1

ASPECTS OF NUMERICAL ANALYSIS

INTERPOLATION AND APPROXIMATION

1.1 Interpolation

There are many situations in which one is given the value of a quantity at certain times and would like to say something about the behaviour at intermediate times. For instance, from readings of an electric meter taken at noon every day one might wish to make deductions about the consumption of electricity at 9 in the morning. This is a problem of *interpolation* in which one attempts to estimate from data at isolated points the form of a function at intervening points. The same problem arises in the use of mathematical tables when the value of a function is required at some point not listed in the table.

When all that we know about a function are the isolated data we can expect that there will be different opinions on its performance in between. Suppose we are given the values of f_1, f_2 and f_3 at $x = x_1$, x_2, and x_3 respectively (see Fig. 1.1). Then, a simple rule would be to join successive values by straight lines and use these lines to tell us the value of f in between. This is an approximation f_0 in which

$$f_0(x) = \frac{x_2 - x}{x_2 - x_1} f_1 + \frac{x - x_1}{x_2 - x_1} f_2 \qquad (x_1 \leqslant x \leqslant x_2)$$

$$= \frac{x_3 - x}{x_3 - x_2} f_2 + \frac{x - x_2}{x_3 - x_2} f_3 \qquad (x_2 \leqslant x \leqslant x_3)$$

and, in general, if f_n is the value at x_n

$$f_0(x) = \frac{x_{n+1} - x}{x_{n+1} - x_n} f_n + \frac{x - x_n}{x_{n+1} - x_n} f_{n+1} \qquad (x_n \leqslant x \leqslant x_{n+1}). \tag{1.1}$$

Of course, some people will say that they are not willing to accept this approximation because the derivatives are not continuous across the data points but, for the moment, let us note that by combining the formulae (1.1) we can obtain the approximation

$$f_0(x) = \sum_{m=1}^{n} B_m(x) f_m \qquad (x_1 \leqslant x \leqslant x_n) \tag{1.2}$$

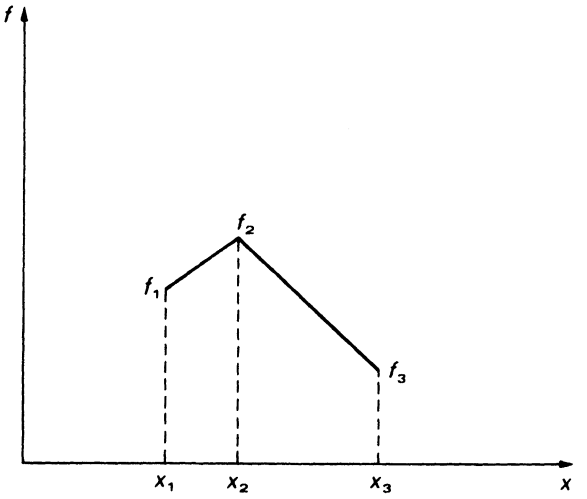

Fig. 1.1. Linear interpolation.

where

$$B_1(x) = \frac{x_2 - x}{x_2 - x_1} \qquad (x_1 \leqslant x \leqslant x_2)$$

$$= 0 \qquad (x_2 \leqslant x \leqslant x_n),$$

$$B_n(x) = \frac{x - x_{n-1}}{x_n - x_{n-1}} \qquad (x_{n-1} \leqslant x \leqslant x_n)$$

$$= 0 \qquad (x_1 \leqslant x \leqslant x_{n-1})$$

and, if $m \neq 1$ or n,

$$B_m(x) = 0 \qquad (x_1 \leqslant x \leqslant x_{m-1})$$

$$= \frac{x - x_{m-1}}{x_m - x_{m-1}} \qquad (x_{m-1} \leqslant x \leqslant x_m)$$

$$= \frac{x_{m+1} - x}{x_{m+1} - x_m} \qquad (x_m \leqslant x \leqslant x_{m+1})$$

$$= 0 \qquad (x_{m+1} \leqslant x \leqslant x_n).$$

Each of $B_1, \ldots, B_n$ vanishes outside a finite interval and has the shape of either a half triangle or full triangle (Fig. 1.2). For this reason the functions $B_1, \ldots, B_n$ are known as *triangle* or *pyramid functions.* So we could call (1.2) an approximation to our function in terms of pyramid functions. It will be necessary to consider more complicated expansions in order to meet some of the conditions encountered.

