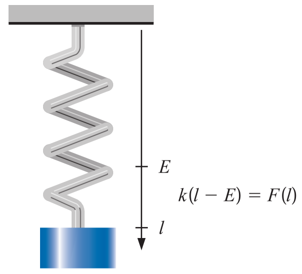
CHAPTER

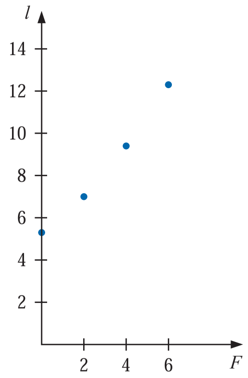
8

Approximation Theory

Introduction

Hooke's law states that when a force is applied to a spring constructed of uniform material, the length of the spring is a linear function of that force. We can write the linear function as $F(l) = k(l - E)$, where $F(l)$ represents the force required to stretch the spring l units, the constant E represents the length of the spring with no force applied, and the constant k is the spring constant.





Suppose we want to determine the spring constant for a spring that has initial length 5.3 in. We apply forces of 2, 4, and 6 lb to the spring and find that its length increases to 7.0, 9.4, and 12.3 in., respectively. A quick examination shows that the points $(0, 5.3)$, $(2, 7.0)$, $(4, 9.4)$, and $(6, 12.3)$ do not quite lie in a straight line. Although we could use a random pair of these data points to approximate the spring constant, it would seem more reasonable to find the line that *best* approximates all the data points to determine the constant. This type of approximation will be considered in this chapter, and this spring application can be found in Exercise 7 of Section 8.1.

Approximation theory involves two general types of problems. One problem arises when a function is given explicitly, but we wish to find a "simpler" type of function, such as a polynomial, to approximate values of the given function. The other problem in approximation theory is concerned with fitting functions to given data and finding the "best" function in a certain class to represent the data.

Both problems have been touched upon in Chapter 3. The nth Taylor polynomial about the number x_0 is an excellent approximation to an $(n + 1)$-times differentiable function f in a small neighborhood of x_0. The Lagrange interpolating polynomials, or, more generally, osculatory polynomials, were discussed both as approximating polynomials and as polynomials to fit certain data. Cubic splines were also discussed in that chapter. In this chapter, limitations to these techniques are considered, and other avenues of approach are discussed.

8.1 Discrete Least Squares Approximation

Consider the problem of estimating the values of a function at nontabulated points, given the experimental data in Table 8.1.

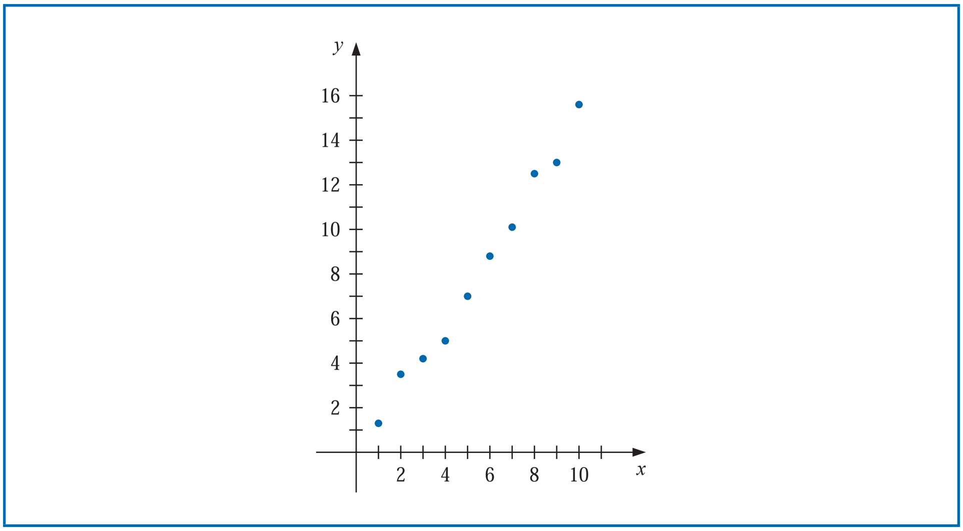
Table 8.1

x_i	y_i	x_i	y_i
1	1.3	6	8.8
2	3.5	7	10.1
3	4.2	8	12.5
4	5.0	9	13.0
5	7.0	10	15.6

Figure 8.1 shows a graph of the values in Table 8.1. From this graph, it appears that the actual relationship between x and y is linear. The likely reason that no line precisely fits the data is because of errors in the data. So it is unreasonable to require that the approximating function agree exactly with the data. In fact, such a function would introduce oscillations that were not originally present. For example, the graph of the ninth-degree interpolating polynomial shown in unconstrained mode for the data in Table 8.1 is obtained in Maple using the commands

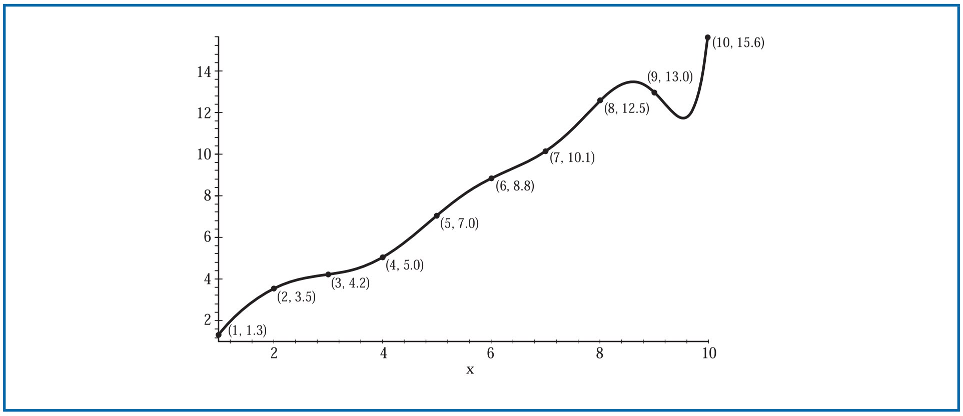
$p := interp([1, 2, 3, 4, 5, 6, 7, 8, 9, 10], [1.3, 3.5, 4.2, 5.0, 7.0, 8.8, 10.1, 12.5, 13.0, 15.6], x):$
$plot(p, x = 1..10)$

Figure 8.1



The plot obtained (with the data points added) is shown in Figure 8.2.

Figure 8.2



This polynomial is clearly a poor predictor of information between a number of the data points. A better approach would be to find the “best” (in some sense) approximating line, even if it does not agree precisely with the data at any point.

Let $a_1x_i + a_0$ denote the ith value on the approximating line and y_i be the ith given y-value. We assume throughout that the independent variables, the x_i, are exact, it is the dependent variables, the y_i, that are suspect. This is a reasonable assumption in most experimental situations.

The problem of finding the equation of the best linear approximation in the absolute sense requires that values of a_0 and a_1 be found to minimize

$$E_\infty(a_0, a_1) = \max_{1\le i\le 10} \{|y_i - (a_1x_i + a_0)|\}.$$

This is commonly called a **minimax** problem and cannot be handled by elementary techniques.

Another approach to determining the best linear approximation involves finding values of a_0 and a_1 to minimize

$$E_1(a_0, a_1) = \sum_{i=1}^{10} |y_i - (a_1x_i + a_0)|.$$

This quantity is called the **absolute deviation**. To minimize a function of two variables, we need to set its partial derivatives to zero and simultaneously solve the resulting equations. In the case of the absolute deviation, we need to find a_0 and a_1 with

$$0 = \frac{\partial}{\partial a_0} \sum_{i=1}^{10} |y_i - (a_1x_i + a_0)| \quad \text{and} \quad 0 = \frac{\partial}{\partial a_1} \sum_{i=1}^{10} |y_i - (a_1x_i + a_0)|.$$

The problem is that the absolute-value function is not differentiable at zero, and we might not be able to find solutions to this pair of equations.

Linear Least Squares

The **least squares** approach to this problem involves determining the best approximating line when the error involved is the sum of the squares of the differences between the y-values on the approximating line and the given y-values. Hence, constants a_0 and a_1 must be found that minimize the least squares error:

$$E_2(a_0, a_1) = \sum_{i=1}^{10} \left[y_i - (a_1x_i + a_0)\right]^2.$$

The least squares method is the most convenient procedure for determining best linear approximations, but there are also important theoretical considerations that favor it. The minimax approach generally assigns too much weight to a bit of data that is badly in error, whereas the absolute deviation method does not give sufficient weight to a point that is considerably out of line with the approximation. The least squares approach puts substantially more weight on a point that is out of line with the rest of the data, but will not permit that point to completely dominate the approximation. An additional reason for considering the least squares approach involves the study of the statistical distribution of error. (See [Lar], pp. 463–481.)

The general problem of fitting the best least squares line to a collection of data $\{(x_i, y_i)\}_{i=1}^m$ involves minimizing the total error,

$$E \equiv E_2(a_0, a_1) = \sum_{i=1}^{m} \left[y_i - (a_1x_i + a_0)\right]^2,$$

with respect to the parameters a_0 and a_1. For a minimum to occur, we need both

$$\frac{\partial E}{\partial a_0} = 0 \quad \text{and} \quad \frac{\partial E}{\partial a_1} = 0,$$

that is,

$$0 = \frac{\partial}{\partial a_0} \sum_{i=1}^{m} \left[(y_i - (a_1 x_i - a_0)\right]^2 = 2 \sum_{i=1}^{m} (y_i - a_1 x_i - a_0)(-1)$$

and

$$0 = \frac{\partial}{\partial a_1} \sum_{i=1}^{m} \left[y_i - (a_1 x_i + a_0)\right]^2 = 2 \sum_{i=1}^{m} (y_i - a_1 x_i - a_0)(-x_i).$$

The word normal as used here implies perpendicular. The normal equations are obtained by finding perpendicular directions to a multidimensional surface.

These equations simplify to the **normal equations**:

$$a_0 \cdot m + a_1 \sum_{i=1}^{m} x_i = \sum_{i=1}^{m} y_i \quad \text{and} \quad a_0 \sum_{i=1}^{m} x_i + a_1 \sum_{i=1}^{m} x_i^2 = \sum_{i=1}^{m} x_i y_i.$$

The solution to this system of equations is

$$a_0 = \frac{\displaystyle\sum_{i=1}^{m} x_i^2 \sum_{i=1}^{m} y_i - \sum_{i=1}^{m} x_i y_i \sum_{i=1}^{m} x_i}{m\left(\displaystyle\sum_{i=1}^{m} x_i^2\right) - \left(\displaystyle\sum_{i=1}^{m} x_i\right)^2} \tag{8.1}$$

and

$$a_1 = \frac{m\displaystyle\sum_{i=1}^{m} x_i y_i - \sum_{i=1}^{m} x_i \sum_{i=1}^{m} y_i}{m\left(\displaystyle\sum_{i=1}^{m} x_i^2\right) - \left(\displaystyle\sum_{i=1}^{m} x_i\right)^2}. \tag{8.2}$$

Example 1 Find the least squares line approximating the data in Table 8.1.

Solution We first extend the table to include x_i^2 and $x_i y_i$ and sum the columns. This is shown in Table 8.2.

Table 8.2

x_i	y_i	x_i^2	$x_i y_i$	$P(x_i) = 1.538x_i - 0.360$
1	1.3	1	1.3	1.18
2	3.5	4	7.0	2.72
3	4.2	9	12.6	4.25
4	5.0	16	20.0	5.79
5	7.0	25	35.0	7.33
6	8.8	36	52.8	8.87
7	10.1	49	70.7	10.41
8	12.5	64	100.0	11.94
9	13.0	81	117.0	13.48
10	15.6	100	156.0	15.02
55	81.0	385	572.4	$E = \sum_{i=1}^{10} (y_i - P(x_i))^2 \approx 2.34$

The normal equations (8.1) and (8.2) imply that

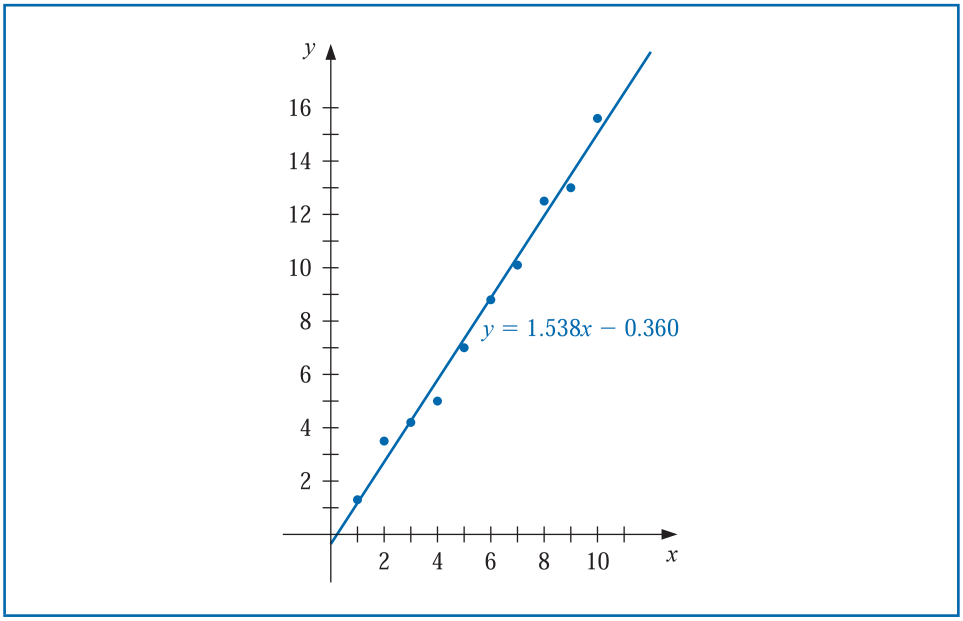
$$a_0 = \frac{385(81) - 55(572.4)}{10(385) - (55)^2} = -0.360$$

and

$$a_1 = \frac{10(572.4) - 55(81)}{10(385) - (55)^2} = 1.538,$$

so $P(x) = 1.538x - 0.360$. The graph of this line and the data points are shown in Figure 8.3. The approximate values given by the least squares technique at the data points are in Table 8.2. ■

Figure 8.3



Polynomial Least Squares

The general problem of approximating a set of data, $\{(x_i, y_i) \mid i = 1, 2, \ldots, m\}$, with an algebraic polynomial

$$P_n(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0,$$

of degree $n < m - 1$, using the least squares procedure is handled similarly. We choose the constants $a_0, a_1, \ldots, a_n$ to minimize the least squares error $E = E_2(a_0, a_1, \ldots, a_n)$, where

$$\begin{aligned} E &= \sum_{i=1}^{m} (y_i - P_n(x_i))^2 \\ &= \sum_{i=1}^{m} y_i^2 - 2 \sum_{i=1}^{m} P_n(x_i) y_i + \sum_{i=1}^{m} (P_n(x_i))^2 \end{aligned}$$

$$= \sum_{i=1}^{m} y_i^2 - 2\sum_{i=1}^{m}\left(\sum_{j=0}^{n} a_j x_i^j\right) y_i + \sum_{i=1}^{m}\left(\sum_{j=0}^{n} a_j x_i^j\right)^2$$

$$= \sum_{i=1}^{m} y_i^2 - 2\sum_{j=0}^{n} a_j\left(\sum_{i=1}^{m} y_i x_i^j\right) + \sum_{j=0}^{n}\sum_{k=0}^{n} a_j a_k\left(\sum_{i=1}^{m} x_i^{j+k}\right).$$

As in the linear case, for E to be minimized it is necessary that $\partial E/\partial a_j = 0$, for each $j = 0, 1, \ldots, n$. Thus, for each j, we must have

$$0 = \frac{\partial E}{\partial a_j} = -2\sum_{i=1}^{m} y_i x_i^j + 2\sum_{k=0}^{n} a_k \sum_{i=1}^{m} x_i^{j+k}.$$

This gives $n+1$ **normal equations** in the $n+1$ unknowns a_j. These are

$$\sum_{k=0}^{n} a_k \sum_{i=1}^{m} x_i^{j+k} = \sum_{i=1}^{m} y_i x_i^j, \quad \text{for each } j = 0, 1, \ldots, n. \tag{8.3}$$

It is helpful to write the equations as follows:

$$a_0 \sum_{i=1}^{m} x_i^0 + a_1 \sum_{i=1}^{m} x_i^1 + a_2 \sum_{i=1}^{m} x_i^2 + \cdots + a_n \sum_{i=1}^{m} x_i^n = \sum_{i=1}^{m} y_i x_i^0,$$

$$a_0 \sum_{i=1}^{m} x_i^1 + a_1 \sum_{i=1}^{m} x_i^2 + a_2 \sum_{i=1}^{m} x_i^3 + \cdots + a_n \sum_{i=1}^{m} x_i^{n+1} = \sum_{i=1}^{m} y_i x_i^1,$$

$$\vdots$$

$$a_0 \sum_{i=1}^{m} x_i^n + a_1 \sum_{i=1}^{m} x_i^{n+1} + a_2 \sum_{i=1}^{m} x_i^{n+2} + \cdots + a_n \sum_{i=1}^{m} x_i^{2n} = \sum_{i=1}^{m} y_i x_i^n.$$

These *normal equations* have a unique solution provided that the x_i are distinct (see Exercise 14).

Example 2 Fit the data in Table 8.3 with the discrete least squares polynomial of degree at most 2.

Table 8.3

i	x_i	y_i
1	0	1.0000
2	0.25	1.2840
3	0.50	1.6487
4	0.75	2.1170
5	1.00	2.7183

Solution For this problem, $n = 2, m = 5$, and the three normal equations are

$$\begin{aligned} 5a_0 + \quad 2.5a_1 + \quad 1.875a_2 &= 8.7680, \\ 2.5a_0 + \ 1.875a_1 + 1.5625a_2 &= 5.4514, \\ 1.875a_0 + 1.5625a_1 + 1.3828a_2 &= 4.4015. \end{aligned}$$

To solve this system using Maple, we first define the equations

$eq1 := 5a0 + 2.5a1 + 1.875a2 = 8.7680:$
$eq2 := 2.5a0 + 1.875a1 + 1.5625a2 = 5.4514:$
$eq3 := 1.875a0 + 1.5625a1 + 1.3828a2 = 4.4015$

and then solve the system with

$solve(\{eq1, eq2, eq3\}, \{a0, a1, a2\})$

This gives

$$\{a_0 = 1.005075519, \quad a_1 = 0.8646758482, \quad a_2 = .8431641518\}$$

Thus the least squares polynomial of degree 2 fitting the data in Table 8.3 is

$$P_2(x) = 1.0051 + 0.86468x + 0.84316x^2,$$

whose graph is shown in Figure 8.4. At the given values of x_i we have the approximations shown in Table 8.4.

Figure 8.4

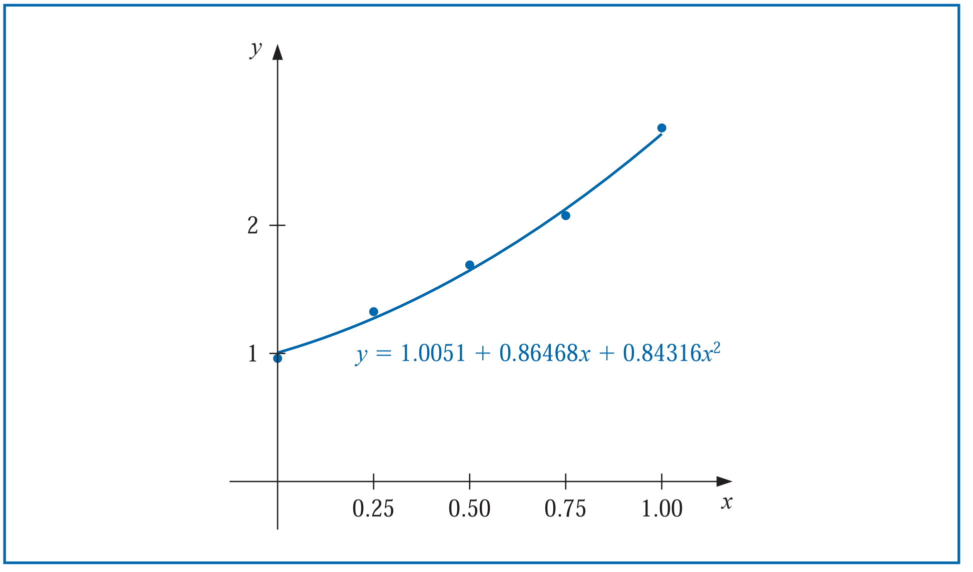


Table 8.4

i	1	2	3	4	5
x_i	0	0.25	0.50	0.75	1.00
y_i	1.0000	1.2840	1.6487	2.1170	2.7183
$P(x_i)$	1.0051	1.2740	1.6482	2.1279	2.7129
$y_i - P(x_i)$	−0.0051	0.0100	0.0004	−0.0109	0.0054

The total error,

$$E = \sum_{i=1}^{5}(y_i - P(x_i))^2 = 2.74 \times 10^{-4},$$

is the least that can be obtained by using a polynomial of degree at most 2. ■

Maple has a function called *LinearFit* within the *Statistics* package which can be used to compute the discrete least squares approximations. To compute the approximation in Example 2 we first load the package and define the data

$with(Statistics)$: $xvals := Vector([0, 0.25, 0.5, 0.75, 1])$: $yvals := Vector([1, 1.284, 1.6487, 2.117, 2.7183])$:

To define the least squares polynomial for this data we enter the command

$P := x \to LinearFit([1, x, x^2], xvals, yvals, x)$: $P(x)$

Maple returns a result which rounded to 5 decimal places is

$$1.00514 + 0.86418x + 0.84366x^2$$

The approximation at a specific value, for example at $x = 1.7$, is found with $P(1.7)$

$$4.91242$$

At times it is appropriate to assume that the data are exponentially related. This requires the approximating function to be of the form

$$y = be^{ax} \tag{8.4}$$

or

$$y = bx^a, \tag{8.5}$$

for some constants a and b. The difficulty with applying the least squares procedure in a situation of this type comes from attempting to minimize

$$E = \sum_{i=1}^{m} (y_i - be^{ax_i})^2, \quad \text{in the case of Eq. (8.4),}$$

or

$$E = \sum_{i=1}^{m} (y_i - bx_i^a)^2, \quad \text{in the case of Eq. (8.5).}$$

The normal equations associated with these procedures are obtained from either

$$0 = \frac{\partial E}{\partial b} = 2\sum_{i=1}^{m} (y_i - be^{ax_i})(-e^{ax_i})$$

and

$$0 = \frac{\partial E}{\partial a} = 2\sum_{i=1}^{m} (y_i - be^{ax_i})(-bx_ie^{ax_i}), \quad \text{in the case of Eq. (8.4);}$$

or

$$0 = \frac{\partial E}{\partial b} = 2\sum_{i=1}^{m} (y_i - bx_i^a)(-x_i^a)$$

and

$$0 = \frac{\partial E}{\partial a} = 2\sum_{i=1}^{m} (y_i - bx_i^a)(-b(\ln x_i)x_i^a), \quad \text{in the case of Eq. (8.5).}$$

No exact solution to either of these systems in a and b can generally be found.

The method that is commonly used when the data are suspected to be exponentially related is to consider the logarithm of the approximating equation:

$$\ln y = \ln b + ax, \quad \text{in the case of Eq. (8.4),}$$

and

$$\ln y = \ln b + a \ln x, \quad \text{in the case of Eq. (8.5).}$$

In either case, a linear problem now appears, and solutions for $\ln b$ and a can be obtained by appropriately modifying the normal equations (8.1) and (8.2).

However, the approximation obtained in this manner is *not* the least squares approximation for the original problem, and this approximation can in some cases differ significantly from the least squares approximation to the original problem. The application in Exercise 13 describes such a problem. This application will be reconsidered as Exercise 11 in Section 10.3, where the exact solution to the exponential least squares problem is approximated by using methods suitable for solving nonlinear systems of equations.

