rules have degrees of precision one and three, respectively.

Integration and summation are linear operations; that is,

$$\int_a^b (\alpha f(x) + \beta g(x))\,dx = \alpha \int_a^b f(x)\,dx + \beta \int_a^b g(x)\,dx$$

and

$$\sum_{i=0}^{n} (\alpha f(x_i) + \beta g(x_i)) = \alpha \sum_{i=0}^{n} f(x_i) + \beta \sum_{i=0}^{n} g(x_i),$$

for each pair of integrable functions f and g and each pair of real constants α and β. This implies (see Exercise 25) that:

The open and closed terminology for methods implies that the open methods use as nodes only points in the open interval, (a, b) to approximate $\int_a^b f(x)\,dx$. The closed methods include the points a and b of the closed interval $[a, b]$ as nodes.

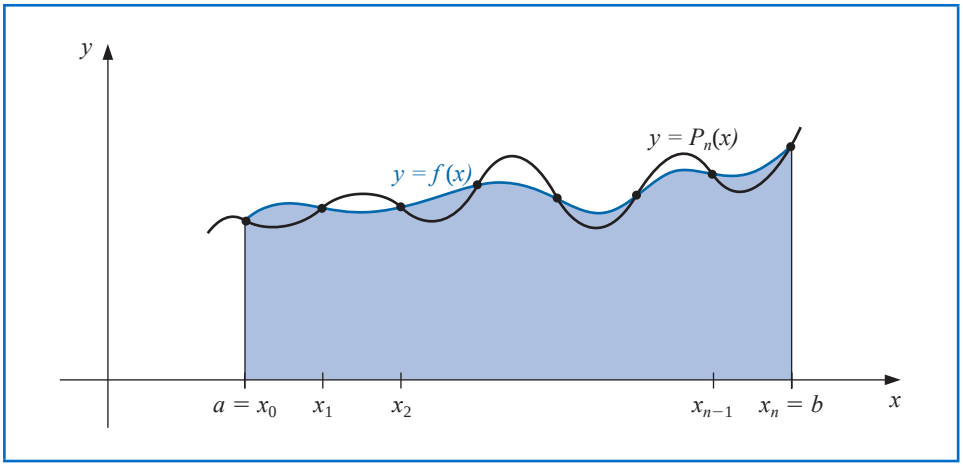
- The degree of precision of a quadrature formula is n if and only if the error is zero for all polynomials of degree $k = 0, 1, \ldots, n$, but is not zero for some polynomial of degree $n + 1$.

The Trapezoidal and Simpson's rules are examples of a class of methods known as Newton-Cotes formulas. There are two types of Newton-Cotes formulas, open and closed.

Closed Newton-Cotes Formulas

The $(n+1)$-*point closed Newton-Cotes formula* uses nodes $x_i = x_0 + ih$, for $i = 0, 1, \ldots, n$, where $x_0 = a$, $x_n = b$ and $h = (b-a)/n$. (See Figure 4.5.) It is called closed because the endpoints of the closed interval $[a, b]$ are included as nodes.

Figure 4.5



The formula assumes the form

$$\int_a^b f(x)\,dx \approx \sum_{i=0}^{n} a_i f(x_i),$$

where

$$a_i = \int_{x_0}^{x_n} L_i(x)\,dx = \int_{x_0}^{x_n} \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{(x - x_j)}{(x_i - x_j)}\,dx.$$

The following theorem details the error analysis associated with the closed Newton-Cotes formulas. For a proof of this theorem, see [IK], p. 313.

Theorem 4.2 Suppose that $\sum_{i=0}^{n} a_i f(x_i)$ denotes the $(n+1)$-point closed Newton-Cotes formula with $x_0 = a$, $x_n = b$, and $h = (b-a)/n$. There exists $\xi \in (a, b)$ for which

$$\int_a^b f(x)\,dx = \sum_{i=0}^{n} a_i f(x_i) + \frac{h^{n+3} f^{(n+2)}(\xi)}{(n+2)!} \int_0^n t^2(t-1)\cdots(t-n)\,dt,$$

if n is even and $f \in C^{n+2}[a, b]$, and

$$\int_a^b f(x)\,dx = \sum_{i=0}^{n} a_i f(x_i) + \frac{h^{n+2} f^{(n+1)}(\xi)}{(n+1)!} \int_0^n t(t-1)\cdots(t-n)\,dt,$$

if n is odd and $f \in C^{n+1}[a, b]$. ■

Roger Cotes (1682–1716) rose from a modest background to become, in 1704, the first Plumian Professor at Cambridge University. He made advances in numerous mathematical areas including numerical methods for interpolation and integration. Newton is reputed to have said of Cotes ...*if he had lived we might have known something.*

Note that when n is an even integer, the degree of precision is $n + 1$, although the interpolation polynomial is of degree at most n. When n is odd, the degree of precision is only n.

Some of the common **closed Newton-Cotes formulas** with their error terms are listed. Note that in each case the unknown value ξ lies in (a, b).

$n = 1$: Trapezoidal rule

$$\int_{x_0}^{x_1} f(x)\,dx = \frac{h}{2}[f(x_0) + f(x_1)] - \frac{h^3}{12} f''(\xi), \quad \text{where} \quad x_0 < \xi < x_1. \tag{4.25}$$

$n = 2$: Simpson's rule

$$\int_{x_0}^{x_2} f(x)\,dx = \frac{h}{3}[f(x_0) + 4f(x_1) + f(x_2)] - \frac{h^5}{90} f^{(4)}(\xi), \quad \text{where} \quad x_0 < \xi < x_2. \tag{4.26}$$

$n = 3$: Simpson's Three-Eighths rule

$$\int_{x_0}^{x_3} f(x)\,dx = \frac{3h}{8}[f(x_0) + 3f(x_1) + 3f(x_2) + f(x_3)] - \frac{3h^5}{80} f^{(4)}(\xi), \tag{4.27}$$

$$\text{where} \quad x_0 < \xi < x_3.$$

$n = 4$:

$$\int_{x_0}^{x_4} f(x)\,dx = \frac{2h}{45}[7f(x_0) + 32f(x_1) + 12f(x_2) + 32f(x_3) + 7f(x_4)] - \frac{8h^7}{945} f^{(6)}(\xi),$$

$$\text{where} \quad x_0 < \xi < x_4. \tag{4.28}$$

Open Newton-Cotes Formulas

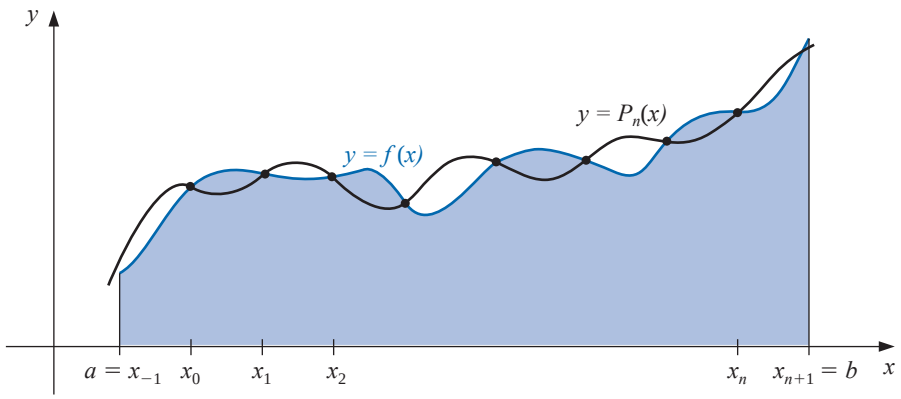
The *open Newton-Cotes formulas* do not include the endpoints of $[a, b]$ as nodes. They use the nodes $x_i = x_0 + ih$, for each $i = 0, 1, \ldots, n$, where $h = (b - a)/(n + 2)$ and $x_0 = a + h$. This implies that $x_n = b - h$, so we label the endpoints by setting $x_{-1} = a$ and $x_{n+1} = b$, as shown in Figure 4.6 on page 200. Open formulas contain all the nodes used for the approximation within the open interval (a, b). The formulas become

$$\int_a^b f(x)\,dx = \int_{x_{-1}}^{x_{n+1}} f(x)\,dx \approx \sum_{i=0}^{n} a_i f(x_i),$$

where

$$a_i = \int_a^b L_i(x)\,dx.$$

Figure 4.6



The following theorem is analogous to Theorem 4.2; its proof is contained in [IK], p. 314.

Theorem 4.3 Suppose that $\sum_{i=0}^{n} a_i f(x_i)$ denotes the $(n+1)$-point open Newton-Cotes formula with $x_{-1} = a$, $x_{n+1} = b$, and $h = (b-a)/(n+2)$. There exists $\xi \in (a, b)$ for which

$$\int_a^b f(x)\,dx = \sum_{i=0}^{n} a_i f(x_i) + \frac{h^{n+3} f^{(n+2)}(\xi)}{(n+2)!} \int_{-1}^{n+1} t^2(t-1)\cdots(t-n)\,dt,$$

if n is even and $f \in C^{n+2}[a, b]$, and

$$\int_a^b f(x)\,dx = \sum_{i=0}^{n} a_i f(x_i) + \frac{h^{n+2} f^{(n+1)}(\xi)}{(n+1)!} \int_{-1}^{n+1} t(t-1)\cdots(t-n)\,dt,$$

if n is odd and $f \in C^{n+1}[a, b]$. ■

Notice, as in the case of the closed methods, we have the degree of precision comparatively higher for the even methods than for the odd methods.

Some of the common **open Newton-Cotes** formulas with their error terms are as follows:

$n = 0$: Midpoint rule

$$\int_{x_{-1}}^{x_1} f(x)\,dx = 2hf(x_0) + \frac{h^3}{3} f''(\xi), \quad \text{where} \quad x_{-1} < \xi < x_1. \tag{4.29}$$

$n = 1$:

$$\int_{x_{-1}}^{x_2} f(x)\,dx = \frac{3h}{2}[f(x_0) + f(x_1)] + \frac{3h^3}{4} f''(\xi), \quad \text{where} \quad x_{-1} < \xi < x_2. \tag{4.30}$$

$n = 2$:

$$\int_{x_{-1}}^{x_3} f(x)\,dx = \frac{4h}{3}[2f(x_0) - f(x_1) + 2f(x_2)] + \frac{14h^5}{45}f^{(4)}(\xi), \tag{4.31}$$

where $x_{-1} < \xi < x_3$.

$n = 3$:

$$\int_{x_{-1}}^{x_4} f(x)\,dx = \frac{5h}{24}[11f(x_0) + f(x_1) + f(x_2) + 11f(x_3)] + \frac{95}{144}h^5 f^{(4)}(\xi), \tag{4.32}$$

where $x_{-1} < \xi < x_4$.

Example 2 Compare the results of the closed and open Newton-Cotes formulas listed as (4.25)–(4.28) and (4.29)–(4.32) when approximating

$$\int_0^{\pi/4} \sin x\,dx = 1 - \sqrt{2}/2 \approx 0.29289322.$$

Solution For the closed formulas we have

$$n = 1: \quad \frac{(\pi/4)}{2}\left[\sin 0 + \sin\frac{\pi}{4}\right] \approx 0.27768018$$

$$n = 2: \quad \frac{(\pi/8)}{3}\left[\sin 0 + 4\sin\frac{\pi}{8} + \sin\frac{\pi}{4}\right] \approx 0.29293264$$

$$n = 3: \quad \frac{3(\pi/12)}{8}\left[\sin 0 + 3\sin\frac{\pi}{12} + 3\sin\frac{\pi}{6} + \sin\frac{\pi}{4}\right] \approx 0.29291070$$

$$n = 4: \quad \frac{2(\pi/16)}{45}\left[7\sin 0 + 32\sin\frac{\pi}{16} + 12\sin\frac{\pi}{8} + 32\sin\frac{3\pi}{16} + 7\sin\frac{\pi}{4}\right] \approx 0.29289318$$

and for the open formulas we have

$$n = 0: \quad 2(\pi/8)\left[\sin\frac{\pi}{8}\right] \approx 0.30055887$$

$$n = 1: \quad \frac{3(\pi/12)}{2}\left[\sin\frac{\pi}{12} + \sin\frac{\pi}{6}\right] \approx 0.29798754$$

$$n = 2: \quad \frac{4(\pi/16)}{3}\left[2\sin\frac{\pi}{16} - \sin\frac{\pi}{8} + 2\sin\frac{3\pi}{16}\right] \approx 0.29285866$$

$$n = 3: \quad \frac{5(\pi/20)}{24}\left[11\sin\frac{\pi}{20} + \sin\frac{\pi}{10} + \sin\frac{3\pi}{20} + 11\sin\frac{\pi}{5}\right] \approx 0.29286923$$

Table 4.8 summarizes these results and shows the approximation errors. ■

Table 4.8

n	0	1	2	3	4
Closed formulas		0.27768018	0.29293264	0.29291070	0.29289318
Error		0.01521303	0.00003942	0.00001748	0.00000004
Open formulas	0.30055887	0.29798754	0.29285866	0.29286923	
Error	0.00766565	0.00509432	0.00003456	0.00002399	

EXERCISE SET 4.3

1. Approximate the following integrals using the Trapezoidal rule.

a. $\int_{0.5}^{1} x^4\,dx$ **b.** $\int_{0}^{0.5} \frac{2}{x-4}\,dx$

c. $\int_{1}^{1.5} x^2 \ln x\,dx$ **d.** $\int_{0}^{1} x^2 e^{-x}\,dx$

e. $\int_{1}^{1.6} \frac{2x}{x^2-4}\,dx$ **f.** $\int_{0}^{0.35} \frac{2}{x^2-4}\,dx$

g. $\int_{0}^{\pi/4} x \sin x\,dx$ **h.** $\int_{0}^{\pi/4} e^{3x} \sin 2x\,dx$

2. Approximate the following integrals using the Trapezoidal rule.

a. $\int_{-0.25}^{0.25} (\cos x)^2\,dx$ **b.** $\int_{-0.5}^{0} x \ln(x+1)\,dx$

c. $\int_{0.75}^{1.3} \left((\sin x)^2 - 2x \sin x + 1\right)\,dx$ **d.** $\int_{e}^{e+1} \frac{1}{x \ln x}\,dx$

3. Find a bound for the error in Exercise 1 using the error formula, and compare this to the actual error.

4. Find a bound for the error in Exercise 2 using the error formula, and compare this to the actual error.

5. Repeat Exercise 1 using Simpson's rule.

6. Repeat Exercise 2 using Simpson's rule.

7. Repeat Exercise 3 using Simpson's rule and the results of Exercise 5.

8. Repeat Exercise 4 using Simpson's rule and the results of Exercise 6.

9. Repeat Exercise 1 using the Midpoint rule.

10. Repeat Exercise 2 using the Midpoint rule.

11. Repeat Exercise 3 using the Midpoint rule and the results of Exercise 9.

12. Repeat Exercise 4 using the Midpoint rule and the results of Exercise 10.

13. The Trapezoidal rule applied to $\int_0^2 f(x)\,dx$ gives the value 4, and Simpson's rule gives the value 2. What is $f(1)$?

14. The Trapezoidal rule applied to $\int_0^2 f(x)\,dx$ gives the value 5, and the Midpoint rule gives the value 4. What value does Simpson's rule give?

15. Find the degree of precision of the quadrature formula

$$\int_{-1}^{1} f(x)\,dx = f\left(-\frac{\sqrt{3}}{3}\right) + f\left(\frac{\sqrt{3}}{3}\right).$$

16. Let $h = (b-a)/3$, $x_0 = a$, $x_1 = a+h$, and $x_2 = b$. Find the degree of precision of the quadrature formula

$$\int_a^b f(x)\,dx = \frac{9}{4}hf(x_1) + \frac{3}{4}hf(x_2).$$

17. The quadrature formula $\int_{-1}^{1} f(x)\,dx = c_0 f(-1) + c_1 f(0) + c_2 f(1)$ is exact for all polynomials of degree less than or equal to 2. Determine c_0, c_1, and c_2.

18. The quadrature formula $\int_0^2 f(x)\,dx = c_0 f(0) + c_1 f(1) + c_2 f(2)$ is exact for all polynomials of degree less than or equal to 2. Determine c_0, c_1, and c_2.

19. Find the constants c_0, c_1, and x_1 so that the quadrature formula

$$\int_0^1 f(x)\,dx = c_0 f(0) + c_1 f(x_1)$$

has the highest possible degree of precision.

20. Find the constants x_0, x_1, and c_1 so that the quadrature formula

$$\int_0^1 f(x)\,dx = \frac{1}{2} f(x_0) + c_1 f(x_1)$$

has the highest possible degree of precision.

21. Approximate the following integrals using formulas (4.25) through (4.32). Are the accuracies of the approximations consistent with the error formulas? Which of parts (d) and (e) give the better approximation?

a. $\int_0^{0.1} \sqrt{1+x}\,dx$ **b.** $\int_0^{\pi/2} (\sin x)^2\,dx$

c. $\int_{1.1}^{1.5} e^x\,dx$ **d.** $\int_1^{10} \frac{1}{x}\,dx$

e. $\int_1^{5.5} \frac{1}{x}\,dx + \int_{5.5}^{10} \frac{1}{x}\,dx$ **f.** $\int_0^1 x^{1/3}\,dx$

22. Given the function f at the following values,

x	1.8	2.0	2.2	2.4	2.6
$f(x)$	3.12014	4.42569	6.04241	8.03014	10.46675

approximate $\int_{1.8}^{2.6} f(x)\,dx$ using all the appropriate quadrature formulas of this section.

23. Suppose that the data of Exercise 22 have round-off errors given by the following table.

x	1.8	2.0	2.2	2.4	2.6
Error in $f(x)$	2×10^{-6}	-2×10^{-6}	-0.9×10^{-6}	-0.9×10^{-6}	2×10^{-6}

Calculate the errors due to round-off in Exercise 22.

24. Derive Simpson's rule with error term by using

$$\int_{x_0}^{x_2} f(x)\,dx = a_0 f(x_0) + a_1 f(x_1) + a_2 f(x_2) + k f^{(4)}(\xi).$$

Find a_0, a_1, and a_2 from the fact that Simpson's rule is exact for $f(x) = x^n$ when $n = 1, 2$, and 3. Then find k by applying the integration formula with $f(x) = x^4$.

25. Prove the statement following Definition 4.1; that is, show that a quadrature formula has degree of precision n if and only if the error $E(P(x)) = 0$ for all polynomials $P(x)$ of degree $k = 0, 1, \ldots, n$, but $E(P(x)) \neq 0$ for some polynomial $P(x)$ of degree $n + 1$.

26. Derive Simpson's three-eighths rule (the closed rule with $n = 3$) with error term by using Theorem 4.2.

27. Derive the open rule with $n = 1$ with error term by using Theorem 4.3.

4.4 Composite Numerical Integration

The Newton-Cotes formulas are generally unsuitable for use over large integration intervals. High-degree formulas would be required, and the values of the coefficients in these formulas are difficult to obtain. Also, the Newton-Cotes formulas are based on interpolatory polynomials that use equally-spaced nodes, a procedure that is inaccurate over large intervals because of the oscillatory nature of high-degree polynomials.

Piecewise approximation is often effective. Recall that this was used for spline interpolation.

In this section, we discuss a *piecewise* approach to numerical integration that uses the low-order Newton-Cotes formulas. These are the techniques most often applied.

Example 1 Use Simpson's rule to approximate $\int_0^4 e^x\,dx$ and compare this to the results obtained by adding the Simpson's rule approximations for $\int_0^2 e^x\,dx$ and $\int_2^4 e^x\,dx$. Compare these approximations to the sum of Simpson's rule for $\int_0^1 e^x\,dx$, $\int_1^2 e^x\,dx$, $\int_2^3 e^x\,dx$, and $\int_3^4 e^x\,dx$.

Solution Simpson's rule on $[0, 4]$ uses $h = 2$ and gives

$$\int_0^4 e^x\,dx \approx \frac{2}{3}(e^0 + 4e^2 + e^4) = 56.76958.$$

The exact answer in this case is $e^4 - e^0 = 53.59815$, and the error -3.17143 is far larger than we would normally accept.

Applying Simpson's rule on each of the intervals $[0, 2]$ and $[2, 4]$ uses $h = 1$ and gives

$$\begin{aligned}\int_0^4 e^x\,dx &= \int_0^2 e^x\,dx + \int_2^4 e^x\,dx\\ &\approx \frac{1}{3}\left(e^0 + 4e + e^2\right) + \frac{1}{3}\left(e^2 + 4e^3 + e^4\right)\\ &= \frac{1}{3}\left(e^0 + 4e + 2e^2 + 4e^3 + e^4\right)\\ &= 53.86385.\end{aligned}$$

The error has been reduced to -0.26570.

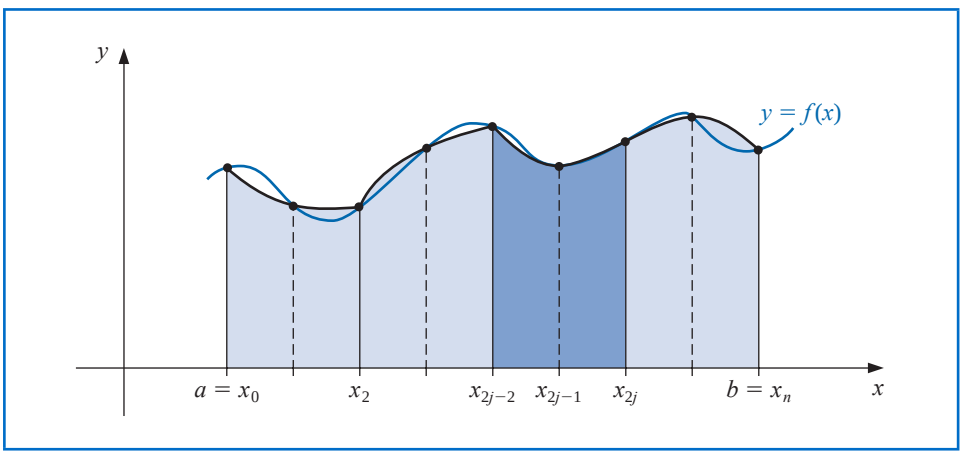
For the integrals on $[0, 1]$,$[1, 2]$,$[3, 4]$, and $[3, 4]$ we use Simpson's rule four times with $h = \frac{1}{2}$ giving

$$\begin{aligned}\int_0^4 e^x\,dx &= \int_0^1 e^x\,dx + \int_1^2 e^x\,dx + \int_2^3 e^x\,dx + \int_3^4 e^x\,dx\\ &\approx \frac{1}{6}\left(e_0 + 4e^{1/2} + e\right) + \frac{1}{6}\left(e + 4e^{3/2} + e^2\right)\\ &\quad + \frac{1}{6}\left(e^2 + 4e^{5/2} + e^3\right) + \frac{1}{6}\left(e^3 + 4e^{7/2} + e^4\right)\\ &= \frac{1}{6}\left(e^0 + 4e^{1/2} + 2e + 4e^{3/2} + 2e^2 + 4e^{5/2} + 2e^3 + 4e^{7/2} + e^4\right)\\ &= 53.61622.\end{aligned}$$

The error for this approximation has been reduced to -0.01807. ■

To generalize this procedure for an arbitrary integral $\int_a^b f(x)\,dx$, choose an even integer n. Subdivide the interval $[a, b]$ into n subintervals, and apply Simpson's rule on each consecutive pair of subintervals. (See Figure 4.7.)