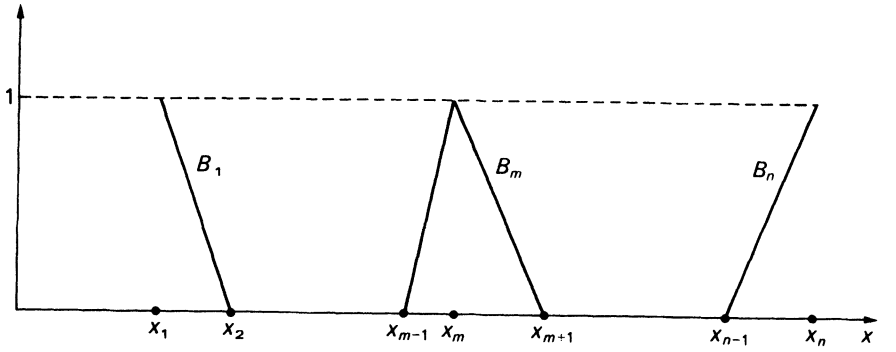

Fig. 1.2. Pyramid functions.

Suppose, now, that we are given the additional data of the values of the derivative of f, say $f'_1, f'_2, \ldots$ at $x = x_1, x_2, \ldots$. It is immediately obvious that the derivatives of f_0 will not agree with the derivatives of f except in rare circumstances. If we are to remedy this we need an approximation between x_i and x_{i+1} which gives the correct derivatives and must therefore satisfy two extra conditions. So our straight lines must be replaced by cubics, if we stick with powers of x for our approximations. Let us try

$$y = a(x - x_i)^3 + b(x - x_i)^2 + c(x - x_i) + d.$$

Then since $y = f_i$ and $y' = f'_i$ when $x = x_i$ we see that $d = f_i$ and $c = f'_i$. The conditions $y = f_{i+1}$, $y' = f'_{i+1}$ at $x = x_{i+1}$ then imply that

$$a(x_{i+1} - x_i)^3 + b(x_{i+1} - x_i)^2 + (x_{i+1} - x_i)f'_i + f_i = f_{i+1},$$
$$3a(x_{i+1} - x_i)^2 + 2b(x_{i+1} - x_i) + f'_i = f'_{i+1}.$$

From these can be deduced

$$a(x_{i+1} - x_i)^3 = (f'_{i+1} + f'_i)(x_{i+1} - x_i) - 2(f_{i+1} - f_i),$$
$$b(x_{i+1} - x_i)^2 = 3(f_{i+1} - f_i) - (f'_{i+1} + 2f'_i)(x_{i+1} - x_i).$$

Therefore our approximation between x_i and x_{i+1} can be expressed as

$$y = \alpha_i(x)f_i + \beta_i(x)f_{i+1} + \gamma_i(x)f'_i + \delta_i(x)f'_{i+1}$$

where

$$\alpha_i(x) = \frac{(x_{i+1} - x)^2}{(x_{i+1} - x_i)^3}\{(x_{i+1} - x_i) + 2(x - x_i)\},$$

$$\beta_i(x) = \frac{(x - x_i)^2}{(x_{i+1} - x_i)^3}\{(x_{i+1} - x_i) + 2(x_{i+1} - x)\},$$