Illustration Consider the collection of data in the first three columns of Table 8.5.

Table 8.5

i	x_i	y_i	$\ln y_i$	x_i^2	$x_i \ln y_i$
1	1.00	5.10	1.629	1.0000	1.629
2	1.25	5.79	1.756	1.5625	2.195
3	1.50	6.53	1.876	2.2500	2.814
4	1.75	7.45	2.008	3.0625	3.514
5	2.00	8.46	2.135	4.0000	4.270
	7.50		9.404	11.875	14.422

If x_i is graphed with $\ln y_i$, the data appear to have a linear relation, so it is reasonable to assume an approximation of the form

$$y = be^{ax}, \quad \text{which implies that} \quad \ln y = \ln b + ax.$$

Extending the table and summing the appropriate columns gives the remaining data in Table 8.5.

Using the normal equations (8.1) and (8.2),

$$a = \frac{(5)(14.422) - (7.5)(9.404)}{(5)(11.875) - (7.5)^2} = 0.5056$$

and

$$\ln b = \frac{(11.875)(9.404) - (14.422)(7.5)}{(5)(11.875) - (7.5)^2} = 1.122.$$

With $\ln b = 1.122$ we have $b = e^{1.122} = 3.071$, and the approximation assumes the form

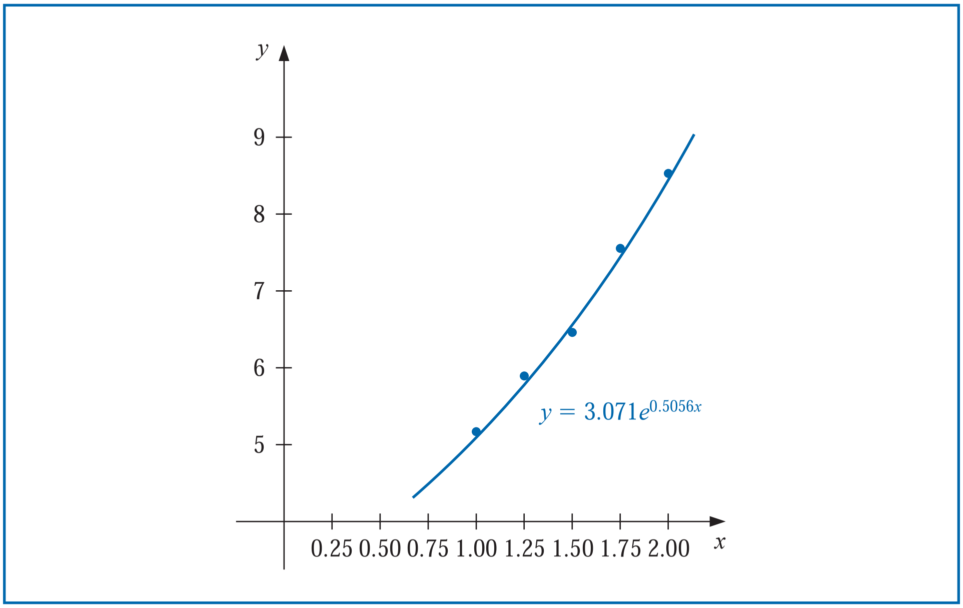
$$y = 3.071e^{0.5056x}.$$

At the data points this gives the values in Table 8.6. (See Figure 8.5.) □

Table 8.6

i	x_i	y_i	$3.071e^{0.5056x_i}$	$\lvert y_i - 3.071e^{0.5056x_i}\rvert$
1	1.00	5.10	5.09	0.01
2	1.25	5.79	5.78	0.01
3	1.50	6.53	6.56	0.03
4	1.75	7.45	7.44	0.01
5	2.00	8.46	8.44	0.02

Figure 8.5



Exponential and other nonlinear discrete least squares approximations can be obtain in the *Statistics* package by using the commands *ExponentialFit* and *NonlinearFit*.

For example, the approximation in the Illustration can be obtained by first defining the data with

$X := Vector([1, 1.25, 1.5, 1.75, 2])$: $Y := Vector([5.1, 5.79, 6.53, 7.45, 8.46])$:

and then issuing the command

ExponentialFit(X, Y, x)

gives the result, rounded to 5 decimal places,

$$3.07249e^{0.50572x}$$

If instead the *NonlinearFit* command is issued, the approximation produced uses methods of Chapter 10 for solving a system of nonlinear equations. The approximation that Maple gives in this case is

$$3.06658(1.66023)^x \approx 3.06658e^{0.50695}.$$

EXERCISE SET 8.1

1. Compute the linear least squares polynomial for the data of Example 2.
2. Compute the least squares polynomial of degree 2 for the data of Example 1, and compare the total error E for the two polynomials.
3. Find the least squares polynomials of degrees 1, 2, and 3 for the data in the following table. Compute the error E in each case. Graph the data and the polynomials.

x_i	1.0	1.1	1.3	1.5	1.9	2.1
y_i	1.84	1.96	2.21	2.45	2.94	3.18

4. Find the least squares polynomials of degrees 1, 2, and 3 for the data in the following table. Compute the error E in each case. Graph the data and the polynomials.

x_i	0	0.15	0.31	0.5	0.6	0.75
y_i	1.0	1.004	1.031	1.117	1.223	1.422

5. Given the data:

x_i	4.0	4.2	4.5	4.7	5.1	5.5	5.9	6.3	6.8	7.1
y_i	102.56	113.18	130.11	142.05	167.53	195.14	224.87	256.73	299.50	326.72

a. Construct the least squares polynomial of degree 1, and compute the error.

b. Construct the least squares polynomial of degree 2, and compute the error.

c. Construct the least squares polynomial of degree 3, and compute the error.

d. Construct the least squares approximation of the form be^{ax}, and compute the error.

e. Construct the least squares approximation of the form bx^a, and compute the error.

6. Repeat Exercise 5 for the following data.

x_i	0.2	0.3	0.6	0.9	1.1	1.3	1.4	1.6
y_i	0.050446	0.098426	0.33277	0.72660	1.0972	1.5697	1.8487	2.5015

7. In the lead example of this chapter, an experiment was described to determine the spring constant k in Hooke's law:

$$F(l) = k(l - E).$$

The function F is the force required to stretch the spring l units, where the constant $E = 5.3$ in. is the length of the unstretched spring.

a. Suppose measurements are made of the length l, in inches, for applied weights $F(l)$, in pounds, as given in the following table.

$F(l)$	l
2	7.0
4	9.4
6	12.3

Find the least squares approximation for k.

b. Additional measurements are made, giving more data:

$F(l)$	l
3	8.3
5	11.3
8	14.4
10	15.9

Compute the new least squares approximation for k. Which of (a) or (b) best fits the total experimental data?

8. The following list contains homework grades and the final-examination grades for 30 numerical analysis students. Find the equation of the least squares line for this data, and use this line to determine the homework grade required to predict minimal A (90%) and D (60%) grades on the final.

Homework	Final	Homework	Final
302	45	323	83
325	72	337	99
285	54	337	70
339	54	304	62
334	79	319	66
322	65	234	51
331	99	337	53
279	63	351	100
316	65	339	67
347	99	343	83
343	83	314	42
290	74	344	79
326	76	185	59
233	57	340	75
254	45	316	45

9. The following table lists the college grade-point averages of 20 mathematics and computer science majors, together with the scores that these students received on the mathematics portion of the ACT (American College Testing Program) test while in high school. Plot these data, and find the equation of the least squares line for this data.

ACT score	Grade-point average	ACT score	Grade-point average
28	3.84	29	3.75
25	3.21	28	3.65
28	3.23	27	3.87
27	3.63	29	3.75
28	3.75	21	1.66
33	3.20	28	3.12
28	3.41	28	2.96
29	3.38	26	2.92
23	3.53	30	3.10
27	2.03	24	2.81

10. The following set of data, presented to the Senate Antitrust Subcommittee, shows the comparative crash-survivability characteristics of cars in various classes. Find the least squares line that approximates these data. (The table shows the percent of accident-involved vehicles in which the most severe injury was fatal or serious.)

Type	Average Weight	Percent Occurrence
1. Domestic luxury regular	4800 lb	3.1
2. Domestic intermediate regular	3700 lb	4.0
3. Domestic economy regular	3400 lb	5.2
4. Domestic compact	2800 lb	6.4
5. Foreign compact	1900 lb	9.6

11. To determine a relationship between the number of fish and the number of species of fish in samples taken for a portion of the Great Barrier Reef, P. Sale and R. Dybdahl [SD] fit a linear least squares polynomial to the following collection of data, which were collected in samples over a 2-year period. Let x be the number of fish in the sample and y be the number of species in the sample.

x	y	x	y	x	y
13	11	29	12	60	14
15	10	30	14	62	21
16	11	31	16	64	21
21	12	36	17	70	24
22	12	40	13	72	17
23	13	42	14	100	23
25	13	55	22	130	34

Determine the linear least squares polynomial for these data.

12. To determine a functional relationship between the attenuation coefficient and the thickness of a sample of taconite, V. P. Singh [Si] fits a collection of data by using a linear least squares polynomial. The following collection of data is taken from a graph in that paper. Find the linear least squares polynomial fitting these data.

Thickness (cm)	Attenuation coefficient (dB/cm)
0.040	26.5
0.041	28.1
0.055	25.2
0.056	26.0
0.062	24.0
0.071	25.0
0.071	26.4
0.078	27.2
0.082	25.6
0.090	25.0
0.092	26.8
0.100	24.8
0.105	27.0
0.120	25.0
0.123	27.3
0.130	26.9
0.140	26.2

13. In a paper dealing with the efficiency of energy utilization of the larvae of the modest sphinx moth (*Pachysphinx modesta*), L. Schroeder [Schr1] used the following data to determine a relation between W, the live weight of the larvae in grams, and R, the oxygen consumption of the larvae in milliliters/hour. For biological reasons, it is assumed that a relationship in the form of $R = bW^a$ exists between W and R.

a. Find the logarithmic linear least squares polynomial by using

$$\ln R = \ln b + a \ln W.$$

b. Compute the error associated with the approximation in part (a):

$$E = \sum_{i=1}^{37} (R_i - bW_i^a)^2.$$

c. Modify the logarithmic least squares equation in part (a) by adding the quadratic term $c(\ln W_i)^2$, and determine the logarithmic quadratic least squares polynomial.

d. Determine the formula for and compute the error associated with the approximation in part (c).

W	R	W	R	W	R	W	R	W	R
0.017	0.154	0.025	0.23	0.020	0.181	0.020	0.180	0.025	0.234
0.087	0.296	0.111	0.357	0.085	0.260	0.119	0.299	0.233	0.537
0.174	0.363	0.211	0.366	0.171	0.334	0.210	0.428	0.783	1.47
1.11	0.531	0.999	0.771	1.29	0.87	1.32	1.15	1.35	2.48
1.74	2.23	3.02	2.01	3.04	3.59	3.34	2.83	1.69	1.44
4.09	3.58	4.28	3.28	4.29	3.40	5.48	4.15	2.75	1.84
5.45	3.52	4.58	2.96	5.30	3.88			4.83	4.66
5.96	2.40	4.68	5.10					5.53	6.94

14. Show that the normal equations (8.3) resulting from discrete least squares approximation yield a symmetric and nonsingular matrix and hence have a unique solution. [*Hint:* Let $A = (a_{ij})$, where

$$a_{ij} = \sum_{k=1}^{m} x_k^{i+j-2}$$

and $x_1, x_2, \ldots, x_m$ are distinct with $n < m - 1$. Suppose A is singular and that $\mathbf{c} \neq \mathbf{0}$ is such that $\mathbf{c}^t A\mathbf{c} = 0$. Show that the nth-degree polynomial whose coefficients are the coordinates of $\mathbf{c}$ has more than n roots, and use this to establish a contradiction.]

8.2 Orthogonal Polynomials and Least Squares Approximation

The previous section considered the problem of least squares approximation to fit a collection of data. The other approximation problem mentioned in the introduction concerns the approximation of functions.

Suppose $f \in C[a, b]$ and that a polynomial $P_n(x)$ of degree at most n is required that will minimize the error

$$\int_a^b [f(x) - P_n(x)]^2 \, dx.$$

To determine a least squares approximating polynomial; that is, a polynomial to minimize this expression, let

$$P_n(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 = \sum_{k=0}^{n} a_k x^k,$$

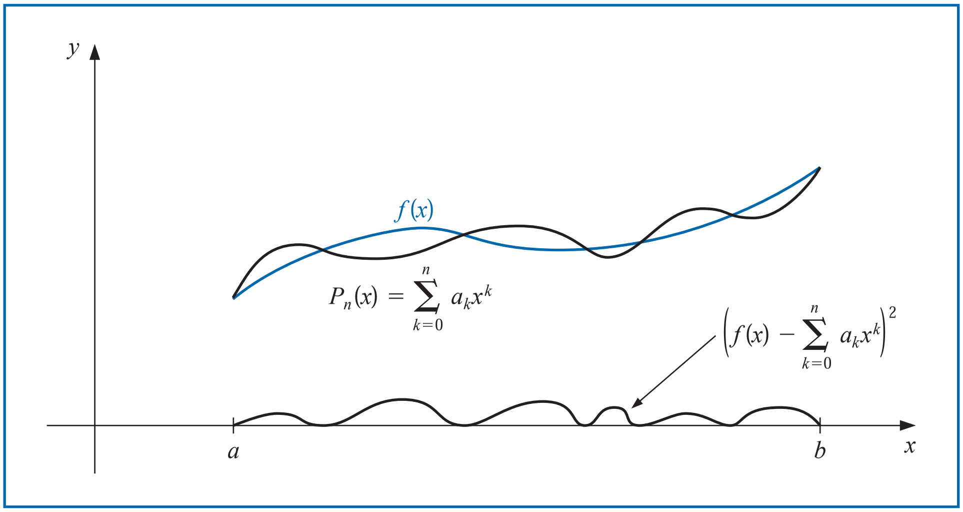
and define, as shown in Figure 8.6,

$$E \equiv E_2(a_0, a_1, \ldots, a_n) = \int_a^b \left(f(x) - \sum_{k=0}^{n} a_k x^k \right)^2 dx.$$

The problem is to find real coefficients $a_0, a_1, \ldots, a_n$ that will minimize E. A necessary condition for the numbers $a_0, a_1, \ldots, a_n$ to minimize E is that

$$\frac{\partial E}{\partial a_j} = 0, \quad \text{for each } j = 0, 1, \ldots, n.$$

Figure 8.6



Since

$$E = \int_a^b [f(x)]^2\,dx - 2\sum_{k=0}^{n} a_k \int_a^b x^k f(x)\,dx + \int_a^b \left(\sum_{k=0}^{n} a_k x^k\right)^2 dx,$$

we have

$$\frac{\partial E}{\partial a_j} = -2\int_a^b x^j f(x)\,dx + 2\sum_{k=0}^{n} a_k \int_a^b x^{j+k}\,dx.$$

Hence, to find $P_n(x)$, the $(n+1)$ linear **normal equations**

$$\sum_{k=0}^{n} a_k \int_a^b x^{j+k}\,dx = \int_a^b x^j f(x)\,dx, \quad \text{for each } j = 0, 1, \ldots, n, \tag{8.6}$$

must be solved for the $(n+1)$ unknowns a_j. The normal equations always have a unique solution provided that $f \in C[a,b]$. (See Exercise 15.)

Example 1 Find the least squares approximating polynomial of degree 2 for the function $f(x) = \sin \pi x$ on the interval [0, 1].

Solution The normal equations for $P_2(x) = a_2x^2 + a_1x + a_0$ are

$$a_0 \int_0^1 1\,dx + a_1 \int_0^1 x\,dx + a_2 \int_0^1 x^2\,dx = \int_0^1 \sin \pi x\,dx,$$

$$a_0 \int_0^1 x\,dx + a_1 \int_0^1 x^2\,dx + a_2 \int_0^1 x^3\,dx = \int_0^1 x \sin \pi x\,dx,$$

$$a_0 \int_0^1 x^2\,dx + a_1 \int_0^1 x^3\,dx + a_2 \int_0^1 x^4\,dx = \int_0^1 x^2 \sin \pi x\,dx.$$

Performing the integration yields

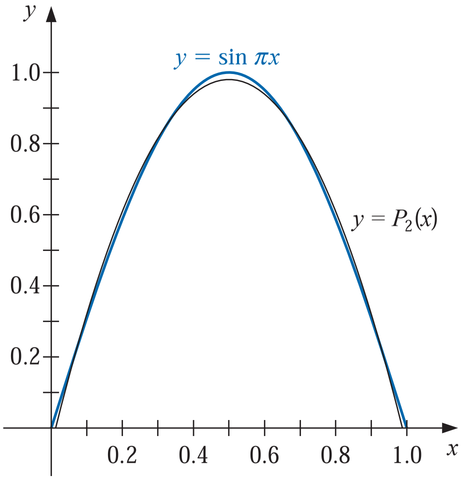
$$a_0 + \frac{1}{2}a_1 + \frac{1}{3}a_2 = \frac{2}{\pi}, \quad \frac{1}{2}a_0 + \frac{1}{3}a_1 + \frac{1}{4}a_2 = \frac{1}{\pi}, \quad \frac{1}{3}a_0 + \frac{1}{4}a_1 + \frac{1}{5}a_2 = \frac{\pi^2 - 4}{\pi^3}.$$

These three equations in three unknowns can be solved to obtain

$$a_0 = \frac{12\pi^2 - 120}{\pi^3} \approx -0.050465 \quad \text{and} \quad a_1 = -a_2 = \frac{720 - 60\pi^2}{\pi^3} \approx 4.12251.$$

Consequently, the least squares polynomial approximation of degree 2 for $f(x) = \sin \pi x$ on $[0, 1]$ is $P_2(x) = -4.12251x^2 + 4.12251x - 0.050465$. (See Figure 8.7.) ■

Figure 8.7



Example 1 illustrates a difficulty in obtaining a least squares polynomial approximation. An $(n + 1) \times (n + 1)$ linear system for the unknowns $a_0, \ldots, a_n$ must be solved, and the coefficients in the linear system are of the form

$$\int_a^b x^{j+k}\, dx = \frac{b^{j+k+1} - a^{j+k+1}}{j + k + 1},$$

a linear system that does not have an easily computed numerical solution. The matrix in the linear system is known as a **Hilbert matrix**, which is a classic example for demonstrating round-off error difficulties. (See Exercise 11 of Section 7.5.)

David Hilbert (1862–1943) was the dominant mathematician at the turn of the 20th century. He is best remembered for giving a talk at the International Congress of Mathematicians in Paris in 1900 in which he posed 23 problems that he thought would be important for mathematicians in the next century.

Another disadvantage is similar to the situation that occurred when the Lagrange polynomials were first introduced in Section 3.1. The calculations that were performed in obtaining the best nth-degree polynomial, $P_n(x)$, do not lessen the amount of work required to obtain $P_{n+1}(x)$, the polynomial of next higher degree.

Linearly Independent Functions

A different technique to obtain least squares approximations will now be considered. This turns out to be computationally efficient, and once $P_n(x)$ is known, it is easy to determine $P_{n+1}(x)$. To facilitate the discussion, we need some new concepts.

Definition 8.1 The set of functions $\{\phi_0, \ldots, \phi_n\}$ is said to be **linearly independent** on $[a, b]$ if, whenever

$$c_0\phi_0(x) + c_1\phi_1(x) + \cdots + c_n\phi_n(x) = 0, \quad \text{for all } x \in [a, b],$$

we have $c_0 = c_1 = \cdots = c_n = 0$. Otherwise the set of functions is said to be **linearly dependent**. ■

Theorem 8.2 Suppose that, for each $j = 0, 1, \ldots, n$, $\phi_j(x)$ is a polynomial of degree j. Then $\{\phi_0, \ldots, \phi_n\}$ is linearly independent on any interval $[a, b]$. ■

Proof Let $c_0, \ldots, c_n$ be real numbers for which

$$P(x) = c_0\phi_0(x) + c_1\phi_1(x) + \cdots + c_n\phi_n(x) = 0, \quad \text{for all } x \in [a, b].$$

The polynomial $P(x)$ vanishes on $[a, b]$, so it must be the zero polynomial, and the coefficients of all the powers of x are zero. In particular, the coefficient of x^n is zero. But $c_n\phi_n(x)$ is the only term in $P(x)$ that contains x^n, so we must have $c_n = 0$. Hence

$$P(x) = \sum_{j=0}^{n-1} c_j\phi_j(x).$$

In this representation of $P(x)$, the only term that contains a power of x^{n-1} is $c_{n-1}\phi_{n-1}(x)$, so this term must also be zero and

$$P(x) = \sum_{j=0}^{n-2} c_j\phi_j(x).$$

In like manner, the remaining constants $c_{n-2}, c_{n-3}, \ldots, c_1, c_0$ are all zero, which implies that $\{\phi_0, \phi_1, \ldots, \phi_n\}$ is linearly independent on $[a, b]$. ■ ■ ■

Example 2 Let $\phi_0(x) = 2$, $\phi_1(x) = x - 3$, and $\phi_2(x) = x^2 + 2x + 7$, and $Q(x) = a_0 + a_1x + a_2x^2$. Show that there exist constants c_0, c_1, and c_2 such that $Q(x) = c_0\phi_0(x) + c_1\phi_1(x) + c_2\phi_2(x)$.

Solution By Theorem 8.2, $\{\phi_0, \phi_1, \phi_2\}$ is linearly independent on any interval $[a, b]$. First note that

$$1 = \frac{1}{2}\phi_0(x), \quad x = \phi_1(x) + 3 = \phi_1(x) + \frac{3}{2}\phi_0(x),$$

and

$$x^2 = \phi_2(x) - 2x - 7 = \phi_2(x) - 2\left[\phi_1(x) + \frac{3}{2}\phi_0(x)\right] - 7\left[\frac{1}{2}\phi_0(x)\right]$$
$$= \phi_2(x) - 2\phi_1(x) - \frac{13}{2}\phi_0(x).$$

Hence

$$Q(x) = a_0\left[\frac{1}{2}\phi_0(x)\right] + a_1\left[\phi_1(x) + \frac{3}{2}\phi_0(x)\right] + a_2\left[\phi_2(x) - 2\phi_1(x) - \frac{13}{2}\phi_0(x)\right]$$
$$= \left(\frac{1}{2}a_0 + \frac{3}{2}a_1 - \frac{13}{2}a_2\right)\phi_0(x) + [a_1 - 2a_2]\,\phi_1(x) + a_2\phi_2(x).$$ ■

The situation illustrated in Example 2 holds in a much more general setting. Let $\prod_n$ denote the **set of all polynomials of degree at most *n***. The following result is used extensively in many applications of linear algebra. Its proof is considered in Exercise 13.

Theorem 8.3 Suppose that $\{\phi_0(x), \phi_1(x), \ldots, \phi_n(x)\}$ is a collection of linearly independent polynomials in $\prod_n$. Then any polynomial in $\prod_n$ can be written uniquely as a linear combination of $\phi_0(x)$, $\phi_1(x), \ldots, \phi_n(x)$. ■

Orthogonal Functions

To discuss general function approximation requires the introduction of the notions of weight functions and orthogonality.

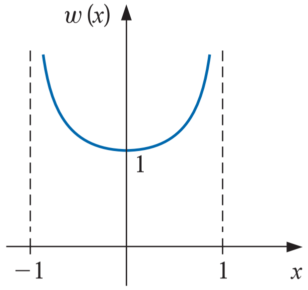
Definition 8.4 An integrable function w is called a **weight function** on the interval I if $w(x) \geq 0$, for all x in I, but $w(x) \not\equiv 0$ on any subinterval of I. ■

The purpose of a weight function is to assign varying degrees of importance to approximations on certain portions of the interval. For example, the weight function

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

places less emphasis near the center of the interval $(-1, 1)$ and more emphasis when $|x|$ is near 1 (see Figure 8.8). This weight function is used in the next section.