Figure 4.7



With $h = (b-a)/n$ and $x_j = a + jh$, for each $j = 0, 1, \ldots, n$, we have

$$\int_a^b f(x)\,dx = \sum_{j=1}^{n/2} \int_{x_{2j-2}}^{x_{2j}} f(x)\,dx$$

$$= \sum_{j=1}^{n/2} \left\{ \frac{h}{3}[f(x_{2j-2}) + 4f(x_{2j-1}) + f(x_{2j})] - \frac{h^5}{90} f^{(4)}(\xi_j) \right\},$$

for some ξ_j with $x_{2j-2} < \xi_j < x_{2j}$, provided that $f \in C^4[a,b]$. Using the fact that for each $j = 1, 2, \ldots, (n/2) - 1$ we have $f(x_{2j})$ appearing in the term corresponding to the interval $[x_{2j-2}, x_{2j}]$ and also in the term corresponding to the interval $[x_{2j}, x_{2j+2}]$, we can reduce this sum to

$$\int_a^b f(x)\,dx = \frac{h}{3}\left[f(x_0) + 2\sum_{j=1}^{(n/2)-1} f(x_{2j}) + 4\sum_{j=1}^{n/2} f(x_{2j-1}) + f(x_n) \right] - \frac{h^5}{90}\sum_{j=1}^{n/2} f^{(4)}(\xi_j).$$

The error associated with this approximation is

$$E(f) = -\frac{h^5}{90}\sum_{j=1}^{n/2} f^{(4)}(\xi_j),$$

where $x_{2j-2} < \xi_j < x_{2j}$, for each $j = 1, 2, \ldots, n/2$.

If $f \in C^4[a,b]$, the Extreme Value Theorem 1.9 implies that $f^{(4)}$ assumes its maximum and minimum in $[a,b]$. Since

$$\min_{x\in[a,b]} f^{(4)}(x) \le f^{(4)}(\xi_j) \le \max_{x\in[a,b]} f^{(4)}(x),$$

we have

$$\frac{n}{2}\min_{x\in[a,b]} f^{(4)}(x) \le \sum_{j=1}^{n/2} f^{(4)}(\xi_j) \le \frac{n}{2}\max_{x\in[a,b]} f^{(4)}(x)$$

and

$$\min_{x\in[a,b]} f^{(4)}(x) \le \frac{2}{n}\sum_{j=1}^{n/2} f^{(4)}(\xi_j) \le \max_{x\in[a,b]} f^{(4)}(x).$$

By the Intermediate Value Theorem 1.11, there is a $\mu \in (a,b)$ such that

$$f^{(4)}(\mu) = \frac{2}{n}\sum_{j=1}^{n/2} f^{(4)}(\xi_j).$$

Thus

$$E(f) = -\frac{h^5}{90}\sum_{j=1}^{n/2} f^{(4)}(\xi_j) = -\frac{h^5}{180} n f^{(4)}(\mu),$$

or, since $h = (b-a)/n$,

$$E(f) = -\frac{(b-a)}{180} h^4 f^{(4)}(\mu).$$

These observations produce the following result.

Theorem 4.4 Let $f \in C^4[a, b]$, n be even, $h = (b - a)/n$, and $x_j = a + jh$, for each $j = 0, 1, \ldots, n$. There exists a $\mu \in (a, b)$ for which the **Composite Simpson's rule** for n subintervals can be written with its error term as

$$\int_a^b f(x)\,dx = \frac{h}{3}\left[f(a) + 2\sum_{j=1}^{(n/2)-1} f(x_{2j}) + 4\sum_{j=1}^{n/2} f(x_{2j-1}) + f(b)\right] - \frac{b-a}{180}h^4 f^{(4)}(\mu).$$

■

Notice that the error term for the Composite Simpson's rule is $O(h^4)$, whereas it was $O(h^5)$ for the standard Simpson's rule. However, these rates are not comparable because for standard Simpson's rule we have h fixed at $h = (b - a)/2$, but for Composite Simpson's rule we have $h = (b - a)/n$, for n an even integer. This permits us to considerably reduce the value of h when the Composite Simpson's rule is used.

Algorithm 4.1 uses the Composite Simpson's rule on n subintervals. This is the most frequently used general-purpose quadrature algorithm.

ALGORITHM
4.1

Composite Simpson's Rule

To approximate the integral $I = \int_a^b f(x)\,dx$:

INPUT endpoints a, b; even positive integer n.

OUTPUT approximation XI to I.

Step 1 Set $h = (b - a)/n$.

Step 2 Set $XI0 = f(a) + f(b)$;
$XI1 = 0$; (*Summation of* $f(x_{2i-1})$.)
$XI2 = 0$. (*Summation of* $f(x_{2i})$.)

Step 3 For $i = 1, \ldots, n - 1$ do Steps 4 and 5.

Step 4 Set $X = a + ih$.

Step 5 If i is even then set $XI2 = XI2 + f(X)$
else set $XI1 = XI1 + f(X)$.

Step 6 Set $XI = h(XI0 + 2 \cdot XI2 + 4 \cdot XI1)/3$.

Step 7 OUTPUT (XI);
STOP.

■

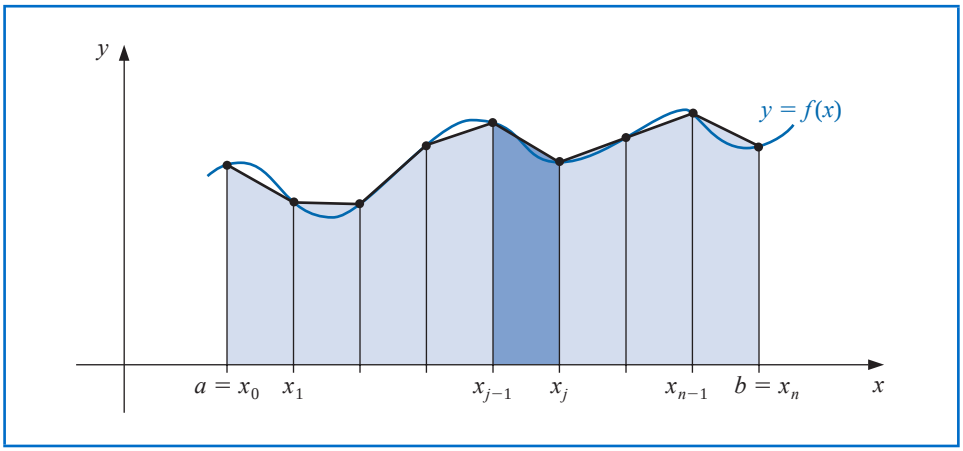
The subdivision approach can be applied to any of the Newton-Cotes formulas. The extensions of the Trapezoidal (see Figure 4.8) and Midpoint rules are given without proof. The Trapezoidal rule requires only one interval for each application, so the integer n can be either odd or even.

Theorem 4.5 Let $f \in C^2[a, b]$, $h = (b - a)/n$, and $x_j = a + jh$, for each $j = 0, 1, \ldots, n$. There exists a $\mu \in (a, b)$ for which the **Composite Trapezoidal rule** for n subintervals can be written with its error term as

$$\int_a^b f(x)\,dx = \frac{h}{2}\left[f(a) + 2\sum_{j=1}^{n-1} f(x_j) + f(b)\right] - \frac{b-a}{12}h^2 f''(\mu).$$

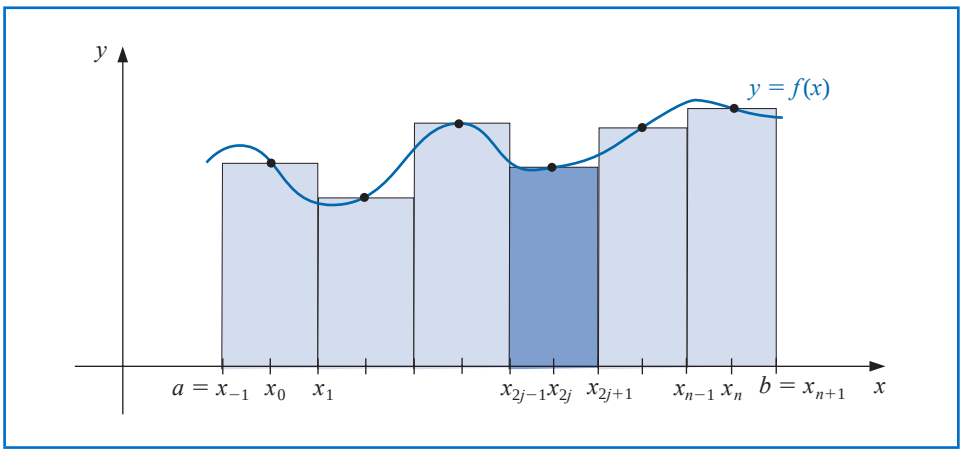
■

Figure 4.8



For the Composite Midpoint rule, n must again be even. (See Figure 4.9.)

Figure 4.9



Theorem 4.6 Let $f \in C^2[a,b]$, n be even, $h = (b-a)/(n+2)$, and $x_j = a + (j+1)h$ for each $j = -1, 0, \ldots, n+1$. There exists a $\mu \in (a,b)$ for which the **Composite Midpoint rule** for $n+2$ subintervals can be written with its error term as

$$\int_a^b f(x)\,dx = 2h\sum_{j=0}^{n/2} f(x_{2j}) + \frac{b-a}{6}h^2 f''(\mu).$$

■

Example 2 Determine values of h that will ensure an approximation error of less than 0.00002 when approximating $\int_0^{\pi} \sin x\,dx$ and employing
(a) Composite Trapezoidal rule and **(b)** Composite Simpson's rule.

Solution **(a)** The error form for the Composite Trapezoidal rule for $f(x) = \sin x$ on $[0, \pi]$ is

$$\left|\frac{\pi h^2}{12} f''(\mu)\right| = \left|\frac{\pi h^2}{12}(-\sin\mu)\right| = \frac{\pi h^2}{12}|\sin\mu|.$$

To ensure sufficient accuracy with this technique we need to have

$$\frac{\pi h^2}{12}|\sin\mu| \le \frac{\pi h^2}{12} < 0.00002.$$

Since $h = \pi/n$ implies that $n = \pi/h$, we need

$$\frac{\pi^3}{12n^2} < 0.00002 \quad \text{which implies that} \quad n > \left(\frac{\pi^3}{12(0.00002)}\right)^{1/2} \approx 359.44.$$

and the Composite Trapezoidal rule requires $n \ge 360$.

(b) The error form for the Composite Simpson's rule for $f(x) = \sin x$ on $[0, \pi]$ is

$$\left|\frac{\pi h^4}{180} f^{(4)}(\mu)\right| = \left|\frac{\pi h^4}{180}\sin\mu\right| = \frac{\pi h^4}{180}|\sin\mu|.$$

To ensure sufficient accuracy with this technique we need to have

$$\frac{\pi h^4}{180}|\sin\mu| \le \frac{\pi h^4}{180} < 0.00002.$$

Using again the fact that $n = \pi/h$ gives

$$\frac{\pi^5}{180n^4} < 0.00002 \quad \text{which implies that} \quad n > \left(\frac{\pi^5}{180(0.00002)}\right)^{1/4} \approx 17.07.$$

So Composite Simpson's rule requires only $n \ge 18$.

Composite Simpson's rule with $n = 18$ gives

$$\int_0^\pi \sin x\,dx \approx \frac{\pi}{54}\left[2\sum_{j=1}^{8}\sin\left(\frac{j\pi}{9}\right) + 4\sum_{j=1}^{9}\sin\left(\frac{(2j-1)\pi}{18}\right)\right] = 2.0000104.$$

This is accurate to within about 10^{-5} because the true value is $-\cos(\pi) - (-\cos(0)) = 2$. ■

Composite Simpson's rule is the clear choice if you wish to minimize computation. For comparison purposes, consider the Composite Trapezoidal rule using $h = \pi/18$ for the integral in Example 2. This approximation uses the same function evaluations as Composite Simpson's rule but the approximation in this case

$$\int_0^\pi \sin x\,dx \approx \frac{\pi}{36}\left[2\sum_{j=1}^{17}\sin\left(\frac{j\pi}{18}\right) + \sin 0 + \sin\pi\right] = \frac{\pi}{36}\left[2\sum_{j=1}^{17}\sin\left(\frac{j\pi}{18}\right)\right] = 1.9949205.$$

is accurate only to about 5×10^{-3}.

Maple contains numerous procedures for numerical integration in the *NumericalAnalysis* subpackage of the *Student* package. First access the library as usual with

with(*Student*[*NumericalAnalysis*])

The command for all methods is *Quadrature* with the options in the call specifying the method to be used. We will use the Trapezoidal method to illustrate the procedure. First define the function and the interval of integration with

$f := x \to \sin(x)$; $a := 0.0$; $b := \pi$

After Maple responds with the function and the interval, enter the command

$$Quadrature(f(x), x = a..b, method = trapezoid, partition = 20, output = value)$$

$$1.995885973$$

The value of the step size h in this instance is the width of the interval $b - a$ divided by the number specified by $partition = 20$.

Simpson's method can be called in a similar manner, except that the step size h is determined by $b - a$ divided by twice the value of *partition*. Hence, the Simpson's rule approximation using the same nodes as those in the Trapezoidal rule is called with

$$Quadrature(f(x), x = a..b, method = simpson, partition = 10, output = value)$$

$$2.000006785$$

Any of the Newton-Cotes methods can be called using the option

$$method = newtoncotes[open, n] \quad \text{or} \quad method = newtoncotes[closed, n]$$

Be careful to correctly specify the number in *partition* when an even number of divisions is required, and when an open method is employed.

Round-Off Error Stability

In Example 2 we saw that ensuring an accuracy of 2×10^{-5} for approximating $\int_0^{\pi} \sin x\, dx$ required 360 subdivisions of $[0, \pi]$ for the Composite Trapezoidal rule and only 18 for Composite Simpson's rule. In addition to the fact that less computation is needed for the Simpson's technique, you might suspect that because of fewer computations this method would also involve less round-off error. However, an important property shared by all the composite integration techniques is a stability with respect to round-off error. That is, the round-off error does not depend on the number of calculations performed.

Numerical integration is expected to be stable, whereas numerical differentiation is unstable.

To demonstrate this rather amazing fact, suppose we apply the Composite Simpson's rule with n subintervals to a function f on $[a, b]$ and determine the maximum bound for the round-off error. Assume that $f(x_i)$ is approximated by $\tilde{f}(x_i)$ and that

$$f(x_i) = \tilde{f}(x_i) + e_i, \quad \text{for each } i = 0, 1, \ldots, n,$$

where e_i denotes the round-off error associated with using $\tilde{f}(x_i)$ to approximate $f(x_i)$. Then the accumulated error, $e(h)$, in the Composite Simpson's rule is

$$\begin{aligned} e(h) &= \left| \frac{h}{3} \left[e_0 + 2 \sum_{j=1}^{(n/2)-1} e_{2j} + 4 \sum_{j=1}^{n/2} e_{2j-1} + e_n \right] \right| \\ &\leq \frac{h}{3} \left[|e_0| + 2 \sum_{j=1}^{(n/2)-1} |e_{2j}| + 4 \sum_{j=1}^{n/2} |e_{2j-1}| + |e_n| \right]. \end{aligned}$$

If the round-off errors are uniformly bounded by ε, then

$$e(h) \leq \frac{h}{3} \left[\varepsilon + 2 \left(\frac{n}{2} - 1 \right) \varepsilon + 4 \left(\frac{n}{2} \right) \varepsilon + \varepsilon \right] = \frac{h}{3} 3n\varepsilon = nh\varepsilon.$$

But $nh = b - a$, so

$$e(h) \leq (b - a)\varepsilon,$$

a bound independent of h (and n). This means that, even though we may need to divide an interval into more parts to ensure accuracy, the increased computation that is required does not increase the round-off error. This result implies that the procedure is stable as h approaches zero. Recall that this was not true of the numerical differentiation procedures considered at the beginning of this chapter.

EXERCISE SET 4.4

1. Use the Composite Trapezoidal rule with the indicated values of n to approximate the following integrals.

a. $\int_1^2 x \ln x \, dx, \quad n = 4$

b. $\int_{-2}^2 x^3 e^x \, dx, \quad n = 4$

c. $\int_0^2 \frac{2}{x^2 + 4} \, dx, \quad n = 6$

d. $\int_0^{\pi} x^2 \cos x \, dx, \quad n = 6$

e. $\int_0^2 e^{2x} \sin 3x \, dx, \quad n = 8$

f. $\int_1^3 \frac{x}{x^2 + 4} \, dx, \quad n = 8$

g. $\int_3^5 \frac{1}{\sqrt{x^2 - 4}} \, dx, \quad n = 8$

h. $\int_0^{3\pi/8} \tan x \, dx, \quad n = 8$

2. Use the Composite Trapezoidal rule with the indicated values of n to approximate the following integrals.

a. $\int_{-0.5}^{0.5} \cos^2 x \, dx, \quad n = 4$

b. $\int_{-0.5}^{0.5} x \ln(x + 1) \, dx, \quad n = 6$

c. $\int_{.75}^{1.75} (\sin^2 x - 2x \sin x + 1) \, dx, \quad n = 8$

d. $\int_e^{e+2} \frac{1}{x \ln x} \, dx, \quad n = 8$

3. Use the Composite Simpson's rule to approximate the integrals in Exercise 1.

4. Use the Composite Simpson's rule to approximate the integrals in Exercise 2.

5. Use the Composite Midpoint rule with $n + 2$ subintervals to approximate the integrals in Exercise 1.

6. Use the Composite Midpoint rule with $n + 2$ subintervals to approximate the integrals in Exercise 2.

7. Approximate $\int_0^2 x^2 \ln(x^2 + 1) \, dx$ using $h = 0.25$. Use

a. Composite Trapezoidal rule.

b. Composite Simpson's rule.

c. Composite Midpoint rule.

8. Approximate $\int_0^2 x^2 e^{-x^2} \, dx$ using $h = 0.25$. Use

a. Composite Trapezoidal rule.

b. Composite Simpson's rule.

c. Composite Midpoint rule.

9. Suppose that $f(0) = 1$, $f(0.5) = 2.5$, $f(1) = 2$, and $f(0.25) = f(0.75) = \alpha$. Find α if the Composite Trapezoidal rule with $n = 4$ gives the value 1.75 for $\int_0^1 f(x) \, dx$.

10. The Midpoint rule for approximating $\int_{-1}^1 f(x) \, dx$ gives the value 12, the Composite Midpoint rule with $n = 2$ gives 5, and Composite Simpson's rule gives 6. Use the fact that $f(-1) = f(1)$ and $f(-0.5) = f(0.5) - 1$ to determine $f(-1)$, $f(-0.5)$, $f(0)$, $f(0.5)$, and $f(1)$.

11. Determine the values of n and h required to approximate

$$\int_0^2 e^{2x} \sin 3x \, dx$$

to within 10^{-4}. Use

a. Composite Trapezoidal rule.

b. Composite Simpson's rule.

c. Composite Midpoint rule.

12. Repeat Exercise 11 for the integral $\int_0^{\pi} x^2 \cos x \, dx$.

13. Determine the values of n and h required to approximate

$$\int_0^2 \frac{1}{x+4} \, dx$$

to within 10^{-5} and compute the approximation. Use

a. Composite Trapezoidal rule.

b. Composite Simpson's rule.

c. Composite Midpoint rule.

14. Repeat Exercise 13 for the integral $\int_1^2 x \ln x \, dx$.

15. Let f be defined by

$$f(x) = \begin{cases} x^3 + 1, & 0 \le x \le 0.1, \\ 1.001 + 0.03(x-0.1) + 0.3(x-0.1)^2 + 2(x-0.1)^3, & 0.1 \le x \le 0.2, \\ 1.009 + 0.15(x-0.2) + 0.9(x-0.2)^2 + 2(x-0.2)^3, & 0.2 \le x \le 0.3. \end{cases}$$

a. Investigate the continuity of the derivatives of f.

b. Use the Composite Trapezoidal rule with $n = 6$ to approximate $\int_0^{0.3} f(x) \, dx$, and estimate the error using the error bound.

c. Use the Composite Simpson's rule with $n = 6$ to approximate $\int_0^{0.3} f(x) \, dx$. Are the results more accurate than in part (b)?

16. Show that the error $E(f)$ for Composite Simpson's rule can be approximated by

$$-\frac{h^4}{180}[f'''(b) - f'''(a)].$$

[*Hint:* $\sum_{j=1}^{n/2} f^{(4)}(\xi_j)(2h)$ is a Riemann Sum for $\int_a^b f^{(4)}(x) \, dx$.]

17. **a.** Derive an estimate for $E(f)$ in the Composite Trapezoidal rule using the method in Exercise 16.

b. Repeat part (a) for the Composite Midpoint rule.

18. Use the error estimates of Exercises 16 and 17 to estimate the errors in Exercise 12.

19. Use the error estimates of Exercises 16 and 17 to estimate the errors in Exercise 14.

20. In multivariable calculus and in statistics courses it is shown that

$$\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-(1/2)(x/\sigma)^2} \, dx = 1,$$

for any positive σ. The function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(1/2)(x/\sigma)^2}$$

is the *normal density function* with *mean* $\mu = 0$ and *standard deviation* σ. The probability that a randomly chosen value described by this distribution lies in $[a, b]$ is given by $\int_a^b f(x) \, dx$. Approximate to within 10^{-5} the probability that a randomly chosen value described by this distribution will lie in

a. $[-\sigma, \sigma]$ **b.** $[-2\sigma, 2\sigma]$ **c.** $[-3\sigma, 3\sigma]$

21. Determine to within 10^{-6} the length of the graph of the ellipse with equation $4x^2 + 9y^2 = 36$.

22. A car laps a race track in 84 seconds. The speed of the car at each 6-second interval is determined by using a radar gun and is given from the beginning of the lap, in feet/second, by the entries in the following table.

Time	0	6	12	18	24	30	36	42	48	54	60	66	72	78	84
Speed	124	134	148	156	147	133	121	109	99	85	78	89	104	116	123

How long is the track?

23. A particle of mass m moving through a fluid is subjected to a viscous resistance R, which is a function of the velocity v. The relationship between the resistance R, velocity v, and time t is given by the equation

$$t = \int_{v(t_0)}^{v(t)} \frac{m}{R(u)}\, du.$$

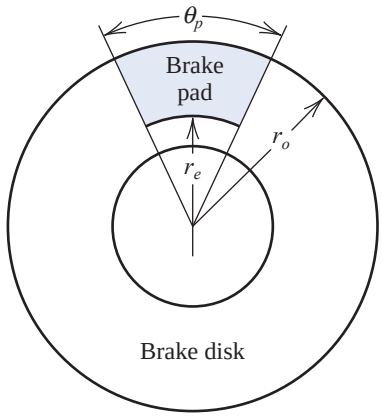
Suppose that $R(v) = -v\sqrt{v}$ for a particular fluid, where R is in newtons and v is in meters/second. If $m = 10$ kg and $v(0) = 10$ m/s, approximate the time required for the particle to slow to $v = 5$ m/s.