$$\gamma_i(x) = \frac{(x_{i+1} - x)^2(x - x_i)}{(x_{i+1} - x_i)^2},$$

$$\delta_i(x) = \frac{(x - x_i)^2(x - x_{i+1})}{(x_{i+1} - x_i)^2}.$$

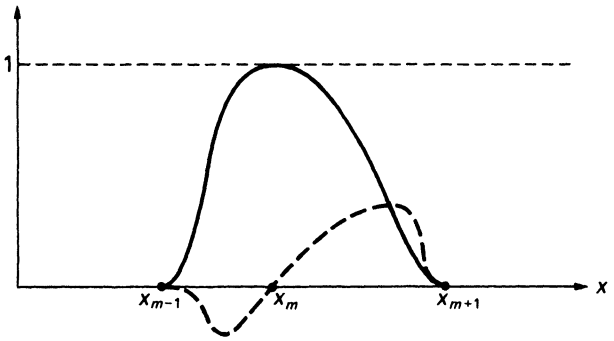

Fig. 1.3. Cubic basis functions: solid curve, $B_m^{(1)}$; broken curve, $B_m^{(2)}$.

By means of these formulae we can construct our approximation over the whole interval as

$$f_0(x) = \sum_{m=1}^{n} \{B_m^{(1)}(x) f_m + B_m^{(2)}(x) f'_m\} \qquad (x_1 \leqslant x \leqslant x_n) \tag{1.3}$$

where

$$\begin{aligned} B_m^{(1)}(x) &= 0 && (x_1 \leqslant x \leqslant x_{m-1}) \\ &= \beta_{m-1}(x) && (x_{m-1} \leqslant x \leqslant x_m) \\ &= \alpha_m(x) && (x_m \leqslant x \leqslant x_{m+1}) \\ &= 0 && (x_{m+1} \leqslant x \leqslant x_n) \end{aligned}$$

and $B_m^{(2)}$ is the same with γ_m and δ_m taking the place of α_m and β_m respectively. These formulae do not hold for $m = 1$ and $m = n$; the necessary modifications are easy to carry out and are left to the reader. The behaviour of $B_m^{(1)}$ and $B_m^{(2)}$ when m is neither 1 nor n is shown in Fig. 1.3.

In both (1.2) and (1.3) the *interpolant* f_0 consists of a series, each term of which is the product of a given value such as f_m or f'_m and a function of x such as B_m or $B_m^{(2)}$. The given values occur only in the coefficients, the functions of x depending only on the points which are selected for observation and not on the values found there. For this reason the functions B_m, $B_m^{(1)}$, and $B_m^{(2)}$ are known as *basis functions*. In the following whenever we have an expansion which has the form of (1.2) or (1.3) we shall call the corresponding B_s basis functions whether or not they are polynomials.

So far we have discussed the two cases in which the B_s are linear and cubic polynomials respectively in the interval (x_{m-1}, x_{m+1}) and zero outside. These are obviously particular instances of the more general situation in which B is a polynomial of degree $2q - 1$ in the interval and zero outside. The general case is known as *piecewise Hermite interpolation*, the adjective piecewise being incorporated to indicate that once we have partitioned our interval at the points

$x_1, x_2, \ldots$ the basis function is required to be zero on all of the sub-intervals except one or two.

Suppose that we are given a function which, together with its first r derivatives, is continuous for $x_1 \leqslant x \leqslant x_n$. Such a function will be signified by writing $f \in C^r[x_1, x_n]$; sometimes C^0 will be denoted by C. Also the brackets will be dropped if there is no ambiguity about which interval is being referred to.

Now, if we have a polynomial of degree $2q - 1$ it will contain $2q$ coefficients which we can adjust. Consequently we can make it satisfy q conditions at $x = x_i$ and q conditions at $x = x_{i+1}$. Thus, if $f \in C^{q-1}[x_i, x_{i+1}]$ we can ask that the polynomials $p_{2q-1}(x)$ satisfy

$$\frac{\mathrm{d}^k f}{\mathrm{d}x^k} = \frac{\mathrm{d}^k p_{2q-1}(x)}{\mathrm{d}x^k}$$

at both $x = x_i$ and $x = x_{i+1}$ for $k = 0, 1, \ldots, q - 1$. In this way we construct a piecewise Hermite interpolant which agrees with a function and its first $q - 1$ derivatives at the points of observation. The corresponding basis functions can be deduced as in the cases $q = 1$ and $q = 2$ which we have already discussed.