Figure 8.8



Suppose $\{\phi_0, \phi_1, \ldots, \phi_n\}$ is a set of linearly independent functions on $[a, b]$ and w is a weight function for $[a, b]$. Given $f \in C[a, b]$, we seek a linear combination

$$P(x) = \sum_{k=0}^{n} a_k \phi_k(x)$$

to minimize the error

$$E = E(a_0, \ldots, a_n) = \int_a^b w(x) \left[f(x) - \sum_{k=0}^{n} a_k \phi_k(x) \right]^2 dx.$$

This problem reduces to the situation considered at the beginning of this section in the special case when $w(x) \equiv 1$ and $\phi_k(x) = x^k$, for each $k = 0, 1, \ldots, n$.

The normal equations associated with this problem are derived from the fact that for each $j = 0, 1, \ldots, n$,

$$0 = \frac{\partial E}{\partial a_j} = 2 \int_a^b w(x) \left[f(x) - \sum_{k=0}^{n} a_k \phi_k(x) \right] \phi_j(x)\, dx.$$

The system of normal equations can be written

$$\int_a^b w(x) f(x) \phi_j(x)\, dx = \sum_{k=0}^{n} a_k \int_a^b w(x) \phi_k(x) \phi_j(x)\, dx, \quad \text{for } j = 0, 1, \ldots, n.$$

If the functions $\phi_0, \phi_1, \ldots, \phi_n$ can be chosen so that

$$\int_a^b w(x) \phi_k(x) \phi_j(x)\, dx = \begin{cases} 0, & \text{when } j \neq k, \\ \alpha_j > 0, & \text{when } j = k, \end{cases} \tag{8.7}$$

then the normal equations will reduce to

$$\int_a^b w(x) f(x) \phi_j(x)\, dx = a_j \int_a^b w(x) [\phi_j(x)]^2\, dx = a_j \alpha_j,$$

for each $j = 0, 1, \ldots, n$. These are easily solved to give

$$a_j = \frac{1}{\alpha_j} \int_a^b w(x) f(x) \phi_j(x)\, dx.$$

Hence the least squares approximation problem is greatly simplified when the functions $\phi_0, \phi_1, \ldots, \phi_n$ are chosen to satisfy the *orthogonality* condition in Eq. (8.7). The remainder of this section is devoted to studying collections of this type.

The word orthogonal means right-angled. So in a sense, orthogonal functions are perpendicular to one another.

Definition 8.5 $\{\phi_0, \phi_1, \ldots, \phi_n\}$ is said to be an **orthogonal set of functions** for the interval $[a, b]$ with respect to the weight function w if

$$\int_a^b w(x)\phi_k(x)\phi_j(x)\,dx = \begin{cases} 0, & \text{when } j \neq k, \\ \alpha_j > 0, & \text{when } j = k. \end{cases}$$

If, in addition, $\alpha_j = 1$ for each $j = 0, 1, \ldots, n$, the set is said to be **orthonormal**. ■

This definition, together with the remarks preceding it, produces the following theorem.

Theorem 8.6 If $\{\phi_0, \ldots, \phi_n\}$ is an orthogonal set of functions on an interval $[a, b]$ with respect to the weight function w, then the least squares approximation to f on $[a, b]$ with respect to w is

$$P(x) = \sum_{j=0}^{n} a_j\phi_j(x),$$

where, for each $j = 0, 1, \ldots, n$,

$$a_j = \frac{\int_a^b w(x)\phi_j(x)f(x)\,dx}{\int_a^b w(x)[\phi_j(x)]^2\,dx} = \frac{1}{\alpha_j}\int_a^b w(x)\phi_j(x)f(x)\,dx.$$ ■

Although Definition 8.5 and Theorem 8.6 allow for broad classes of orthogonal functions, we will consider only orthogonal sets of polynomials. The next theorem, which is based on the **Gram-Schmidt process**, describes how to construct orthogonal polynomials on $[a, b]$ with respect to a weight function w.

Theorem 8.7 The set of polynomial functions $\{\phi_0, \phi_1, \ldots, \phi_n\}$ defined in the following way is orthogonal on $[a, b]$ with respect to the weight function w.

$$\phi_0(x) \equiv 1, \quad \phi_1(x) = x - B_1, \quad \text{for each } x \text{ in } [a, b],$$

where

$$B_1 = \frac{\int_a^b xw(x)[\phi_0(x)]^2\,dx}{\int_a^b w(x)[\phi_0(x)]^2\,dx},$$

and when $k \geq 2$,

$$\phi_k(x) = (x - B_k)\phi_{k-1}(x) - C_k\phi_{k-2}(x), \quad \text{for each } x \text{ in } [a, b],$$

where

$$B_k = \frac{\int_a^b xw(x)[\phi_{k-1}(x)]^2\,dx}{\int_a^b w(x)[\phi_{k-1}(x)]^2\,dx}$$

and

$$C_k = \frac{\int_a^b xw(x)\phi_{k-1}(x)\phi_{k-2}(x)\,dx}{\int_a^b w(x)[\phi_{k-2}(x)]^2\,dx}.$$ ■

Erhard Schmidt (1876–1959) received his doctorate under the supervision of David Hilbert in 1905 for a problem involving integral equations. Schmidt published a paper in 1907 in which he gave what is now called the Gram-Schmidt process for constructing an orthonormal basis for a set of functions. This generalized results of Jorgen Pedersen Gram (1850–1916) who considered this problem when studying least squares. Laplace, however, presented a similar process much earlier than either Gram or Schmidt.

Theorem 8.7 provides a recursive procedure for constructing a set of orthogonal polynomials. The proof of this theorem follows by applying mathematical induction to the degree of the polynomial $\phi_n(x)$.

Corollary 8.8 For any $n > 0$, the set of polynomial functions $\{\phi_0, \ldots, \phi_n\}$ given in Theorem 8.7 is linearly independent on $[a, b]$ and

$$\int_a^b w(x)\phi_n(x)Q_k(x)\,dx = 0,$$

for any polynomial $Q_k(x)$ of degree $k < n$. ■

Proof For each $k = 0, 1, \ldots, n$, $\phi_k(x)$ is a polynomial of degree k. So Theorem 8.2 implies that $\{\phi_0, \ldots, \phi_n\}$ is a linearly independent set.

Let $Q_k(x)$ be a polynomial of degree $k < n$. By Theorem 8.3 there exist numbers $c_0, \ldots, c_k$ such that

$$Q_k(x) = \sum_{j=0}^{k} c_j\phi_j(x).$$

Because ϕ_n is orthogonal to ϕ_j for each $j = 0, 1, \ldots, k$ we have

$$\int_a^b w(x)Q_k(x)\phi_n(x)\,dx = \sum_{j=0}^{k} c_j \int_a^b w(x)\phi_j(x)\phi_n(x)\,dx = \sum_{j=0}^{k} c_j \cdot 0 = 0.$$ ■ ■ ■

Illustration The set of **Legendre polynomials**, $\{P_n(x)\}$, is orthogonal on $[-1, 1]$ with respect to the weight function $w(x) \equiv 1$. The classical definition of the Legendre polynomials requires that $P_n(1) = 1$ for each n, and a recursive relation is used to generate the polynomials when $n \geq 2$. This normalization will not be needed in our discussion, and the least squares approximating polynomials generated in either case are essentially the same.

Using the Gram-Schmidt process with $P_0(x) \equiv 1$ gives

$$B_1 = \frac{\int_{-1}^{1} x\,dx}{\int_{-1}^{1}\,dx} = 0 \quad \text{and} \quad P_1(x) = (x - B_1)P_0(x) = x.$$

Also,

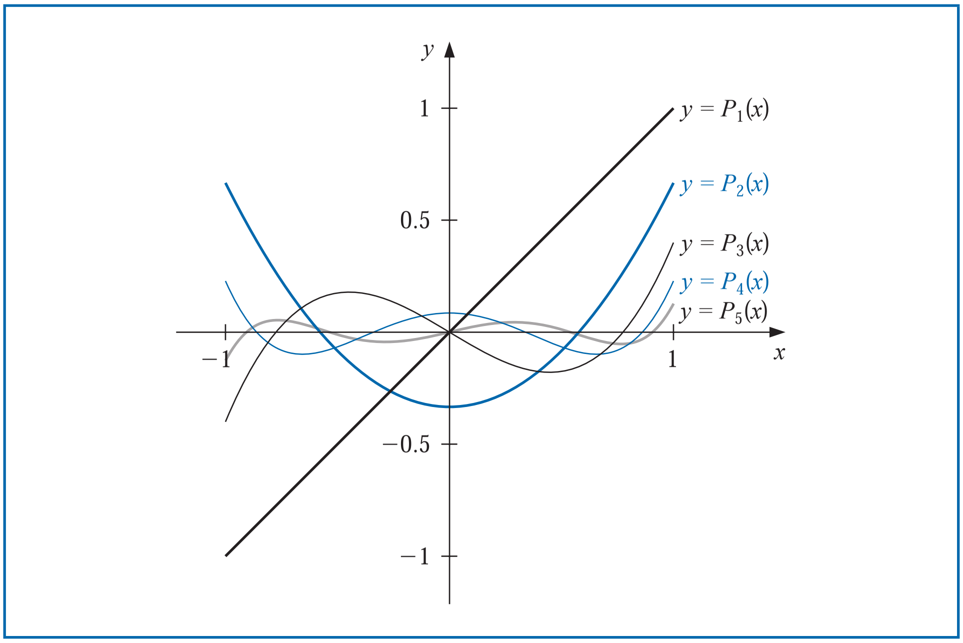
$$B_2 = \frac{\int_{-1}^{1} x^3\,dx}{\int_{-1}^{1} x^2\,dx} = 0 \quad \text{and} \quad C_2 = \frac{\int_{-1}^{1} x^2\,dx}{\int_{-1}^{1} 1\,dx} = \frac{1}{3},$$

so

$$P_2(x) = (x - B_2)P_1(x) - C_2P_0(x) = (x - 0)x - \frac{1}{3} \cdot 1 = x^2 - \frac{1}{3}.$$

The higher-degree Legendre polynomials shown in Figure 8.9 are derived in the same manner. Although the integration can be tedious, it is not difficult with a Computer Algebra System.

Figure 8.9



For example, the Maple command *int* is used to compute the integrals B_3 and C_3:

$$B3 := \frac{int\left(x\left(x^2-\frac{1}{3}\right)^2, x=-1..1\right)}{int\left(\left(x^2-\frac{1}{3}\right)^2, x=-1..1\right)}; \quad C3 := \frac{int\left(x\left(x^2-\frac{1}{3}\right), x=-1..1\right)}{int(x^2, x=-1..1)}$$

$$0$$

$$\frac{4}{15}$$

Thus

$$P_3(x) = xP_2(x) - \frac{4}{15}P_1(x) = x^3 - \frac{1}{3}x - \frac{4}{15}x = x^3 - \frac{3}{5}x.$$

The next two Legendre polynomials are

$$P_4(x) = x^4 - \frac{6}{7}x^2 + \frac{3}{35} \quad \text{and} \quad P_5(x) = x^5 - \frac{10}{9}x^3 + \frac{5}{21}x. \qquad \square$$

The Legendre polynomials were introduced in Section 4.7, where their roots, given on page 232, were used as the nodes in Gaussian quadrature.

EXERCISE SET 8.2

1. Find the linear least squares polynomial approximation to $f(x)$ on the indicated interval if

a. $f(x) = x^2 + 3x + 2, \quad [0, 1];$ **b.** $f(x) = x^3, \quad [0, 2];$

c. $f(x) = \frac{1}{x}, \quad [1, 3];$ **d.** $f(x) = e^x, \quad [0, 2];$

e. $f(x) = \frac{1}{2}\cos x + \frac{1}{3}\sin 2x, \quad [0, 1];$ **f.** $f(x) = x \ln x, \quad [1, 3].$

2. Find the linear least squares polynomial approximation on the interval $[-1, 1]$ for the following functions.

a. $f(x) = x^2 - 2x + 3$ **b.** $f(x) = x^3$

c. $f(x) = \frac{1}{x+2}$ **d.** $f(x) = e^x$

e. $f(x) = \frac{1}{2}\cos x + \frac{1}{3}\sin 2x$ **f.** $f(x) = \ln(x + 2)$

3. Find the least squares polynomial approximation of degree two to the functions and intervals in Exercise 1.

4. Find the least squares polynomial approximation of degree 2 on the interval $[-1, 1]$ for the functions in Exercise 3.

5. Compute the error E for the approximations in Exercise 3.

6. Compute the error E for the approximations in Exercise 4.

7. Use the Gram-Schmidt process to construct $\phi_0(x)$, $\phi_1(x)$, $\phi_2(x)$, and $\phi_3(x)$ for the following intervals.

a. $[0, 1]$ **b.** $[0, 2]$ **c.** $[1, 3]$

8. Repeat Exercise 1 using the results of Exercise 7.

9. Obtain the least squares approximation polynomial of degree 3 for the functions in Exercise 1 using the results of Exercise 7.

10. Repeat Exercise 3 using the results of Exercise 7.

11. Use the Gram-Schmidt procedure to calculate L_1, L_2, and L_3, where $\{L_0(x), L_1(x), L_2(x), L_3(x)\}$ is an orthogonal set of polynomials on $(0, \infty)$ with respect to the weight functions $w(x) = e^{-x}$ and $L_0(x) \equiv 1$. The polynomials obtained from this procedure are called the **Laguerre polynomials**.

12. Use the Laguerre polynomials calculated in Exercise 11 to compute the least squares polynomials of degree one, two, and three on the interval $(0, \infty)$ with respect to the weight function $w(x) = e^{-x}$ for the following functions:

a. $f(x) = x^2$ **b.** $f(x) = e^{-x}$ **c.** $f(x) = x^3$ **d.** $f(x) = e^{-2x}$

13. Suppose $\{\phi_0, \phi_1, \ldots, \phi_n\}$ is any linearly independent set in $\prod_n$. Show that for any element $Q \in \prod_n$, there exist unique constants $c_0, c_1, \ldots, c_n$, such that

$$Q(x) = \sum_{k=0}^{n} c_k \phi_k(x).$$

14. Show that if $\{\phi_0, \phi_1, \ldots, \phi_n\}$ is an orthogonal set of functions on $[a, b]$ with respect to the weight function w, then $\{\phi_0, \phi_1, \ldots, \phi_n\}$ is a linearly independent set.

15. Show that the normal equations (8.6) have a unique solution. [*Hint:* Show that the only solution for the function $f(x) \equiv 0$ is $a_j = 0, j = 0, 1, \ldots, n$. Multiply Eq. (8.6) by a_j, and sum over all j. Interchange the integral sign and the summation sign to obtain $\int_a^b [P(x)]^2 dx = 0$. Thus, $P(x) \equiv 0$, so $a_j = 0$, for $j = 0, \ldots, n$. Hence, the coefficient matrix is nonsingular, and there is a unique solution to Eq. (8.6).]

8.3 Chebyshev Polynomials and Economization of Power Series

The Chebyshev polynomials $\{T_n(x)\}$ are orthogonal on $(-1, 1)$ with respect to the weight function $w(x) = (1 - x^2)^{-1/2}$. Although they can be derived by the method in the previous section, it is easier to give their definition and then show that they satisfy the required orthogonality properties.

For $x \in [-1, 1]$, define

$$T_n(x) = \cos[n \arccos x], \quad \text{for each } n \geq 0. \tag{8.8}$$

It might not be obvious from this definition that for each n, $T_n(x)$ is a polynomial in x, but we will now show this. First note that

$$T_0(x) = \cos 0 = 1 \quad \text{and} \quad T_1(x) = \cos(\arccos x) = x.$$

For $n \geq 1$, we introduce the substitution $\theta = \arccos x$ to change this equation to

$$T_n(\theta(x)) \equiv T_n(\theta) = \cos(n\theta), \quad \text{where } \theta \in [0, \pi].$$

A recurrence relation is derived by noting that

$$T_{n+1}(\theta) = \cos(n+1)\theta = \cos\theta\, \cos(n\theta) - \sin\theta\, \sin(n\theta)$$

and

$$T_{n-1}(\theta) = \cos(n-1)\theta = \cos\theta\, \cos(n\theta) + \sin\theta\, \sin(n\theta)$$

Adding these equations gives

$$T_{n+1}(\theta) = 2\cos\theta\, \cos(n\theta) - T_{n-1}(\theta).$$

Returning to the variable $x = \cos\theta$, we have, for $n \geq 1$,

$$T_{n+1}(x) = 2x\cos(n \arccos x) - T_{n-1}(x),$$

that is,

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x). \tag{8.9}$$

Because $T_0(x) = 1$ and $T_1(x) = x$, the recurrence relation implies that the next three Chebyshev polynomials are

$$T_2(x) = 2xT_1(x) - T_0(x) = 2x^2 - 1,$$
$$T_3(x) = 2xT_2(x) - T_1(x) = 4x^3 - 3x,$$

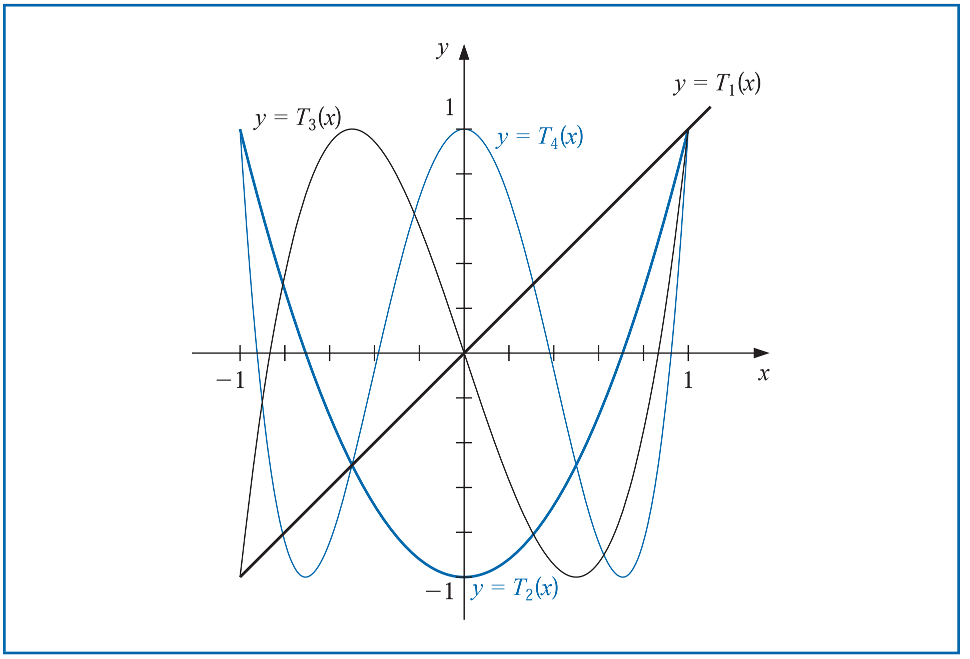
and

$$T_4(x) = 2xT_3(x) - T_2(x) = 8x^4 - 8x^2 + 1.$$

The recurrence relation also implies that when $n \geq 1$, $T_n(x)$ is a polynomial of degree n with leading coefficient 2^{n-1}. The graphs of T_1, T_2, T_3, and T_4 are shown in Figure 8.10.

Pafnuty Lvovich Chebyshev (1821–1894) did exceptional mathematical work in many areas, including applied mathematics, number theory, approximation theory, and probability. In 1852 he traveled from St. Petersburg to visit mathematicians in France, England, and Germany. Lagrange and Legendre had studied individual sets of orthogonal polynomials, but Chebyshev was the first to see the important consequences of studying the theory in general. He developed the Chebyshev polynomials to study least squares approximation and probability, then applied his results to interpolation, approximate quadrature, and other areas.

Figure 8.10



To show the orthogonality of the Chebyshev polynomials with respect to the weight function $w(x) = (1 - x^2)^{-1/2}$, consider

$$\int_{-1}^{1} \frac{T_n(x)T_m(x)}{\sqrt{1-x^2}}\, dx = \int_{-1}^{1} \frac{\cos(n \arccos x)\cos(m \arccos x)}{\sqrt{1-x^2}}\, dx.$$

Reintroducing the substitution $\theta = \arccos x$ gives

$$d\theta = -\frac{1}{\sqrt{1-x^2}}\, dx$$

and

$$\int_{-1}^{1} \frac{T_n(x)T_m(x)}{\sqrt{1-x^2}}\, dx = -\int_{\pi}^{0} \cos(n\theta)\cos(m\theta)\, d\theta = \int_{0}^{\pi} \cos(n\theta)\cos(m\theta)\, d\theta.$$

Suppose $n \neq m$. Since

$$\cos(n\theta)\cos(m\theta) = \frac{1}{2}[\cos(n+m)\theta + \cos(n-m)\theta],$$

we have

$$\int_{-1}^{1} \frac{T_n(x)T_m(x)}{\sqrt{1-x^2}}\, dx = \frac{1}{2}\int_{0}^{\pi} \cos((n+m)\theta)\, d\theta + \frac{1}{2}\int_{0}^{\pi} \cos((n-m)\theta)\, d\theta$$

$$= \left[\frac{1}{2(n+m)}\sin((n+m)\theta) + \frac{1}{2(n-m)}\sin((n-m)\theta)\right]_{0}^{\pi} = 0.$$

By a similar technique (see Exercise 9), we also have

$$\int_{-1}^{1} \frac{[T_n(x)]^2}{\sqrt{1-x^2}}\, dx = \frac{\pi}{2}, \quad \text{for each } n \geq 1. \qquad (8.10)$$

The Chebyshev polynomials are used to minimize approximation error. We will see how they are used to solve two problems of this type:

- an optimal placing of interpolating points to minimize the error in Lagrange interpolation;
- a means of reducing the degree of an approximating polynomial with minimal loss of accuracy.

The next result concerns the zeros and extreme points of $T_n(x)$.

Theorem 8.9 The Chebyshev polynomial $T_n(x)$ of degree $n \geq 1$ has n simple zeros in $[-1, 1]$ at

$$\bar{x}_k = \cos\left(\frac{2k-1}{2n}\pi\right), \quad \text{for each } k = 1, 2, \ldots, n.$$

Moreover, $T_n(x)$ assumes its absolute extrema at

$$\bar{x}'_k = \cos\left(\frac{k\pi}{n}\right) \quad \text{with} \quad T_n(\bar{x}'_k) = (-1)^k, \quad \text{for each} \quad k = 0, 1, \ldots, n.$$ ■

Proof Let

$$\bar{x}_k = \cos\left(\frac{2k-1}{2n}\pi\right), \quad \text{for } k = 1, 2, \ldots, n.$$

Then

$$T_n(\bar{x}_k) = \cos(n \arccos \bar{x}_k) = \cos\left(n \arccos\left(\cos\left(\frac{2k-1}{2n}\pi\right)\right)\right) = \cos\left(\frac{2k-1}{2}\pi\right) = 0.$$

But the $\bar{x}_k$ are distinct (see Exercise 10) and $T_n(x)$ is a polynomial of degree n, so all the zeros of $T_n(x)$ must have this form.

To show the second statement, first note that

$$T'_n(x) = \frac{d}{dx}[\cos(n \arccos x)] = \frac{n \sin(n \arccos x)}{\sqrt{1-x^2}},$$

and that, when $k = 1, 2, \ldots, n-1$,

$$T'_n(\bar{x}'_k) = \frac{n \sin\left(n \arccos\left(\cos\left(\frac{k\pi}{n}\right)\right)\right)}{\sqrt{1-\left[\cos\left(\frac{k\pi}{n}\right)\right]^2}} = \frac{n \sin(k\pi)}{\sin\left(\frac{k\pi}{n}\right)} = 0.$$

Since $T_n(x)$ is a polynomial of degree n, its derivative $T'_n(x)$ is a polynomial of degree $(n-1)$, and all the zeros of $T'_n(x)$ occur at these $n-1$ distinct points (that they are distinct is considered in Exercise 11). The only other possibilities for extrema of $T_n(x)$ occur at the endpoints of the interval $[-1, 1]$; that is, at $\bar{x}'_0 = 1$ and at $\bar{x}'_n = -1$.