24. To simulate the thermal characteristics of disk brakes (see the following figure), D. A. Secrist and R. W. Hornbeck [SH] needed to approximate numerically the "area averaged lining temperature," T, of the brake pad from the equation

$$T = \frac{\displaystyle\int_{r_e}^{r_0} T(r) r\theta_p\, dr}{\displaystyle\int_{r_e}^{r_0} r\theta_p\, dr},$$

where r_e represents the radius at which the pad-disk contact begins, r_0 represents the outside radius of the pad-disk contact, θ_p represents the angle subtended by the sector brake pads, and $T(r)$ is the temperature at each point of the pad, obtained numerically from analyzing the heat equation (see Section 12.2). Suppose $r_e = 0.308$ ft, $r_0 = 0.478$ ft, $\theta_p = 0.7051$ radians, and the temperatures given in the following table have been calculated at the various points on the disk. Approximate T.

r (ft)	$T(r)$ (°F)	r (ft)	$T(r)$ (°F)	r (ft)	$T(r)$ (°F)
0.308	640	0.376	1034	0.444	1204
0.325	794	0.393	1064	0.461	1222
0.342	885	0.410	1114	0.478	1239
0.359	943	0.427	1152		



25. Find an approximation to within 10^{-4} of the value of the integral considered in the application opening this chapter:

$$\int_0^{48} \sqrt{1 + (\cos x)^2}\, dx.$$

26. The equation

$$\int_0^x \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt = 0.45$$

can be solved for x by using Newton's method with

$$f(x) = \int_0^x \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt - 0.45$$

and

$$f'(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

To evaluate f at the approximation p_k, we need a quadrature formula to approximate

$$\int_0^{p_k} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt.$$

a. Find a solution to $f(x) = 0$ accurate to within 10^{-5} using Newton's method with $p_0 = 0.5$ and the Composite Simpson's rule.

b. Repeat (a) using the Composite Trapezoidal rule in place of the Composite Simpson's rule.

4.5 Romberg Integration

In this section we will illustrate how Richardson extrapolation applied to results from the Composite Trapezoidal rule can be used to obtain high accuracy approximations with little computational cost.

In Section 4.4 we found that the Composite Trapezoidal rule has a truncation error of order $O(h^2)$. Specifically, we showed that for $h = (b-a)/n$ and $x_j = a + jh$ we have

$$\int_a^b f(x)\, dx = \frac{h}{2}\left[f(a) + 2\sum_{j=1}^{n-1} f(x_j) + f(b)\right] - \frac{(b-a)f''(\mu)}{12}h^2.$$

for some number μ in (a, b).

By an alternative method it can be shown (see [RR], pp. 136–140), that if $f \in C^\infty[a, b]$, the Composite Trapezoidal rule can also be written with an error term in the form

$$\int_a^b f(x)\, dx = \frac{h}{2}\left[f(a) + 2\sum_{j=1}^{n-1} f(x_j) + f(b)\right] + K_1h^2 + K_2h^4 + K_3h^6 + \cdots, \quad (4.33)$$

where each K_i is a constant that depends only on $f^{(2i-1)}(a)$ and $f^{(2i-1)}(b)$.

Recall from Section 4.2 that Richardson extrapolation can be performed on any approximation procedure whose truncation error is of the form

$$\sum_{j=1}^{m-1} K_j h^{\alpha_j} + O(h^{\alpha_m}),$$

for a collection of constants K_j and when $\alpha_1 < \alpha_2 < \alpha_3 < \cdots < \alpha_m$. In that section we gave demonstrations to illustrate how effective this techniques is when the approximation procedure has a truncation error with only even powers of h, that is, when the truncation error has the form.

$$\sum_{j=1}^{m-1} K_j h^{2j} + O(h^{2m}).$$

Werner Romberg (1909–2003) devised this procedure for improving the accuracy of the Trapezoidal rule by eliminating the successive terms in the asymptotic expansion in 1955.

Because the Composite Trapezoidal rule has this form, it is an obvious candidate for extrapolation. This results in a technique known as **Romberg integration**.

To approximate the integral $\int_a^b f(x)\,dx$ we use the results of the Composite Trapezoidal rule with $n = 1, 2, 4, 8, 16, \ldots$, and denote the resulting approximations, respectively, by $R_{1,1}, R_{2,1}, R_{3,1}$, etc. We then apply extrapolation in the manner given in Section 4.2, that is, we obtain $O(h^4)$ approximations $R_{2,2}, R_{3,2}, R_{4,2}$, etc., by

$$R_{k,2} = R_{k,1} + \frac{1}{3}(R_{k,1} - R_{k-1,1}), \quad \text{for } k = 2, 3, \ldots$$

Then $O(h^6)$ approximations $R_{3,3}, R_{4,3}, R_{5,3}$, etc., by

$$R_{k,3} = R_{k,2} + \frac{1}{15}(R_{k,2} - R_{k-1,2}), \quad \text{for } k = 3, 4, \ldots.$$

In general, after the appropriate $R_{k,j-1}$ approximations have been obtained, we determine the $O(h^{2j})$ approximations from

$$R_{k,j} = R_{k,j-1} + \frac{1}{4^{j-1} - 1}(R_{k,j-1} - R_{k-1,j-1}), \quad \text{for } k = j, j+1, \ldots$$

Example 1 Use the Composite Trapezoidal rule to find approximations to $\int_0^{\pi} \sin x\,dx$ with $n = 1$, 2, 4, 8, and 16. Then perform Romberg extrapolation on the results.

The Composite Trapezoidal rule for the various values of n gives the following approximations to the true value 2.

$$R_{1,1} = \frac{\pi}{2}[\sin 0 + \sin \pi] = 0;$$

$$R_{2,1} = \frac{\pi}{4}\left[\sin 0 + 2\sin\frac{\pi}{2} + \sin \pi\right] = 1.57079633;$$

$$R_{3,1} = \frac{\pi}{8}\left[\sin 0 + 2\left(\sin\frac{\pi}{4} + \sin\frac{\pi}{2} + \sin\frac{3\pi}{4}\right) + \sin \pi\right] = 1.89611890;$$

$$R_{4,1} = \frac{\pi}{16}\left[\sin 0 + 2\left(\sin\frac{\pi}{8} + \sin\frac{\pi}{4} + \cdots + \sin\frac{3\pi}{4} + \sin\frac{7\pi}{8}\right) + \sin \pi\right] = 1.97423160;$$

$$R_{5,1} = \frac{\pi}{32}\left[\sin 0 + 2\left(\sin\frac{\pi}{16} + \sin\frac{\pi}{8} + \cdots + \sin\frac{7\pi}{8} + \sin\frac{15\pi}{16}\right) + \sin \pi\right] = 1.99357034.$$

The $O(h^4)$ approximations are

$$R_{2,2} = R_{2,1} + \frac{1}{3}(R_{2,1} - R_{1,1}) = 2.09439511; \quad R_{3,2} = R_{3,1} + \frac{1}{3}(R_{3,1} - R_{2,1}) = 2.00455976;$$

$$R_{4,2} = R_{4,1} + \frac{1}{3}(R_{4,1} - R_{3,1}) = 2.00026917; \quad R_{5,2} = R_{5,1} + \frac{1}{3}(R_{5,1} - R_{4,1}) = 2.00001659;$$

The $O(h^6)$ approximations are

$$R_{3,3} = R_{3,2} + \frac{1}{15}(R_{3,2} - R_{2,2}) = 1.99857073; \quad R_{4,3} = R_{4,2} + \frac{1}{15}(R_{4,2} - R_{3,2}) = 1.99998313;$$

$$R_{5,3} = R_{5,2} + \frac{1}{15}(R_{5,2} - R_{4,2}) = 1.99999975.$$

The two $O(h^8)$ approximations are

$$R_{4,4} = R_{4,3} + \frac{1}{63}(R_{4,3} - R_{3,3}) = 2.00000555; \quad R_{5,4} = R_{5,3} + \frac{1}{63}(R_{5,3} - R_{4,3}) = 2.00000001,$$

and the final $O(h^{10})$ approximation is

$$R_{5,5} = R_{5,4} + \frac{1}{255}(R_{5,4} - R_{4,4}) = 1.99999999.$$

These results are shown in Table 4.9. ■

Table 4.9

0				
1.57079633	2.09439511			
1.89611890	2.00455976	1.99857073		
1.97423160	2.00026917	1.99998313	2.00000555	
1.99357034	2.00001659	1.99999975	2.00000001	1.99999999

Notice that when generating the approximations for the Composite Trapezoidal rule approximations in Example 1, each consecutive approximation included all the functions evaluations from the previous approximation. That is, $R_{1,1}$ used evaluations at 0 and π, $R_{2,1}$ used these evaluations and added an evaluation at the intermediate point $\pi/2$. Then $R_{3,1}$ used the evaluations of $R_{2,1}$ and added two additional intermediate ones at $\pi/4$ and $3\pi/4$. This pattern continues with $R_{4,1}$ using the same evaluations as $R_{3,1}$ but adding evaluations at the 4 intermediate points $\pi/8$, $3\pi/8$, $5\pi/8$, and $7\pi/8$, and so on.

This evaluation procedure for Composite Trapezoidal rule approximations holds for an integral on any interval $[a, b]$. In general, the Composite Trapezoidal rule denoted $R_{k+1,1}$ uses the same evaluations as $R_{k,1}$ but adds evaluations at the 2^{k-2} intermediate points. Efficient calculation of these approximations can therefore be done in a recursive manner.

To obtain the Composite Trapezoidal rule approximations for $\int_a^b f(x)\,dx$, let $h_k = (b-a)/m_k = (b-a)/2^{k-1}$. Then

$$R_{1,1} = \frac{h_1}{2}[f(a) + f(b)] = \frac{(b-a)}{2}[f(a) + f(b)];$$

and

$$R_{2,1} = \frac{h_2}{2}[f(a) + f(b) + 2f(a + h_2)].$$

By reexpressing this result for $R_{2,1}$ we can incorporate the previously determined approximation $R_{1,1}$

$$R_{2,1} = \frac{(b-a)}{4}\left[f(a) + f(b) + 2f\left(a + \frac{(b-a)}{2}\right)\right] = \frac{1}{2}[R_{1,1} + h_1 f(a + h_2)].$$

In a similar manner we can write

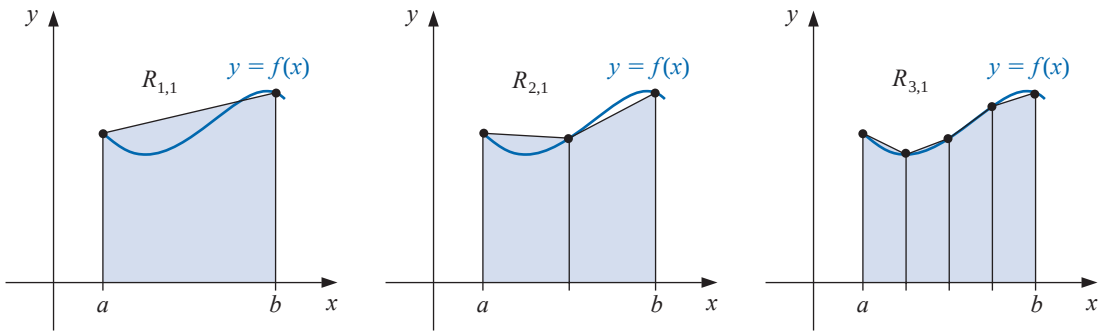
$$R_{3,1} = \frac{1}{2}\{R_{2,1} + h_2[f(a + h_3) + f(a + 3h_3)]\};$$

and, in general (see Figure 4.10 on page 216), we have

$$R_{k,1} = \frac{1}{2}\left[R_{k-1,1} + h_{k-1}\sum_{i=1}^{2^{k-2}} f\,(a + (2i-1)h_k)\right], \tag{4.34}$$

for each $k = 2, 3, \ldots, n$. (See Exercises 14 and 15.)

Figure 4.10



Extrapolation then is used to produce $O(h_k^{2j})$ approximations by

$$R_{k,j} = R_{k,j-1} + \frac{1}{4^{j-1} - 1}(R_{k,j-1} - R_{k-1,j-1}), \quad \text{for } k = j, j+1, \ldots$$

as shown in Table 4.10.

Table 4.10

k	$O(h_k^2)$	$O(h_k^4)$	$O(h_k^6)$	$O(h_k^8)$		$O(h_k^{2n})$
1	$R_{1,1}$					
2	$R_{2,1}$	$R_{2,2}$				
3	$R_{3,1}$	$R_{3,2}$	$R_{3,3}$			
4	$R_{4,1}$	$R_{4,2}$	$R_{4,3}$	$R_{4,4}$		
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	
n	$R_{n,1}$	$R_{n,2}$	$R_{n,3}$	$R_{n,4}$	$\cdots$	$R_{n,n}$

The effective method to construct the Romberg table makes use of the highest order of approximation at each step. That is, it calculates the entries row by row, in the order $R_{1,1}$, $R_{2,1}$, $R_{2,2}$, $R_{3,1}$, $R_{3,2}$, $R_{3,3}$, etc. This also permits an entire new row in the table to be calculated by doing only one additional application of the Composite Trapezoidal rule. It then uses a simple averaging on the previously calculated values to obtain the remaining entries in the row. Remember

- Calculate the Romberg table one complete row at a time.

Example 2 Add an additional extrapolation row to Table 4.10 to approximate $\int_0^{\pi} \sin x \, dx$.

Solution To obtain the additional row we need the trapezoidal approximation

$$R_{6,1} = \frac{1}{2}\left[R_{5,1} + \frac{\pi}{16}\sum_{k=1}^{2^4} \sin\frac{(2k-1)\pi}{32}\right] = 1.99839336.$$

The values in Table 4.10 give

$$R_{6,2} = R_{6,1} + \frac{1}{3}(R_{6,1} - R_{5,1}) = 1.99839336 + \frac{1}{3}(1.99839336 - 1.99357035)$$
$$= 2.00000103;$$
$$R_{6,3} = R_{6,2} + \frac{1}{15}(R_{6,2} - R_{5,2}) = 2.00000103 + \frac{1}{15}(2.00000103 - 2.00001659)$$
$$= 2.00000000;$$
$$R_{6,4} = R_{6,3} + \frac{1}{63}(R_{6,3} - R_{5,3}) = 2.00000000;$$
$$R_{6,5} = R_{6,4} + \frac{1}{255}(R_{6,4} - R_{5,4}) = 2.00000000;$$

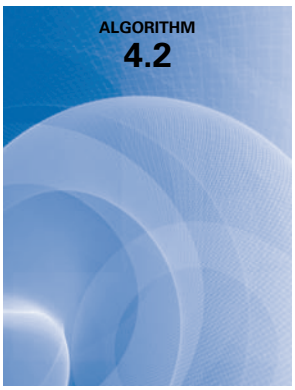
and $R_{6,6} = R_{6,5} + \frac{1}{1023}(R_{6,5} - R_{5,5}) = 2.00000000$. The new extrapolation table is shown in Table 4.11. ■

Table 4.11

0					
1.57079633	2.09439511				
1.89611890	2.00455976	1.99857073			
1.97423160	2.00026917	1.99998313	2.00000555		
1.99357034	2.00001659	1.99999975	2.00000001	1.99999999	
1.99839336	2.00000103	2.00000000	2.00000000	2.00000000	2.00000000

Notice that all the extrapolated values except for the first (in the first row of the second column) are more accurate than the best composite trapezoidal approximation (in the last row of the first column). Although there are 21 entries in Table 4.11, only the six in the left column require function evaluations since these are the only entries generated by the Composite Trapezoidal rule; the other entries are obtained by an averaging process. In fact, because of the recurrence relationship of the terms in the left column, the only function evaluations needed are those to compute the final Composite Trapezoidal rule approximation. In general, $R_{k,1}$ requires $1 + 2^{k-1}$ function evaluations, so in this case $1 + 2^5 = 33$ are needed.

Algorithm 4.2 uses the recursive procedure to find the initial Composite Trapezoidal Rule approximations and computes the results in the table row by row.



Romberg

To approximate the integral $I = \int_a^b f(x)\,dx$, select an integer $n > 0$.

INPUT endpoints a, b; integer n.

OUTPUT an array R. (*Compute R by rows; only the last 2 rows are saved in storage.*)

Step 1 Set $h = b - a$;
$R_{1,1} = \frac{h}{2}(f(a) + f(b))$.

Step 2 OUTPUT $(R_{1,1})$.

Step 3 For $i = 2, \ldots, n$ do Steps 4–8.

Step 4 Set $R_{2,1} = \frac{1}{2}\left[R_{1,1} + h\sum_{k=1}^{2^{i-2}} f(a + (k - 0.5)h)\right]$.

(*Approximation from Trapezoidal method.*)

Step 5 For $j = 2, \ldots, i$

set $R_{2,j} = R_{2,j-1} + \frac{R_{2,j-1} - R_{1,j-1}}{4^{j-1} - 1}$. (*Extrapolation.*)

Step 6 OUTPUT ($R_{2,j}$ for $j = 1, 2, \ldots, i$).

Step 7 Set $h = h/2$.

Step 8 For $j = 1, 2, \ldots, i$ set $R_{1,j} = R_{2,j}$. (*Update row 1 of R.*)

Step 9 STOP. ■

Algorithm 4.2 requires a preset integer n to determine the number of rows to be generated. We could also set an error tolerance for the approximation and generate n, within some upper bound, until consecutive diagonal entries $R_{n-1,n-1}$ and $R_{n,n}$ agree to within the tolerance. To guard against the possibility that two consecutive row elements agree with each other but not with the value of the integral being approximated, it is common to generate approximations until not only $|R_{n-1,n-1} - R_{n,n}|$ is within the tolerance, but also $|R_{n-2,n-2} - R_{n-1,n-1}|$. Although not a universal safeguard, this will ensure that two differently generated sets of approximations agree within the specified tolerance before $R_{n,n}$, is accepted as sufficiently accurate.

Romberg integration can be performed with the *Quadrature* command in the *NumericalAnalysis* subpackage of Maple's *Student* package. For example, after loading the package and defining the function and interval, the command

$$Quadrature(f(x), x = a..b, method = romberg_6, output = information)$$

produces the values shown in Table 4.11 together with the information that 6 applications of the Trapezoidal rule were used and 33 function evaluations were required.

Romberg integration applied to a function f on the interval $[a, b]$ relies on the assumption that the Composite Trapezoidal rule has an error term that can be expressed in the form of Eq. (4.33); that is, we must have $f \in C^{2k+2}[a, b]$ for the kth row to be generated. General-purpose algorithms using Romberg integration include a check at each stage to ensure that this assumption is fulfilled. These methods are known as *cautious Romberg algorithms* and are described in [Joh]. This reference also describes methods for using the Romberg technique as an adaptive procedure, similar to the adaptive Simpson's rule that will be discussed in Section 4.6.

The adjective *cautious* used in the description of a numerical method indicates that a check is incorporated to determine if the continuity hypotheses are likely to be true.

EXERCISE SET 4.5

1. Use Romberg integration to compute $R_{3,3}$ for the following integrals.

a. $\int_1^{1.5} x^2 \ln x \, dx$

b. $\int_0^1 x^2 e^{-x} \, dx$

c. $\int_0^{0.35} \frac{2}{x^2 - 4} \, dx$

d. $\int_0^{\pi/4} x^2 \sin x \, dx$

e. $\int_0^{\pi/4} e^{3x} \sin 2x \, dx$ **f.** $\int_1^{1.6} \frac{2x}{x^2 - 4} \, dx$

g. $\int_3^{3.5} \frac{x}{\sqrt{x^2 - 4}} \, dx$ **h.** $\int_0^{\pi/4} (\cos x)^2 \, dx$

2. Use Romberg integration to compute $R_{3,3}$ for the following integrals.

a. $\int_{-1}^{1} (\cos x)^2 \, dx$ **b.** $\int_{-0.75}^{0.75} x \ln(x + 1) \, dx$

c. $\int_1^4 \left((\sin x)^2 - 2x \sin x + 1\right) \, dx$ **d.** $\int_e^{2e} \frac{1}{x \ln x} \, dx$

3. Calculate $R_{4,4}$ for the integrals in Exercise 1.

4. Calculate $R_{4,4}$ for the integrals in Exercise 2.

5. Use Romberg integration to approximate the integrals in Exercise 1 to within 10^{-6}. Compute the Romberg table until either $|R_{n-1,n-1} - R_{n,n}| < 10^{-6}$, or $n = 10$. Compare your results to the exact values of the integrals.

6. Use Romberg integration to approximate the integrals in Exercise 2 to within 10^{-6}. Compute the Romberg table until either $|R_{n-1,n-1} - R_{n,n}| < 10^{-6}$, or $n = 10$. Compare your results to the exact values of the integrals.

7. Use the following data to approximate $\int_1^5 f(x) \, dx$ as accurately as possible.

x	1	2	3	4	5
$f(x)$	2.4142	2.6734	2.8974	3.0976	3.2804

8. Romberg integration is used to approximate

$$\int_0^1 \frac{x^2}{1 + x^3} \, dx.$$

If $R_{11} = 0.250$ and $R_{22} = 0.2315$, what is R_{21}?

9. Romberg integration is used to approximate

$$\int_2^3 f(x) \, dx.$$

If $f(2) = 0.51342$, $f(3) = 0.36788$, $R_{31} = 0.43687$, and $R_{33} = 0.43662$, find $f(2.5)$.

10. Romberg integration for approximating $\int_0^1 f(x) \, dx$ gives $R_{11} = 4$ and $R_{22} = 5$. Find $f(1/2)$.

11. Romberg integration for approximating $\int_a^b f(x) \, dx$ gives $R_{11} = 8$, $R_{22} = 16/3$, and $R_{33} = 208/45$. Find R_{31}.

12. Use Romberg integration to compute the following approximations to

$$\int_0^{48} \sqrt{1 + (\cos x)^2} \, dx.$$

[*Note:* The results in this exercise are most interesting if you are using a device with between seven- and nine-digit arithmetic.]

a. Determine $R_{1,1}$, $R_{2,1}$, $R_{3,1}$, $R_{4,1}$, and $R_{5,1}$, and use these approximations to predict the value of the integral.

b. Determine $R_{2,2}$, $R_{3,3}$, $R_{4,4}$, and $R_{5,5}$, and modify your prediction.

c. Determine $R_{6,1}$, $R_{6,2}$, $R_{6,3}$, $R_{6,4}$, $R_{6,5}$, and $R_{6,6}$, and modify your prediction.

d. Determine $R_{7,7}$, $R_{8,8}$, $R_{9,9}$, and $R_{10,10}$, and make a final prediction.

e. Explain why this integral causes difficulty with Romberg integration and how it can be reformulated to more easily determine an accurate approximation.

13. Show that the approximation obtained from $R_{k,2}$ is the same as that given by the Composite Simpson's rule described in Theorem 4.4 with $h = h_k$.