Of course, even if f or one of its derivatives is not continuous between the points of observation we can use the same interpolant so long as there is continuity near the points of observation. This is an example of approximating a discontinuity by something continuous. Whether it is valuable or not will depend upon the circumstances.

One plain disadvantage of this type of interpolation when $q > 1$ is its involvement of the derivatives of f and the steadily increasing complexity of the equations to be solved as q grows. One way of avoiding the derivatives of f is to ask that the derivative of the interpolant be continuous at the points $x_1, x_2, \ldots, x_{n-1}$ but not to impose the additional restriction that it has the same value as the derivative of f. So we can reduce the order of the polynomial to 2 and try

$$y = a_i(x - x_i)^2 + b_i(x - x_i) + c_i.$$

To satisfy $y = f_i$ at $x = x_i$ and $y = f_{i+1}$ at $x = x_{i+1}$ we need $c_i = f_i$ and

$$a_i(x_{i+1} - x_i)^2 + b_i(x_{i+1} - x_i) = f_{i+1} - f_i.$$

If we substitute for b_i from this relation we obtain

$$y = a_i(x - x_i)(x - x_{i+1}) + f_i + (f_{i+1} - f_i)(x - x_i)/(x_{i+1} - x_i). \quad (1.4)$$

The constant a_i is at our disposal but must be such that the derivative of y is the same as x approaches x_i from above or below. Hence

$$a_i(x_i - x_{i+1}) + \frac{f_{i+1} - f_i}{x_{i+1} - x_i} = a_{i-1}(x_i - x_{i-1}) + \frac{f_i - f_{i-1}}{x_i - x_{i-1}}. \quad (1.5)$$

If $x_{i+1} - x_i = x_i - x_{i-1} = h$ this simplifies to

$$a_{i-1} + a_i = \frac{1}{h^2}(f_{i+1} - 2f_i + f_{i-1}). \tag{1.6}$$

These equations hold at the $n-2$ points $x_2, \ldots, x_{n-1}$. Since there are $n-1$ coefficients a_i it follows that one can be chosen arbitrarily and then the remainder are known from (1.5) and (1.6) as appropriate. It will be noticed that the second derivative of y is $2a_i$ so that choosing one of the a_i is equivalent to specifying the second derivative of the interpolant in a sub-interval.

An approximation of the form (1.4) subject to (1.5) or (1.6) is known as a *quadratic spline* and $x_1, \ldots, x_n$ are known as its *nodes* or *nodal points* or *knots*.

The quadratic is the simplest of the splines. If we demand that the first and second derivative be continuous at the internal nodal points we are led to a *cubic spline*. It is easiest to work with the second derivative of the spline. Since it will be a linear function we can ensure its continuity by adopting the form (1.1) i.e.

$$\frac{\mathrm{d}^2 y}{\mathrm{d}x^2} = b_i \frac{x_{i+1} - x}{x_{i+1} - x_i} + b_{i+1} \frac{x - x_i}{x_{i+1} - x_i}$$

for each of the intervals (x_i, x_{i+1}). The coefficients b_i will then be values of the second derivative of the spline at the nodal points. For simplicity, it will now be assumed that the nodal points are equally spaced so that $x_{i+1} - x_i = h$ for $i = 1, \ldots, n-1$. Then an integration gives

$$\frac{\mathrm{d}y}{\mathrm{d}x} = -\frac{1}{2}\frac{b_i}{h}(x_{i+1} - x)^2 + \frac{1}{2}\frac{b_{i+1}}{h}(x - x_i)^2 + c_i$$

and

$$y = \frac{1}{6}\frac{b_i}{h}(x_{i+1} - x)^3 + \frac{1}{6}\frac{b_{i+1}}{h}(x - x_i)^3 + c_i(x - x_i) + d_i.$$