For any $k = 0, 1, \ldots, n$ we have

$$T_n(\bar{x}'_k) = \cos\left(n \arccos\left(\cos\left(\frac{k\pi}{n}\right)\right)\right) = \cos(k\pi) = (-1)^k.$$

So a maximum occurs at each even value of k and a minimum at each odd value. ■ ■ ■

The monic (polynomials with leading coefficient 1) Chebyshev polynomials $\tilde{T}_n(x)$ are derived from the Chebyshev polynomials $T_n(x)$ by dividing by the leading coefficient 2^{n-1}. Hence

$$\tilde{T}_0(x) = 1 \quad \text{and} \quad \tilde{T}_n(x) = \frac{1}{2^{n-1}} T_n(x), \quad \text{for each } n \geq 1. \tag{8.11}$$

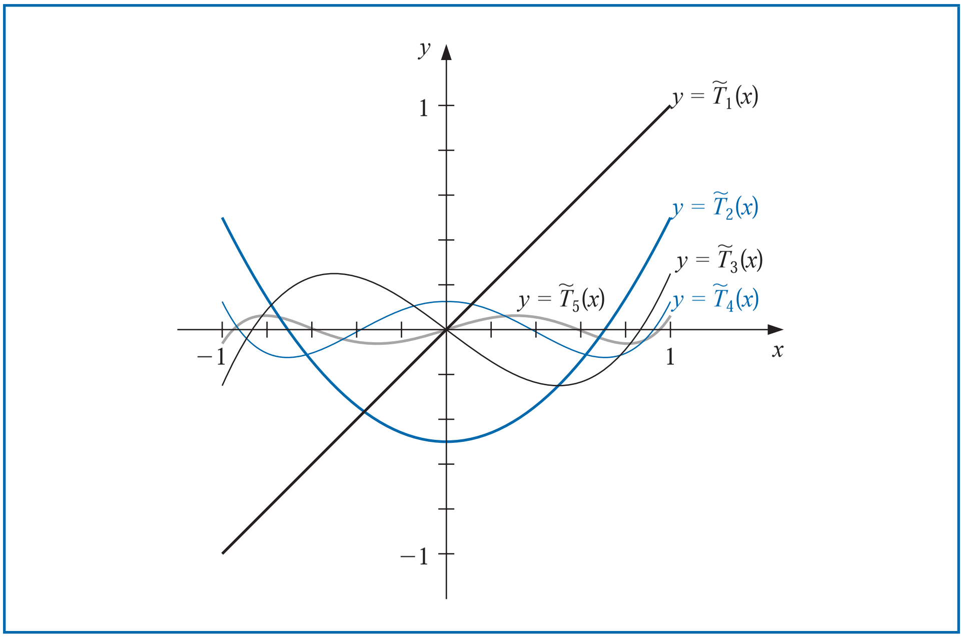
The recurrence relationship satisfied by the Chebyshev polynomials implies that

$$\tilde{T}_2(x) = x\tilde{T}_1(x) - \frac{1}{2}\tilde{T}_0(x) \quad \text{and} \tag{8.12}$$

$$\tilde{T}_{n+1}(x) = x\tilde{T}_n(x) - \frac{1}{4}\tilde{T}_{n-1}(x), \quad \text{for each } n \geq 2.$$

The graphs of $\tilde{T}_1$, $\tilde{T}_2$, $\tilde{T}_3$, $\tilde{T}_4$, and $\tilde{T}_5$ are shown in Figure 8.11.

Figure 8.11



Because $\tilde{T}_n(x)$ is just a multiple of $T_n(x)$, Theorem 8.9 implies that the zeros of $\tilde{T}_n(x)$ also occur at

$$\bar{x}_k = \cos\left(\frac{2k-1}{2n}\pi\right), \quad \text{for each } k = 1, 2, \ldots, n,$$

and the extreme values of $\tilde{T}_n(x)$, for $n \geq 1$, occur at

$$\bar{x}'_k = \cos\left(\frac{k\pi}{n}\right), \quad \text{with} \quad \tilde{T}_n(\bar{x}'_k) = \frac{(-1)^k}{2^{n-1}}, \quad \text{for each } k = 0, 1, 2, \ldots, n. \tag{8.13}$$

Let $\tilde{\prod}_n$ denote **the set of all monic polynomials of degree *n***. The relation expressed in Eq. (8.13) leads to an important minimization property that distinguishes $\tilde{T}_n(x)$ from the other members of $\tilde{\prod}_n$.

Theorem 8.10 The polynomials of the form $\tilde{T}_n(x)$, when $n \geq 1$, have the property that

$$\frac{1}{2^{n-1}} = \max_{x\in[-1,1]} |\tilde{T}_n(x)| \leq \max_{x\in[-1,1]} |P_n(x)|, \quad \text{for all } P_n(x) \in \tilde{\prod_n}.$$

Moreover, equality occurs only if $P_n \equiv \tilde{T}_n$. ■

Proof Suppose that $P_n(x) \in \widetilde{\prod}_n$ and that

$$\max_{x\in[-1,1]} |P_n(x)| \le \frac{1}{2^{n-1}} = \max_{x\in[-1,1]} |\tilde{T}_n(x)|.$$

Let $Q = \tilde{T}_n - P_n$. Then $\tilde{T}_n(x)$ and $P_n(x)$ are both monic polynomials of degree n, so $Q(x)$ is a polynomial of degree at most $(n-1)$. Moreover, at the $n+1$ extreme points $\bar{x}'_k$ of $\tilde{T}_n(x)$, we have

$$Q(\bar{x}'_k) = \tilde{T}_n(\bar{x}'_k) - P_n(\bar{x}'_k) = \frac{(-1)^k}{2^{n-1}} - P_n(\bar{x}'_k).$$

However

$$|P_n(\bar{x}'_k)| \le \frac{1}{2^{n-1}}, \quad \text{for each } k = 0, 1, \ldots, n,$$

so we have

$$Q(\bar{x}'_k) \le 0, \quad \text{when } k \text{ is odd} \quad \text{and} \quad Q(\bar{x}'_k) \ge 0, \quad \text{when } k \text{ is even.}$$

Since Q is continuous, the Intermediate Value Theorem implies that for each $j = 0, 1, \ldots, n-1$ the polynomial $Q(x)$ has at least one zero between $\bar{x}'_j$ and $\bar{x}'_{j+1}$. Thus, Q has at least n zeros in the interval $[-1, 1]$. But the degree of $Q(x)$ is less than n, so $Q \equiv 0$. This implies that $P_n \equiv \tilde{T}_n$. ■ ■ ■

Minimizing Lagrange Interpolation Error

Theorem 8.10 can be used to answer the question of where to place interpolating nodes to minimize the error in Lagrange interpolation. Theorem 3.3 on page 112 applied to the interval $[-1, 1]$ states that, if $x_0, \ldots, x_n$ are distinct numbers in the interval $[-1, 1]$ and if $f \in C^{n+1}[-1, 1]$, then, for each $x \in [-1, 1]$, a number $\xi(x)$ exists in $(-1, 1)$ with

$$f(x) - P(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x-x_0)(x-x_1)\cdots(x-x_n),$$

where $P(x)$ is the Lagrange interpolating polynomial. Generally, there is no control over $\xi(x)$, so to minimize the error by shrewd placement of the nodes $x_0, \ldots, x_n$, we choose $x_0, \ldots, x_n$ to minimize the quantity

$$|(x-x_0)(x-x_1)\cdots(x-x_n)|$$

throughout the interval $[-1, 1]$.

Since $(x-x_0)(x-x_1)\cdots(x-x_n)$ is a monic polynomial of degree $(n+1)$, we have just seen that the minimum is obtained when

$$(x-x_0)(x-x_1)\cdots(x-x_n) = \tilde{T}_{n+1}(x).$$

The maximum value of $|(x-x_0)(x-x_1)\cdots(x-x_n)|$ is smallest when x_k is chosen for each $k = 0, 1, \ldots, n$ to be the $(k+1)$st zero of $\tilde{T}_{n+1}$. Hence we choose x_k to be

$$\bar{x}_{k+1} = \cos\left(\frac{2k+1}{2(n+1)}\pi\right).$$

Because $\max_{x\in[-1,1]} |\tilde{T}_{n+1}(x)| = 2^{-n}$, this also implies that

$$\frac{1}{2^n} = \max_{x\in[-1,1]} |(x-\bar{x}_1)\cdots(x-\bar{x}_{n+1})| \le \max_{x\in[-1,1]} |(x-x_0)\cdots(x-x_n)|,$$

for any choice of $x_0, x_1, \ldots, x_n$ in the interval $[-1, 1]$. The next corollary follows from these observations.

Corollary 8.11 Suppose that $P(x)$ is the interpolating polynomial of degree at most n with nodes at the zeros of $T_{n+1}(x)$. Then

$$\max_{x\in[-1,1]} |f(x) - P(x)| \le \frac{1}{2^n(n+1)!} \max_{x\in[-1,1]} |f^{(n+1)}(x)|, \quad \text{for each } f \in C^{n+1}[-1,1]. \qquad \blacksquare$$

Minimizing Approximation Error on Arbitrary Intervals

The technique for choosing points to minimize the interpolating error is extended to a general closed interval $[a, b]$ by using the change of variables

$$\tilde{x} = \frac{1}{2}[(b-a)x + a + b]$$

to transform the numbers $\bar{x}_k$ in the interval $[-1, 1]$ into the corresponding number $\tilde{x}_k$ in the interval $[a, b]$, as shown in the next example.

Example 1 Let $f(x) = xe^x$ on [0, 1.5]. Compare the values given by the Lagrange polynomial with four equally-spaced nodes with those given by the Lagrange polynomial with nodes given by zeros of the fourth Chebyshev polynomial.

Solution The equally-spaced nodes $x_0 = 0, x_1 = 0.5, x_2 = 1$, and $x_3 = 1.5$ give

$$L_0(x) = -1.3333x^3 + 4.0000x^2 - 3.6667x + 1,$$

$$L_1(x) = 4.0000x^3 - 10.000x^2 + 6.0000x,$$

$$L_2(x) = -4.0000x^3 + 8.0000x^2 - 3.0000x,$$

$$L_3(x) = 1.3333x^3 - 2.000x^2 + 0.66667x,$$

which produces the polynomial

$$P_3(x) = L_0(x)(0) + L_1(x)(0.5e^{0.5}) + L_2(x)e^1 + L_3(x)(1.5e^{1.5}) = 1.3875x^3$$
$$+ 0.057570x^2 + 1.2730x.$$

For the second interpolating polynomial, we shift the zeros $\bar{x}_k = \cos((2k+1)/8)\pi$, for $k = 0, 1, 2, 3$, of $\tilde{T}_4$ from $[-1, 1]$ to $[0, 1.5]$, using the linear transformation

$$\tilde{x}_k = \frac{1}{2}[(1.5 - 0)\bar{x}_k + (1.5 + 0)] = 0.75 + 0.75\bar{x}_k.$$

Because

$$\bar{x}_0 = \cos\frac{\pi}{8} = 0.92388, \quad \bar{x}_1 = \cos\frac{3\pi}{8} = 0.38268,$$
$$\bar{x}_2 = \cos\frac{5\pi}{8} = -0.38268, \quad \text{and}\bar{x}_4 = \cos\frac{7\pi}{8} = -0.92388,$$

we have

$$\tilde{x}_0 = 1.44291, \quad \tilde{x}_1 = 1.03701, \quad \tilde{x}_2 = 0.46299, \quad \text{and} \quad \tilde{x}_3 = 0.05709.$$

The Lagrange coefficient polynomials for this set of nodes are

$$\tilde{L}_0(x) = 1.8142x^3 - 2.8249x^2 + 1.0264x - 0.049728,$$

$$\tilde{L}_1(x) = -4.3799x^3 + 8.5977x^2 - 3.4026x + 0.16705,$$

$$\tilde{L}_2(x) = 4.3799x^3 - 11.112x^2 + 7.1738x - 0.37415,$$

$$\tilde{L}_3(x) = -1.8142x^3 + 5.3390x^2 - 4.7976x + 1.2568.$$

The functional values required for these polynomials are given in the last two columns of Table 8.7. The interpolation polynomial of degree at most 3 is

$$\tilde{P}_3(x) = 1.3811x^3 + 0.044652x^2 + 1.3031x - 0.014352.$$

Table 8.7

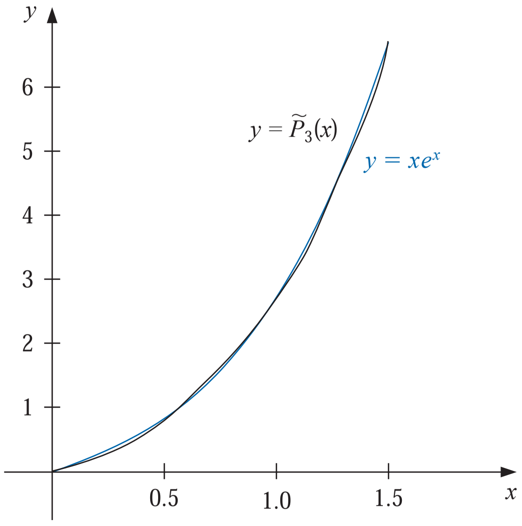
x	$f(x) = xe^x$	$\tilde{x}$	$f(\tilde{x}) = xe^x$
$x_0 = 0.0$	0.00000	$\tilde{x}_0 = 1.44291$	6.10783
$x_1 = 0.5$	0.824361	$\tilde{x}_1 = 1.03701$	2.92517
$x_2 = 1.0$	2.71828	$\tilde{x}_2 = 0.46299$	0.73560
$x_3 = 1.5$	6.72253	$\tilde{x}_3 = 0.05709$	0.060444

For comparison, Table 8.8 lists various values of x, together with the values of $f(x), P_3(x)$, and $\tilde{P}_3(x)$. It can be seen from this table that, although the error using $P_3(x)$ is less than using $\tilde{P}_3(x)$ near the middle of the table, the maximum error involved with using $\tilde{P}_3(x)$, 0.0180, is considerably less than when using $P_3(x)$, which gives the error 0.0290. (See Figure 8.12.) ■

Table 8.8

x	$f(x) = xe^x$	$P_3(x)$	$\lvert xe^x - P_3(x)\rvert$	$\tilde{P}_3(x)$	$\lvert xe^x - \tilde{P}_3(x)\rvert$
0.15	0.1743	0.1969	0.0226	0.1868	0.0125
0.25	0.3210	0.3435	0.0225	0.3358	0.0148
0.35	0.4967	0.5121	0.0154	0.5064	0.0097
0.65	1.245	1.233	0.012	1.231	0.014
0.75	1.588	1.572	0.016	1.571	0.017
0.85	1.989	1.976	0.013	1.974	0.015
1.15	3.632	3.650	0.018	3.644	0.012
1.25	4.363	4.391	0.028	4.382	0.019
1.35	5.208	5.237	0.029	5.224	0.016

Figure 8.12



Reducing the Degree of Approximating Polynomials

Chebyshev polynomials can also be used to reduce the degree of an approximating polynomial with a minimal loss of accuracy. Because the Chebyshev polynomials have a minimum maximum-absolute value that is spread uniformly on an interval, they can be used to reduce the degree of an approximation polynomial without exceeding the error tolerance.

Consider approximating an arbitrary nth-degree polynomial

$$P_n(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

on $[-1, 1]$ with a polynomial of degree at most $n-1$. The object is to choose $P_{n-1}(x)$ in $\prod_{n-1}$ so that

$$\max_{x\in[-1,1]} |P_n(x) - P_{n-1}(x)|$$

is as small as possible.

We first note that $(P_n(x) - P_{n-1}(x))/a_n$ is a monic polynomial of degree n, so applying Theorem 8.10 gives

$$\max_{x\in[-1,1]} \left| \frac{1}{a_n}(P_n(x) - P_{n-1}(x)) \right| \geq \frac{1}{2^{n-1}}.$$

Equality occurs precisely when

$$\frac{1}{a_n}(P_n(x) - P_{n-1}(x)) = \tilde{T}_n(x).$$

This means that we should choose

$$P_{n-1}(x) = P_n(x) - a_n \tilde{T}_n(x),$$

and with this choice we have the minimum value of

$$\max_{x\in[-1,1]} |P_n(x) - P_{n-1}(x)| = |a_n| \max_{x\in[-1,1]} \left| \frac{1}{a_n}(P_n(x) - P_{n-1}(x)) \right| = \frac{|a_n|}{2^{n-1}}.$$

Illustration The function $f(x) = e^x$ is approximated on the interval $[-1, 1]$ by the fourth Maclaurin polynomial

$$P_4(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24},$$

which has truncation error

$$|R_4(x)| = \frac{|f^{(5)}(\xi(x))||x^5|}{120} \leq \frac{e}{120} \approx 0.023, \quad \text{for } -1 \leq x \leq 1.$$

Suppose that an error of 0.05 is tolerable and that we would like to reduce the degree of the approximating polynomial while staying within this bound.

The polynomial of degree 3 or less that best uniformly approximates $P_4(x)$ on $[-1, 1]$ is

$$P_3(x) = P_4(x) - a_4\tilde{T}_4(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} - \frac{1}{24}\left(x^4 - x^2 + \frac{1}{8}\right)$$

$$= \frac{191}{192} + x + \frac{13}{24}x^2 + \frac{1}{6}x^3.$$

With this choice, we have

$$|P_4(x) - P_3(x)| = |a_4\tilde{T}_4(x)| \le \frac{1}{24}\cdot\frac{1}{2^3} = \frac{1}{192} \le 0.0053.$$

Adding this error bound to the bound for the Maclaurin truncation error gives

$$0.023 + 0.0053 = 0.0283,$$

which is within the permissible error of 0.05.

The polynomial of degree 2 or less that best uniformly approximates $P_3(x)$ on $[-1, 1]$ is

$$\begin{aligned} P_2(x) &= P_3(x) - \frac{1}{6}\tilde{T}_3(x) \\ &= \frac{191}{192} + x + \frac{13}{24}x^2 + \frac{1}{6}x^3 - \frac{1}{6}(x^3 - \frac{3}{4}x) = \frac{191}{192} + \frac{9}{8}x + \frac{13}{24}x^2. \end{aligned}$$

However,

$$|P_3(x) - P_2(x)| = \left|\frac{1}{6}\tilde{T}_3(x)\right| = \frac{1}{6}\left(\frac{1}{2}\right)^2 = \frac{1}{24} \approx 0.042,$$

which—when added to the already accumulated error bound of 0.0283—exceeds the tolerance of 0.05. Consequently, the polynomial of least degree that best approximates e^x on $[-1, 1]$ with an error bound of less than 0.05 is

$$P_3(x) = \frac{191}{192} + x + \frac{13}{24}x^2 + \frac{1}{6}x^3.$$

Table 8.9 lists the function and the approximating polynomials at various points in $[-1, 1]$. Note that the tabulated entries for P_2 are well within the tolerance of 0.05, even though the error bound for $P_2(x)$ exceeded the tolerance. □

Table 8.9

x	e^x	$P_4(x)$	$P_3(x)$	$P_2(x)$	$\lvert e^x - P_2(x)\rvert$
−0.75	0.47237	0.47412	0.47917	0.45573	0.01664
−0.25	0.77880	0.77881	0.77604	0.74740	0.03140
0.00	1.00000	1.00000	0.99479	0.99479	0.00521
0.25	1.28403	1.28402	1.28125	1.30990	0.02587
0.75	2.11700	2.11475	2.11979	2.14323	0.02623

EXERCISE SET 8.3

1. Use the zeros of $\tilde{T}_3$ to construct an interpolating polynomial of degree 2 for the following functions on the interval $[-1, 1]$.

   a. $f(x) = e^x$ **b.** $f(x) = \sin x$ **c.** $f(x) = \ln(x + 2)$ **d.** $f(x) = x^4$

2. Use the zeros of $\tilde{T}_4$ to construct an interpolating polynomial of degree 3 for the functions in Exercise 1.
3. Find a bound for the maximum error of the approximation in Exercise 1 on the interval $[-1, 1]$.
4. Repeat Exercise 3 for the approximations computed in Exercise 3.

5. Use the zeros of $\tilde{T}_3$ and transformations of the given interval to construct an interpolating polynomial of degree 2 for the following functions.

   a. $f(x) = \dfrac{1}{x}, \quad [1, 3]$

   b. $f(x) = e^{-x}, \quad [0, 2]$

   c. $f(x) = \dfrac{1}{2}\cos x + \dfrac{1}{3}\sin 2x, \quad [0, 1]$

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

6. Find the sixth Maclaurin polynomial for xe^x, and use Chebyshev economization to obtain a lesser-degree polynomial approximation while keeping the error less than 0.01 on $[-1, 1]$.

7. Find the sixth Maclaurin polynomial for $\sin x$, and use Chebyshev economization to obtain a lesser-degree polynomial approximation while keeping the error less than 0.01 on $[-1, 1]$.

8. Show that for any positive integers i and j with $i > j$, we have $T_i(x)T_j(x) = \frac{1}{2}[T_{i+j}(x) + T_{i-j}(x)]$.

9. Show that for each Chebyshev polynomial $T_n(x)$, we have

$$\int_{-1}^{1} \frac{[T_n(x)]^2}{\sqrt{1-x^2}}\,dx = \frac{\pi}{2}.$$

10. Show that for each n, the Chebyshev polynomial $T_n(x)$ has n distinct zeros in $(-1, 1)$.

11. Show that for each n, the derivative of the Chebyshev polynomial $T_n(x)$ has $n - 1$ distinct zeros in $(-1, 1)$.

8.4 Rational Function Approximation

The class of algebraic polynomials has some distinct advantages for use in approximation:

- There are a sufficient number of polynomials to approximate any continuous function on a closed interval to within an arbitrary tolerance;

- Polynomials are easily evaluated at arbitrary values; and

- The derivatives and integrals of polynomials exist and are easily determined.

The disadvantage of using polynomials for approximation is their tendency for oscillation. This often causes error bounds in polynomial approximation to significantly exceed the average approximation error, because error bounds are determined by the maximum approximation error. We now consider methods that spread the approximation error more evenly over the approximation interval. These techniques involve rational functions.

A **rational function** r of degree N has the form

$$r(x) = \frac{p(x)}{q(x)},$$

where $p(x)$ and $q(x)$ are polynomials whose degrees sum to N.

Every polynomial is a rational function (simply let $q(x) \equiv 1$), so approximation by rational functions gives results that are no worse than approximation by polynomials. However, rational functions whose numerator and denominator have the same or nearly the same degree often produce approximation results superior to polynomial methods for the same amount of computation effort. (This statement is based on the assumption that the amount of computation effort required for division is approximately the same as for multiplication.)

Rational functions have the added advantage of permitting efficient approximation of functions with infinite discontinuities near, but outside, the interval of approximation. Polynomial approximation is generally unacceptable in this situation.

Padé Approximation

Henri Padé (1863–1953) gave a systematic study of what we call today Padé approximations in his doctoral thesis in 1892. He proved results on their general structure and also clearly set out the connection between Padé approximations and continued fractions. These ideas, however, had been studied by Daniel Bernoulli (1700–1782) and others as early as 1730. James Stirling (1692–1770) gave a similar method in *Methodus differentialis* published in the same year, and Euler used Padé-type approximation to find the sum of a series.

Suppose r is a rational function of degree $N = n + m$ of the form

$$r(x) = \frac{p(x)}{q(x)} = \frac{p_0 + p_1x + \cdots + p_nx^n}{q_0 + q_1x + \cdots + q_mx^m},$$

that is used to approximate a function f on a closed interval I containing zero. For r to be defined at zero requires that $q_0 \neq 0$. In fact, we can assume that $q_0 = 1$, for if this is not the case we simply replace $p(x)$ by $p(x)/q_0$ and $q(x)$ by $q(x)/q_0$. Consequently, there are $N + 1$ parameters $q_1, q_2, \ldots, q_m, p_0, p_1, \ldots, p_n$ available for the approximation of f by r.