14. Show that, for any k,

$$\sum_{i=1}^{2^{k-1}-1} f\left(a+\frac{i}{2}h_{k-1}\right) = \sum_{i=1}^{2^{k-2}} f\left(a+\left(i-\frac{1}{2}\right)h_{k-1}\right) + \sum_{i=1}^{2^{k-2}-1} f(a+ih_{k-1}).$$

15. Use the result of Exercise 14 to verify Eq. (4.34); that is, show that for all k,

$$R_{k,1} = \frac{1}{2}\left[R_{k-1,1} + h_{k-1}\sum_{i=1}^{2^{k-2}} f\left(a+\left(i-\frac{1}{2}\right)h_{k-1}\right)\right].$$

16. In Exercise 26 of Section 1.1, a Maclaurin series was integrated to approximate erf(1), where erf(x) is the normal distribution error function defined by

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\,dt.$$

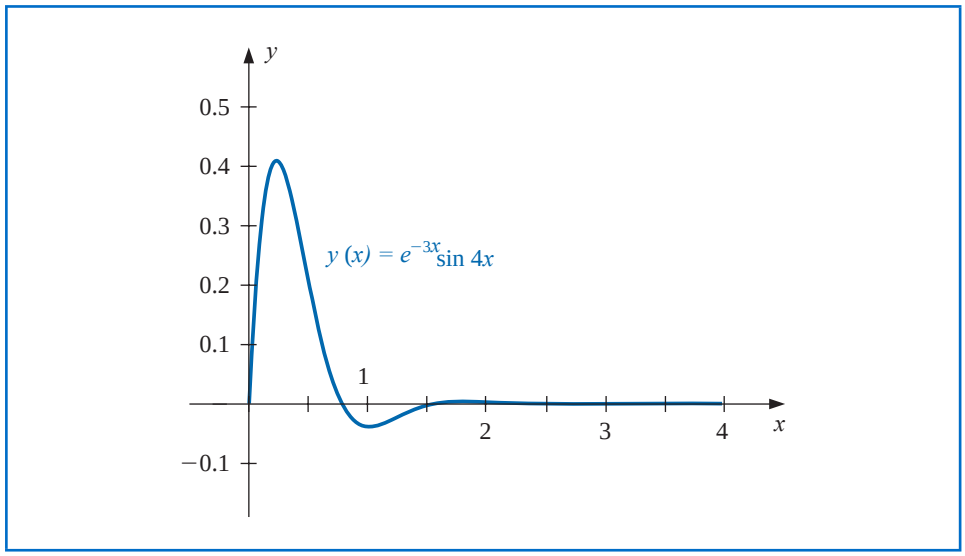
Approximate erf(1) to within 10^{-7}.

4.6 Adaptive Quadrature Methods

The composite formulas are very effective in most situations, but they suffer occasionally because they require the use of equally-spaced nodes. This is inappropriate when integrating a function on an interval that contains both regions with large functional variation and regions with small functional variation.

Illustration The unique solution to the differential equation $y'' + 6y' + 25 = 0$ that additionally satisfies $y(0) = 0$ and $y'(0) = 4$ is $y(x) = e^{-3x}\sin 4x$. Functions of this type are common in mechanical engineering because they describe certain features of spring and shock absorber systems, and in electrical engineering because they are common solutions to elementary circuit problems. The graph of $y(x)$ for x in the interval $[0, 4]$ is shown in Figure 4.11.

Figure 4.11



Suppose that we need the integral of $y(x)$ on $[0, 4]$. The graph indicates that the integral on $[3, 4]$ must be very close to 0, and on $[2, 3]$ would also not be expected to be large. However, on $[0, 2]$ there is significant variation of the function and it is not at all clear what the integral is on this interval. This is an example of a situation where composite integration would be inappropriate. A very low order method could be used on $[2, 4]$, but a higher-order method would be necessary on $[0, 2]$. □

The question we will consider in this section is:

- How can we determine what technique should be applied on various portions of the interval of integration, and how accurate can we expect the final approximation to be?

We will see that under quite reasonable conditions we can answer this question and also determine approximations that satisfy given accuracy requirements.

If the approximation error for an integral on a given interval is to be evenly distributed, a smaller step size is needed for the large-variation regions than for those with less variation. An efficient technique for this type of problem should predict the amount of functional variation and adapt the step size as necessary. These methods are called **Adaptive quadrature methods**. Adaptive methods are particularly popular for inclusion in professional software packages because, in addition to being efficient, they generally provide approximations that are within a given specified tolerance.

In this section we consider an Adaptive quadrature method and see how it can be used to reduce approximation error and also to predict an error estimate for the approximation that does not rely on knowledge of higher derivatives of the function. The method we discuss is based on the Composite Simpson's rule, but the technique is easily modified to use other composite procedures.

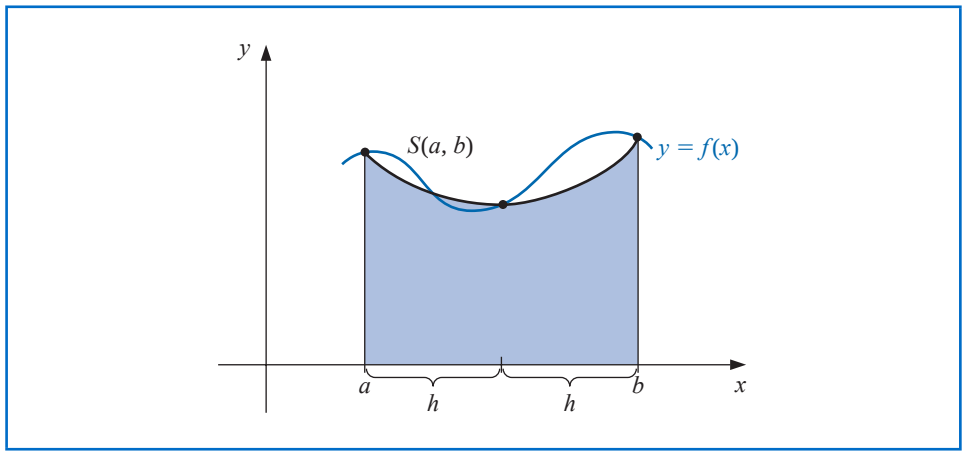
Suppose that we want to approximate $\int_a^b f(x)\,dx$ to within a specified tolerance $\varepsilon > 0$. The first step is to apply Simpson's rule with step size $h = (b - a)/2$. This produces (see Figure 4.12)

$$\int_a^b f(x)\,dx = S(a, b) - \frac{h^5}{90} f^{(4)}(\xi), \quad \text{for some } \xi \text{ in } (a, b), \tag{4.35}$$

where we denote the Simpson's rule approximation on $[a, b]$ by

$$S(a, b) = \frac{h}{3}[f(a) + 4f(a + h) + f(b)].$$

Figure 4.12



The next step is to determine an accuracy approximation that does not require $f^{(4)}(\xi)$. To do this, we apply the Composite Simpson's rule with $n = 4$ and step size $(b-a)/4 = h/2$, giving

$$\int_a^b f(x)\,dx = \frac{h}{6}\left[f(a) + 4f\left(a + \frac{h}{2}\right) + 2f(a+h) + 4f\left(a + \frac{3h}{2}\right) + f(b)\right] - \left(\frac{h}{2}\right)^4 \frac{(b-a)}{180} f^{(4)}(\tilde{\xi}), \tag{4.36}$$

for some $\tilde{\xi}$ in (a, b). To simplify notation, let

$$S\left(a, \frac{a+b}{2}\right) = \frac{h}{6}\left[f(a) + 4f\left(a + \frac{h}{2}\right) + f(a+h)\right]$$

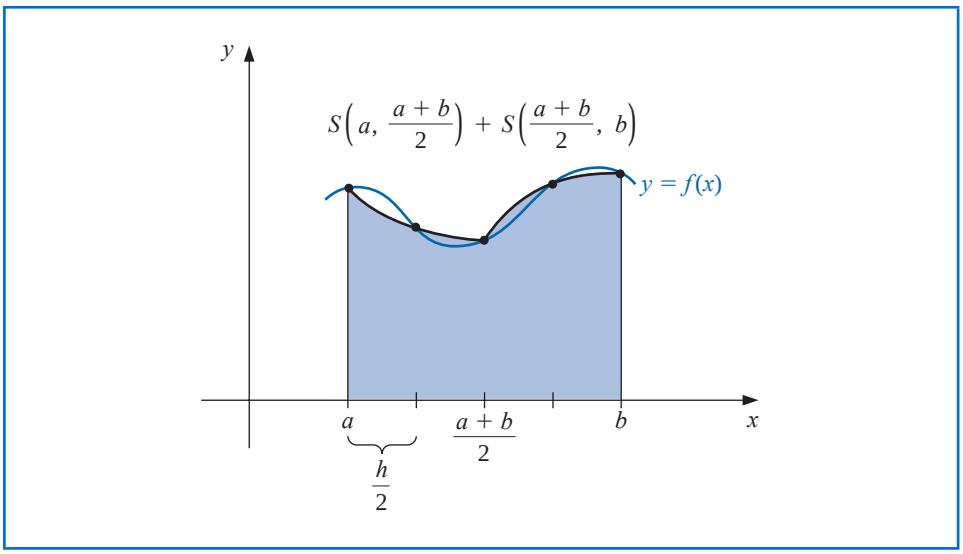
and

$$S\left(\frac{a+b}{2}, b\right) = \frac{h}{6}\left[f(a+h) + 4f\left(a + \frac{3h}{2}\right) + f(b)\right].$$

Then Eq. (4.36) can be rewritten (see Figure 4.13) as

$$\int_a^b f(x)\,dx = S\left(a, \frac{a+b}{2}\right) + S\left(\frac{a+b}{2}, b\right) - \frac{1}{16}\left(\frac{h^5}{90}\right) f^{(4)}(\tilde{\xi}). \tag{4.37}$$

Figure 4.13



The error estimation is derived by assuming that $\xi \approx \tilde{\xi}$ or, more precisely, that $f^{(4)}(\xi) \approx f^{(4)}(\tilde{\xi})$, and the success of the technique depends on the accuracy of this assumption. If it is accurate, then equating the integrals in Eqs. (4.35) and (4.37) gives

$$S\left(a, \frac{a+b}{2}\right) + S\left(\frac{a+b}{2}, b\right) - \frac{1}{16}\left(\frac{h^5}{90}\right) f^{(4)}(\xi) \approx S(a,b) - \frac{h^5}{90} f^{(4)}(\xi),$$

so

$$\frac{h^5}{90} f^{(4)}(\xi) \approx \frac{16}{15}\left[S(a,b) - S\left(a, \frac{a+b}{2}\right) - S\left(\frac{a+b}{2}, b\right)\right].$$

Using this estimate in Eq. (4.37) produces the error estimation

$$\left|\int_a^b f(x)\,dx - S\left(a, \frac{a+b}{2}\right) - S\left(\frac{a+b}{2}, b\right)\right| \approx \frac{1}{16}\left(\frac{h^5}{90}\right) f^{(4)}(\xi)$$
$$\approx \frac{1}{15}\left|S(a,b) - S\left(a, \frac{a+b}{2}\right) - S\left(\frac{a+b}{2}, b\right)\right|.$$

This implies that $S(a,(a+b)/2)+S((a+b)/2,b)$ approximates $\int_a^b f(x)\,dx$ about 15 times better than it agrees with the computed value $S(a,b)$. Thus, if

$$\left|S(a,b) - S\left(a, \frac{a+b}{2}\right) - S\left(\frac{a+b}{2}, b\right)\right| < 15\varepsilon, \tag{4.38}$$

we expect to have

$$\left|\int_a^b f(x)\,dx - S\left(a, \frac{a+b}{2}\right) - S\left(\frac{a+b}{2}, b\right)\right| < \varepsilon, \tag{4.39}$$

and

$$S\left(a, \frac{a+b}{2}\right) + S\left(\frac{a+b}{2}, b\right)$$

is assumed to be a sufficiently accurate approximation to $\int_a^b f(x)\,dx$.

Example 1 Check the accuracy of the error estimate given in (4.38) and (4.39) when applied to the integral

$$\int_0^{\pi/2} \sin x\,dx = 1.$$

by comparing

$$\frac{1}{15}\left|S\left(0, \frac{\pi}{2}\right) - S\left(0, \frac{\pi}{4}\right) - S\left(\frac{\pi}{4}, \frac{\pi}{2}\right)\right| \quad \text{to} \quad \left|\int_0^{\pi/2} \sin x\,dx - S\left(0, \frac{\pi}{4}\right) - S\left(\frac{\pi}{4}, \frac{\pi}{2}\right)\right|.$$

Solution We have

$$S\left(0, \frac{\pi}{2}\right) = \frac{\pi/4}{3}\left[\sin 0 + 4\sin\frac{\pi}{4} + \sin\frac{\pi}{2}\right] = \frac{\pi}{12}(2\sqrt{2}+1) = 1.002279878$$

and

$$S\left(0, \frac{\pi}{4}\right) + S\left(\frac{\pi}{4}, \frac{\pi}{2}\right) = \frac{\pi/8}{3}\left[\sin 0 + 4\sin\frac{\pi}{8} + 2\sin\frac{\pi}{4} + 4\sin\frac{3\pi}{8} + \sin\frac{\pi}{2}\right]$$
$$= 1.000134585.$$

So

$$\left|S\left(0, \frac{\pi}{2}\right) - S\left(0, \frac{\pi}{4}\right) - S\left(\frac{\pi}{4}, \frac{\pi}{2}\right)\right| = |1.002279878 - 1.000134585| = 0.002145293.$$

The estimate for the error obtained when using $S(a,(a+b))+S((a+b),b)$ to approximate $\int_a^b f(x)\,dx$ is consequently

$$\frac{1}{15}\left|S\left(0, \frac{\pi}{2}\right) - S\left(0, \frac{\pi}{4}\right) - S\left(\frac{\pi}{4}, \frac{\pi}{2}\right)\right| = 0.000143020,$$

which closely approximates the actual error

$$\left| \int_0^{\pi/2} \sin x \, dx - 1.000134585 \right| = 0.000134585,$$

even though $D_x^4 \sin x = \sin x$ varies significantly in the interval $(0, \pi/2)$. ■

When the approximations in (4.38) differ by more than 15ε, we can apply the Simpson's rule technique individually to the subintervals $[a, (a+b)/2]$ and $[(a+b)/2, b]$. Then we use the error estimation procedure to determine if the approximation to the integral on each subinterval is within a tolerance of $\varepsilon/2$. If so, we sum the approximations to produce an approximation to $\int_a^b f(x)\,dx$ within the tolerance ε.

If the approximation on one of the subintervals fails to be within the tolerance $\varepsilon/2$, then that subinterval is itself subdivided, and the procedure is reapplied to the two subintervals to determine if the approximation on each subinterval is accurate to within $\varepsilon/4$. This halving procedure is continued until each portion is within the required tolerance.

Problems can be constructed for which this tolerance will never be met, but the technique is usually successful, because each subdivision typically increases the accuracy of the approximation by a factor of 16 while requiring an increased accuracy factor of only 2.

Algorithm 4.3 details this Adaptive quadrature procedure for Simpson's rule, although some technical difficulties arise that require the implementation to differ slightly from the preceding discussion. For example, in Step 1 the tolerance has been set at 10ε rather than the 15ε figure in Inequality (4.38). This bound is chosen conservatively to compensate for error in the assumption $f^{(4)}(\xi) \approx f^{(4)}(\tilde{\xi})$. In problems where $f^{(4)}$ is known to be widely varying, this bound should be decreased even further.

It is a good idea to include a margin of safety when it is impossible to verify accuracy assumptions.

The procedure listed in the algorithm first approximates the integral on the leftmost subinterval in a subdivision. This requires the efficient storing and recalling of previously computed functional evaluations for the nodes in the right half subintervals. Steps 3, 4, and 5 contain a stacking procedure with an indicator to keep track of the data that will be required for calculating the approximation on the subinterval immediately adjacent and to the right of the subinterval on which the approximation is being generated. The method is easier to implement using a recursive programming language.

Adaptive Quadrature

To approximate the integral $I = \int_a^b f(x)\,dx$ to within a given tolerance:

INPUT endpoints a, b; tolerance TOL; limit N to number of levels.

OUTPUT approximation APP or message that N is exceeded.

Step 1 Set $APP = 0$;
$i = 1$;
$TOL_i = 10\,TOL$;
$a_i = a$;
$h_i = (b-a)/2$;
$FA_i = f(a)$;
$FC_i = f(a + h_i)$;
$FB_i = f(b)$;
$S_i = h_i(FA_i + 4FC_i + FB_i)/3$; (*Approximation from Simpson's method for entire interval.*)
$L_i = 1$.

Step 2 While $i > 0$ do Steps 3–5.

Step 3 Set $FD = f(a_i + h_i/2)$;
$FE = f(a_i + 3h_i/2)$;
$S1 = h_i(FA_i + 4FD + FC_i)/6$; (*Approximations from Simpson's method for halves of subintervals.*)
$S2 = h_i(FC_i + 4FE + FB_i)/6$;
$v_1 = a_i$; (*Save data at this level.*)
$v_2 = FA_i$;
$v_3 = FC_i$;
$v_4 = FB_i$;
$v_5 = h_i$;
$v_6 = TOL_i$;
$v_7 = S_i$;
$v_8 = L_i$.

Step 4 Set $i = i - 1$. (*Delete the level.*)

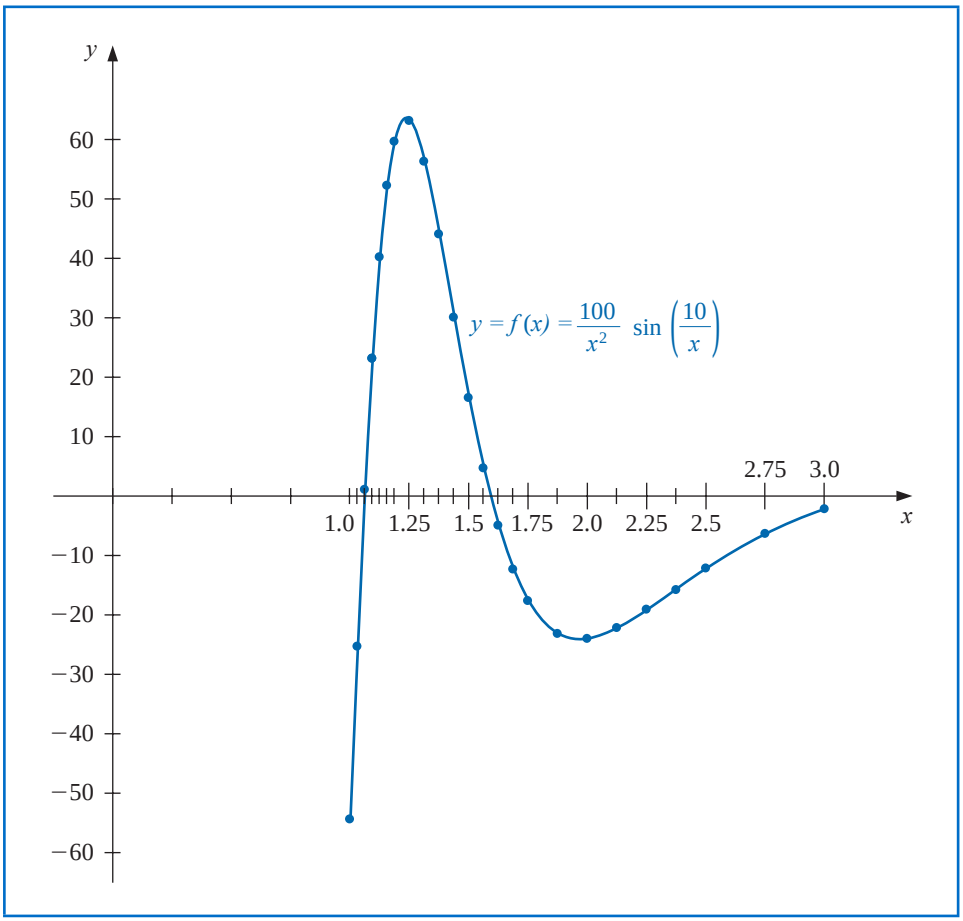
Step 5 If $|S1 + S2 - v_7| < v_6$
then set $APP = APP + (S1 + S2)$
else
if $(v_8 \geq N)$
then
OUTPUT ('LEVEL EXCEEDED'); (*Procedure fails.*)
STOP.
else (*Add one level.*)
set $i = i + 1$; (*Data for right half subinterval.*)
$a_i = v_1 + v_5$;
$FA_i = v_3$;
$FC_i = FE$;
$FB_i = v_4$;
$h_i = v_5/2$;
$TOL_i = v_6/2$;
$S_i = S2$;
$L_i = v_8 + 1$;
set $i = i + 1$; (*Data for left half subinterval.*)
$a_i = v_1$;
$FA_i = v_2$;
$FC_i = FD$;
$FB_i = v_3$;
$h_i = h_{i-1}$;
$TOL_i = TOL_{i-1}$;
$S_i = S1$;
$L_i = L_{i-1}$.

Step 6 OUTPUT (APP); (*APP approximates I to within TOL.*)
STOP. ■

Illustration The graph of the function $f(x) = (100/x^2)\sin(10/x)$ for x in $[1, 3]$ is shown in Figure 4.14. Using the Adaptive Quadrature Algorithm 4.3 with tolerance 10^{-4} to approximate $\int_1^3 f(x)\,dx$ produces -1.426014, a result that is accurate to within 1.1×10^{-5}. The approximation required that Simpson's rule with $n = 4$ be performed on the 23 subintervals whose

endpoints are shown on the horizontal axis in Figure 4.14. The total number of functional evaluations required for this approximation is 93.

Figure 4.14



The largest value of h for which the standard Composite Simpson's rule gives 10^{-4} accuracy is $h = 1/88$. This application requires 177 function evaluations, nearly twice as many as Adaptive quadrature. □

Adaptive quadrature can be performed with the *Quadrature* command in the *NumericalAnalysis* subpackage of Maple's *Student* package. In this situation the option *adaptive* = *true* is used. For example, to produce the values in the Illustration we first load the package and define the function and interval with

$$f := x \to \frac{100}{x^2} \cdot \sin\left(\frac{10}{x}\right);\ a := 1.0;\ b := 3.0$$

Then give the *NumericalAnalysis* command

$$Quadrature(f(x), x = a..b, adaptive = true, method = [simpson, 10^{-4}], output = information)$$

This produces the approximation -1.42601481 and a table that lists all the intervals on which Simpson's rule was employed and whether the appropriate tolerance was satisfied (indicated by the word `PASS`) or was not satisfied (indicated by the word `fail`). It also gives what Maple thinks is the correct value of the integral to the decimal places listed, in this case -1.42602476. Then it gives the absolute and relative errors, 9.946×10^{-6} and 6.975×10^{-4}, respectively, assuming that its correct value is accurate.

EXERCISE SET 4.6

1. Compute the Simpson's rule approximations $S(a, b)$, $S(a, (a + b)/2)$, and $S((a + b)/2, b)$ for the following integrals, and verify the estimate given in the approximation formula.

a. $\displaystyle\int_1^{1.5} x^2 \ln x \, dx$

b. $\displaystyle\int_0^1 x^2 e^{-x} \, dx$

c. $\displaystyle\int_0^{0.35} \frac{2}{x^2 - 4} \, dx$

d. $\displaystyle\int_0^{\pi/4} x^2 \sin x \, dx$

e. $\displaystyle\int_0^{\pi/4} e^{3x} \sin 2x \, dx$

f. $\displaystyle\int_1^{1.6} \frac{2x}{x^2 - 4} \, dx$

g. $\displaystyle\int_3^{3.5} \frac{x}{\sqrt{x^2 - 4}} \, dx$

h. $\displaystyle\int_0^{\pi/4} (\cos x)^2 \, dx$

2. Use Adaptive quadrature to find approximations to within 10^{-3} for the integrals in Exercise 1. Do not use a computer program to generate these results.