To make $\mathrm{d}y/\mathrm{d}x$ continuous at $x = x_i$ we must have

$$-\tfrac{1}{2}b_i h + c_i = \tfrac{1}{2}b_i h + c_{i-1}$$

while $y = f_i, f_{i+1}$ at $x = x_i, x_{i+1}$ necessitate

$$f_i = \tfrac{1}{6}b_i h^2 + d_i,$$

$$f_{i+1} = \tfrac{1}{6}b_{i+1}h^2 + c_i h + d_i.$$

From the last two equations we deduce that the cubic spline can be written as

$$y = \frac{1}{6}\frac{b_i}{h}(x_{i+1} - x)^3 + \frac{1}{6}\frac{b_{i+1}}{h}(x - x_i)^3 + \left(\frac{f_i}{h} - \frac{hb_i}{6}\right)(x_{i+1} - x)$$
$$+ \left(\frac{f_{i+1}}{h} - \frac{hb_{i+1}}{6}\right)(x - x_i) \tag{1.7}$$

provided that

$$b_{i+1} + 4b_i + b_{i-1} = \frac{6}{h^2}(f_{i+1} - 2f_i + f_{i-1}) \tag{1.8}$$

for $i = 2, \ldots, n-1$. There are now n coefficients available so that two can be selected arbitrarily and the rest are then determined by (1.8). Often, the choice $b_1 = b_n = 0$ is made.

These formulae can be combined so as to express the interpolant in terms of basis functions. However, it is more convenient to proceed in a different way. Let $S_i(x)$ denote the spline in (x_i, x_{i+1}). Then S_i'' and S_{i-1}'' must agree at $x = x_i$ so that

$$S_i'' = S_{i-1}'' + 6\beta_i(x - x_i)/(x_{i+1} - x_i)^3. \tag{1.9}$$

where β_i has to be found. Define the function x_+ by

$$\begin{aligned} x_+ &= x \qquad (x > 0) \\ &= 0 \qquad (x \leqslant 0). \end{aligned}$$

Thus $[(3)_+\}^3 = 27$, $\{(-3)_+\}^3 = 0$ whereas $(t-5)_+ = t - 5$ if $t > 5$ but 0 if $t \leqslant 5$. Then applying (1.9) for $i = 2, \ldots, n$ and using an equally spaced partition with $x_{i+1} = ih$ we see that

$$S'' = 2\beta_0/h^2 + 6\beta_1 x/h^3 + \sum_{i=2}^{n} 6\beta_i\{x - (i-1)h\}_+/h^3 \qquad (0 \leqslant x \leqslant nh)$$

where the first two terms represent S_1''. After two integrations we obtain

$$S = \alpha_0 + \alpha_1(x/h) + \beta_0(x/h)^2 + \beta_1(x/h)^3 + \sum_{i=2}^{n} \beta_i\{x - (i-1)h\}_+^3/h^3. \tag{1.10}$$

By construction S and its first two derivatives are continuous: we make it take the value f_i at $x = ih$ by requiring that

$$\alpha_0 = f_0,$$

$$\alpha_1 + \beta_0 + \beta_1 = f_1 - f_0,$$

$$\alpha_0 + \alpha_1 m + \beta_0 m^2 + \beta_1 m^3 + \sum_{i=2}^{m} \beta_i(m - i + 1)^3 = f_m \qquad (n \geqslant m \geqslant 2). \tag{1.11}$$