The **Padé approximation technique**, is the extension of Taylor polynomial approximation to rational functions. It chooses the $N + 1$ parameters so that $f^{(k)}(0) = r^{(k)}(0)$, for each $k = 0, 1, \ldots, N$. When $n = N$ and $m = 0$, the Padé approximation is simply the Nth Maclaurin polynomial.

Consider the difference

$$f(x) - r(x) = f(x) - \frac{p(x)}{q(x)} = \frac{f(x)q(x) - p(x)}{q(x)} = \frac{f(x)\sum_{i=0}^{m} q_ix^i - \sum_{i=0}^{n} p_ix^i}{q(x)},$$

and suppose f has the Maclaurin series expansion $f(x) = \sum_{i=0}^{\infty} a_ix^i$. Then

$$f(x) - r(x) = \frac{\sum_{i=0}^{\infty} a_ix^i \sum_{i=0}^{m} q_ix^i - \sum_{i=0}^{n} p_ix^i}{q(x)}. \tag{8.14}$$

The object is to choose the constants $q_1, q_2, \ldots, q_m$ and $p_0, p_1, \ldots, p_n$ so that

$$f^{(k)}(0) - r^{(k)}(0) = 0, \quad \text{for each } k = 0, 1, \ldots, N.$$

In Section 2.4 (see, in particular, Exercise 10 on page 86) we found that this is equivalent to $f - r$ having a zero of multiplicity $N + 1$ at $x = 0$. As a consequence, we choose $q_1, q_2, \ldots, q_m$ and $p_0, p_1, \ldots, p_n$ so that the numerator on the right side of Eq. (8.14),

$$(a_0 + a_1x + \cdots)(1 + q_1x + \cdots + q_mx^m) - (p_0 + p_1x + \cdots + p_nx^n), \tag{8.15}$$

has no terms of degree less than or equal to N.

To simplify notation, we define $p_{n+1} = p_{n+2} = \cdots = p_N = 0$ and $q_{m+1} = q_{m+2} = \cdots = q_N = 0$. We can then express the coefficient of x^k in expression (8.15) more compactly as

$$\left(\sum_{i=0}^{k} a_iq_{k-i}\right) - p_k.$$

The rational function for Padé approximation results from the solution of the $N + 1$ linear equations

$$\sum_{i=0}^{k} a_iq_{k-i} = p_k, \quad k = 0, 1, \ldots, N$$

in the $N + 1$ unknowns $q_1, q_2, \ldots, q_m, p_0, p_1, \ldots, p_n$.

Example 1 The Maclaurin series expansion for e^{-x} is

$$\sum_{i=0}^{\infty} \frac{(-1)^i}{i!}x^i.$$

Find the Padé approximation to e^{-x} of degree 5 with $n = 3$ and $m = 2$.

Solution To find the Padé approximation we need to choose p_0, p_1, p_2, p_3, q_1, and q_2 so that the coefficients of x^k for $k = 0, 1, \ldots, 5$ are 0 in the expression

$$\left(1 - x + \frac{x^2}{2} - \frac{x^3}{6} + \cdots\right)(1 + q_1x + q_2x^2) - (p_0 + p_1x + p_2x^2 + p_3x^3).$$

Expanding and collecting terms produces

$$\begin{aligned}
x^5 &: & -\frac{1}{120} + \frac{1}{24}q_1 - \frac{1}{6}q_2 &= 0; & \quad x^2 &: & \frac{1}{2} - q_1 + q_2 &= p_2;\\
x^4 &: & \frac{1}{24} - \frac{1}{6}q_1 + \frac{1}{2}q_2 &= 0; & \quad x^1 &: & -1 + q_1 &= p_1;\\
x^3 &: & -\frac{1}{6} + \frac{1}{2}q_1 - q_2 &= p_3; & \quad x^0 &: & 1 &= p_0.
\end{aligned}$$

To solve the system in Maple, we use the following commands:

$eq\,1 := -1 + q1 = p1:$
$eq\,2 := \frac{1}{2} - q1 + q2 = p2:$
$eq\,3 := -\frac{1}{6} + \frac{1}{2}q1 - q2 = p3:$
$eq\,4 := \frac{1}{24} - \frac{1}{6}q1 + \frac{1}{2}q2 = 0:$
$eq\,5 := -\frac{1}{120} + \frac{1}{24}q1 - \frac{1}{6}q2 = 0:$
$solve(\{eq1, eq2, eq3, eq4, eq5\}, \{q1, q2, p1, p2, p3\})$

This gives

$$\left\{p_1 = -\frac{3}{5},\ p_2 = \frac{3}{20},\ p_3 = -\frac{1}{60},\ q_1 = \frac{2}{5},\ q_2 = \frac{1}{20}\right\}$$

So the Padé approximation is

$$r(x) = \frac{1 - \frac{3}{5}x + \frac{3}{20}x^2 - \frac{1}{60}x^3}{1 + \frac{2}{5}x + \frac{1}{20}x^2}.$$

Table 8.10 lists values of $r(x)$ and $P_5(x)$, the fifth Maclaurin polynomial. The Padé approximation is clearly superior in this example. ■

Table 8.10

x	e^{-x}	$P_5(x)$	$\lvert e^{-x} - P_5(x)\rvert$	$r(x)$	$\lvert e^{-x} - r(x)\rvert$
0.2	0.81873075	0.81873067	8.64×10^{-8}	0.81873075	7.55×10^{-9}
0.4	0.67032005	0.67031467	5.38×10^{-6}	0.67031963	4.11×10^{-7}
0.6	0.54881164	0.54875200	5.96×10^{-5}	0.54880763	4.00×10^{-6}
0.8	0.44932896	0.44900267	3.26×10^{-4}	0.44930966	1.93×10^{-5}
1.0	0.36787944	0.36666667	1.21×10^{-3}	0.36781609	6.33×10^{-5}

Maple can also be used directly to compute a Padé approximation. We first compute the Maclaurin series with the call

$series(exp(-x), x)$

to obtain

$$1 - x + \frac{1}{2}x^2 - \frac{1}{6}x^3 + \frac{1}{24}x^4 - \frac{1}{120}x^5 + O(x^6)$$

The Padé approximation $r(x)$ with $n = 3$ and $m = 2$ is found using the command

$r := x \to convert(\%, ratpoly, 3, 2);$

where the % refers to the result of the preceding calculation, namely, the series. The Maple result is

$$x \to \frac{1 - \frac{3}{5}x + \frac{3}{20}x^2 - \frac{1}{60}x^3}{1 + \frac{2}{5}x + \frac{1}{20}x^2}$$

We can then compute, for example, $r(0.8)$ by entering

$r(0.8)$

which produces the approximation 0.4493096647 to $e^{-0.8} = 0.449328964$.

Algorithm 8.1 implements the Padé approximation technique.

ALGORITHM 8.1

Padé Rational Approximation

To obtain the rational approximation

$$r(x) = \frac{p(x)}{q(x)} = \frac{\sum_{i=0}^{n} p_i x^i}{\sum_{j=0}^{m} q_j x^j}$$

for a given function $f(x)$:

```
INPUT   nonnegative integers $m$ and $n$.
OUTPUT  coefficients $q_0, q_1, \ldots, q_m$ and $p_0, p_1, \ldots, p_n$.
Step 1  Set $N = m + n$.
Step 2  For $i = 0, 1, \ldots, N$ set $a_i = \frac{f^{(i)}(0)}{i!}$.
        (The coefficients of the Maclaurin polynomial are $a_0, \ldots, a_N$, which could be
        input instead of calculated.)
Step 3  Set $q_0 = 1$;
            $p_0 = a_0$.
Step 4  For $i = 1, 2, \ldots, N$ do Steps 5–10. (Set up a linear system with matrix $B$.)
    Step 5  For $j = 1, 2, \ldots, i - 1$
                if $j \le n$ then set $b_{i,j} = 0$.
    Step 6  If $i \le n$ then set $b_{i,i} = 1$.
    Step 7  For $j = i + 1, i + 2, \ldots, N$ set $b_{i,j} = 0$.
    Step 8  For $j = 1, 2, \ldots, i$
                if $j \le m$ then set $b_{i,n+j} = -a_{i-j}$.
    Step 9  For $j = n + i + 1, n + i + 2, \ldots, N$ set $b_{i,j} = 0$.
    Step 10  Set $b_{i,N+1} = a_i$.
(Steps 11–22 solve the linear system using partial pivoting.)
Step 11  For $i = n + 1, n + 2, \ldots, N - 1$ do Steps 12–18.
    Step 12  Let $k$ be the smallest integer with $i \le k \le N$ and $|b_{k,i}| = \max_{i \le j \le N} |b_{j,i}|$.
                 (Find pivot element.)
```

```
Step 13  If $b_{k,i} = 0$ then OUTPUT ("The system is singular ");
                   STOP.
Step 14  If $k \neq i$ then    (Interchange row $i$ and row $k$.)
           for $j = i, i+1, \ldots, N+1$ set
                 $b_{COPY} = b_{i,j}$;
                 $b_{i,j} = b_{k,j}$;
                 $b_{k,j} = b_{COPY}$.
Step 15  For $j = i+1, i+2, \ldots, N$ do Steps 16–18.    (Perform elimination.)
    Step 16  Set $xm = \frac{b_{j,i}}{b_{i,i}}$.
    Step 17  For $k = i+1, i+2, \ldots, N+1$
               set $b_{j,k} = b_{j,k} - xm \cdot b_{i,k}$.
    Step 18  Set $b_{j,i} = 0$.
Step 19  If $b_{N,N} = 0$ then OUTPUT ("The system is singular");
                   STOP.
Step 20  If $m > 0$ then set $q_m = \frac{b_{N,N+1}}{b_{N,N}}$.    (Start backward substitution.)
Step 21  For $i = N-1, N-2, \ldots, n+1$ set $q_{i-n} = \frac{b_{i,N+1} - \sum_{j=i+1}^{N} b_{i,j}q_{j-n}}{b_{i,i}}$.
Step 22  For $i = n, n-1, \ldots, 1$ set $p_i = b_{i,N+1} - \sum_{j=n+1}^{N} b_{i,j}q_{j-n}$.
Step 23  OUTPUT $(q_0, q_1, \ldots, q_m, p_0, p_1, \ldots, p_n)$;
         STOP.    (The procedure was successful.)
```

■

Continued Fraction Approximation

It is interesting to compare the number of arithmetic operations required for calculations of $P_5(x)$ and $r(x)$ in Example 1. Using nested multiplication, $P_5(x)$ can be expressed as

$$P_5(x) = \left(\left(\left(\left(-\frac{1}{120}x + \frac{1}{24}\right)x - \frac{1}{6}\right)x + \frac{1}{2}\right)x - 1\right)x + 1.$$

Assuming that the coefficients of $1, x, x^2, x^3, x^4$, and x^5 are represented as decimals, a single calculation of $P_5(x)$ in nested form requires five multiplications and five additions/subtractions.

Using nested multiplication, $r(x)$ is expressed as

$$r(x) = \frac{\left(\left(-\frac{1}{60}x + \frac{3}{20}\right)x - \frac{3}{5}\right)x + 1}{\left(\frac{1}{20}x + \frac{2}{5}\right)x + 1},$$

so a single calculation of $r(x)$ requires five multiplications, five additions/subtractions, and one division. Hence, computational effort appears to favor the polynomial approximation. However, by reexpressing $r(x)$ by continued division, we can write

$$r(x) = \frac{1 - \frac{3}{5}x + \frac{3}{20}x^2 - \frac{1}{60}x^3}{1 + \frac{2}{5}x + \frac{1}{20}x^2}$$

$$= \frac{-\frac{1}{3}x^3 + 3x^2 - 12x + 20}{x^2 + 8x + 20}$$

$$= -\frac{1}{3}x + \frac{17}{3} + \frac{(-\frac{152}{3}x - \frac{280}{3})}{x^2 + 8x + 20}$$

$$= -\frac{1}{3}x + \frac{17}{3} + \frac{-\frac{152}{3}}{\left(\frac{x^2+8x+20}{x+(35/19)}\right)}$$

or

$$r(x) = -\frac{1}{3}x + \frac{17}{3} + \frac{-\frac{152}{3}}{\left(x + \frac{117}{19} + \frac{3125/361}{(x+(35/19))}\right)}. \qquad (8.16)$$

Written in this form, a single calculation of $r(x)$ requires one multiplication, five additions/subtractions, and two divisions. If the amount of computation required for division is approximately the same as for multiplication, the computational effort required for an evaluation of the polynomial $P_5(x)$ significantly exceeds that required for an evaluation of the rational function $r(x)$.

Using continued fractions for rational approximation is a subject that has its roots in the works of Christopher Clavius (1537–1612). It was employed in the 18th and 19th centuries by, for example, Euler, Lagrange, and Hermite.

Expressing a rational function approximation in a form such as Eq. (8.16) is called **continued-fraction** approximation. This is a classical approximation technique of current interest because of the computational efficiency of this representation. It is, however, a specialized technique that we will not discuss further. A rather extensive treatment of this subject and of rational approximation in general can be found in [RR], pp. 285–322.

Although the rational-function approximation in Example 1 gave results superior to the polynomial approximation of the same degree, note that the approximation has a wide variation in accuracy. The approximation at 0.2 is accurate to within 8×10^{-9}, but at 1.0 the approximation and the function agree only to within 7×10^{-5}. This accuracy variation is expected because the Padé approximation is based on a Taylor polynomial representation of e^{-x}, and the Taylor representation has a wide variation of accuracy in [0.2, 1.0].

Chebyshev Rational Function Approximation

To obtain more uniformly accurate rational-function approximations we use Chebyshev polynomials, a class that exhibits more uniform behavior. The general Chebyshev rational-function approximation method proceeds in the same manner as Padé approximation, except that each x^k term in the Padé approximation is replaced by the kth-degree Chebyshev polynomial $T_k(x)$.

Suppose we want to approximate the function f by an Nth-degree rational function r written in the form

$$r(x) = \frac{\sum_{k=0}^{n} p_k T_k(x)}{\sum_{k=0}^{m} q_k T_k(x)}, \quad \text{where } N = n + m \text{ and } q_0 = 1.$$

Writing $f(x)$ in a series involving Chebyshev polynomials as

$$f(x) = \sum_{k=0}^{\infty} a_k T_k(x),$$

gives

$$f(x) - r(x) = \sum_{k=0}^{\infty} a_k T_k(x) - \frac{\sum_{k=0}^{n} p_k T_k(x)}{\sum_{k=0}^{m} q_k T_k(x)}$$

or

$$f(x) - r(x) = \frac{\sum_{k=0}^{\infty} a_k T_k(x) \sum_{k=0}^{m} q_k T_k(x) - \sum_{k=0}^{n} p_k T_k(x)}{\sum_{k=0}^{m} q_k T_k(x)}. \tag{8.17}$$

The coefficients $q_1, q_2, \ldots, q_m$ and $p_0, p_1, \ldots, p_n$ are chosen so that the numerator on the right-hand side of this equation has zero coefficients for $T_k(x)$ when $k = 0, 1, \ldots, N$. This implies that the series

$$\begin{aligned}&(a_0T_0(x) + a_1T_1(x) + \cdots)(T_0(x) + q_1T_1(x) + \cdots + q_mT_m(x))\\ &\quad - (p_0T_0(x) + p_1T_1(x) + \cdots + p_nT_n(x))\end{aligned}$$

has no terms of degree less than or equal to N.

Two problems arise with the Chebyshev procedure that make it more difficult to implement than the Padé method. One occurs because the product of the polynomial $q(x)$ and the series for $f(x)$ involves products of Chebyshev polynomials. This problem is resolved by making use of the relationship

$$T_i(x)T_j(x) = \frac{1}{2}\left[T_{i+j}(x) + T_{|i-j|}(x)\right]. \tag{8.18}$$

(See Exercise 8 of Section 8.3.) The other problem is more difficult to resolve and involves the computation of the Chebyshev series for $f(x)$. In theory, this is not difficult for if

$$f(x) = \sum_{k=0}^{\infty} a_k T_k(x),$$

then the orthogonality of the Chebyshev polynomials implies that

$$a_0 = \frac{1}{\pi}\int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}}\,dx \quad \text{and} \quad a_k = \frac{2}{\pi}\int_{-1}^{1} \frac{f(x)T_k(x)}{\sqrt{1-x^2}}\,dx, \quad \text{where } k \geq 1.$$

Practically, however, these integrals can seldom be evaluated in closed form, and a numerical integration technique is required for each evaluation.

Example 2 The first five terms of the Chebyshev expansion for e^{-x} are

$$\begin{aligned}\tilde{P}_5(x) = {}& 1.266066T_0(x) - 1.130318T_1(x) + 0.271495T_2(x) - 0.044337T_3(x)\\ &+ 0.005474T_4(x) - 0.000543T_5(x).\end{aligned}$$

Determine the Chebyshev rational approximation of degree 5 with $n = 3$ and $m = 2$.

Solution Finding this approximation requires choosing p_0, p_1, p_2, p_3, q_1, and q_2 so that for $k = 0$, 1, 2, 3, 4, and 5, the coefficients of $T_k(x)$ are 0 in the expansion

$$\tilde{P}_5(x)[T_0(x) + q_1T_1(x) + q_2T_2(x)] - [p_0T_0(x) + p_1T_1(x) + p_2T_2(x) + p_3T_3(x)].$$

Using the relation (8.18) and collecting terms gives the equations

$$\begin{aligned}
T_0:&\quad 1.266066 - 0.565159q_1 + 0.1357485q_2 = p_0,\\
T_1:&\quad -1.130318 + 1.401814q_1 - 0.587328q_2 = p_1,\\
T_2:&\quad 0.271495 - 0.587328q_1 + 1.268803q_2 = p_2,\\
T_3:&\quad -0.044337 + 0.138485q_1 - 0.565431q_2 = p_3,\\
T_4:&\quad 0.005474 - 0.022440q_1 + 0.135748q_2 = 0,\\
T_5:&\quad -0.000543 + 0.002737q_1 - 0.022169q_2 = 0.
\end{aligned}$$

The solution to this system produces the rational function

$$r_T(x) = \frac{1.055265T_0(x) - 0.613016T_1(x) + 0.077478T_2(x) - 0.004506T_3(x)}{T_0(x) + 0.378331T_1(x) + 0.022216T_2(x)}.$$

We found at the beginning of Section 8.3 that

$$T_0(x) = 1,\ T_1(x) = x,\ T_2(x) = 2x^2 - 1,\ T_3(x) = 4x^3 - 3x.$$

Using these to convert to an expression involving powers of x gives

$$r_T(x) = \frac{0.977787 - 0.599499x + 0.154956x^2 - 0.018022x^3}{0.977784 + 0.378331x + 0.044432x^2}.$$

Table 8.11 lists values of $r_T(x)$ and, for comparison purposes, the values of $r(x)$ obtained in Example 1. Note that the approximation given by $r(x)$ is superior to that of $r_T(x)$ for $x = 0.2$ and 0.4, but that the maximum error for $r(x)$ is 6.33×10^{-5} compared to 9.13×10^{-6} for $r_T(x)$. ■

Table 8.11

x	e^{-x}	$r(x)$	$\lvert e^{-x} - r(x)\rvert$	$r_T(x)$	$\lvert e^{-x} - r_T(x)\rvert$
0.2	0.81873075	0.81873075	7.55×10^{-9}	0.81872510	5.66×10^{-6}
0.4	0.67032005	0.67031963	4.11×10^{-7}	0.67031310	6.95×10^{-6}
0.6	0.54881164	0.54880763	4.00×10^{-6}	0.54881292	1.28×10^{-6}
0.8	0.44932896	0.44930966	1.93×10^{-5}	0.44933809	9.13×10^{-6}
1.0	0.36787944	0.36781609	6.33×10^{-5}	0.36787155	7.89×10^{-6}

The Chebyshev approximation can be generated using Algorithm 8.2.

ALGORITHM 8.2

Chebyshev Rational Approximation

To obtain the rational approximation

$$r_T(x) = \frac{\sum_{k=0}^{n} p_k T_k(x)}{\sum_{k=0}^{m} q_k T_k(x)}$$

for a given function $f(x)$:

```
INPUT nonnegative integers $m$ and $n$.

OUTPUT coefficients $q_0, q_1, \ldots, q_m$ and $p_0, p_1, \ldots, p_n$.

Step 1 Set $N = m + n$.

Step 2 Set $a_0 = \frac{2}{\pi}\int_0^{\pi} f(\cos\theta)\,d\theta$;   (The coefficient $a_0$ is doubled for computational efficiency.)
       For $k = 1, 2, \ldots, N + m$ set
           $a_k = \frac{2}{\pi}\int_0^{\pi} f(\cos\theta)\cos k\theta\,d\theta$.
       (The integrals can be evaluated using a numerical integration procedure or the
       coefficients can be input directly.)

Step 3 Set $q_0 = 1$.

Step 4 For $i = 0, 1, \ldots, N$ do Steps 5–9.   (Set up a linear system with matrix B.)
```

```
    Step 5  For $j = 0, 1, \dots, i$
                if $j \le n$ then set $b_{i,j} = 0$.
    Step 6  If $i \le n$ then set $b_{i,i} = 1$.
    Step 7  For $j = i + 1, i + 2, \dots, n$ set $b_{i,j} = 0$.
    Step 8  For $j = n + 1, n + 2, \dots, N$
                if $i \ne 0$ then set $b_{i,j} = -\frac{1}{2}(a_{i+j-n} + a_{|i-j+n|})$
                       else set $b_{i,j} = -\frac{1}{2}a_{j-n}$.
    Step 9  If $i \ne 0$ then set $b_{i,N+1} = a_i$
                   else set $b_{i,N+1} = \frac{1}{2}a_i$.
(Steps 10–21 solve the linear system using partial pivoting.)
Step 10  For $i = n + 1, n + 2, \dots, N - 1$ do Steps 11–17.
    Step 11  Let $k$ be the smallest integer with $i \le k \le N$ and
             $|b_{k,i}| = \max_{i \le j \le N} |b_{j,i}|$.   (Find pivot element.)
    Step 12  If $b_{k,i} = 0$ then OUTPUT ("The system is singular");
                           STOP.
    Step 13  If $k \ne i$ then    (Interchange row i and row k.)
                 for $j = i, i + 1, \dots, N + 1$ set
                     $b_{COPY} = b_{i,j}$;
                     $b_{i,j} = b_{k,j}$;
                     $b_{k,j} = b_{COPY}$.
    Step 14  For $j = i + 1, i + 2, \dots, N$ do Steps 15–17.   (Perform elimination.)
        Step 15  Set $xm = \frac{b_{j,i}}{b_{i,i}}$.
        Step 16  For $k = i + 1, i + 2, \dots, N + 1$
                     set $b_{j,k} = b_{j,k} - xm \cdot b_{i,k}$.
        Step 17  Set $b_{j,i} = 0$.
Step 18  If $b_{N,N} = 0$ then OUTPUT ("The system is singular");
                         STOP.
Step 19  If $m > 0$ then set $q_m = \frac{b_{N,N+1}}{b_{N,N}}$.   (Start backward substitution.)
Step 20  For $i = N - 1, N - 2, \dots, n + 1$ set $q_{i-n} = \frac{b_{i,N+1} - \sum_{j=i+1}^{N} b_{i,j}q_{j-n}}{b_{i,i}}$.
Step 21  For $i = n, n - 1, \dots, 0$ set $p_i = b_{i,N+1} - \sum_{j=n+1}^{N} b_{i,j}q_{j-n}$.
Step 22  OUTPUT $(q_0, q_1, \dots, q_m, p_0, p_1, \dots, p_n)$;
         STOP.   (The procedure was successful.)
```

We can obtain both the Chebyshev series expansion and the Chebyshev rational approximation using Maple using the *orthopoly* and *numapprox* packages. Load the packages and then enter the command

$g := chebyshev(e^{-x}, x, 0.00001)$

The parameter 0.000001 tells Maple to truncate the series when the remaining coefficients divided by the largest coefficient is smaller that 0.000001. Maple returns

$$1.266065878T(0,x) - 1.130318208T(1,x) + .2714953396T(2,x) - 0.04433684985T(3,x)$$
$$+ 0.005474240442T(4,x) - 0.0005429263119T(5,x) + 0.00004497732296T(6,x)$$
$$- 0.000003198436462T(7,x)$$

The approximation to $e^{-0.8} = 0.449328964$ is found with

$evalf(subs(x = .8, g))$

$$0.4493288893$$

To obtain the Chebyshev rational approximation enter

$gg := convert(chebyshev(e^{-x}, x, 0.00001), ratpoly, 3, 2)$

resulting in

$$gg := \frac{0.9763521942 - 0.5893075371x + 0.1483579430x^2 - 0.01643823341x^3}{0.9763483269 + 0.3870509565x + 0.04730334625x^2}$$

We can evaluate $g(0.8)$ by

$evalf(subs(x = 0.8, g))$

which gives 0.4493317577 as an approximation to $e^{-0.8} = 0.449328964$.