3. Use Adaptive quadrature to approximate the following integrals to within 10^{-5}.

a. $\displaystyle\int_1^3 e^{2x} \sin 3x \, dx$

b. $\displaystyle\int_1^3 e^{3x} \sin 2x \, dx$

c. $\displaystyle\int_0^5 \left(2x \cos(2x) - (x - 2)^2\right) \, dx$

d. $\displaystyle\int_0^5 \left(4x \cos(2x) - (x - 2)^2\right) \, dx$

4. Use Adaptive quadrature to approximate the following integrals to within 10^{-5}.

a. $\displaystyle\int_0^{\pi} (\sin x + \cos x) \, dx$

b. $\displaystyle\int_1^2 (x + \sin 4x) \, dx$

c. $\displaystyle\int_{-1}^1 x \sin 4x \, dx$

d. $\displaystyle\int_0^{\pi/2} (6 \cos 4x + 4 \sin 6x) e^x \, dx$

5. Use Simpson's Composite rule with $n = 4, 6, 8, \ldots$, until successive approximations to the following integrals agree to within 10^{-6}. Determine the number of nodes required. Use the Adaptive Quadrature Algorithm to approximate the integral to within 10^{-6}, and count the number of nodes. Did Adaptive quadrature produce any improvement?

a. $\displaystyle\int_0^{\pi} x \cos x^2 \, dx$

b. $\displaystyle\int_0^{\pi} x \sin x^2 \, dx$

c. $\displaystyle\int_0^{\pi} x^2 \cos x \, dx$

d. $\displaystyle\int_0^{\pi} x^2 \sin x \, dx$

6. Sketch the graphs of $\sin(1/x)$ and $\cos(1/x)$ on $[0.1, 2]$. Use Adaptive quadrature to approximate the following integrals to within 10^{-3}.

a. $\displaystyle\int_{0.1}^2 \sin \frac{1}{x} \, dx$

b. $\displaystyle\int_{0.1}^2 \cos \frac{1}{x} \, dx$

7. The differential equation

$$mu''(t) + ku(t) = F_0 \cos \omega t$$

describes a spring-mass system with mass m, spring constant k, and no applied damping. The term $F_0 \cos \omega t$ describes a periodic external force applied to the system. The solution to the equation when the system is initially at rest ($u'(0) = u(0) = 0$) is

$$u(t) = \frac{F_0}{m(\omega_0^2 - \omega^2)}(\cos \omega t - \cos \omega_0 t), \quad \text{where} \quad \omega_0 = \sqrt{\frac{k}{m}} \neq \omega.$$

Sketch the graph of u when $m = 1$, $k = 9$, $F_0 = 1$, $\omega = 2$, and $t \in [0, 2\pi]$. Approximate $\int_0^{2\pi} u(t)\,dt$ to within 10^{-4}.

8. If the term $cu'(t)$ is added to the left side of the motion equation in Exercise 7, the resulting differential equation describes a spring-mass system that is damped with damping constant $c \neq 0$. The solution to this equation when the system is initially at rest is

$$u(t) = c_1 e^{r_1 t} + c_2 e^{r_2 t} + \frac{F_0}{c^2\omega^2 + m^2(\omega_0^2 - \omega^2)^2}\left(c\omega \sin \omega t + m\left(\omega_0^2 - \omega^2\right)\cos \omega t\right),$$

where

$$r_1 = \frac{-c + \sqrt{c^2 - 4\omega_0^2 m^2}}{2m} \quad \text{and} \quad r_2 = \frac{-c - \sqrt{c^2 - 4\omega_0^2 m^2}}{2m}.$$

a. Let $m = 1$, $k = 9$, $F_0 = 1$, $c = 10$, and $\omega = 2$. Find the values of c_1 and c_2 so that $u(0) = u'(0) = 0$.

b. Sketch the graph of $u(t)$ for $t \in [0, 2\pi]$ and approximate $\int_0^{2\pi} u(t)\,dt$ to within 10^{-4}.

9. Let $T(a, b)$ and $T(a, \frac{a+b}{2}) + T(\frac{a+b}{2}, b)$ be the single and double applications of the Trapezoidal rule to $\int_a^b f(x)\,dx$. Derive the relationship between

$$\left|T(a, b) - T\left(a, \frac{a+b}{2}\right) - T\left(\frac{a+b}{2}, b\right)\right|$$

and

$$\left|\int_a^b f(x)\,dx - T\left(a, \frac{a+b}{2}\right) - T\left(\frac{a+b}{2}, b\right)\right|.$$

10. The study of light diffraction at a rectangular aperture involves the Fresnel integrals

$$c(t) = \int_0^t \cos \frac{\pi}{2}\omega^2\,d\omega \quad \text{and} \quad s(t) = \int_0^t \sin \frac{\pi}{2}\omega^2\,d\omega.$$

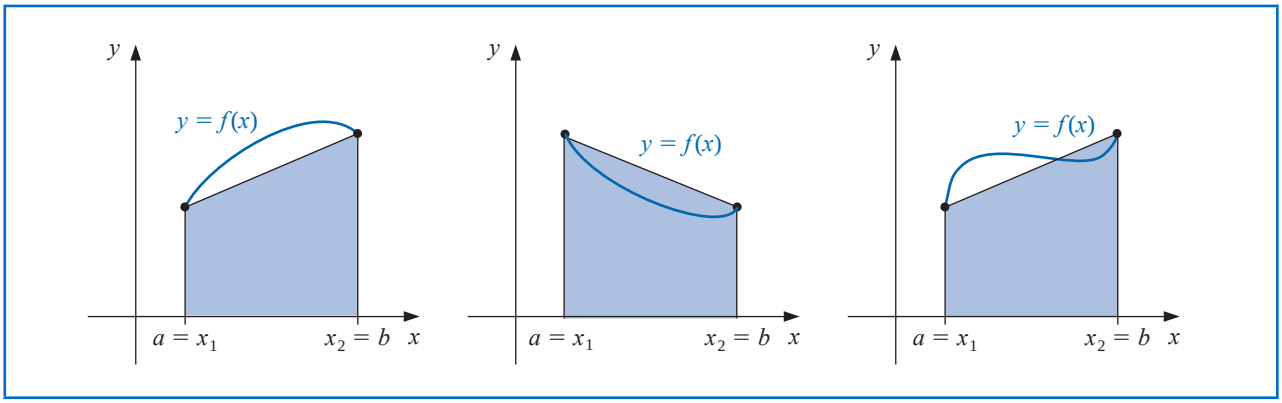
Construct a table of values for $c(t)$ and $s(t)$ that is accurate to within 10^{-4} for values of $t = 0.1, 0.2, \ldots, 1.0$.

4.7 Gaussian Quadrature

The Newton-Cotes formulas in Section 4.3 were derived by integrating interpolating polynomials. The error term in the interpolating polynomial of degree n involves the $(n + 1)$st derivative of the function being approximated, so a Newton-Cotes formula is exact when approximating the integral of any polynomial of degree less than or equal to n.

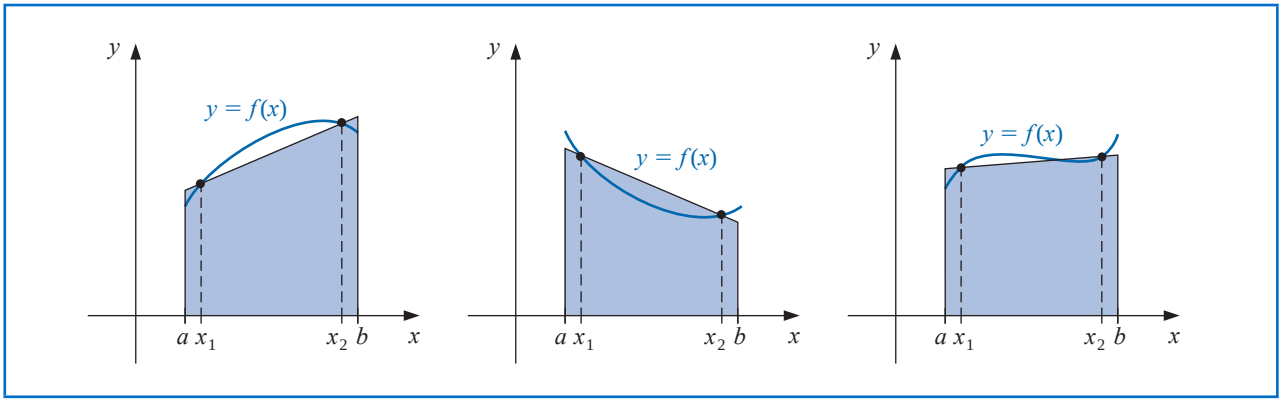
All the Newton-Cotes formulas use values of the function at equally-spaced points. This restriction is convenient when the formulas are combined to form the composite rules we considered in Section 4.4, but it can significantly decrease the accuracy of the approximation. Consider, for example, the Trapezoidal rule applied to determine the integrals of the functions whose graphs are shown in Figure 4.15.

Figure 4.15



The Trapezoidal rule approximates the integral of the function by integrating the linear function that joins the endpoints of the graph of the function. But this is not likely the best line for approximating the integral. Lines such as those shown in Figure 4.16 would likely give much better approximations in most cases.

Figure 4.16



Gauss demonstrated his method of efficient numerical integration in a paper that was presented to the Göttingen Society in 1814. He let the nodes as well as the coefficients of the function evaluations be parameters in the summation formula and found the optimal placement of the nodes. Goldstine [Golds], pp 224–232, has an interesting description of his development.

Gaussian quadrature chooses the points for evaluation in an optimal, rather than equally-spaced, way. The nodes $x_1, x_2, \ldots, x_n$ in the interval $[a, b]$ and coefficients $c_1, c_2, \ldots, c_n$, are chosen to minimize the expected error obtained in the approximation

$$\int_a^b f(x)\,dx \approx \sum_{i=1}^{n} c_i f(x_i).$$

To measure this accuracy, we assume that the best choice of these values produces the exact result for the largest class of polynomials, that is, the choice that gives the greatest degree of precision.

The coefficients $c_1, c_2, \ldots, c_n$ in the approximation formula are arbitrary, and the nodes $x_1, x_2, \ldots, x_n$ are restricted only by the fact that they must lie in $[a, b]$, the interval of integration. This gives us $2n$ parameters to choose. If the coefficients of a polynomial are

considered parameters, the class of polynomials of degree at most $2n - 1$ also contains $2n$ parameters. This, then, is the largest class of polynomials for which it is reasonable to expect a formula to be exact. With the proper choice of the values and constants, exactness on this set can be obtained.

To illustrate the procedure for choosing the appropriate parameters, we will show how to select the coefficients and nodes when $n = 2$ and the interval of integration is $[-1, 1]$. We will then discuss the more general situation for an arbitrary choice of nodes and coefficients and show how the technique is modified when integrating over an arbitrary interval.

Suppose we want to determine c_1, c_2, x_1, and x_2 so that the integration formula

$$\int_{-1}^{1} f(x)\,dx \approx c_1 f(x_1) + c_2 f(x_2)$$

gives the exact result whenever $f(x)$ is a polynomial of degree $2(2) - 1 = 3$ or less, that is, when

$$f(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3,$$

for some collection of constants, a_0, a_1, a_2, and a_3. Because

$$\int (a_0 + a_1 x + a_2 x^2 + a_3 x^3)\,dx = a_0 \int 1\,dx + a_1 \int x\,dx + a_2 \int x^2\,dx + a_3 \int x^3\,dx,$$

this is equivalent to showing that the formula gives exact results when $f(x)$ is 1, x, x^2, and x^3. Hence, we need c_1, c_2, x_1, and x_2, so that

$$c_1 \cdot 1 + c_2 \cdot 1 = \int_{-1}^{1} 1\,dx = 2, \qquad c_1 \cdot x_1 + c_2 \cdot x_2 = \int_{-1}^{1} x\,dx = 0,$$

$$c_1 \cdot x_1^2 + c_2 \cdot x_2^2 = \int_{-1}^{1} x^2\,dx = \frac{2}{3}, \quad \text{and} \quad c_1 \cdot x_1^3 + c_2 \cdot x_2^3 = \int_{-1}^{1} x^3\,dx = 0.$$

A little algebra shows that this system of equations has the unique solution

$$c_1 = 1, \quad c_2 = 1, \quad x_1 = -\frac{\sqrt{3}}{3}, \quad \text{and} \quad x_2 = \frac{\sqrt{3}}{3},$$

which gives the approximation formula

$$\int_{-1}^{1} f(x)\,dx \approx f\left(\frac{-\sqrt{3}}{3}\right) + f\left(\frac{\sqrt{3}}{3}\right). \tag{4.40}$$

This formula has degree of precision 3, that is, it produces the exact result for every polynomial of degree 3 or less.

Legendre Polynomials

The technique we have described could be used to determine the nodes and coefficients for formulas that give exact results for higher-degree polynomials, but an alternative method obtains them more easily. In Sections 8.2 and 8.3 we will consider various collections of orthogonal polynomials, functions that have the property that a particular definite integral of the product of any two of them is 0. The set that is relevant to our problem is the Legendre polynomials, a collection $\{P_0(x), P_1(x), \ldots, P_n(x), \ldots,\}$ with properties:

(1) For each n, $P_n(x)$ is a monic polynomial of degree n.

(2) $\int_{-1}^{1} P(x)P_n(x)\,dx = 0$ whenever $P(x)$ is a polynomial of degree less than n.

Recall that *monic* polynomials have leading coefficient 1.

The first few Legendre polynomials are

$$P_0(x) = 1, \quad P_1(x) = x, \quad P_2(x) = x^2 - \frac{1}{3},$$

$$P_3(x) = x^3 - \frac{3}{5}x, \quad \text{and} \quad P_4(x) = x^4 - \frac{6}{7}x^2 + \frac{3}{35}.$$

Adrien-Marie Legendre (1752–1833) introduced this set of polynomials in 1785. He had numerous priority disputes with Gauss, primarily due to Gauss' failure to publish many of his original results until long after he had discovered them.

The roots of these polynomials are distinct, lie in the interval $(-1, 1)$, have a symmetry with respect to the origin, and, most importantly, are the correct choice for determining the parameters that give us the nodes and coefficients for our quadrature method.

The nodes $x_1, x_2, \ldots, x_n$ needed to produce an integral approximation formula that gives exact results for any polynomial of degree less than $2n$ are the roots of the nth-degree Legendre polynomial. This is established by the following result.

Theorem 4.7 Suppose that $x_1, x_2, \ldots, x_n$ are the roots of the nth Legendre polynomial $P_n(x)$ and that for each $i = 1, 2, \ldots, n$, the numbers c_i are defined by

$$c_i = \int_{-1}^{1} \prod_{\substack{j=1 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j}\,dx.$$

If $P(x)$ is any polynomial of degree less than $2n$, then

$$\int_{-1}^{1} P(x)\,dx = \sum_{i=1}^{n} c_i P(x_i).$$

■

Proof Let us first consider the situation for a polynomial $P(x)$ of degree less than n. Rewrite $P(x)$ in terms of $(n-1)$st Lagrange coefficient polynomials with nodes at the roots of the nth Legendre polynomial $P_n(x)$. The error term for this representation involves the nth derivative of $P(x)$. Since $P(x)$ is of degree less than n, the nth derivative of $P(x)$ is 0, and this representation of is exact. So

$$P(x) = \sum_{i=1}^{n} P(x_i)L_i(x) = \sum_{i=1}^{n} \prod_{\substack{j=1 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j} P(x_i)$$

and

$$\int_{-1}^{1} P(x)\,dx = \int_{-1}^{1} \left[\sum_{i=1}^{n} \prod_{\substack{j=1 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j} P(x_i) \right] dx$$

$$= \sum_{i=1}^{n} \left[\int_{-1}^{1} \prod_{\substack{j=1 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j}\,dx \right] P(x_i) = \sum_{i=1}^{n} c_i P(x_i).$$

Hence the result is true for polynomials of degree less than n.

Now consider a polynomial $P(x)$ of degree at least n but less than $2n$. Divide $P(x)$ by the nth Legendre polynomial $P_n(x)$. This gives two polynomials $Q(x)$ and $R(x)$, each of degree less than n, with

$$P(x) = Q(x)P_n(x) + R(x).$$

Note that x_i is a root of $P_n(x)$ for each $i = 1, 2, \ldots, n$, so we have

$$P(x_i) = Q(x_i)P_n(x_i) + R(x_i) = R(x_i).$$

We now invoke the unique power of the Legendre polynomials. First, the degree of the polynomial $Q(x)$ is less than n, so (by Legendre property **(2)**),

$$\int_{-1}^{1} Q(x)P_n(x)\,dx = 0.$$

Then, since $R(x)$ is a polynomial of degree less than n, the opening argument implies that

$$\int_{-1}^{1} R(x)\,dx = \sum_{i=1}^{n} c_i R(x_i).$$

Putting these facts together verifies that the formula is exact for the polynomial $P(x)$:

$$\int_{-1}^{1} P(x)\,dx = \int_{-1}^{1} [Q(x)P_n(x) + R(x)]\,dx = \int_{-1}^{1} R(x)\,dx = \sum_{i=1}^{n} c_i R(x_i) = \sum_{i=1}^{n} c_i P(x_i).$$

■ ■ ■

The constants c_i needed for the quadrature rule can be generated from the equation in Theorem 4.7, but both these constants and the roots of the Legendre polynomials are extensively tabulated. Table 4.12 lists these values for $n = 2, 3, 4$, and 5.

Table 4.12

n	Roots $r_{n,i}$	Coefficients $c_{n,i}$
2	0.5773502692	1.0000000000
	−0.5773502692	1.0000000000
3	0.7745966692	0.5555555556
	0.0000000000	0.8888888889
	−0.7745966692	0.5555555556
4	0.8611363116	0.3478548451
	0.3399810436	0.6521451549
	−0.3399810436	0.6521451549
	−0.8611363116	0.3478548451
5	0.9061798459	0.2369268850
	0.5384693101	0.4786286705
	0.0000000000	0.5688888889
	−0.5384693101	0.4786286705
	−0.9061798459	0.2369268850

Example 1 Approximate $\int_{-1}^{1} e^x \cos x\,dx$ using Gaussian quadrature with $n = 3$.

Solution The entries in Table 4.12 give us

$$\begin{aligned}\int_{-1}^{1} e^x \cos x\,dx &\approx 0.\overline{5}e^{0.774596692}\cos 0.774596692 \\ &\quad + 0.\overline{8}\cos 0 + 0.\overline{5}e^{-0.774596692}\cos(-0.774596692) \\ &= 1.9333904.\end{aligned}$$

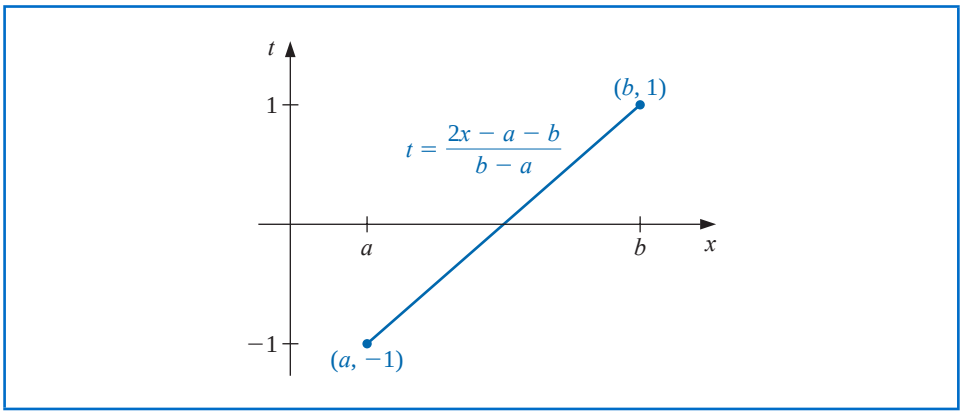
Integration by parts can be used to show that the true value of the integral is 1.9334214, so the absolute error is less than 3.2×10^{-5}. ■

Gaussian Quadrature on Arbitrary Intervals

An integral $\int_a^b f(x)\,dx$ over an arbitrary $[a, b]$ can be transformed into an integral over $[-1, 1]$ by using the change of variables (see Figure 4.17):

$$t = \frac{2x - a - b}{b - a} \iff x = \frac{1}{2}[(b - a)t + a + b].$$

Figure 4.17



This permits Gaussian quadrature to be applied to any interval $[a, b]$, because

$$\int_a^b f(x)\,dx = \int_{-1}^{1} f\left(\frac{(b - a)t + (b + a)}{2}\right)\frac{(b - a)}{2}\,dt. \tag{4.41}$$

Example 2 Consider the integral $\int_1^3 x^6 - x^2 \sin(2x)\,dx = 317.3442466$.

(a) Compare the results for the closed Newton-Cotes formula with $n = 1$, the open Newton-Cotes formula with $n = 1$, and Gaussian Quadrature when $n = 2$.

(b) Compare the results for the closed Newton-Cotes formula with $n = 2$, the open Newton-Cotes formula with $n = 2$, and Gaussian Quadrature when $n = 3$.

Solution **(a)** Each of the formulas in this part requires 2 evaluations of the function $f(x) = x^6 - x^2 \sin(2x)$. The Newton-Cotes approximations are

$$\textbf{Closed } \boldsymbol{n = 1}: \quad \frac{2}{2}[f(1) + f(3)] = 731.6054420;$$

$$\textbf{Open } \boldsymbol{n = 1}: \quad \frac{3(2/3)}{2}[f(5/3) + f(7/3)] = 188.7856682.$$

Gaussian quadrature applied to this problem requires that the integral first be transformed into a problem whose interval of integration is $[-1, 1]$. Using Eq. (4.41) gives

$$\int_1^3 x^6 - x^2 \sin(2x)\,dx = \int_{-1}^{1} (t + 2)^6 - (t + 2)^2 \sin(2(t + 2))\,dt.$$

Gaussian quadrature with $n = 2$ then gives

$$\int_1^3 x^6 - x^2 \sin(2x)\,dx \approx f(-0.5773502692 + 2) + f(0.5773502692 + 2) = 306.8199344;$$

(b) Each of the formulas in this part requires 3 function evaluations. The Newton-Cotes approximations are

$$\textbf{Closed } n = 2: \quad \frac{(1)}{3}[f(1) + 4f(2) + f(3)] = 333.2380940;$$

$$\textbf{Open } n = 2: \quad \frac{4(1/2)}{3}[2f(1.5) - f(2) + 2f(2.5)] = 303.5912023.$$

Gaussian quadrature with $n = 3$, once the transformation has been done, gives

$$\int_1^3 x^6 - x^2 \sin(2x)\,dx \approx 0.\overline{5}f(-0.7745966692 + 2) + 0.\overline{8}f(2) + 0.\overline{5}f(0.7745966692 + 2) = 317.2641516.$$

The Gaussian quadrature results are clearly superior in each instance. ■

Maple has Composite Gaussian Quadrature in the *NumericalAnalysis* subpackage of Maple's *Student* package. The default for the number of partitions in the command is 10, so the results in Example 2 would be found for $n = 2$ with

$f := x^6 - x^2 \sin(2x); a := 1; b := 3:$
$Quadrature(f(x), x = a..b, method = gaussian[2], partition = 1, output = information)$

which returns the approximation, what Maple assumes is the exact value of the integral, the absolute, and relative errors in the approximations, and the number of function evaluations.