The Chebyshev method does not produce the best rational function approximation in the sense of the approximation whose maximum approximation error is minimal. The method can, however, be used as a starting point for an iterative method known as the second Remez' algorithm that converges to the best approximation. A discussion of the techniques involved with this procedure and an improvement on this algorithm can be found in [RR], pp. 292–305, or in [Pow], pp. 90–92.

In 1930, Evgeny Remez (1896–1975) developed general computational methods of Chebyshev approximation for polynomials. He later developed a similar algorithm for the rational approximation of continuous functions defined on an interval with a prescribed degree of accuracy. His work encompassed various areas of approximation theory as well as the methods for approximating the solutions of differential equations.

EXERCISE SET 8.4

1. Determine all degree 2 Padé approximations for $f(x) = e^{2x}$. Compare the results at $x_i = 0.2i$, for $i = 1, 2, 3, 4, 5$, with the actual values $f(x_i)$.
2. Determine all degree 3 Padé approximations for $f(x) = x\ln(x+1)$. Compare the results at $x_i = 0.2i$, for $i = 1, 2, 3, 4, 5$, with the actual values $f(x_i)$.
3. Determine the Padé approximation of degree 5 with $n = 2$ and $m = 3$ for $f(x) = e^x$. Compare the results at $x_i = 0.2i$, for $i = 1, 2, 3, 4, 5$, with those from the fifth Maclaurin polynomial.
4. Repeat Exercise 3 using instead the Padé approximation of degree 5 with $n = 3$ and $m = 2$. Compare the results at each x_i with those computed in Exercise 3.
5. Determine the Padé approximation of degree 6 with $n = m = 3$ for $f(x) = \sin x$. Compare the results at $x_i = 0.1i$, for $i = 0, 1, \ldots, 5$, with the exact results and with the results of the sixth Maclaurin polynomial.
6. Determine the Padé approximations of degree 6 with (a) $n = 2, m = 4$ and (b) $n = 4$, $m = 2$ for $f(x) = \sin x$. Compare the results at each x_i to those obtained in Exercise 5.
7. Table 8.10 lists results of the Padé approximation of degree 5 with $n = 3$ and $m = 2$, the fifth Maclaurin polynomial, and the exact values of $f(x) = e^{-x}$ when $x_i = 0.2i$, for $i = 1, 2, 3, 4,$

and 5. Compare these results with those produced from the other Padé approximations of degree five.

a. $n = 0, m = 5$ **b.** $n = 1, m = 4$ **c.** $n = 3, m = 2$ **d.** $n = 4, m = 1$

8. Express the following rational functions in continued-fraction form:

a. $\dfrac{x^2 + 3x + 2}{x^2 - x + 1}$ **b.** $\dfrac{4x^2 + 3x - 7}{2x^3 + x^2 - x + 5}$

c. $\dfrac{2x^3 - 3x^2 + 4x - 5}{x^2 + 2x + 4}$ **d.** $\dfrac{2x^3 + x^2 - x + 3}{3x^3 + 2x^2 - x + 1}$

9. Find all the Chebyshev rational approximations of degree 2 for $f(x) = e^{-x}$. Which give the best approximations to $f(x) = e^{-x}$ at $x = 0.25$, 0.5, and 1?

10. Find all the Chebyshev rational approximations of degree 3 for $f(x) = \cos x$. Which give the best approximations to $f(x) = \cos x$ at $x = \pi/4$ and $\pi/3$?

11. Find the Chebyshev rational approximation of degree 4 with $n = m = 2$ for $f(x) = \sin x$. Compare the results at $x_i = 0.1i$, for $i = 0, 1, 2, 3, 4, 5$, from this approximation with those obtained in Exercise 5 using a sixth-degree Padé approximation.

12. Find all Chebyshev rational approximations of degree 5 for $f(x) = e^x$. Compare the results at $x_i = 0.2i$, for $i = 1, 2, 3, 4, 5$, with those obtained in Exercises 3 and 4.

13. To accurately approximate $f(x) = e^x$ for inclusion in a mathematical library, we first restrict the domain of f. Given a real number x, divide by $\ln\sqrt{10}$ to obtain the relation

$$x = M \cdot \ln\sqrt{10} + s,$$

where M is an integer and s is a real number satisfying $|s| \le \frac{1}{2}\ln\sqrt{10}$.

a. Show that $e^x = e^s \cdot 10^{M/2}$.

b. Construct a rational function approximation for e^s using $n = m = 3$. Estimate the error when $0 \le |s| \le \frac{1}{2}\ln\sqrt{10}$.

c. Design an implementation of e^x using the results of part (a) and (b) and the approximations

$$\frac{1}{\ln\sqrt{10}} = 0.8685889638 \quad \text{and} \quad \sqrt{10} = 3.162277660.$$

14. To accurately approximate $\sin x$ and $\cos x$ for inclusion in a mathematical library, we first restrict their domains. Given a real number x, divide by π to obtain the relation

$$|x| = M\pi + s, \quad \text{where } M \text{ is an integer and } |s| \le \frac{\pi}{2}.$$

a. Show that $\sin x = \text{sgn}(x) \cdot (-1)^M \cdot \sin s$.

b. Construct a rational approximation to $\sin s$ using $n = m = 4$. Estimate the error when $0 \le |s| \le \pi/2$.

c. Design an implementation of $\sin x$ using parts (a) and (b).

d. Repeat part (c) for $\cos x$ using the fact that $\cos x = \sin(x + \pi/2)$.

8.5 Trigonometric Polynomial Approximation

The use of series of sine and cosine functions to represent arbitrary functions had its beginnings in the 1750s with the study of the motion of a vibrating string. This problem was considered by Jean d'Alembert and then taken up by the foremost mathematician of the time, Leonhard Euler. But it was Daniel Bernoulli who first advocated the use of the infinite sums of sine and cosines as a solution to the problem, sums that we now know as Fourier series. In the early part of the 19th century, Jean Baptiste Joseph Fourier used these series to study the flow of heat and developed quite a complete theory of the subject.

During the late 17th and early 18th centuries, the Bernoulli family produced no less than 8 important mathematicians and physicists. Daniel Bernoulli's most important work involved the pressure, density, and velocity of fluid flow, which produced what is known as the *Bernoulli principle.*

The first observation in the development of Fourier series is that, for each positive integer n, the set of functions $\{\phi_0, \phi_1, \ldots, \phi_{2n-1}\}$, where

$$\phi_0(x) = \frac{1}{2},$$

$$\phi_k(x) = \cos kx, \quad \text{for each } k = 1, 2, \ldots, n,$$

and

$$\phi_{n+k}(x) = \sin kx, \quad \text{for each } k = 1, 2, \ldots, n-1,$$

is an orthogonal set on $[-\pi, \pi]$ with respect to $w(x) \equiv 1$. This orthogonality follows from the fact that for every integer j, the integrals of $\sin jx$ and $\cos jx$ over $[-\pi, \pi]$ are 0, and we can rewrite products of sine and cosine functions as sums by using the three trigonometric identities

$$\begin{aligned}\sin t_1 \sin t_2 &= \frac{1}{2}[\cos(t_1 - t_2) - \cos(t_1 + t_2)],\\ \cos t_1 \cos t_2 &= \frac{1}{2}[\cos(t_1 - t_2) + \cos(t_1 + t_2)],\\ \sin t_1 \cos t_2 &= \frac{1}{2}[\sin(t_1 - t_2) + \sin(t_1 + t_2)].\end{aligned} \tag{8.19}$$

Orthogonal Trigonometric Polynomials

Let $\mathcal{T}_n$ denote the set of all linear combinations of the functions $\phi_0, \phi_1, \ldots, \phi_{2n-1}$. This set is called the set of **trigonometric polynomials** of degree less than or equal to n. (Some sources also include an additional function in the set, $\phi_{2n}(x) = \sin nx$.)

For a function $f \in C[-\pi, \pi]$, we want to find the *continuous least squares* approximation by functions in $\mathcal{T}_n$ in the form

$$S_n(x) = \frac{a_0}{2} + a_n \cos nx + \sum_{k=1}^{n-1}(a_k \cos kx + b_k \sin kx).$$

Since the set of functions $\{\phi_0, \phi_1, \ldots, \phi_{2n-1}\}$ is orthogonal on $[-\pi, \pi]$ with respect to $w(x) \equiv 1$, it follows from Theorem 8.6 on page 515 and the equations in (8.19) that the appropriate selection of coefficients is

$$a_k = \frac{\int_{-\pi}^{\pi} f(x) \cos kx\,dx}{\int_{-\pi}^{\pi} (\cos kx)^2\,dx} = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x) \cos kx\,dx, \quad \text{for each } k = 0, 1, 2, \ldots, n, \tag{8.20}$$

and

Joseph Fourier (1768–1830) published his theory of trigonometric series in *Théorie analytique de la chaleur* to solve the problem of steady state heat distribution in a solid.

$$b_k = \frac{\int_{-\pi}^{\pi} f(x) \sin kx\,dx}{\int_{-\pi}^{\pi} (\sin kx)^2\,dx} = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x) \sin kx\,dx, \quad \text{for each } k = 1, 2, \ldots, n-1. \tag{8.21}$$

The limit of $S_n(x)$ when $n \to \infty$ is called the **Fourier series** of f. Fourier series are used to describe the solution of various ordinary and partial-differential equations that occur in physical situations.

Example 1 Determine the trigonometric polynomial from $\mathcal{T}_n$ that approximates

$$f(x) = |x|, \quad \text{for } -\pi < x < \pi.$$

Solution We first need to find the coefficients

$$a_0 = \frac{1}{\pi}\int_{-\pi}^{\pi} |x|\,dx = -\frac{1}{\pi}\int_{-\pi}^{0} x\,dx + \frac{1}{\pi}\int_{0}^{\pi} x\,dx = \frac{2}{\pi}\int_{0}^{\pi} x\,dx = \pi,$$

$$a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} |x|\cos kx\,dx = \frac{2}{\pi}\int_{0}^{\pi} x\cos kx\,dx = \frac{2}{\pi k^2}\left[(-1)^k - 1\right],$$

for each $k = 1, 2, \ldots, n$, and

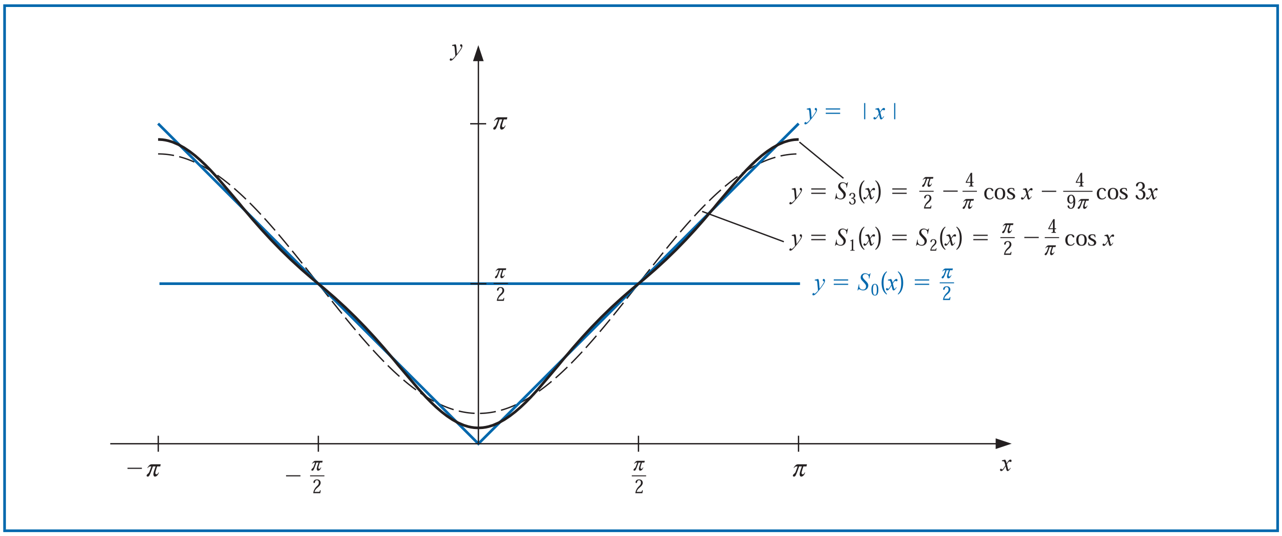
$$b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} |x|\sin kx\,dx = 0, \quad \text{for each } k = 1, 2, \ldots, n-1.$$

That the b_k's are all 0 follows from the fact that $g(x) = |x|\sin kx$ is an odd function for each k, and the integral of a continuous odd function over an interval of the form $[-a, a]$ is 0. (See Exercises 13 and 14.) The trigonometric polynomial from $\mathcal{T}_n$ approximating f is therefore,

$$S_n(x) = \frac{\pi}{2} + \frac{2}{\pi}\sum_{k=1}^{n} \frac{(-1)^k - 1}{k^2}\cos kx.$$

The first few trigonometric polynomials for $f(x) = |x|$ are shown in Figure 8.13. ■

Figure 8.13



The Fourier series for f is

$$S(x) = \lim_{n\to\infty} S_n(x) = \frac{\pi}{2} + \frac{2}{\pi}\sum_{k=1}^{\infty} \frac{(-1)^k - 1}{k^2}\cos kx.$$

Since $|\cos kx| \leq 1$ for every k and x, the series converges, and $S(x)$ exists for all real numbers x.

Discrete Trigonometric Approximation

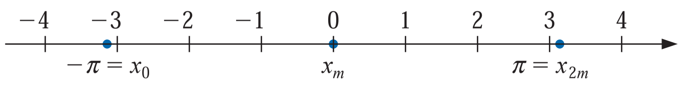
There is a discrete analog that is useful for the *discrete least squares* approximation and the interpolation of large amounts of data.

Suppose that a collection of $2m$ paired data points $\{(x_j, y_j)\}_{j=0}^{2m-1}$ is given, with the first elements in the pairs equally partitioning a closed interval. For convenience, we assume that the interval is $[-\pi, \pi]$, so, as shown in Figure 8.14,

$$x_j = -\pi + \left(\frac{j}{m}\right)\pi, \quad \text{for each } j = 0, 1, \dots, 2m-1. \tag{8.22}$$

If it is not $[-\pi, \pi]$, a simple linear transformation could be used to transform the data into this form.

Figure 8.14



The goal in the discrete case is to determine the trigonometric polynomial $S_n(x)$ in $\mathcal{T}_n$ that will minimize

$$E(S_n) = \sum_{j=0}^{2m-1} [y_j - S_n(x_j)]^2.$$

To do this we need to choose the constants $a_0, a_1, \dots, a_n, b_1, b_2, \dots, b_{n-1}$ to minimize

$$E(S_n) = \sum_{j=0}^{2m-1} \left\{ y_j - \left[\frac{a_0}{2} + a_n \cos nx_j + \sum_{k=1}^{n-1} (a_k \cos kx_j + b_k \sin kx_j) \right] \right\}^2. \tag{8.23}$$

The determination of the constants is simplified by the fact that the set $\{\phi_0, \phi_1, \dots, \phi_{2n-1}\}$ is orthogonal with respect to summation over the equally spaced points $\{x_j\}_{j=0}^{2m-1}$ in $[-\pi, \pi]$. By this we mean that for each $k \neq l$,

$$\sum_{j=0}^{2m-1} \phi_k(x_j)\phi_l(x_j) = 0. \tag{8.24}$$

To show this orthogonality, we use the following lemma.

Lemma 8.12 Suppose that the integer r is not a multiple of $2m$. Then

- $\displaystyle\sum_{j=0}^{2m-1} \cos rx_j = 0 \quad \text{and} \quad \sum_{j=0}^{2m-1} \sin rx_j = 0.$

Moreover, if r is not a multiple of m, then

- $\displaystyle\sum_{j=0}^{2m-1} (\cos rx_j)^2 = m \quad \text{and} \quad \sum_{j=0}^{2m-1} (\sin rx_j)^2 = m.$ ■

Euler first used the symbol i in 1794 to represent $\sqrt{-1}$ in his memoir *De Formulis Differentialibus Angularibus.*

Proof *Euler's Formula* states that with $i^2 = -1$, we have, for every real number z,

$$e^{iz} = \cos z + i \sin z. \tag{8.25}$$

Applying this result gives

$$\sum_{j=0}^{2m-1} \cos rx_j + i \sum_{j=0}^{2m-1} \sin rx_j = \sum_{j=0}^{2m-1} \left(\cos rx_j + i \sin rx_j\right) = \sum_{j=0}^{2m-1} e^{irx_j}.$$

But

$$e^{irx_j} = e^{ir(-\pi + j\pi/m)} = e^{-ir\pi} \cdot e^{irj\pi/m},$$

so

$$\sum_{j=0}^{2m-1} \cos rx_j + i \sum_{j=0}^{2m-1} \sin rx_j = e^{-ir\pi} \sum_{j=0}^{2m-1} e^{irj\pi/m}.$$

Since $\sum_{j=0}^{2m-1} e^{irj\pi/m}$ is a geometric series with first term 1 and ratio $e^{ir\pi/m} \neq 1$, we have

$$\sum_{j=0}^{2m-1} e^{irj\pi/m} = \frac{1 - (e^{ir\pi/m})^{2m}}{1 - e^{ir\pi/m}} = \frac{1 - e^{2ir\pi}}{1 - e^{ir\pi/m}}.$$

But $e^{2ir\pi} = \cos 2r\pi + i \sin 2r\pi = 1$, so $1 - e^{2ir\pi} = 0$ and

$$\sum_{j=0}^{2m-1} \cos rx_j + i \sum_{j=0}^{2m-1} \sin rx_j = e^{-ir\pi} \sum_{j=0}^{2m-1} e^{irj\pi/m} = 0.$$

This implies that both the real and imaginary parts are zero, so

$$\sum_{j=0}^{2m-1} \cos rx_j = 0 \quad \text{and} \quad \sum_{j=0}^{2m-1} \sin rx_j = 0.$$

In addition, if r is not a multiple of m, these sums imply that

$$\sum_{j=0}^{2m-1} (\cos rx_j)^2 = \sum_{j=0}^{2m-1} \frac{1}{2}\left(1 + \cos 2rx_j\right) = \frac{1}{2}\left[2m + \sum_{j=0}^{2m-1} \cos 2rx_j\right] = \frac{1}{2}(2m + 0) = m$$

and, similarly, that

$$\sum_{j=0}^{2m-1} (\sin rx_j)^2 = \sum_{j=0}^{2m-1} \frac{1}{2}\left(1 - \cos 2rx_j\right) = m. \qquad \blacksquare\ \blacksquare\ \blacksquare$$

We can now show the orthogonality stated in (8.24). Consider, for example, the case

$$\sum_{j=0}^{2m-1} \phi_k(x_j)\phi_{n+l}(x_j) = \sum_{j=0}^{2m-1} (\cos kx_j)(\sin lx_j).$$

Since

$$\cos kx_j \sin lx_j = \frac{1}{2}[\sin(l+k)x_j + \sin(l-k)x_j]$$

and $(l+k)$ and $(l-k)$ are both integers that are not multiples of $2m$, Lemma 8.12 implies that

$$\sum_{j=0}^{2m-1} (\cos kx_j)(\sin lx_j) = \frac{1}{2}\left[\sum_{j=0}^{2m-1} \sin(l+k)x_j + \sum_{j=0}^{2m-1} \sin(l-k)x_j\right] = \frac{1}{2}(0+0) = 0.$$

This technique is used to show that the orthogonality condition is satisfied for any pair of the functions and to produce the following result.

Theorem 8.13 The constants in the summation

$$S_n(x) = \frac{a_0}{2} + a_n \cos nx + \sum_{k=1}^{n-1}(a_k \cos kx + b_k \sin kx)$$

that minimize the least squares sum

$$E(a_0, \ldots, a_n, b_1, \ldots, b_{n-1}) = \sum_{j=0}^{2m-1} (y_j - S_n(x_j))^2$$

are

- $a_k = \dfrac{1}{m} \displaystyle\sum_{j=0}^{2m-1} y_j \cos kx_j, \quad \text{for each } k = 0, 1, \ldots, n,$

and

- $b_k = \dfrac{1}{m} \displaystyle\sum_{j=0}^{2m-1} y_j \sin kx_j, \quad \text{for each } k = 1, 2, \ldots, n-1.$ ■

The theorem is proved by setting the partial derivatives of E with respect to the a_k's and the b_k's to zero, as was done in Sections 8.1 and 8.2, and applying the orthogonality to simplify the equations. For example,

$$0 = \frac{\partial E}{\partial b_k} = 2\sum_{j=0}^{2m-1}[y_j - S_n(x_j)](-\sin kx_j),$$

so

$$\begin{aligned} 0 &= \sum_{j=0}^{2m-1} y_j \sin kx_j - \sum_{j=0}^{2m-1} S_n(x_j) \sin kx_j \\ &= \sum_{j=0}^{2m-1} y_j \sin kx_j - \frac{a_0}{2}\sum_{j=0}^{2m-1} \sin kx_j - a_n \sum_{j=0}^{2m-1} \sin kx_j \cos nx_j \\ &\quad - \sum_{l=1}^{n-1} a_l \sum_{j=0}^{2m-1} \sin kx_j \cos lx_j - \sum_{\substack{l=1,\\ l\neq k}}^{n-1} b_l \sum_{j=0}^{2m-1} \sin kx_j \sin lx_j - b_k \sum_{j=0}^{2m-1} (\sin kx_j)^2. \end{aligned}$$

The orthogonality implies that all but the first and last sums on the right side are zero, and Lemma 8.12 states the final sum is m. Hence

$$0 = \sum_{j=0}^{2m-1} y_j \sin kx_j - mb_k,$$

which implies that

$$b_k = \frac{1}{m} \sum_{j=0}^{2m-1} y_j \sin kx_j.$$

The result for the a_k's is similar but need an additional step to determine a_0 (See Exercise 17.)

Example 2 Find $S_2(x)$, the discrete least squares trigonometric polynomial of degree 2 for $f(x) = 2x^2 - 9$ when x is in $[-\pi, \pi]$.