The result when $n = 3$ is, of course, obtained by replacing the statement *method* = *gaussian*[2] with *method* = *gaussian*[3].

EXERCISE SET 4.7

1. Approximate the following integrals using Gaussian quadrature with $n = 2$, and compare your results to the exact values of the integrals.

a. $\displaystyle\int_1^{1.5} x^2 \ln x\,dx$

b. $\displaystyle\int_0^1 x^2 e^{-x}\,dx$

c. $\displaystyle\int_0^{0.35} \frac{2}{x^2 - 4}\,dx$

d. $\displaystyle\int_0^{\pi/4} x^2 \sin x\,dx$

e. $\displaystyle\int_0^{\pi/4} e^{3x} \sin 2x\,dx$

f. $\displaystyle\int_1^{1.6} \frac{2x}{x^2 - 4}\,dx$

g. $\displaystyle\int_3^{3.5} \frac{x}{\sqrt{x^2 - 4}}\,dx$

h. $\displaystyle\int_0^{\pi/4} (\cos x)^2\,dx$

2. Repeat Exercise 1 with $n = 3$.

3. Repeat Exercise 1 with $n = 4$.

4. Repeat Exercise 1 with $n = 5$.

5. Determine constants a, b, c, and d that will produce a quadrature formula

$$\int_{-1}^1 f(x)\,dx = af(-1) + bf(1) + cf'(-1) + df'(1)$$

that has degree of precision 3.

6. Determine constants a, b, c, and d that will produce a quadrature formula

$$\int_{-1}^1 f(x)\,dx = af(-1) + bf(0) + cf(1) + df'(-1) + ef'(1)$$

that has degree of precision 4.

7. Verify the entries for the values of $n = 2$ and 3 in Table 4.12 on page 232 by finding the roots of the respective Legendre polynomials, and use the equations preceding this table to find the coefficients associated with the values.
8. Show that the formula $Q(P) = \sum_{i=1}^{n} c_i P(x_i)$ cannot have degree of precision greater than $2n - 1$, regardless of the choice of $c_1, \ldots, c_n$ and $x_1, \ldots, x_n$. [*Hint:* Construct a polynomial that has a double root at each of the x_i's.]
9. Apply Maple's Composite Gaussian Quadrature routine to approximate $\int_{-1}^{1} x^2 e^x \, dx$ in the following manner.
   a. Use Gaussian Quadrature with $n = 8$ on the single interval $[-1, 1]$.
   b. Use Gaussian Quadrature with $n = 4$ on the intervals $[-1, 0]$ and $[0, 1]$.
   c. Use Gaussian Quadrature with $n = 2$ on the intervals $[-1, -0.5]$, $[-0.5, 0]$, $[0, 0.5]$ and $[0.5, 1]$.
   d. Give an explanation for the accuracy of the results.

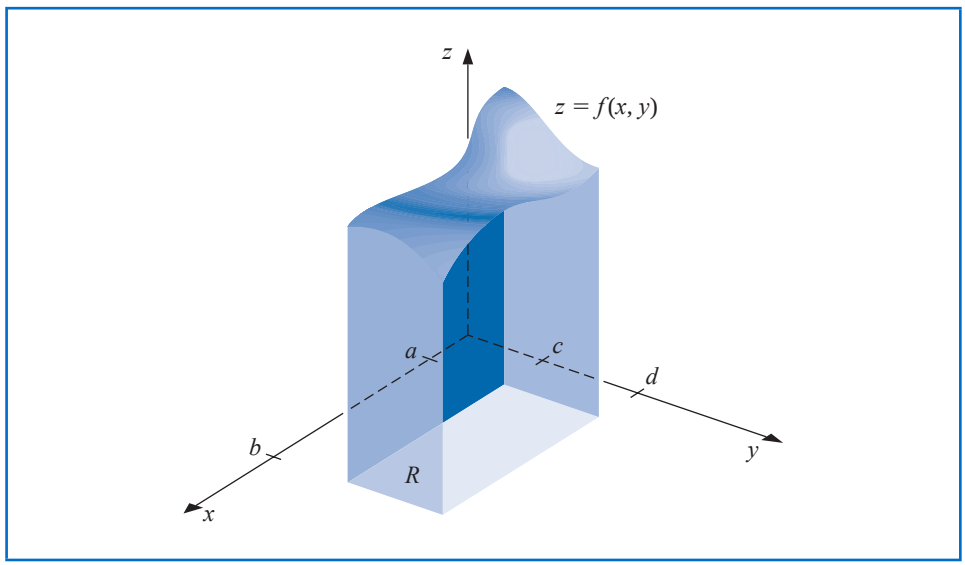
4.8 Multiple Integrals

The techniques discussed in the previous sections can be modified for use in the approximation of multiple integrals. Consider the double integral

$$\iint_R f(x, y) \, dA,$$

where $R = \{(x, y) \mid a \le x \le b, c \le y \le d\}$, for some constants a, b, c, and d, is a rectangular region in the plane. (See Figure 4.18.)

Figure 4.18



The following illustration shows how the Composite Trapezoidal rule using two subintervals in each coordinate direction would be applied to this integral.

Illustration Writing the double integral as an iterated integral gives

$$\iint_R f(x, y) \, dA = \int_a^b \left(\int_c^d f(x, y) \, dy \right) dx.$$

To simplify notation, let $k = (d-c)/2$ and $h = (b-a)/2$. Apply the Composite Trapezoidal rule to the interior integral to obtain

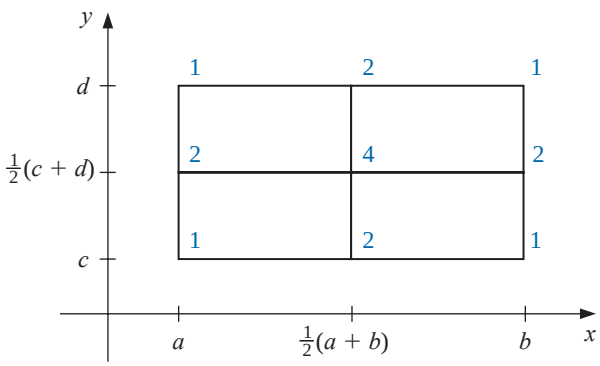
$$\int_c^d f(x,y)\,dy \approx \frac{k}{2}\left[f(x,c) + f(x,d) + 2f\left(x, \frac{c+d}{2}\right)\right].$$

This approximation is of order $O\left((d-c)^3\right)$. Then apply the Composite Trapezoidal rule again to approximate the integral of this function of x:

$$\begin{aligned}
\int_a^b \left(\int_c^d f(x,y)\,dy\right) dx &\approx \int_a^b \left(\frac{d-c}{4}\right)\left[f(x,c) + 2f\left(x, \frac{c+d}{2}\right) + f(d)\right] dx \\
&= \frac{b-a}{4}\left(\frac{d-c}{4}\right)\left[f(a,c) + 2f\left(a, \frac{c+d}{2}\right) + f(a,d)\right] \\
&\quad + \frac{b-a}{4}\left(2\left(\frac{d-c}{4}\right)\left[f\left(\frac{a+b}{2}, c\right)\right.\right. \\
&\quad \left.\left. + 2f\left(\frac{a+b}{2}, \frac{c+d}{2}\right) + \left(\frac{a+b}{2}, d\right)\right]\right) \\
&\quad + \frac{b-a}{4}\left(\frac{d-c}{4}\right)\left[f(b,c) + 2f\left(b, \frac{c+d}{2}\right) + f(b,d)\right] \\
&= \frac{(b-a)(d-c)}{16}\left[f(a,c) + f(a,d) + f(b,c) + f(b,d)\right. \\
&\quad + 2\left(f\left(\frac{a+b}{2}, c\right) + f\left(\frac{a+b}{2}, d\right) + f\left(a, \frac{c+d}{2}\right)\right. \\
&\quad \left.\left. + f\left(b, \frac{c+d}{2}\right)\right) + 4f\left(\frac{a+b}{2}, \frac{c+d}{2}\right)\right]
\end{aligned}$$

This approximation is of order $O\left((b-a)(d-c)\left[(b-a)^2 + (d-c)^2\right]\right)$. Figure 4.19 shows a grid with the number of functional evaluations at each of the nodes used in the approximation. □

Figure 4.19



As the illustration shows, the procedure is quite straightforward. But the number of function evaluations grows with the square of the number required for a single integral. In a practical situation we would not expect to use a method as elementary as the Composite Trapezoidal rule. Instead we will employ the Composite Simpson's rule to illustrate the general approximation technique, although any other composite formula could be used in its place.

To apply the Composite Simpson's rule, we divide the region R by partitioning both $[a, b]$ and $[c, d]$ into an even number of subintervals. To simplify the notation, we choose even integers n and m and partition $[a, b]$ and $[c, d]$ with the evenly spaced mesh points $x_0, x_1, \ldots, x_n$ and $y_0, y_1, \ldots, y_m$, respectively. These subdivisions determine step sizes $h = (b - a)/n$ and $k = (d - c)/m$. Writing the double integral as the iterated integral

$$\iint_R f(x, y)\, dA = \int_a^b \left(\int_c^d f(x, y)\, dy \right) dx,$$

we first use the Composite Simpson's rule to approximate

$$\int_c^d f(x, y)\, dy,$$

treating x as a constant.

Let $y_j = c + jk$, for each $j = 0, 1, \ldots, m$. Then

$$\int_c^d f(x, y)\, dy = \frac{k}{3}\left[f(x, y_0) + 2 \sum_{j=1}^{(m/2)-1} f(x, y_{2j}) + 4 \sum_{j=1}^{m/2} f(x, y_{2j-1}) + f(x, y_m) \right] - \frac{(d - c)k^4}{180} \frac{\partial^4 f}{\partial y^4}(x, \mu),$$

for some μ in (c, d). Thus

$$\begin{aligned} \int_a^b \int_c^d f(x, y)\, dy\, dx = \frac{k}{3}\Bigg[& \int_a^b f(x, y_0)\, dx + 2 \sum_{j=1}^{(m/2)-1} \int_a^b f(x, y_{2j})\, dx \\ & + 4 \sum_{j=1}^{m/2} \int_a^b f(x, y_{2j-1})\, dx + \int_a^b f(x, y_m)\, dx \Bigg] \\ & - \frac{(d - c)k^4}{180} \int_a^b \frac{\partial^4 f}{\partial y^4}(x, \mu)\, dx. \end{aligned}$$

Composite Simpson's rule is now employed on the integrals in this equation. Let $x_i = a + ih$, for each $i = 0, 1, \ldots, n$. Then for each $j = 0, 1, \ldots, m$, we have

$$\int_a^b f(x, y_j)\, dx = \frac{h}{3}\left[f(x_0, y_j) + 2 \sum_{i=1}^{(n/2)-1} f(x_{2i}, y_j) + 4 \sum_{i=1}^{n/2} f(x_{2i-1}, y_j) + f(x_n, y_j) \right] - \frac{(b - a)h^4}{180} \frac{\partial^4 f}{\partial x^4}(\xi_j, y_j),$$

for some ξ_j in (a, b). The resulting approximation has the form

$$\begin{aligned}\int_a^b \int_c^d f(x,y)\,dy\,dx \approx \frac{hk}{9}\Bigg\{&\Bigg[f(x_0,y_0)+2\sum_{i=1}^{(n/2)-1} f(x_{2i},y_0)\\&+4\sum_{i=1}^{n/2} f(x_{2i-1},y_0)+f(x_n,y_0)\Bigg]\\&+2\Bigg[\sum_{j=1}^{(m/2)-1} f(x_0,y_{2j})+2\sum_{j=1}^{(m/2)-1}\sum_{i=1}^{(n/2)-1} f(x_{2i},y_{2j})\\&+4\sum_{j=1}^{(m/2)-1}\sum_{i=1}^{n/2} f(x_{2i-1},y_{2j})+\sum_{j=1}^{(m/2)-1} f(x_n,y_{2j})\Bigg]\\&+4\Bigg[\sum_{j=1}^{m/2} f(x_0,y_{2j-1})+2\sum_{j=1}^{m/2}\sum_{i=1}^{(n/2)-1} f(x_{2i},y_{2j-1})\\&+4\sum_{j=1}^{m/2}\sum_{i=1}^{n/2} f(x_{2i-1},y_{2j-1})+\sum_{j=1}^{m/2} f(x_n,y_{2j-1})\Bigg]\\&+\Bigg[f(x_0,y_m)+2\sum_{i=1}^{(n/2)-1} f(x_{2i},y_m)+4\sum_{i=1}^{n/2} f(x_{2i-1},y_m)+f(x_n,y_m)\Bigg]\Bigg\}.\end{aligned}$$

The error term E is given by

$$\begin{aligned}E=\frac{-k(b-a)h^4}{540}\Bigg[&\frac{\partial^4 f}{\partial x^4}(\xi_0,y_0)+2\sum_{j=1}^{(m/2)-1}\frac{\partial^4 f}{\partial x^4}(\xi_{2j},y_{2j})+4\sum_{j=1}^{m/2}\frac{\partial^4 f}{\partial x^4}(\xi_{2j-1},y_{2j-1})\\&+\frac{\partial^4 f}{\partial x^4}(\xi_m,y_m)\Bigg]-\frac{(d-c)k^4}{180}\int_a^b \frac{\partial^4 f}{\partial y^4}(x,\mu)\,dx.\end{aligned}$$

If $\partial^4 f/\partial x^4$ is continuous, the Intermediate Value Theorem 1.11 can be repeatedly applied to show that the evaluation of the partial derivatives with respect to x can be replaced by a common value and that

$$E=\frac{-k(b-a)h^4}{540}\left[3m\frac{\partial^4 f}{\partial x^4}(\overline{\eta},\overline{\mu})\right]-\frac{(d-c)k^4}{180}\int_a^b \frac{\partial^4 f}{\partial y^4}(x,\mu)\,dx,$$

for some $(\overline{\eta}, \overline{\mu})$ in R. If $\partial^4 f/\partial y^4$ is also continuous, the Weighted Mean Value Theorem for Integrals 1.13 implies that

$$\int_a^b \frac{\partial^4 f}{\partial y^4}(x,\mu)\,dx=(b-a)\frac{\partial^4 f}{\partial y^4}(\hat{\eta},\hat{\mu}),$$

for some $(\hat{\eta}, \hat{\mu})$ in R. Because $m = (d - c)/k$, the error term has the form

$$E=\frac{-k(b-a)h^4}{540}\left[3m\frac{\partial^4 f}{\partial x^4}(\overline{\eta},\overline{\mu})\right]-\frac{(d-c)(b-a)}{180}k^4\frac{\partial^4 f}{\partial y^4}(\hat{\eta},\hat{\mu})$$

which simplifies to

$$E=-\frac{(d-c)(b-a)}{180}\left[h^4\frac{\partial^4 f}{\partial x^4}(\overline{\eta},\overline{\mu})+k^4\frac{\partial^4 f}{\partial y^4}(\hat{\eta},\hat{\mu})\right],$$

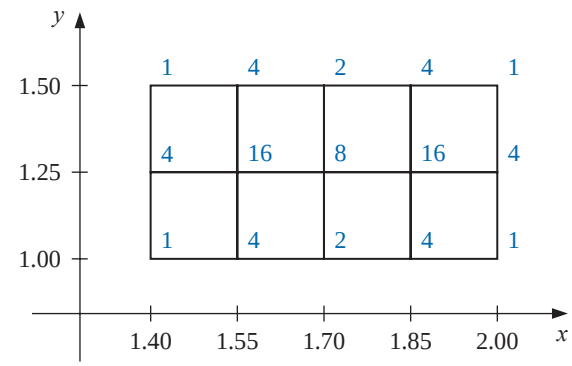
for some $(\overline{\eta}, \overline{\mu})$ and $(\hat{\eta}, \hat{\mu})$ in R.

Example 1 Use Composite Simpson's rule with $n = 4$ and $m = 2$ to approximate

$$\int_{1.4}^{2.0} \int_{1.0}^{1.5} \ln(x + 2y)\, dy\, dx,$$

Solution The step sizes for this application are $h = (2.0 - 1.4)/4 = 0.15$ and $k = (1.5 - 1.0)/2 = 0.25$. The region of integration R is shown in Figure 4.20, together with the nodes (x_i, y_j), where $i = 0, 1, 2, 3, 4$ and $j = 0, 1, 2$. It also shows the coefficients $w_{i,j}$ of $f(x_i, y_i) = \ln(x_i + 2y_i)$ in the sum that gives the Composite Simpson's rule approximation to the integral.

Figure 4.20



The approximation is

$$\int_{1.4}^{2.0} \int_{1.0}^{1.5} \ln(x + 2y)\, dy\, dx \approx \frac{(0.15)(0.25)}{9} \sum_{i=0}^{4} \sum_{j=0}^{2} w_{i,j} \ln(x_i + 2y_j)$$

$$= 0.4295524387.$$

We have

$$\frac{\partial^4 f}{\partial x^4}(x, y) = \frac{-6}{(x + 2y)^4} \quad \text{and} \quad \frac{\partial^4 f}{\partial y^4}(x, y) = \frac{-96}{(x + 2y)^4},$$

and the maximum values of the absolute values of these partial derivatives occur on R when $x = 1.4$ and $y = 1.0$. So the error is bounded by

$$|E| \le \frac{(0.5)(0.6)}{180} \left[(0.15)^4 \max_{(x,y) \text{in} R} \frac{6}{(x + 2y)^4} + (0.25)^4 \max_{(x,y) \text{in} R} \frac{96}{(x + 2y)^4} \right] \le 4.72 \times 10^{-6}.$$

The actual value of the integral to ten decimal places is

$$\int_{1.4}^{2.0} \int_{1.0}^{1.5} \ln(x + 2y)\, dy\, dx = 0.4295545265,$$

so the approximation is accurate to within 2.1×10^{-6}. ■

The same techniques can be applied for the approximation of triple integrals as well as higher integrals for functions of more than three variables. The number of functional evaluations required for the approximation is the product of the number of functional evaluations required when the method is applied to each variable.

Gaussian Quadrature for Double Integral Approximation

To reduce the number of functional evaluations, more efficient methods such as Gaussian quadrature, Romberg integration, or Adaptive quadrature can be incorporated in place of the Newton-Cotes formulas. The following example illustrates the use of Gaussian quadrature for the integral considered in Example 1.

Example 2 Use Gaussian quadrature with $n = 3$ in both dimensions to approximate the integral

$$\int_{1.4}^{2.0} \int_{1.0}^{1.5} \ln(x + 2y)\, dy\, dx.$$

Solution Before employing Gaussian quadrature to approximate this integral, we need to transform the region of integration

$$R = \{ (x, y) \mid 1.4 \le x \le 2.0, 1.0 \le y \le 1.5 \}$$

into

$$\hat{R} = \{ (u, v) \mid -1 \le u \le 1, -1 \le v \le 1 \}.$$

The linear transformations that accomplish this are

$$u = \frac{1}{2.0 - 1.4}(2x - 1.4 - 2.0) \quad \text{and} \quad v = \frac{1}{1.5 - 1.0}(2y - 1.0 - 1.5),$$

or, equivalently, $x = 0.3u + 1.7$ and $y = 0.25v + 1.25$. Employing this change of variables gives an integral on which Gaussian quadrature can be applied:

$$\int_{1.4}^{2.0} \int_{1.0}^{1.5} \ln(x + 2y)\, dy\, dx = 0.075 \int_{-1}^{1} \int_{-1}^{1} \ln(0.3u + 0.5v + 4.2)\, dv\, du.$$

The Gaussian quadrature formula for $n = 3$ in both u and v requires that we use the nodes

$$u_1 = v_1 = r_{3,2} = 0, \quad u_0 = v_0 = r_{3,1} = -0.7745966692,$$

and

$$u_2 = v_2 = r_{3,3} = 0.7745966692.$$

The associated weights are $c_{3,2} = 0.\overline{8}$ and $c_{3,1} = c_{3,3} = 0.\overline{5}$. (These are given in Table 4.12 on page 232.) The resulting approximation is

$$\int_{1.4}^{2.0} \int_{1.0}^{1.5} \ln(x + 2y)\, dy\, dx \approx 0.075 \sum_{i=1}^{3} \sum_{j=1}^{3} c_{3,i} c_{3,j} \ln(0.3r_{3,i} + 0.5r_{3,j} + 4.2)$$

$$= 0.4295545313.$$

Although this result requires only 9 functional evaluations compared to 15 for the Composite Simpson's rule considered in Example 1, it is accurate to within 4.8×10^{-9}, compared to 2.1×10^{-6} accuracy in Example 1. ■

Non-Rectangular Regions

The use of approximation methods for double integrals is not limited to integrals with rectangular regions of integration. The techniques previously discussed can be modified to approximate double integrals of the form

$$\int_a^b \int_{c(x)}^{d(x)} f(x, y)\, dy\, dx \tag{4.42}$$

or

$$\int_c^d \int_{a(y)}^{b(y)} f(x, y)\, dx\, dy. \tag{4.43}$$

In fact, integrals on regions not of this type can also be approximated by performing appropriate partitions of the region. (See Exercise 10.)

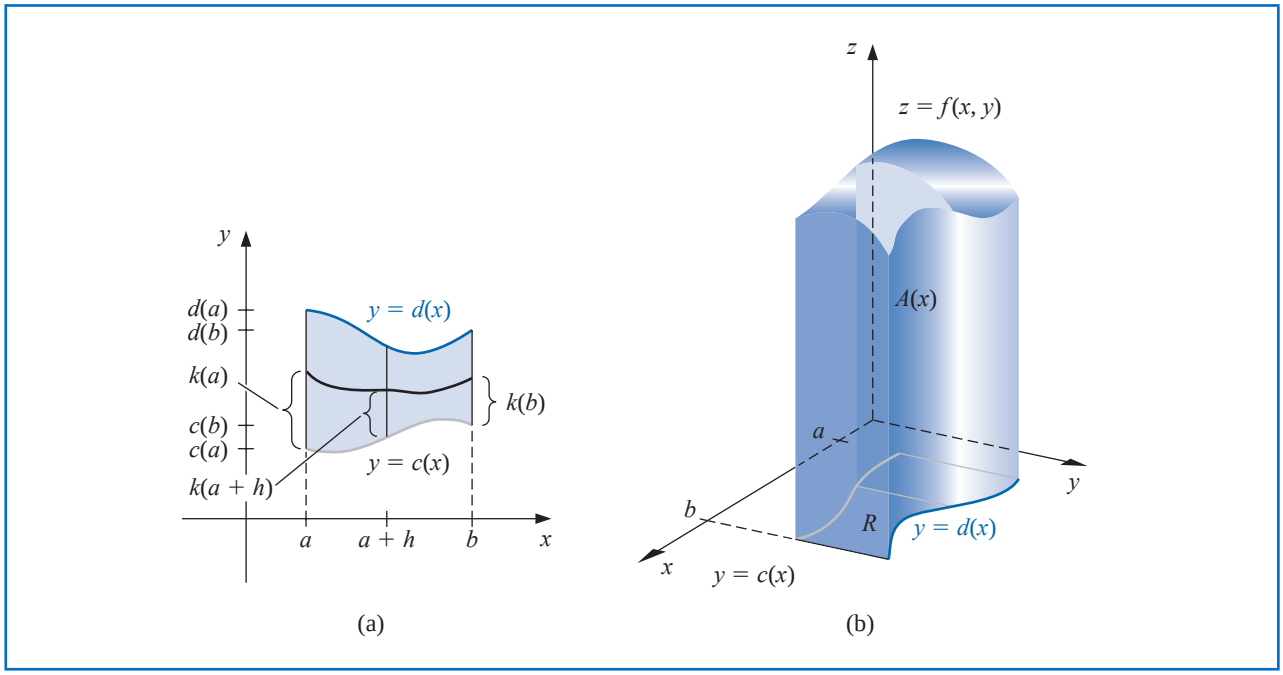
To describe the technique involved with approximating an integral in the form

$$\int_a^b \int_{c(x)}^{d(x)} f(x, y)\, dy\, dx,$$

we will use the basic Simpson's rule to integrate with respect to both variables. The step size for the variable x is $h = (b - a)/2$, but the step size for y varies with x (see Figure 4.21) and is written

$$k(x) = \frac{d(x) - c(x)}{2}.$$

Figure 4.21



This gives

$$\int_a^b \int_{c(x)}^{d(x)} f(x,y)\,dy\,dx \approx \int_a^b \frac{k(x)}{3}[f(x,c(x)) + 4f(x,c(x)+k(x)) + f(x,d(x))]\,dx$$

$$\approx \frac{h}{3}\left\{\frac{k(a)}{3}[f(a,c(a)) + 4f(a,c(a)+k(a)) + f(a,d(a))]\right.$$

$$+ \frac{4k(a+h)}{3}[f(a+h,c(a+h)) + 4f(a+h,c(a+h)$$

$$+ k(a+h)) + f(a+h,d(a+h))]$$

$$\left. + \frac{k(b)}{3}[f(b,c(b)) + 4f(b,c(b)+k(b)) + f(b,d(b))]\right\}.$$

Algorithm 4.4 applies the Composite Simpson's rule to an integral in the form (4.42). Integrals in the form (4.43) can, of course, be handled similarly.

ALGORITHM
4.4

Simpson's Double Integral

To approximate the integral

$$I = \int_a^b \int_{c(x)}^{d(x)} f(x,y)\,dy\,dx:$$

INPUT endpoints a, b: even positive integers m, n.

OUTPUT approximation J to I.

Step 1 Set $h = (b-a)/n$;
$J_1 = 0$; (*End terms.*)
$J_2 = 0$; (*Even terms.*)
$J_3 = 0$. (*Odd terms.*)

Step 2 For $i = 0, 1, \ldots, n$ do Steps 3–8.

Step 3 Set $x = a + ih$; (*Composite Simpson's method for x.*)
$HX = (d(x) - c(x))/m$;
$K_1 = f(x,c(x)) + f(x,d(x))$; (*End terms.*)
$K_2 = 0$; (*Even terms.*)
$K_3 = 0$. (*Odd terms.*)

Step 4 For $j = 1, 2, \ldots, m-1$ do Step 5 and 6.

Step 5 Set $y = c(x) + jHX$;
$Q = f(x,y)$.

Step 6 If j is even then set $K_2 = K_2 + Q$
else set $K_3 = K_3 + Q$.

Step 7 Set $L = (K_1 + 2K_2 + 4K_3)HX/3$.

$\left(L \approx \int_{c(x_i)}^{d(x_i)} f(x_i,y)\,dy \quad \textit{by the Composite Simpson's method.}\right)$

Step 8 If $i = 0$ or $i = n$ then set $J_1 = J_1 + L$
else if i is even then set $J_2 = J_2 + L$
else set $J_3 = J_3 + L$.

Step 9 Set $J = h(J_1 + 2J_2 + 4J_3)/3$.

Step 10 OUTPUT (J);
STOP. ■

To apply Gaussian quadrature to the double integral

$$\int_a^b \int_{c(x)}^{d(x)} f(x,y)\,dy\,dx,$$

first requires transforming, for each x in $[a,b]$, the variable y in the interval $[c(x),d(x)]$ into the variable t in the interval $[-1,1]$. This linear transformation gives

$$f(x,y) = f\left(x, \frac{(d(x)-c(x))t + d(x) + c(x)}{2}\right) \quad \text{and} \quad dy = \frac{d(x)-c(x)}{2}\,dt.$$

The reduced calculation makes it generally worthwhile to apply Gaussian quadrature rather than a Simpson's technique when approximating double integrals.

Then, for each x in $[a,b]$, we apply Gaussian quadrature to the resulting integral

$$\int_{c(x)}^{d(x)} f(x,y)\,dy = \int_{-1}^{1} f\left(x, \frac{(d(x)-c(x))t + d(x) + c(x)}{2}\right) dt$$

to produce

$$\int_a^b \int_{c(x)}^{d(x)} f(x,y)\,dy\,dx \approx \int_a^b \frac{d(x)-c(x)}{2} \sum_{j=1}^{n} c_{n,j} f\left(x, \frac{(d(x)-c(x))r_{n,j} + d(x) + c(x)}{2}\right) dx,$$

where, as before, the roots $r_{n,j}$ and coefficients $c_{n,j}$ come from Table 4.12 on page 232. Now the interval $[a,b]$ is transformed to $[-1,1]$, and Gaussian quadrature is applied to approximate the integral on the right side of this equation. The details are given in Algorithm 4.5.

ALGORITHM
4.5

Gaussian Double Integral

To approximate the integral

$$\int_a^b \int_{c(x)}^{d(x)} f(x,y)\,dy\,dx :$$

INPUT endpoints a, b; positive integers m, n.
(*The roots $r_{i,j}$ and coefficients $c_{i,j}$ need to be available for $i = \max\{m,n\}$ and for $1 \le j \le i$.*)

OUTPUT approximation J to I.

Step 1 Set $h_1 = (b-a)/2$;
$h_2 = (b+a)/2$;
$J = 0$.

Step 2 For $i = 1, 2, \ldots, m$ do Steps 3–5.

Step 3 Set $JX = 0$;
$x = h_1 r_{m,i} + h_2$;
$d_1 = d(x)$;
$c_1 = c(x)$;
$k_1 = (d_1 - c_1)/2$;
$k_2 = (d_1 + c_1)/2$.

Step 4 For $j = 1, 2, \ldots, n$ do
set $y = k_1 r_{n,j} + k_2$;
$Q = f(x, y)$;
$JX = JX + c_{n,j} Q$.

Step 5 Set $J = J + c_{m,i} k_1 JX$.

Step 6 Set $J = h_1 J$.

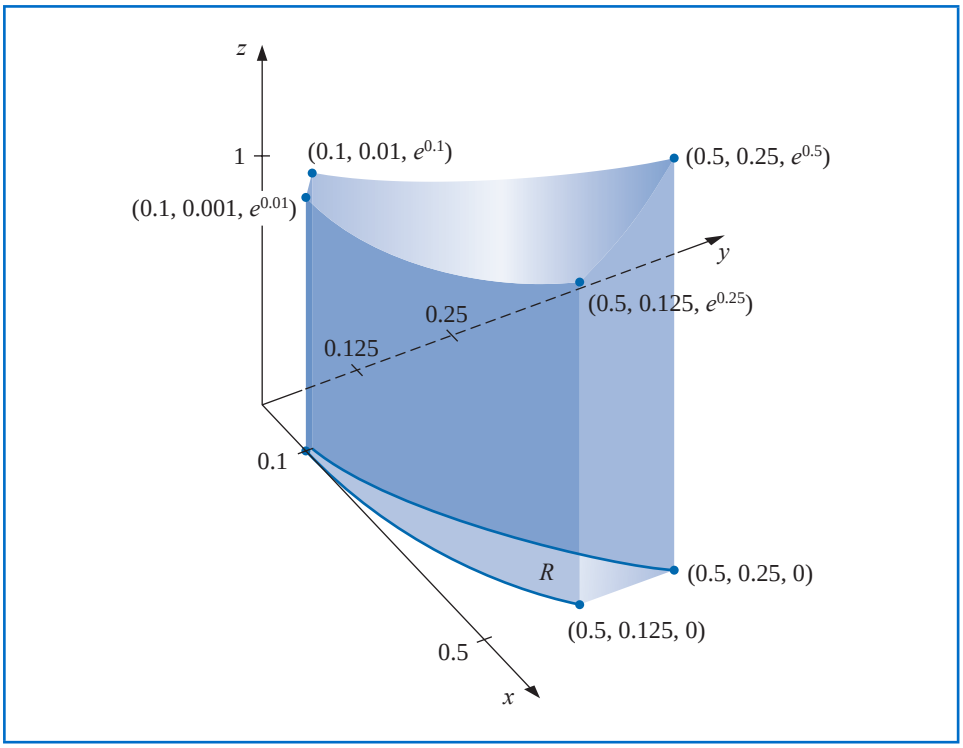
Step 7 OUTPUT (J);
STOP. ■

Illustration The volume of the solid in Figure 4.22 is approximated by applying Simpson's Double Integral Algorithm with $n = m = 10$ to

$$\int_{0.1}^{0.5} \int_{x^3}^{x^2} e^{y/x}\, dy\, dx.$$

This requires 121 evaluations of the function $f(x, y) = e^{y/x}$ and produces the value 0.0333054, which approximates the volume of the solid shown in Figure 4.22 to nearly seven decimal places. Applying the Gaussian Quadrature Algorithm with $n = m = 5$ requires only 25 function evaluations and gives the approximation 0.03330556611, which is accurate to 11 decimal places. □

Figure 4.22



Triple Integral Approximation

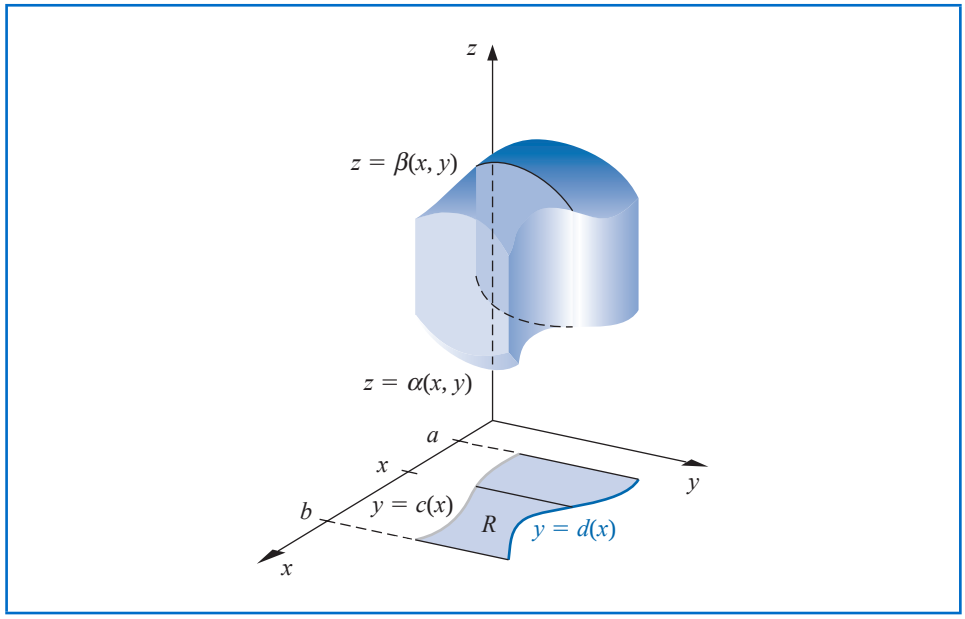
Triple integrals of the form

$$\int_a^b \int_{c(x)}^{d(x)} \int_{\alpha(x,y)}^{\beta(x,y)} f(x, y, z)\, dz\, dy\, dx$$

(see Figure 4.23) are approximated in a similar manner. Because of the number of calculations involved, Gaussian quadrature is the method of choice. Algorithm 4.6 implements this procedure.

The reduced calculation makes it almost always worthwhile to apply Gaussian quadrature rather than a Simpson's technique when approximating triple or higher integrals.

Figure 4.23



ALGORITHM
4.6

Gaussian Triple Integral

To approximate the integral

$$\int_a^b \int_{c(x)}^{d(x)} \int_{\alpha(x,y)}^{\beta(x,y)} f(x, y, z)\, dz\, dy\, dx :$$

INPUT endpoints a, b; positive integers m, n, p.
(*The roots* $r_{i,j}$ *and coefficients* $c_{i,j}$ *need to be available for* $i = \max\{n, m, p\}$ *and for* $1 \le j \le i$.)

OUTPUT approximation J to I.

Step 1 Set $h_1 = (b - a)/2$;
$h_2 = (b + a)/2$;
$J = 0$.

Step 2 For $i = 1, 2, \ldots, m$ do Steps 3–8.

Step 3 Set $JX = 0$;
$x = h_1 r_{m,i} + h_2$;
$d_1 = d(x)$;
$c_1 = c(x)$;
$k_1 = (d_1 - c_1)/2$;
$k_2 = (d_1 + c_1)/2$.

Step 4 For $j = 1, 2, \ldots, n$ do Steps 5–7.

Step 5 Set $JY = 0$;
$y = k_1 r_{n,j} + k_2$;
$\beta_1 = \beta(x, y)$;
$\alpha_1 = \alpha(x, y)$;
$l_1 = (\beta_1 - \alpha_1)/2$;
$l_2 = (\beta_1 + \alpha_1)/2$.

Step 6 For $k = 1, 2, \ldots, p$ do
set $z = l_1 r_{p,k} + l_2$;
$Q = f(x, y, z)$;
$JY = JY + c_{p,k} Q$.

Step 7 Set $JX = JX + c_{n,j} l_1 JY$.

Step 8 Set $J = J + c_{m,i} k_1 JX$.

Step 9 Set $J = h_1 J$.

Step 10 OUTPUT (J);
STOP. ■

The following example requires the evaluation of four triple integrals.

Illustration The center of a mass of a solid region D with density function σ occurs at

$$(\bar{x}, \bar{y}, \bar{z}) = \left(\frac{M_{yz}}{M}, \frac{M_{xz}}{M}, \frac{M_{xy}}{M} \right),$$

where

$$M_{yz} = \iiint_D x\sigma(x, y, z)\, dV, \quad M_{xz} = \iiint_D y\sigma(x, y, z)\, dV$$

and

$$M_{xy} = \iiint_D z\sigma(x, y, z)\, dV$$

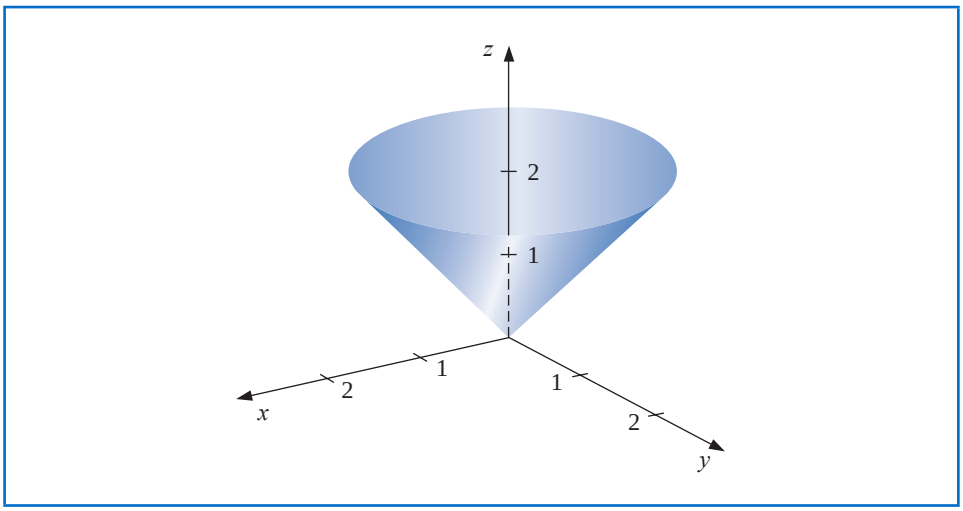
are the moments about the coordinate planes and the mass of D is

$$M = \iiint_D \sigma(x, y, z)\, dV.$$

The solid shown in Figure 4.24 is bounded by the upper nappe of the cone $z^2 = x^2 + y^2$ and the plane $z = 2$. Suppose that this solid has density function given by

$$\sigma(x, y, z) = \sqrt{x^2 + y^2}.$$

Figure 4.24



Applying the Gaussian Triple Integral Algorithm 4.6 with $n = m = p = 5$ requires 125 function evaluations per integral and gives the following approximations:

$$M = \int_{-2}^{2} \int_{-\sqrt{4-x^2}}^{\sqrt{4-x^2}} \int_{\sqrt{x^2+y^2}}^{2} \sqrt{x^2+y^2}\, dz\, dy\, dx$$

$$= 4 \int_{0}^{2} \int_{0}^{\sqrt{4-x^2}} \int_{\sqrt{x^2+y^2}}^{2} \sqrt{x^2+y^2}\, dz\, dy\, dx \approx 8.37504476,$$

$$M_{yz} = \int_{-2}^{2} \int_{-\sqrt{4-x^2}}^{\sqrt{4-x^2}} \int_{\sqrt{x^2+y^2}}^{2} x\sqrt{x^2+y^2}\, dz\, dy\, dx \approx -5.55111512 \times 10^{-17},$$

$$M_{xz} = \int_{-2}^{2} \int_{-\sqrt{4-x^2}}^{\sqrt{4-x^2}} \int_{\sqrt{x^2+y^2}}^{2} y\sqrt{x^2+y^2}\, dz\, dy\, dx \approx -8.01513675 \times 10^{-17},$$

$$M_{xy} = \int_{-2}^{2} \int_{-\sqrt{4-x^2}}^{\sqrt{4-x^2}} \int_{\sqrt{x^2+y^2}}^{2} z\sqrt{x^2+y^2}\, dz\, dy\, dx \approx 13.40038156.$$

This implies that the approximate location of the center of mass is

$$(\bar{x}, \bar{y}, \bar{z}) = (0, 0, 1.60003701).$$

These integrals are quite easy to evaluate directly. If you do this, you will find that the exact center of mass occurs at $(0, 0, 1.6)$. □

Multiple integrals can be evaluated in Maple using the *MultInt* command in the *MultivariateCalculus* subpackage of the *Student* package. For example, to evaluate the multiple integral

$$\int_{2}^{4} \int_{x-1}^{x+6} \int_{-2}^{4+y^2} x^2 + y^2 + z\, dz\, dy\, dx$$

we first load the package and define the function with

with(*Student*[*MultivariateCalculus*]): $f := (x, y, z) \to x^2 + y^2 + z$

Then issue the command

MultiInt($f(x, y, z), z = -2..4 + y^2,\ y = x - 1..x + 6,\ x = 2..4$)

which produces the result

1.995885970

EXERCISE SET 4.8

1. Use Algorithm 4.4 with $n = m = 4$ to approximate the following double integrals, and compare the results to the exact answers.

a. $\int_{2.1}^{2.5} \int_{1.2}^{1.4} xy^2 \, dy \, dx$

b. $\int_{0}^{0.5} \int_{0}^{0.5} e^{y-x} \, dy \, dx$

c. $\int_{2}^{2.2} \int_{x}^{2x} (x^2 + y^3) \, dy \, dx$

d. $\int_{1}^{1.5} \int_{0}^{x} (x^2 + \sqrt{y}) \, dy \, dx$

2. Find the smallest values for $n = m$ so that Algorithm 4.4 can be used to approximate the integrals in Exercise 1 to within 10^{-6} of the actual value.

3. Use Algorithm 4.4 with (i) $n = 4, m = 8$, (ii) $n = 8, m = 4$, and (iii) $n = m = 6$ to approximate the following double integrals, and compare the results to the exact answers.

a. $\int_{0}^{\pi/4} \int_{\sin x}^{\cos x} (2y \sin x + \cos^2 x) \, dy \, dx$

b. $\int_{1}^{e} \int_{1}^{x} \ln xy \, dy \, dx$

c. $\int_{0}^{1} \int_{x}^{2x} (x^2 + y^3) \, dy \, dx$

d. $\int_{0}^{1} \int_{x}^{2x} (y^2 + x^3) \, dy \, dx$

e. $\int_{0}^{\pi} \int_{0}^{x} \cos x \, dy \, dx$

f. $\int_{0}^{\pi} \int_{0}^{x} \cos y \, dy \, dx$

g. $\int_{0}^{\pi/4} \int_{0}^{\sin x} \frac{1}{\sqrt{1 - y^2}} \, dy \, dx$

h. $\int_{-\pi}^{3\pi/2} \int_{0}^{2\pi} (y \sin x + x \cos y) \, dy \, dx$

4. Find the smallest values for $n = m$ so that Algorithm 4.4 can be used to approximate the integrals in Exercise 3 to within 10^{-6} of the actual value.

5. Use Algorithm 4.5 with $n = m = 2$ to approximate the integrals in Exercise 1, and compare the results to those obtained in Exercise 1.

6. Find the smallest values of $n = m$ so that Algorithm 4.5 can be used to approximate the integrals in Exercise 1 to within 10^{-6}. Do not continue beyond $n = m = 5$. Compare the number of functional evaluations required to the number required in Exercise 2.

7. Use Algorithm 4.5 with (i) $n = m = 3$, (ii) $n = 3, m = 4$, (iii) $n = 4, m = 3$, and (iv) $n = m = 4$ to approximate the integrals in Exercise 3.

8. Use Algorithm 4.5 with $n = m = 5$ to approximate the integrals in Exercise 3. Compare the number of functional evaluations required to the number required in Exercise 4.

9. Use Algorithm 4.4 with $n = m = 14$ and Algorithm 4.5 with $n = m = 4$ to approximate

$$\iint_R e^{-(x+y)} \, dA,$$

for the region R in the plane bounded by the curves $y = x^2$ and $y = \sqrt{x}$.

10. Use Algorithm 4.4 to approximate

$$\iint_R \sqrt{xy + y^2}\, dA,$$

where R is the region in the plane bounded by the lines $x + y = 6$, $3y - x = 2$, and $3x - y = 2$. First partition R into two regions R_1 and R_2 on which Algorithm 4.4 can be applied. Use $n = m = 6$ on both R_1 and R_2.

11. A plane lamina is a thin sheet of continuously distributed mass. If σ is a function describing the density of a lamina having the shape of a region R in the xy-plane, then the center of the mass of the lamina $(\bar{x}, \bar{y})$ is

$$\bar{x} = \frac{\iint_R x\sigma(x, y)\, dA}{\iint_R \sigma(x, y)\, dA}, \quad \bar{y} = \frac{\iint_R y\sigma(x, y)\, dA}{\iint_R \sigma(x, y)\, dA}.$$

Use Algorithm 4.4 with $n = m = 14$ to find the center of mass of the lamina described by $R = \{(x, y) \mid 0 \le x \le 1, 0 \le y \le \sqrt{1 - x^2}\}$ with the density function $\sigma(x, y) = e^{-(x^2+y^2)}$. Compare the approximation to the exact result.