Solution We have $m = 2(2) - 1 = 3$, so the nodes are

$$x_j = \pi + \frac{j}{m}\pi \quad \text{and} \quad y_j = f(x_j) = 2x_j^2 - 9, \quad \text{for } j = 0, 1, 2, 3, 4, 5.$$

The trigonometric polynomial is

$$S_2(x) = \frac{1}{2}a_0 + a_2 \cos 2x + (a_1 \cos x + b_1 \sin x),$$

where

$$a_k = \frac{1}{3} \sum_{j=0}^{5} y_j \cos kx_j, \text{ for } k = 0, 1, 2, \quad \text{and} \quad b_1 = \frac{1}{3} \sum_{j=0}^{5} y_j \sin x_j.$$

The coefficients are

$$a_0 = \frac{1}{3}\left(f(-\pi) + f\left(-\frac{2\pi}{3}\right) + f\left(-\frac{\pi}{3}\right) f(0) + f\left(\frac{\pi}{3}\right) + f\left(\frac{2\pi}{3}\right)\right) = -4.10944566,$$

$$a_1 = \frac{1}{3}\left(f(-\pi)\cos(-\pi) + f\left(-\frac{2\pi}{3}\right)\cos\left(-\frac{2\pi}{3}\right) + f\left(-\frac{\pi}{3}\right)\cos\left(-\frac{\pi}{3}\right) f(0)\cos 0\right.$$
$$\left. + f\left(\frac{\pi}{3}\right)\cos\left(\frac{\pi}{3}\right) + f\left(\frac{2\pi}{3}\right)\cos\left(\frac{2\pi}{3}\right)\right) = -8.77298169,$$

$$a_2 = \frac{1}{3}\left(f(-\pi)\cos(-2\pi) + f\left(-\frac{2\pi}{3}\right)\cos\left(-\frac{4\pi}{3}\right) + f\left(-\frac{\pi}{3}\right)\cos\left(-\frac{2\pi}{3}\right) f(0)\cos 0\right.$$
$$\left. + f\left(\frac{\pi}{3}\right)\cos\left(\frac{2\pi}{3}\right) + f\left(\frac{2\pi}{3}\right)\cos\left(\frac{4\pi}{3}\right)\right) = 2.92432723,$$

and

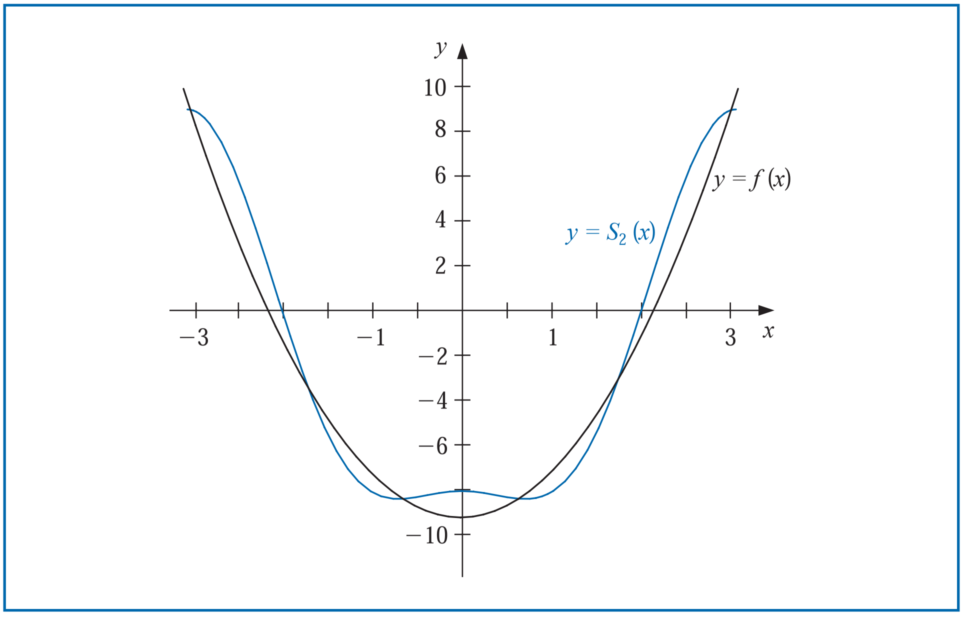
$$b_1 = \frac{1}{3}\left(f(-\pi)\sin(-\pi) + f\left(-\frac{2\pi}{3}\right)\sin\left(-\frac{\pi}{3}\right) + f\left(-\frac{\pi}{3}\right)\left(-\frac{\pi}{3}\right) f(0)\sin 0\right.$$
$$\left. + f\left(\frac{\pi}{3}\right)\left(\frac{\pi}{3}\right) + f\left(\frac{2\pi}{3}\right)\left(\frac{2\pi}{3}\right)\right) = 0.$$

Thus

$$S_2(x) = \frac{1}{2}(-4.10944562) - 8.77298169 \cos x + 2.92432723 \cos 2x.$$

Figure 8.15 shows $f(x)$ and the discrete least squares trigonometric polynomial $S_2(x)$. ■

Figure 8.15



The next example gives an illustration of finding a least-squares approximation for a function that is defined on a closed interval other than $[-\pi, \pi]$.

Example 3 Find the discrete least squares approximation $S_3(x)$ for

$$f(x) = x^4 - 3x^3 + 2x^2 - \tan x(x-2)$$

using the data $\{(x_j, y_j)\}_{j=0}^{9}$, where $x_j = j/5$ and $y_j = f(x_j)$.

Solution We first need the linear transformation from $[0, 2]$ to $[-\pi, \pi]$ given by

$$z_j = \pi(x_j - 1).$$

Then the transformed data have the form

$$\left\{\left(z_j, f\left(1 + \frac{z_j}{\pi}\right)\right)\right\}_{j=0}^{9}.$$

The least squares trigonometric polynomial is consequently,

$$S_3(z) = \left[\frac{a_0}{2} + a_3 \cos 3z + \sum_{k=1}^{2}(a_k \cos kz + b_k \sin kz)\right],$$

where

$$a_k = \frac{1}{5}\sum_{j=0}^{9} f\left(1 + \frac{z_j}{\pi}\right)\cos kz_j, \quad \text{for } k = 0, 1, 2, 3,$$

and

$$b_k = \frac{1}{5}\sum_{j=0}^{9} f\left(1 + \frac{z_j}{\pi}\right)\sin kz_j, \quad \text{for } k = 1, 2.$$

Evaluating these sums produces the approximation

$$S_3(z) = 0.76201 + 0.77177 \cos z + 0.017423 \cos 2z + 0.0065673 \cos 3z$$
$$- 0.38676 \sin z + 0.047806 \sin 2z,$$

and converting back to the variable x gives

$$S_3(x) = 0.76201 + 0.77177 \cos \pi(x-1) + 0.017423 \cos 2\pi(x-1)$$
$$+ 0.0065673 \cos 3\pi(x-1) - 0.38676 \sin \pi(x-1) + 0.047806 \sin 2\pi(x-1).$$

Table 8.12 lists values of $f(x)$ and $S_3(x)$. ■

Table 8.12

x	$f(x)$	$S_3(x)$	$\lvert f(x) - S_3(x)\rvert$
0.125	0.26440	0.24060	2.38×10^{-2}
0.375	0.84081	0.85154	1.07×10^{-2}
0.625	1.36150	1.36248	9.74×10^{-4}
0.875	1.61282	1.60406	8.75×10^{-3}
1.125	1.36672	1.37566	8.94×10^{-3}
1.375	0.71697	0.71545	1.52×10^{-3}
1.625	0.07909	0.06929	9.80×10^{-3}
1.875	−0.14576	−0.12302	2.27×10^{-2}

EXERCISE SET 8.5

1. Find the continuous least squares trigonometric polynomial $S_2(x)$ for $f(x) = x^2$ on $[-\pi, \pi]$.
2. Find the continuous least squares trigonometric polynomial $S_n(x)$ for $f(x) = x$ on $[-\pi, \pi]$.
3. Find the continuous least squares trigonometric polynomial $S_3(x)$ for $f(x) = e^x$ on $[-\pi, \pi]$.
4. Find the general continuous least squares trigonometric polynomial $S_n(x)$ for $f(x) = e^x$ on $[-\pi, \pi]$.
5. Find the general continuous least squares trigonometric polynomial $S_n(x)$ for

$$f(x) = \begin{cases} 0, & \text{if } -\pi < x \le 0, \\ 1, & \text{if } 0 < x < \pi. \end{cases}$$

6. Find the general continuous least squares trigonometric polynomial $S_n(x)$ in for

$$f(x) = \begin{cases} -1, & \text{if } -\pi < x < 0. \\ 1, & \text{if } 0 \le x \le \pi. \end{cases}$$

7. Determine the discrete least squares trigonometric polynomial $S_n(x)$ on the interval $[-\pi, \pi]$ for the following functions, using the given values of m and n:
   a. $f(x) = \cos 2x,\ m = 4, n = 2$ **b.** $f(x) = \cos 3x,\ m = 4, n = 2$
   c. $f(x) = \sin \frac{x}{2} + 2 \cos \frac{x}{3},\ m = 6, n = 3$ **d.** $f(x) = x^2 \cos x,\ m = 6, n = 3$
8. Compute the error $E(S_n)$ for each of the functions in Exercise 7.
9. Determine the discrete least squares trigonometric polynomial $S_3(x)$, using $m = 4$ for $f(x) = e^x \cos 2x$ on the interval $[-\pi, \pi]$. Compute the error $E(S_3)$.
10. Repeat Exercise 9 using $m = 8$. Compare the values of the approximating polynomials with the values of f at the points $\xi_j = -\pi + 0.2j\pi$, for $0 \le j \le 10$. Which approximation is better?

11. Let $f(x) = 2\tan x - \sec 2x$, for $2 \leq x \leq 4$. Determine the discrete least squares trigonometric polynomials $S_n(x)$, using the values of n and m as follows, and compute the error in each case.

a. $n = 3, \quad m = 6$ **b.** $n = 4, \quad m = 6$

12. **a.** Determine the discrete least squares trigonometric polynomial $S_4(x)$, using $m = 16$, for $f(x) = x^2 \sin x$ on the interval $[0, 1]$.

b. Compute $\int_0^1 S_4(x)\,dx$.

c. Compare the integral in part (b) to $\int_0^1 x^2 \sin x\,dx$.

13. Show that for any continuous odd function f defined on the interval $[-a, a]$, we have $\int_{-a}^{a} f(x)\,dx = 0$.

14. Show that for any continuous even function f defined on the interval $[-a, a]$, we have $\int_{-a}^{a} f(x)\,dx = 2\int_0^a f(x)\,dx$.

15. Show that the functions $\phi_0(x) = 1/2, \phi_1(x) = \cos x, \ldots, \phi_n(x) = \cos nx, \phi_{n+1}(x) = \sin x, \ldots, \phi_{2n-1}(x) = \sin(n-1)x$ are orthogonal on $[-\pi, \pi]$ with respect to $w(x) \equiv 1$.

16. In Example 1 the Fourier series was determined for $f(x) = |x|$. Use this series and the assumption that it represents f at zero to find the value of the convergent infinite series $\sum_{k=0}^{\infty}(1/(2k+1)^2)$.

17. Show that the form of the constants a_k for $k = 0, \ldots, n$ in Theorem 8.13 is correct as stated.

8.6 Fast Fourier Transforms

In the latter part of Section 8.5, we determined the form of the discrete least squares polynomial of degree n on the $2m$ data points $\{(x_j, y_j)\}_{j=0}^{2m-1}$, where $x_j = -\pi + (j/m)\pi$, for each $j = 0, 1, \ldots, 2m-1$.

The *interpolatory* trigonometric polynomial in $\mathcal{T}_m$ on these $2m$ data points is nearly the same as the least squares polynomial. This is because the least squares trigonometric polynomial minimizes the error term

$$E(S_m) = \sum_{j=0}^{2m-1} \left(y_j - S_m(x_j)\right)^2,$$

and for the interpolatory trigonometric polynomial, this error is 0, hence minimized, when the $S_m(x_j) = y_j$, for each $j = 0, 1, \ldots, 2m-1$.

A modification is needed to the form of the polynomial, however, if we want the coefficients to assume the same form as in the least squares case. In Lemma 8.12 we found that if r is not a multiple of m, then

$$\sum_{j=0}^{2m-1} (\cos rx_j)^2 = m.$$

Interpolation requires computing instead

$$\sum_{j=0}^{2m-1} (\cos mx_j)^2,$$

which (see Exercise 8) has the value $2m$. This requires the interpolatory polynomial to be written as

$$S_m(x) = \frac{a_0 + a_m \cos mx}{2} + \sum_{k=1}^{m-1} (a_k \cos kx + b_k \sin kx), \qquad (8.26)$$

if we want the form of the constants a_k and b_k to agree with those of the discrete least squares polynomial; that is,

- $a_k = \dfrac{1}{m}\displaystyle\sum_{j=0}^{2m-1} y_j \cos kx_j, \quad$ for each $k = 0, 1, \ldots, m$, and

- $b_k = \dfrac{1}{m}\displaystyle\sum_{j=0}^{2m-1} y_j \sin kx_j \quad$ for each $k = 1, 2, \ldots, m-1$.

The interpolation of large amounts of equally-spaced data by trigonometric polynomials can produce very accurate results. It is the appropriate approximation technique in areas involving digital filters, antenna field patterns, quantum mechanics, optics, and in numerous simulation problems. Until the middle of the 1960s, however, the method had not been extensively applied due to the number of arithmetic calculations required for the determination of the constants in the approximation.

The interpolation of $2m$ data points by the direct-calculation technique requires approximately $(2m)^2$ multiplications and $(2m)^2$ additions. The approximation of many thousands of data points is not unusual in areas requiring trigonometric interpolation, so the direct methods for evaluating the constants require multiplication and addition operations numbering in the millions. The roundoff error associated with this number of calculations generally dominates the approximation.

In 1965, a paper by J. W. Cooley and J. W. Tukey in the journal *Mathematics of Computation* [CT] described a different method of calculating the constants in the interpolating trigonometric polynomial. This method requires only $O(m \log_2 m)$ multiplications and $O(m \log_2 m)$ additions, provided m is chosen in an appropriate manner. For a problem with thousands of data points, this reduces the number of calculations from millions to thousands. The method had actually been discovered a number of years before the Cooley-Tukey paper appeared but had gone largely unnoticed. ([Brigh], pp. 8–9, contains a short, but interesting, historical summary of the method.)

The method described by Cooley and Tukey is known either as the **Cooley-Tukey algorithm** or the **fast Fourier transform (FFT) algorithm** and has led to a revolution in the use of interpolatory trigonometric polynomials. The method consists of organizing the problem so that the number of data points being used can be easily factored, particularly into powers of two.

Instead of directly evaluating the constants a_k and b_k, the fast Fourier transform procedure computes the complex coefficients c_k in

$$\frac{1}{m}\sum_{k=0}^{2m-1} c_k e^{ikx}, \tag{8.27}$$

where

$$c_k = \sum_{j=0}^{2m-1} y_j e^{ik\pi j/m}, \quad \text{for each } k = 0, 1, \ldots, 2m-1. \tag{8.28}$$

Leonhard Euler first gave this formula in 1748 in *Introductio in analysin infinitorum*, which made the ideas of Johann Bernoulli more precise. This work bases the calculus on the theory of elementary functions rather than curves.

Once the constants c_k have been determined, a_k and b_k can be recovered by using *Euler's Formula*,

$$e^{iz} = \cos z + i \sin z.$$

For each $k = 0, 1, \ldots, m$ we have

$$\begin{aligned}\frac{1}{m}c_k(-1)^k = \frac{1}{m}c_k e^{-i\pi k} &= \frac{1}{m}\sum_{j=0}^{2m-1} y_j e^{ik\pi j/m} e^{-i\pi k} = \frac{1}{m}\sum_{j=0}^{2m-1} y_j e^{ik(-\pi+(\pi j/m))}\\ &= \frac{1}{m}\sum_{j=0}^{2m-1} y_j \left(\cos k\left(-\pi + \frac{\pi j}{m}\right) + i\sin k\left(-\pi + \frac{\pi j}{m}\right)\right)\\ &= \frac{1}{m}\sum_{j=0}^{2m-1} y_j(\cos kx_j + i\sin kx_j).\end{aligned}$$

So, given c_k we have

$$a_k + ib_k = \frac{(-1)^k}{m}c_k. \tag{8.29}$$

For notational convenience, b_0 and b_m are added to the collection, but both are 0 and do not contribute to the resulting sum.

The operation-reduction feature of the fast Fourier transform results from calculating the coefficients c_k in clusters, and uses as a basic relation the fact that for any integer n,

$$e^{n\pi i} = \cos n\pi + i\sin n\pi = (-1)^n.$$

Suppose $m = 2^p$ for some positive integer p. For each $k = 0, 1, \ldots, m-1$ we have

$$c_k + c_{m+k} = \sum_{j=0}^{2m-1} y_j e^{ik\pi j/m} + \sum_{j=0}^{2m-1} y_j e^{i(m+k)\pi j/m} = \sum_{j=0}^{2m-1} y_j e^{ik\pi j/m}(1 + e^{\pi ij}).$$

But

$$1 + e^{i\pi j} = \begin{cases} 2, & \text{if } j \text{ is even,}\\ 0, & \text{if } j \text{ is odd,}\end{cases}$$

so there are only m nonzero terms to be summed.

If j is replaced by $2j$ in the index of the sum, we can write the sum as

$$c_k + c_{m+k} = 2\sum_{j=0}^{m-1} y_{2j} e^{ik\pi(2j)/m};$$

that is,

$$c_k + c_{m+k} = 2\sum_{j=0}^{m-1} y_{2j} e^{ik\pi j/(m/2)}. \tag{8.30}$$

In a similar manner,

$$c_k - c_{m+k} = 2e^{ik\pi/m}\sum_{j=0}^{m-1} y_{2j+1} e^{ik\pi j/(m/2)}. \tag{8.31}$$

Since c_k and c_{m+k} can both be recovered from Eqs. (8.30) and (8.31), these relations determine all the coefficients c_k. Note also that the sums in Eqs. (8.30) and (8.31) are of the same form as the sum in Eq. (8.28), except that the index m has been replaced by $m/2$.

There are $2m$ coefficients $c_0, c_1, \ldots, c_{2m-1}$ to be calculated. Using the basic formula (8.28) requires $2m$ complex multiplications per coefficient, for a total of $(2m)^2$ operations. Equation (8.30) requires m complex multiplications for each $k = 0, 1, \ldots, m-1$, and (8.31) requires $m+1$ complex multiplications for each $k = 0, 1, \ldots, m-1$. Using these equations to compute $c_0, c_1, \ldots, c_{2m-1}$ reduces the number of complex multiplications from $(2m)^2 = 4m^2$ to

$$m \cdot m + m(m+1) = 2m^2 + m.$$

The sums in (8.30) and (8.31) have the same form as the original and m is a power of 2, so the reduction technique can be reapplied to the sums in (8.30) and (8.31). Each of these is replaced by two sums from $j = 0$ to $j = (m/2) - 1$. This reduces the $2m^2$ portion of the sum to

$$2\left[\frac{m}{2} \cdot \frac{m}{2} + \frac{m}{2} \cdot \left(\frac{m}{2} + 1\right)\right] = m^2 + m.$$

So a total of

$$(m^2 + m) + m = m^2 + 2m$$

complex multiplications are now needed, instead of $(2m)^2$.

Applying the technique one more time gives us 4 sums each with $m/4$ terms and reduces the m^2 portion of this total to

$$4\left[\left(\frac{m}{4}\right)^2 + \frac{m}{4}\left(\frac{m}{4} + 1\right)\right] = \frac{m^2}{2} + m,$$

for a new total of $(m^2/2) + 3m$ complex multiplications. Repeating the process r times reduces the total number of required complex multiplications to

$$\frac{m^2}{2^{r-2}} + mr.$$

The process is complete when $r = p + 1$, because we then have $m = 2^p$ and $2m = 2^{p+1}$. As a consequence, after $r = p + 1$ reductions of this type, the number of complex multiplications is reduced from $(2m)^2$ to

$$\frac{(2^p)^2}{2^{p-1}} + m(p+1) = 2m + pm + m = 3m + m\log_2 m = O(m \log_2 m).$$

Because of the way the calculations are arranged, the number of required complex additions is comparable.

To illustrate the significance of this reduction, suppose we have $m = 2^{10} = 1024$. The direct calculation of the c_k, for $k = 0, 1, \ldots, 2m-1$, would require

$$(2m)^2 = (2048)^2 \approx 4{,}200{,}000$$

calculations. The fast Fourier transform procedure reduces the number of calculations to

$$3(1024) + 1024 \log_2 1024 \approx 13{,}300.$$

Illustration Consider the fast Fourier transform technique applied to $8 = 2^3$ data points $\{(x_j, y_j)\}_{j=0}^{7}$, where $x_j = -\pi + j\pi/4$, for each $j = 0, 1, \ldots, 7$. In this case $2m = 8$, so $m = 4 = 2^2$ and $p = 2$.

From Eq. (8.26) we have

$$S_4(x) = \frac{a_0 + a_4 \cos 4x}{2} + \sum_{k=1}^{3} (a_k \cos kx + b_k \sin kx),$$

where

$$a_k = \frac{1}{4} \sum_{j=0}^{7} y_j \cos kx_j \quad \text{and} \quad b_k = \frac{1}{4} \sum_{j=0}^{7} y_j \sin kx_j, \quad k = 0, 1, 2, 3, 4.$$

Define the Fourier transform as

$$\frac{1}{4} \sum_{j=0}^{7} c_k e^{ikx},$$

where

$$c_k = \sum_{j=0}^{7} y_j e^{ik\pi j/4}, \quad \text{for } k = 0, 1, \ldots, 7.$$

Then by Eq. (8.31), for $k = 0, 1, 2, 3, 4$, we have

$$\frac{1}{4} c_k e^{-ik\pi} = a_k + ib_k.$$

By direct calculation, the complex constants c_k are given by

$$\begin{aligned}
c_0 =& y_0 + y_1 + y_2 + y_3 + y_4 + y_5 + y_6 + y_7;\\
c_1 =& y_0 + \left(\frac{i+1}{\sqrt{2}}\right) y_1 + iy_2 + \left(\frac{i-1}{\sqrt{2}}\right) y_3 - y_4 - \left(\frac{i+1}{\sqrt{2}}\right) y_5 - iy_6 - \left(\frac{i-1}{\sqrt{2}}\right) y_7;\\
c_2 =& y_0 + iy_1 - y_2 - iy_3 + y_4 + iy_5 - y_6 - iy_7;\\
c_3 =& y_0 + \left(\frac{i-1}{\sqrt{2}}\right) y_1 - iy_2 + \left(\frac{i+1}{\sqrt{2}}\right) y_3 - y_4 - \left(\frac{i-1}{\sqrt{2}}\right) y_5 + iy_6 - \left(\frac{i+1}{\sqrt{2}}\right) y_7;\\
c_4 =& y_0 - y_1 + y_2 - y_3 + y_4 - y_5 + y_6 - y_7;\\
c_5 =& y_0 - \left(\frac{i+1}{\sqrt{2}}\right) y_1 + iy_2 - \left(\frac{i-1}{\sqrt{2}}\right) y_3 - y_4 + \left(\frac{i+1}{\sqrt{2}}\right) y_5 - iy_6 + \left(\frac{i-1}{\sqrt{2}}\right) y_7;\\
c_6 =& y_0 - iy_1 - y_2 + iy_3 + y_4 - iy_5 - y_6 + iy_7;\\
c_7 =& y_0 - \left(\frac{i-1}{\sqrt{2}}\right) y_1 - iy_2 - \left(\frac{i+1}{\sqrt{2}}\right) y_3 - y_4 + \left(\frac{i-1}{\sqrt{2}}\right) y_5 + iy_6 + \left(\frac{i+1}{\sqrt{2}}\right) y_7.
\end{aligned}$$

Because of the small size of the collection of data points, many of the coefficients of the y_j in these equations are 1 or -1. This frequency will decrease in a larger application, so to count the computational operations accurately, multiplication by 1 or -1 will be included, even though it would not be necessary in this example. With this understanding, 64 multiplications/divisions and 56 additions/subtractions are required for the direct computation of $c_0, c_1, \ldots, c_7$.

To apply the fast Fourier transform procedure with $r = 1$, we first define

$$\begin{aligned} d_0 &= \frac{c_0 + c_4}{2} = y_0 + y_2 + y_4 + y_6; & d_4 &= \frac{c_2 + c_6}{2} = y_0 - y_2 + y_4 - y_6;\\ d_1 &= \frac{c_0 - c_4}{2} = y_1 + y_3 + y_5 + y_7; & d_5 &= \frac{c_2 - c_6}{2} = i(y_1 - y_3 + y_5 - y_7);\\ d_2 &= \frac{c_1 + c_5}{2} = y_0 + iy_2 - y_4 - iy_6; & d_6 &= \frac{c_3 + c_7}{2} = y_0 - iy_2 - y_4 + iy_6;\\ d_3 &= \frac{c_1 - c_5}{2} & d_7 &= \frac{c_3 - c_7}{2}\\ &= \left(\frac{i+1}{\sqrt{2}}\right)(y_1 + iy_3 - y_5 - iy_7); & &= \left(\frac{i-1}{\sqrt{2}}\right)(y_1 - iy_3 - y_5 + iy_7). \end{aligned}$$

We then define, for $r = 2$,

$$\begin{aligned} e_0 &= \frac{d_0 + d_4}{2} = y_0 + y_4; & e_4 &= \frac{d_2 + d_6}{2} = y_0 - y_4;\\ e_1 &= \frac{d_0 - d_4}{2} = y_2 + y_6; & e_5 &= \frac{d_2 - d_6}{2} = i(y_2 - y_6);\\ e_2 &= \frac{id_1 + d_5}{2} = i(y_1 + y_5); & e_6 &= \frac{id_3 + d_7}{2} = \left(\frac{i-1}{\sqrt{2}}\right)(y_1 - y_5);\\ e_3 &= \frac{id_1 - d_5}{2} = i(y_3 + y_7); & e_7 &= \frac{id_3 - d_7}{2} = i\left(\frac{i-1}{\sqrt{2}}\right)(y_3 - y_7). \end{aligned}$$

Finally, for $r = p + 1 = 3$, we define

$$\begin{aligned} f_0 &= \frac{e_0 + e_4}{2} = y_0; & f_4 &= \frac{((i+1)/\sqrt{2})e_2 + e_6}{2} = \left(\frac{i-1}{\sqrt{2}}\right)y_1;\\ f_1 &= \frac{e_0 - e_4}{2} = y_4; & f_5 &= \frac{((i+1)/\sqrt{2})e_2 - e_6}{2} = \left(\frac{i-1}{\sqrt{2}}\right)y_5;\\ f_2 &= \frac{ie_1 + e_5}{2} = iy_2; & f_6 &= \frac{((i-1)/\sqrt{2})e_3 + e_7}{2} = \left(\frac{-i-1}{\sqrt{2}}\right)y_3;\\ f_3 &= \frac{ie_1 - e_5}{2} = iy_6; & f_7 &= \frac{((i-1)/\sqrt{2})e_3 - e_7}{2} = \left(\frac{-i-1}{\sqrt{2}}\right)y_7. \end{aligned}$$

The $c_0, \ldots, c_7$, $d_0, \ldots, d_7$, $e_0, \ldots, e_7$, and $f_0, \ldots, f_7$ are independent of the particular data points; they depend only on the fact that $m = 4$. For each m there is a unique set of constants $\{c_k\}_{k=0}^{2m-1}$, $\{d_k\}_{k=0}^{2m-1}$, $\{e_k\}_{k=0}^{2m-1}$, and $\{f_k\}_{k=0}^{2m-1}$. This portion of the work is not needed for a particular application, only the following calculations are required:

The f_k:

$$f_0 = y_0; \quad f_1 = y_4; \quad f_2 = iy_2; \quad f_3 = iy_6;$$

$$f_4 = \left(\frac{i-1}{\sqrt{2}}\right)y_1; \quad f_5 = \left(\frac{i-1}{\sqrt{2}}\right)y_5; \quad f_6 = -\left(\frac{i+1}{\sqrt{2}}\right)y_3; \quad f_7 = -\left(\frac{i+1}{\sqrt{2}}\right)y_7.$$

The e_k:

$$e_0 = f_0 + f_1;\ \ e_1 = -i(f_2 + f_3);\ \ e_2 = -\left(\frac{i-1}{\sqrt{2}}\right)(f_4 + f_5);$$

$$e_3 = -\left(\frac{i+1}{\sqrt{2}}\right)(f_6 + f_7);\ \ e_4 = f_0 - f_1;\ \ e_5 = f_2 - f_3;\ \ e_6 = f_4 - f_5;\ \ e_7 = f_6 - f_7.$$

The d_k:

$$d_0 = e_0 + e_1; \quad d_1 = -i(e_2 + e_3); \quad d_2 = e_4 + e_5; \quad d_3 = -i(e_6 + e_7);$$

$$d_4 = e_0 - e_1; \quad d_5 = e_2 - e_3; \quad d_6 = e_4 - e_5; \quad d_7 = e_6 - e_7.$$

The c_k:

$$c_0 = d_0 + d_1; \quad c_1 = d_2 + d_3; \quad c_2 = d_4 + d_5; \quad c_3 = d_6 + d_7;$$

$$c_4 = d_0 - d_1; \quad c_5 = d_2 - d_3; \quad c_6 = d_4 - d_5; \quad c_7 = d_6 - d_7.$$

Computing the constants $c_0, c_1, \ldots, c_7$ in this manner requires the number of operations shown in Table 8.13. Note again that multiplication by 1 or -1 has been included in the count, even though this does not require computational effort.

Table 8.13

Step	Multiplications/divisions	Additions/subtractions
(The f_k:)	8	0
(The e_k:)	8	8
(The d_k:)	8	8
(The c_k:)	0	8
Total	24	24

The lack of multiplications/divisions when finding the c_k reflects the fact that for any m, the coefficients $\{c_k\}_{k=0}^{2m-1}$ are computed from $\{d_k\}_{k=0}^{2m-1}$ in the same manner:

$$c_k = d_{2k} + d_{2k+1} \quad \text{and} \quad c_{k+m} = d_{2k} - d_{2k+1}, \quad \text{for } k = 0, 1, \ldots, m-1,$$

so no complex multiplication is involved.

In summary, the direct computation of the coefficients $c_0, c_1, \ldots, c_7$ requires 64 multiplications/divisions and 56 additions/subtractions. The fast Fourier transform technique reduces the computations to 24 multiplications/divisions and 24 additions/subtractions. □

Algorithm 8.3 performs the fast Fourier transform when $m = 2^p$ for some positive integer p. Modifications of the technique can be made when m takes other forms.

ALGORITHM
8.3

Fast Fourier Transform

To compute the coefficients in the summation

$$\frac{1}{m}\sum_{k=0}^{2m-1} c_k e^{ikx} = \frac{1}{m}\sum_{k=0}^{2m-1} c_k(\cos kx + i\sin kx), \quad \text{where } i = \sqrt{-1},$$

for the data $\{(x_j, y_j)\}_{j=0}^{2m-1}$ where $m = 2^p$ and $x_j = -\pi + j\pi/m$ for $j = 0, 1, \ldots, 2m-1$:

```
INPUT   $m, p; y_0, y_1, \ldots, y_{2m-1}$.

OUTPUT   complex numbers $c_0, \ldots, c_{2m-1}$; real numbers $a_0, \ldots, a_m; b_1, \ldots, b_{m-1}$.

Step 1   Set $M = m$;
             $q = p$;
             $\zeta = e^{\pi i/m}$.

Step 2   For $j = 0, 1, \ldots, 2m-1$ set $c_j = y_j$.

Step 3   For $j = 1, 2, \ldots, M$     set $\xi_j = \zeta^j$;
                                        $\xi_{j+M} = -\xi_j$.

Step 4   Set $K = 0$;
             $\xi_0 = 1$.

Step 5   For $L = 1, 2, \ldots, p+1$ do Steps 6–12.
    Step 6   While $K < 2m-1$ do Steps 7–11.
        Step 7   For $j = 1, 2, \ldots, M$ do Steps 8–10.
            Step 8   Let $K = k_p \cdot 2^p + k_{p-1} \cdot 2^{p-1} + \cdots + k_1 \cdot 2 + k_0$;
                     (*Decompose* $k$.)
                     set $K_1 = K/2^q = k_p \cdot 2^{p-q} + \cdots + k_{q+1} \cdot 2 + k_q$;
                         $K_2 = k_q \cdot 2^p + k_{q+1} \cdot 2^{p-1} + \cdots + k_p \cdot 2^q$.
            Step 9   Set $\eta = c_{K+M}\xi_{K_2}$;
                         $c_{K+M} = c_K - \eta$;
                         $c_K = c_K + \eta$.
            Step 10   Set $K = K + 1$.
        Step 11   Set $K = K + M$.
    Step 12   Set $K = 0$;
                  $M = M/2$;
                  $q = q - 1$.

Step 13   While $K < 2m-1$ do Steps 14–16.
    Step 14   Let $K = k_p \cdot 2^p + k_{p-1} \cdot 2^{p-1} + \cdots + k_1 \cdot 2 + k_0$;   (*Decompose* $k$.)
              set $j = k_0 \cdot 2^p + k_1 \cdot 2^{p-1} + \cdots + k_{p-1} \cdot 2 + k_p$.
    Step 15   If $j > K$ then interchange $c_j$ and $c_k$.
    Step 16   Set $K = K + 1$.

Step 17   Set $a_0 = c_0/m$;
              $a_m = \text{Re}(e^{-i\pi m} c_m/m)$.

Step 18   For $j = 1, \ldots, m-1$ set $a_j = \text{Re}(e^{-i\pi j} c_j/m)$;
                                       $b_j = \text{Im}(e^{-i\pi j} c_j/m)$.

Step 19   OUTPUT $(c_0, \ldots, c_{2m-1}; a_0, \ldots, a_m; b_1, \ldots, b_{m-1})$;
          STOP.
```

■

Example 1 Find the interpolating trigonometric polynomial of degree 2 on $[-\pi, \pi]$ for the data $\left\{(x_j, f(x_j))\right\}_{j=0}^{3}$, where

$$a_k = \frac{1}{2}\sum_{j=0}^{3} f(x_j)\cos(kx_j) \quad \text{for } k = 0, 1, 2 \quad \text{and} \quad b_1 = \frac{1}{2}\sum_{j=0}^{3} f(x_j)\sin(x_j).$$

Solution We have

$$a_0 = \frac{1}{2}\left(f(-\pi) + f\left(-\frac{\pi}{2}\right) + f(0) + f\left(\frac{\pi}{2}\right)\right) = -3.19559339,$$

$$a_1 = \frac{1}{2}\left(f(-\pi)\cos(-\pi) + f\left(-\frac{\pi}{2}\right)\cos\left(-\frac{\pi}{2}\right) + f(0)\cos 0 + f\left(\frac{\pi}{2}\right)\right)\cos\left(\frac{\pi}{2}\right)$$

$$= -9.86960441,$$

$$a_2 = \frac{1}{2}\left(f(-\pi)\cos(-2\pi) + f\left(-\frac{\pi}{2}\right)\cos(-\pi) + f(0)\cos 0 + f\left(\frac{\pi}{2}\right)\right)\cos(\pi)$$

$$= 4.93480220,$$

and

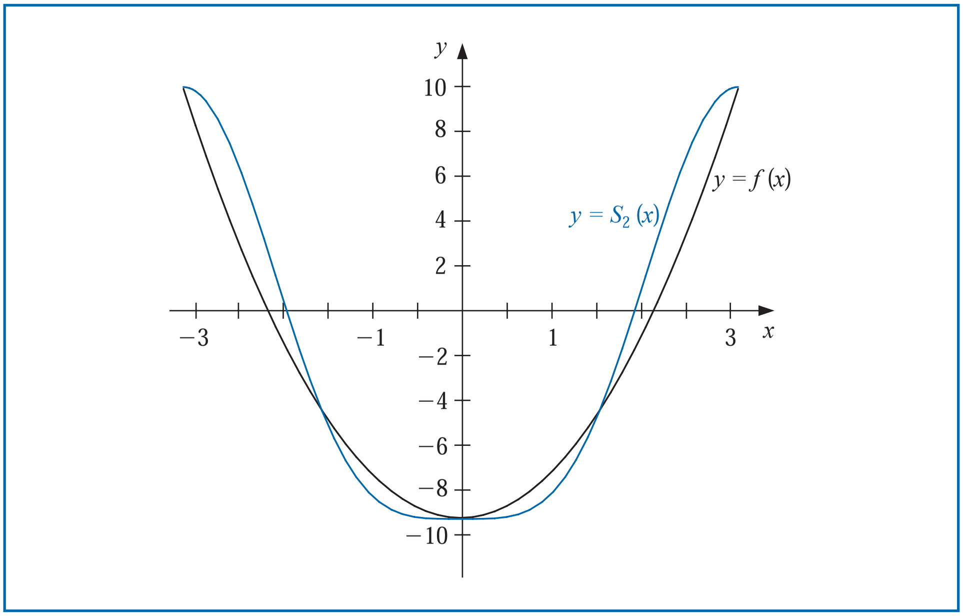
$$b_1 = \frac{1}{2}\left(f(-\pi)\sin(-\pi) + f\left(-\frac{\pi}{2}\right)\sin\left(-\frac{\pi}{2}\right) + f(0)\sin 0 + f\left(\frac{\pi}{2}\right)\sin\left(\frac{\pi}{2}\right)\right) = 0.$$

So

$$S_2(x) = \frac{1}{2}(-3.19559339 + 4.93480220\cos 2x) - 9.86960441\cos x.$$

Figure 8.16 shows $f(x)$ and the interpolating trigonometric polynomial $S_2(x)$. ■

Figure 8.16



The next example gives an illustration of finding an interpolating trigonometric polynomial for a function that is defined on a closed interval other than $[-\pi, \pi]$.

Example 2 Determine the trigonometric interpolating polynomial of degree 4 on $[0, 2]$ for the data $\{(j/4, f(j/4))\}_{j=0}^{7}$, where $f(x) = x^4 - 3x^3 + 2x^2 - \tan x(x-2)$.

Solution We first need to transform the interval $[0, 2]$ to $[-\pi, \pi]$. This is given by

$$z_j = \pi(x_j - 1),$$

so that the input data to Algorithm 8.3 are

$$\left\{z_j, f\left(1 + \frac{z_j}{\pi}\right)\right\}_{j=0}^{7}.$$

The interpolating polynomial in z is

$$S_4(z) = 0.761979 + 0.771841 \cos z + 0.0173037 \cos 2z + 0.00686304 \cos 3z$$
$$- 0.000578545 \cos 4z - 0.386374 \sin z + 0.0468750 \sin 2z - 0.0113738 \sin 3z.$$

The trigonometric polynomial $S_4(x)$ on $[0, 2]$ is obtained by substituting $z = \pi(x-1)$ into $S_4(z)$. The graphs of $y = f(x)$ and $y = S_4(x)$ are shown in Figure 8.17. Values of $f(x)$ and $S_4(x)$ are given in Table 8.14. ■

Figure 8.17

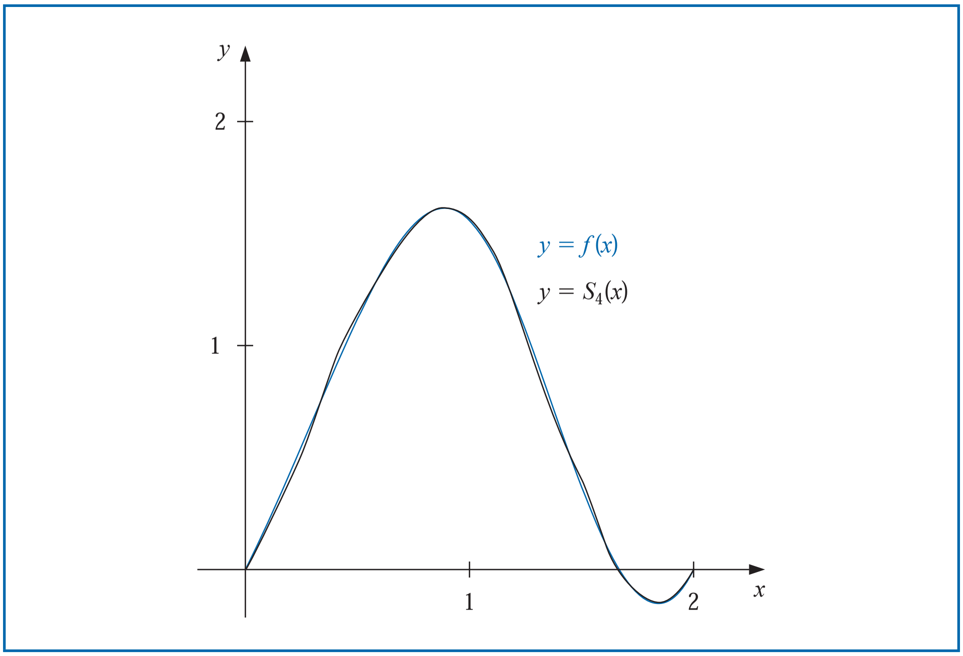


Table 8.14

x	$f(x)$	$S_4(x)$	$\lvert f(x) - S_4(x)\rvert$
0.125	0.26440	0.25001	1.44×10^{-2}
0.375	0.84081	0.84647	5.66×10^{-3}
0.625	1.36150	1.35824	3.27×10^{-3}
0.875	1.61282	1.61515	2.33×10^{-3}
1.125	1.36672	1.36471	2.02×10^{-3}
1.375	0.71697	0.71931	2.33×10^{-3}
1.625	0.07909	0.07496	4.14×10^{-3}
1.875	−0.14576	−0.13301	1.27×10^{-2}

More details on the verification of the validity of the fast Fourier transform procedure can be found in [Ham], which presents the method from a mathematical approach, or in [Brac], where the presentation is based on methods more likely to be familiar to engineers. [AHU], pp. 252–269, is a good reference for a discussion of the computational aspects of the method. Modification of the procedure for the case when m is not a power of 2 can be found in [Win]. A presentation of the techniques and related material from the point of view of applied abstract algebra is given in [Lau, pp. 438–465].

EXERCISE SET 8.6

1. Determine the trigonometric interpolating polynomial $S_2(x)$ of degree 2 on $[-\pi, \pi]$ for the following functions, and graph $f(x) - S_2(x)$:

a. $f(x) = \pi(x - \pi)$ **b.** $f(x) = x(\pi - x)$

c. $f(x) = |x|$ **d.** $f(x) = \begin{cases} -1, & -\pi \le x \le 0 \\ 1, & 0 < x \le \pi \end{cases}$

2. Determine the trigonometric interpolating polynomial of degree 4 for $f(x) = x(\pi - x)$ on the interval $[-\pi, \pi]$ using:

a. Direct calculation; **b.** The Fast Fourier Transform Algorithm.

3. Use the Fast Fourier Transform Algorithm to compute the trigonometric interpolating polynomial of degree 4 on $[-\pi, \pi]$ for the following functions.

a. $f(x) = \pi(x - \pi)$ **b.** $f(x) = |x|$

c. $f(x) = \cos \pi x - 2 \sin \pi x$ **d.** $f(x) = x \cos x^2 + e^x \cos e^x$

4. **a.** Determine the trigonometric interpolating polynomial $S_4(x)$ of degree 4 for $f(x) = x^2 \sin x$ on the interval $[0, 1]$.

b. Compute $\int_0^1 S_4(x)\, dx$.

c. Compare the integral in part (b) to $\int_0^1 x^2 \sin x\, dx$.

5. Use the approximations obtained in Exercise 3 to approximate the following integrals, and compare your results to the actual values.

a. $\displaystyle\int_{-\pi}^{\pi} \pi(x - \pi)\, dx$ **b.** $\displaystyle\int_{-\pi}^{\pi} |x|\, dx$

c. $\displaystyle\int_{-\pi}^{\pi} (\cos \pi x - 2 \sin \pi x)\, dx$ **d.** $\displaystyle\int_{-\pi}^{\pi} (x \cos x^2 + e^x \cos e^x)\, dx$

6. Use the Fast Fourier Transform Algorithm to determine the trigonometric interpolating polynomial of degree 16 for $f(x) = x^2 \cos x$ on $[-\pi, \pi]$.

7. Use the Fast Fourier Transform Algorithm to determine the trigonometric interpolating polynomial of degree 64 for $f(x) = x^2 \cos x$ on $[-\pi, \pi]$.

8. Use a trigonometric identity to show that $\sum_{j=0}^{2m-1} (\cos m x_j)^2 = 2m$.

9. Show that $c_0, \ldots, c_{2m-1}$ in Algorithm 8.3 are given by

$$\begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ \vdots \\ c_{2m-1} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \zeta & \zeta^2 & \cdots & \zeta^{2m-1} \\ 1 & \zeta^2 & \zeta^4 & \cdots & \zeta^{4m-2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \zeta^{2m-1} & \zeta^{4m-2} & \cdots & \zeta^{(2m-1)^2} \end{bmatrix} \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_{2m-1} \end{bmatrix},$$

where $\zeta = e^{\pi i/m}$.

10. In the discussion preceding Algorithm 8.3, an example for $m = 4$ was explained. Define vectors $\mathbf{c}$, $\mathbf{d}$, $\mathbf{e}$, f, and $\mathbf{y}$ as

$$\mathbf{c} = (c_0, \ldots, c_7)^t,\ \mathbf{d} = (d_0, \ldots, d_7)^t,\ \mathbf{e} = (e_0, \ldots, e_7)^t,\ \mathbf{f} = (f_0, \ldots, f_7)^t,\ \mathbf{y} = (y_0, \ldots, y_7)^t.$$

Find matrices A, B, C, and D so that $\mathbf{c} = A\mathbf{d}$, $\mathbf{d} = B\mathbf{e}$, $\mathbf{e} = C\,\mathbf{f}$, and $\mathbf{f} = D\mathbf{y}$.

8.7 Survey of Methods and Software

In this chapter we have considered approximating data and functions with elementary functions. The elementary functions used were polynomials, rational functions, and trigonometric polynomials. We considered two types of approximations, discrete and continuous. Discrete approximations arise when approximating a finite set of data with an elementary function. Continuous approximations are used when the function to be approximated is known.

Discrete least squares techniques are recommended when the function is specified by giving a set of data that may not exactly represent the function. Least squares fit of data can take the form of a linear or other polynomial approximation or even an exponential form. These approximations are computed by solving sets of normal equations, as given in Section 8.1.

If the data are periodic, a trigonometric least squares fit may be appropriate. Because of the orthonormality of the trigonometric basis functions, the least squares trigonometric approximation does not require the solution of a linear system. For large amounts of periodic data, interpolation by trigonometric polynomials is also recommended. An efficient method of computing the trigonometric interpolating polynomial is given by the fast Fourier transform.

When the function to be approximated can be evaluated at any required argument, the approximations seek to minimize an integral instead of a sum. The continuous least squares polynomial approximations were considered in Section 8.2. Efficient computation of least squares polynomials lead to orthonormal sets of polynomials, such as the Legendre and Chebyshev polynomials. Approximation by rational functions was studied in Section 8.4, where Padé approximation as a generalization of the Maclaurin polynomial and its extension to Chebyshev rational approximation were presented. Both methods allow a more uniform method of approximation than polynomials. Continuous least squares approximation by trigonometric functions was discussed in Section 8.5, especially as it relates to Fourier series.

The IMSL Library provides a number of routines for approximation including

1. Linear least squares fit of data with statistics;
2. Discrete least squares fit of data with the user's choice of basis functions;
3. Cubic spline least squares approximation;
4. Rational weighted Chebyshev approximation;
5. Fast Fourier transform fit of data.

The NAG Library provides routines that include computing

1. Least square polynomial approximation using a technique to minimize round-off error;
2. Cubic spline least squares approximation;

3. Best fit in the l_1 sense;

4. Best fit in the l_∞ sense;

5. Fast Fourier transform fit of data.

The netlib library contains a routine to compute the polynomial least squares approximation to a discrete set of points, and a routine to evaluate this polynomial and any of its derivatives at a given point.

For further information on the general theory of approximation theory see Powell [Pow], Davis [Da], or Cheney [Ch]. A good reference for methods of least squares is Lawson and Hanson [LH], and information about Fourier transforms can be found in Van Loan [Van] and in Briggs and Hanson [BH].