12. Repeat Exercise 11 using Algorithm 4.5 with $n = m = 5$.

13. The area of the surface described by $z = f(x, y)$ for (x, y) in R is given by

$$\iint_R \sqrt{[f_x(x, y)]^2 + [f_y(x, y)]^2 + 1}\, dA.$$

Use Algorithm 4.4 with $n = m = 8$ to find an approximation to the area of the surface on the hemisphere $x^2 + y^2 + z^2 = 9, z \ge 0$ that lies above the region in the plane described by $R = \{(x, y) \mid 0 \le x \le 1, 0 \le y \le 1\}$.

14. Repeat Exercise 13 using Algorithm 4.5 with $n = m = 4$.

15. Use Algorithm 4.6 with $n = m = p = 2$ to approximate the following triple integrals, and compare the results to the exact answers.

a. $\displaystyle\int_0^1 \int_1^2 \int_0^{0.5} e^{x+y+z}\, dz\, dy\, dx$

b. $\displaystyle\int_0^1 \int_x^1 \int_0^y y^2 z\, dz\, dy\, dx$

c. $\displaystyle\int_0^1 \int_{x^2}^x \int_{x-y}^{x+y} y\, dz\, dy\, dx$

d. $\displaystyle\int_0^1 \int_{x^2}^x \int_{x-y}^{x+y} z\, dz\, dy\, dx$

e. $\displaystyle\int_0^\pi \int_0^x \int_0^{xy} \frac{1}{y} \sin\frac{z}{y}\, dz\, dy\, dx$

f. $\displaystyle\int_0^1 \int_0^1 \int_{-xy}^{xy} e^{x^2+y^2}\, dz\, dy\, dx$

16. Repeat Exercise 15 using $n = m = p = 3$.

17. Repeat Exercise 15 using $n = m = p = 4$ and $n = m = p = 5$.

18. Use Algorithm 4.6 with $n = m = p = 4$ to approximate

$$\iiint_S xy \sin(yz)\, dV,$$

where S is the solid bounded by the coordinate planes and the planes $x = \pi$, $y = \pi/2$, $z = \pi/3$. Compare this approximation to the exact result.

19. Use Algorithm 4.6 with $n = m = p = 5$ to approximate

$$\iiint_S \sqrt{xyz}\, dV,$$

where S is the region in the first octant bounded by the cylinder $x^2 + y^2 = 4$, the sphere $x^2 + y^2 + z^2 = 4$, and the plane $x + y + z = 8$. How many functional evaluations are required for the approximation?

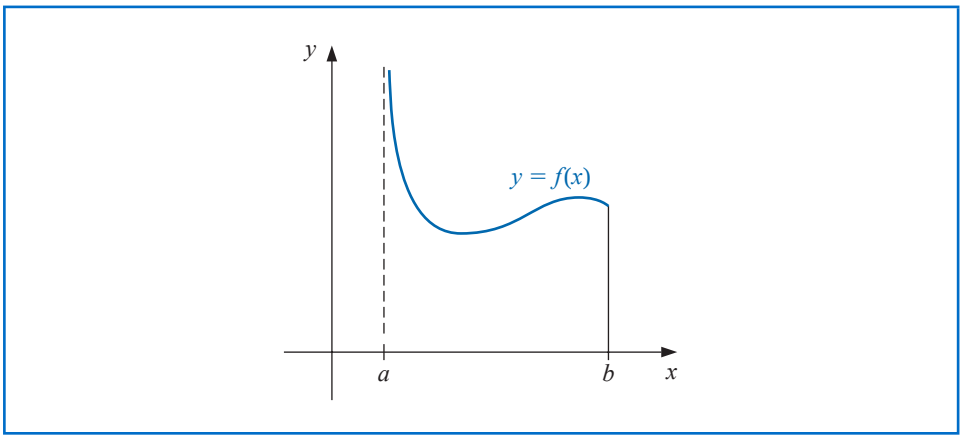
4.9 Improper Integrals

Improper integrals result when the notion of integration is extended either to an interval of integration on which the function is unbounded or to an interval with one or more infinite endpoints. In either circumstance, the normal rules of integral approximation must be modified.

Left Endpoint Singularity

We will first consider the situation when the integrand is unbounded at the left endpoint of the interval of integration, as shown in Figure 4.25. In this case we say that f has a **singularity** at the endpoint a. We will then show how other improper integrals can be reduced to problems of this form.

Figure 4.25



It is shown in calculus that the improper integral with a singularity at the left endpoint,

$$\int_a^b \frac{dx}{(x-a)^p},$$

converges if and only if $0 < p < 1$, and in this case, we define

$$\int_a^b \frac{1}{(x-a)^p}\,dx = \lim_{M \to a^+} \frac{(x-a)^{1-p}}{1-p}\bigg|_{x=M}^{x=b} = \frac{(b-a)^{1-p}}{1-p}.$$

Example 1 Show that the improper integral $\displaystyle\int_0^1 \frac{1}{\sqrt{x}}\,dx$ converges but $\displaystyle\int_0^1 \frac{1}{x^2}\,dx$ diverges.

Solution For the first integral we have

$$\int_0^1 \frac{1}{\sqrt{x}}\,dx = \lim_{M \to 0^+} \int_M^1 x^{-1/2}\,dx = \lim_{M \to 0^+} 2x^{1/2}\Big|_{x=M}^{x=1} = 2 - 0 = 2,$$

but the second integral

$$\int_0^1 \frac{1}{x^2}\,dx = \lim_{M \to 0^+} \int_M^1 x^{-2}\,dx = \lim_{M \to 0^+} -x^{-1}\Big|_{x=M}^{x=1}$$

is unbounded.

If f is a function that can be written in the form

$$f(x) = \frac{g(x)}{(x-a)^p},$$

where $0 < p < 1$ and g is continuous on $[a, b]$, then the improper integral

$$\int_a^b f(x)\,dx$$

also exists. We will approximate this integral using the Composite Simpson's rule, provided that $g \in C^5[a, b]$. In that case, we can construct the fourth Taylor polynomial, $P_4(x)$, for g about a,

$$P_4(x) = g(a) + g'(a)(x-a) + \frac{g''(a)}{2!}(x-a)^2 + \frac{g'''(a)}{3!}(x-a)^3 + \frac{g^{(4)}(a)}{4!}(x-a)^4,$$

and write

$$\int_a^b f(x)\,dx = \int_a^b \frac{g(x) - P_4(x)}{(x-a)^p}\,dx + \int_a^b \frac{P_4(x)}{(x-a)^p}\,dx. \tag{4.44}$$

Because $P(x)$ is a polynomial, we can exactly determine the value of

$$\int_a^b \frac{P_4(x)}{(x-a)^p}\,dx = \sum_{k=0}^{4} \int_a^b \frac{g^{(k)}(a)}{k!}(x-a)^{k-p}\,dx = \sum_{k=0}^{4} \frac{g^{(k)}(a)}{k!(k+1-p)}(b-a)^{k+1-p}. \tag{4.45}$$

This is generally the dominant portion of the approximation, especially when the Taylor polynomial $P_4(x)$ agrees closely with $g(x)$ throughout the interval $[a, b]$.

To approximate the integral of f, we must add to this value the approximation of

$$\int_a^b \frac{g(x) - P_4(x)}{(x-a)^p}\,dx.$$

To determine this, we first define

$$G(x) = \begin{cases} \frac{g(x)-P_4(x)}{(x-a)^p}, & \text{if } a < x \le b, \\ 0, & \text{if } x = a. \end{cases}$$

This gives us a continuous function on $[a, b]$. In fact, $0 < p < 1$ and $P_4^{(k)}(a)$ agrees with $g^{(k)}(a)$ for each $k = 0, 1, 2, 3, 4$, so we have $G \in C^4[a, b]$. This implies that the Composite Simpson's rule can be applied to approximate the integral of G on $[a, b]$. Adding this approximation to the value in Eq. (4.45) gives an approximation to the improper integral of f on $[a, b]$, within the accuracy of the Composite Simpson's rule approximation.

Example 2 Use Composite Simpson's rule with $h = 0.25$ to approximate the value of the improper integral

$$\int_0^1 \frac{e^x}{\sqrt{x}}\,dx.$$

Solution The fourth Taylor polynomial for e^x about $x = 0$ is

$$P_4(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24},$$

so the dominant portion of the approximation to $\int_0^1 \frac{e^x}{\sqrt{x}}\,dx$ is

$$\begin{aligned}\int_0^1 \frac{P_4(x)}{\sqrt{x}}\,dx &= \int_0^1 \left(x^{-1/2} + x^{1/2} + \frac{1}{2}x^{3/2} + \frac{1}{6}x^{5/2} + \frac{1}{24}x^{7/2}\right) dx \\ &= \lim_{M\to 0^+}\left[2x^{1/2} + \frac{2}{3}x^{3/2} + \frac{1}{5}x^{5/2} + \frac{1}{21}x^{7/2} + \frac{1}{108}x^{9/2}\right]_M^1 \\ &= 2 + \frac{2}{3} + \frac{1}{5} + \frac{1}{21} + \frac{1}{108} \approx 2.9235450.\end{aligned}$$

For the second portion of the approximation to $\int_0^1 \frac{e^x}{\sqrt{x}}\,dx$ we need to approximate $\int_0^1 G(x)\,dx$, where

$$G(x) = \begin{cases} \frac{1}{\sqrt{x}}\,(e^x - P_4(x)), & \text{if } 0 < x \le 1, \\ 0, & \text{if } x = 0. \end{cases}$$

Table 4.13

x	$G(x)$
0.00	0
0.25	0.0000170
0.50	0.0004013
0.75	0.0026026
1.00	0.0099485

Table 4.13 lists the values needed for the Composite Simpson's rule for this approximation.

Using these data and the Composite Simpson's rule gives

$$\begin{aligned}\int_0^1 G(x)\,dx &\approx \frac{0.25}{3}[0 + 4(0.0000170) + 2(0.0004013) + 4(0.0026026) + 0.0099485] \\ &= 0.0017691.\end{aligned}$$

Hence

$$\int_0^1 \frac{e^x}{\sqrt{x}}\,dx \approx 2.9235450 + 0.0017691 = 2.9253141.$$

This result is accurate to within the accuracy of the Composite Simpson's rule approximation for the function G. Because $|G^{(4)}(x)| < 1$ on $[0, 1]$, the error is bounded by

$$\frac{1-0}{180}(0.25)^4 = 0.0000217.$$ ■

Right Endpoint Singularity

To approximate the improper integral with a singularity at the right endpoint, we could develop a similar technique but expand in terms of the right endpoint b instead of the left endpoint a. Alternatively, we can make the substitution

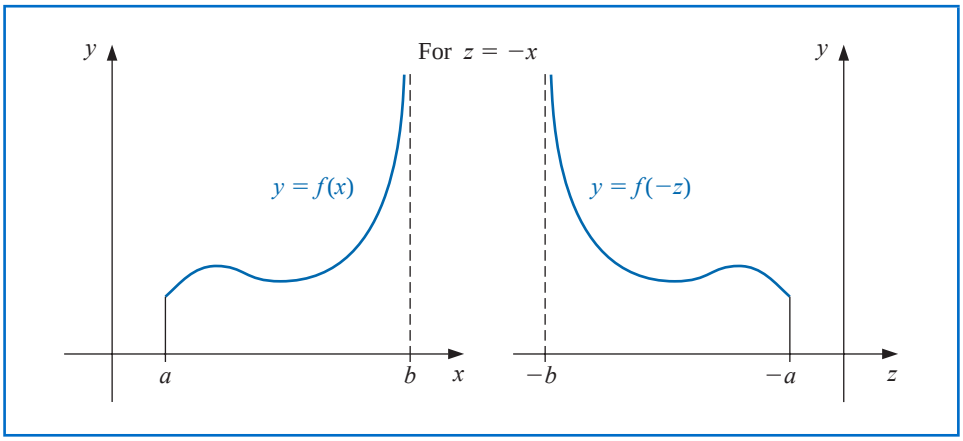
$$z = -x, \qquad dz = -\,dx$$

to change the improper integral into one of the form

$$\int_a^b f(x)\,dx = \int_{-b}^{-a} f(-z)\,dz, \tag{4.46}$$

which has its singularity at the left endpoint. Then we can apply the left endpoint singularity technique we have already developed. (See Figure 4.26.)

Figure 4.26



An improper integral with a singularity at c, where $a < c < b$, is treated as the sum of improper integrals with endpoint singularities since

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx.$$

Infinite Singularity

The other type of improper integral involves infinite limits of integration. The basic integral of this type has the form

$$\int_a^\infty \frac{1}{x^p}\,dx,$$

for $p > 1$. This is converted to an integral with left endpoint singularity at 0 by making the integration substitution

$$t = x^{-1}, \quad dt = -x^{-2}\,dx, \quad \text{so} \quad dx = -x^2\,dt = -t^{-2}\,dt.$$

Then

$$\int_a^\infty \frac{1}{x^p}\,dx = \int_{1/a}^0 -\frac{t^p}{t^2}\,dt = \int_0^{1/a} \frac{1}{t^{2-p}}\,dt.$$

In a similar manner, the variable change $t = x^{-1}$ converts the improper integral $\int_a^\infty f(x)\,dx$ into one that has a left endpoint singularity at zero:

$$\int_a^\infty f(x)\,dx = \int_0^{1/a} t^{-2} f\left(\frac{1}{t}\right)\,dt. \tag{4.47}$$

It can now be approximated using a quadrature formula of the type described earlier.

Example 3 Approximate the value of the improper integral

$$I = \int_1^\infty x^{-3/2} \sin\frac{1}{x}\,dx.$$

Solution We first make the variable change $t = x^{-1}$, which converts the infinite singularity into one with a left endpoint singularity. Then

$$dt = -x^{-2}\,dx, \quad \text{so} \quad dx = -x^2\,dt = -\frac{1}{t^2}\,dt,$$

and

$$I = \int_{x=1}^{x=\infty} x^{-3/2} \sin\frac{1}{x}\,dx = \int_{t=1}^{t=0} \left(\frac{1}{t}\right)^{-3/2} \sin t \left(-\frac{1}{t^2}\,dt\right) = \int_0^1 t^{-1/2} \sin t\,dt.$$

The fourth Taylor polynomial, $P_4(t)$, for $\sin t$ about 0 is

$$P_4(t) = t - \frac{1}{6}t^3,$$

so

$$G(t) = \begin{cases} \dfrac{\sin t - t + \frac{1}{6}t^3}{t^{1/2}}, & \text{if } 0 < t \le 1 \\ 0, & \text{if } t = 0 \end{cases}$$

is in $C^4[0, 1]$, and we have

$$\begin{aligned} I &= \int_0^1 t^{-1/2}\left(t - \frac{1}{6}t^3\right) dt + \int_0^1 \frac{\sin t - t + \frac{1}{6}t^3}{t^{1/2}}\,dt \\ &= \left[\frac{2}{3}t^{3/2} - \frac{1}{21}t^{7/2}\right]_0^1 + \int_0^1 \frac{\sin t - t + \frac{1}{6}t^3}{t^{1/2}}\,dt \\ &= 0.61904761 + \int_0^1 \frac{\sin t - t + \frac{1}{6}t^3}{t^{1/2}}\,dt. \end{aligned}$$

The result from the Composite Simpson's rule with $n = 16$ for the remaining integral is 0.0014890097. This gives a final approximation of

$$I = 0.0014890097 + 0.61904761 = 0.62053661,$$

which is accurate to within 4.0×10^{-8}. ■

EXERCISE SET 4.9

1. Use Simpson's Composite rule and the given values of n to approximate the following improper integrals.

a. $\displaystyle\int_0^1 x^{-1/4} \sin x\,dx, \quad n = 4$

b. $\displaystyle\int_0^1 \frac{e^{2x}}{\sqrt[5]{x^2}}\,dx, \quad n = 6$

c. $\displaystyle\int_1^2 \frac{\ln x}{(x-1)^{1/5}}\,dx, \quad n = 8$

d. $\displaystyle\int_0^1 \frac{\cos 2x}{x^{1/3}}\,dx, \quad n = 6$

2. Use the Composite Simpson's rule and the given values of n to approximate the following improper integrals.

a. $\displaystyle\int_0^1 \frac{e^{-x}}{\sqrt{1-x}}\,dx, \quad n = 6$

b. $\displaystyle\int_0^2 \frac{xe^x}{\sqrt[3]{(x-1)^2}}\,dx, \quad n = 8$

3. Use the transformation $t = x^{-1}$ and then the Composite Simpson's rule and the given values of n to approximate the following improper integrals.

a. $\displaystyle\int_1^\infty \frac{1}{x^2+9}\,dx, \quad n = 4$

b. $\displaystyle\int_1^\infty \frac{1}{1+x^4}\,dx, \quad n = 4$

c. $\displaystyle\int_1^\infty \frac{\cos x}{x^3}\,dx, \quad n = 6$

d. $\displaystyle\int_1^\infty x^{-4}\sin x\,dx, \quad n = 6$

4. The improper integral $\int_0^\infty f(x)\,dx$ cannot be converted into an integral with finite limits using the substitution $t = 1/x$ because the limit at zero becomes infinite. The problem is resolved by first writing $\int_0^\infty f(x)\,dx = \int_0^1 f(x)\,dx + \int_1^\infty f(x)\,dx$. Apply this technique to approximate the following improper integrals to within 10^{-6}.

a. $\displaystyle\int_0^\infty \frac{1}{1+x^4}\,dx$

b. $\displaystyle\int_0^\infty \frac{1}{(1+x^2)^3}\,dx$

5. Suppose a body of mass m is traveling vertically upward starting at the surface of the earth. If all resistance except gravity is neglected, the escape velocity v is given by

$$v^2 = 2gR\int_1^\infty z^{-2}\,dz, \quad \text{where } z = \frac{x}{R},$$

$R = 3960$ miles is the radius of the earth, and $g = 0.00609$ mi/s^2 is the force of gravity at the earth's surface. Approximate the escape velocity v.

6. The Laguerre polynomials $\{L_0(x), L_1(x)\ldots\}$ form an orthogonal set on $[0,\infty)$ and satisfy $\int_0^\infty e^{-x}L_i(x)L_j(x)\,dx = 0$, for $i \neq j$. (See Section 8.2.) The polynomial $L_n(x)$ has n distinct zeros $x_1, x_2, \ldots, x_n$ in $[0,\infty)$. Let

$$c_{n,i} = \int_0^\infty e^{-x}\prod_{\substack{j=1\\ j\neq i}}^{n}\frac{x-x_j}{x_i-x_j}\,dx.$$

Show that the quadrature formula

$$\int_0^\infty f(x)e^{-x}\,dx = \sum_{i=1}^{n} c_{n,i}f(x_i)$$

has degree of precision $2n-1$. (Hint: Follow the steps in the proof of Theorem 4.7.)

7. The Laguerre polynomials $L_0(x) = 1$, $L_1(x) = 1-x$, $L_2(x) = x^2-4x+2$, and $L_3(x) = -x^3+9x^2-18x+6$ are derived in Exercise 11 of Section 8.2. As shown in Exercise 6, these polynomials are useful in approximating integrals of the form

$$\int_0^\infty e^{-x}f(x)\,dx = 0.$$

a. Derive the quadrature formula using $n = 2$ and the zeros of $L_2(x)$.

b. Derive the quadrature formula using $n = 3$ and the zeros of $L_3(x)$.

8. Use the quadrature formulas derived in Exercise 7 to approximate the integral

$$\int_0^\infty \sqrt{x}e^{-x}\,dx.$$

9. Use the quadrature formulas derived in Exercise 7 to approximate the integral

$$\int_{-\infty}^\infty \frac{1}{1+x^2}\,dx.$$

4.10 Survey of Methods and Software

In this chapter we considered approximating integrals of functions of one, two, or three variables, and approximating the derivatives of a function of a single real variable.

The Midpoint rule, Trapezoidal rule, and Simpson's rule were studied to introduce the techniques and error analysis of quadrature methods. Composite Simpson's rule is easy to use and produces accurate approximations unless the function oscillates in a subinterval of the interval of integration. Adaptive quadrature can be used if the function is suspected of oscillatory behavior. To minimize the number of nodes while maintaining accuracy, we used Gaussian quadrature. Romberg integration was introduced to take advantage of the easily applied Composite Trapezoidal rule and extrapolation.

Most software for integrating a function of a single real variable is based either on the adaptive approach or extremely accurate Gaussian formulas. Cautious Romberg integration is an adaptive technique that includes a check to make sure that the integrand is smoothly behaved over subintervals of the integral of integration. This method has been successfully used in software libraries. Multiple integrals are generally approximated by extending good adaptive methods to higher dimensions. Gaussian-type quadrature is also recommended to decrease the number of function evaluations.

The main routines in both the IMSL and NAG Libraries are based on *QUADPACK: A Subroutine Package for Automatic Integration* by R. Piessens, E. de Doncker-Kapenga, C. W. Uberhuber, and D. K. Kahaner published by Springer-Verlag in 1983 [PDUK].

The IMSL Library contains an adaptive integration scheme based on the 21-point Gaussian-Kronrod rule using the 10-point Gaussian rule for error estimation. The Gaussian rule uses the ten points $x_1, \ldots, x_{10}$ and weights $w_1, \ldots, w_{10}$ to give the quadrature formula $\sum_{i=1}^{10} w_i f(x_i)$ to approximate $\int_a^b f(x)\,dx$. The additional points $x_{11}, \ldots, x_{21}$, and the new weights $v_1, \ldots, v_{21}$, are then used in the Kronrod formula $\sum_{i=1}^{21} v_i f(x_i)$. The results of the two formulas are compared to eliminate error. The advantage in using $x_1, \ldots, x_{10}$ in each formula is that f needs to be evaluated only at 21 points. If independent 10- and 21-point Gaussian rules were used, 31 function evaluations would be needed. This procedure permits endpoint singularities in the integrand.

Other IMSL subroutines allow for endpoint singularities, user-specified singularities, and infinite intervals of integration. In addition, there are routines for applying Gauss-Kronrod rules to integrate a function of two variables, and a routine to use Gaussian quadrature to integrate a function of n variables over n intervals of the form $[a_i, b_i]$.

The NAG Library includes a routine to compute the integral of f over the interval $[a, b]$ using an adaptive method based on Gaussian Quadrature using Gauss 10-point and Kronrod 21-point rules. It also has a routine to approximate an integral using a family of Gaussian-type formulas based on 1, 3, 5, 7, 15, 31, 63, 127, and 255 nodes. These interlacing high-precision rules are due to Patterson [Pat] and are used in an adaptive manner. NAG includes many other subroutines for approximating integrals.

MATLAB has a routine to approximate a definite integral using an adaptive Simpson's rule, and another to approximate the definite integral using an adaptive eight-panel Newton-Cotes rule.

Although numerical differentiation is unstable, derivative approximation formulas are needed for solving differential equations. The NAG Library includes a subroutine for the numerical differentiation of a function of one real variable with differentiation to the fourteenth derivative being possible. IMSL has a function that uses an adaptive change in step size for finite differences to approximate the first, second, or third, derivative of f at x to within a given tolerance. IMSL also includes a subroutine to compute the derivatives of a function defined on a set of points using quadratic interpolation. Both packages allow the

differentiation and integration of interpolatory cubic splines constructed by the subroutines mentioned in Section 3.5.

For further reading on numerical integration we recommend the books by Engels [E] and by Davis and Rabinowitz [DR]. For more information on Gaussian quadrature see Stroud and Secrest [StS]. Books on multiple integrals include those by Stroud [Stro] and by Sloan and Joe [SJ].