Let us use the *central difference* operator δ, defined so that

$$\delta f(x) = f(x + \tfrac{1}{2}h) - f(x - \tfrac{1}{2}h).$$

Then $\delta f(\frac{1}{2}h) = f(h) - f(0)$ or $\delta f_{1/2} = f_1 - f_0$. Similarly

$$\delta^2 f_m = f_{m+1} - 2f_m + f_{m-1}.$$

Our equations can now be written as

$$\alpha_0 = f_0, \quad \alpha_1 + \beta_0 + \beta_1 = \delta f_{1/2}, \quad 2\beta_0 + 6\beta_1 + \beta_2 = \delta^2 f_1$$
$$6\beta_1 + 5\beta_2 + \beta_3 = \delta^3 f_{3/2}, \quad \beta_{m+2} + 4\beta_{m+1} + \beta_m = \delta^4 f_m$$
$$(m = 2, \ldots, n-2) \quad (1.12)$$

which are $(n+1)$ equations governing the $(n+3)$ coefficients $\alpha_0, \alpha_1, \beta_0, \ldots, \beta_n$. Two of these coefficients may be chosen arbitrarily and then the others found from (1.11) or (1.12).

Once the eqns (1.11) or (1.12) have been solved the coefficients in (1.10) are linear combinations of the values f_i of f at $x = ih$. Accordingly, (1.10) can be rewritten in the form

$$S = \sum_{i=0}^{n} f_i C_i(x) \tag{1.13}$$

where the polynomials $C_i(x)$ can be determined. Clearly $C_i(jh) = 0$ $(j \neq i)$ and $C_i(ih) = 1$ for $i, j = 0, 1, \ldots, n$. The functions $C_i(x)$ are known as *cardinal splines*. They can be regarded as basic functions for (1.13) but they are not satisfactory for many practical applications because they are non-zero over most of the interval.

To overcome this difficulty cubic splines which vanish identically outside an interval of length $4h$ have been constructed. Consider the function B_i^s defined by

$$B_i^s(x) = \tfrac{1}{4}[(x-i+2)_+^3 - 4(x-i+1)_+^3 + 6(x-i)_+^3 - 4(x-i-1)_+^3 + (x-i-2)_+^3]. \tag{1.14}$$

Notice firstly that B_i^s vanishes identically for $x \leqslant i-2$ and is also identically zero for $x \geqslant i+2$. Also, since the first two derivatives of x_+^3 are continuous, the first two derivatives of B_i^s are continuous and, in addition, vanish identically for $x \leqslant i-2$ and $x \geqslant i+2$. Thus the B_i^s are splines which are non-zero only for the interval $i-2 < x < i+2$; they are known as *cubic B-splines* and each forms a bell-shaped curve.

Special consideration may have to be given to the B-splines to be used at the ends of intervals. Often one will wish them to be lop-sided in order not to stray outside the given interval; sometimes taking half a bell is satisfactory. (There is additional information about B-splines in §6.8.)

One reason why splines may be preferred to the polynomial approximations described earlier in this section is that the latter are subject to the *Runge phenomenon*. If one is given a function and, in a definite interval, one seeks to improve the approximation by increasing the number n of points where the given function and approximant agree, one finds that, although the separation between the points of agreement decreases, the maximum difference between the given function and approximant increases and, in fact, becomes infinite as $n \to \infty$ if the length of the interval exceeds a certain quantity. By using different

polynomials in adjacent intervals as when splines are employed this difficulty can be overcome.

It is, of course, possible once the splines have been constructed with specified knots to ask that the given function be matched not at the knots but at some data points chosen in some convenient way. For quadratic splines the error between the given function and approximant tends to have a ripple on it when the data points coincide with the knots. If, however, the data points are midway between the knots the ripples die away, effectively by a factor of 6, as can be seen from the parabolic shape of cardinal spines. (For further information on splines see Ahlberg, Nilson, and Walsh (1967). Extensive tables of coefficients are given by Sard and Weintraub (1971).)

1.2 Inverse interpolation

Frequently, the problem of determining where a function takes a specified value is met. In other words, given y find an approximate value of x such that $f(x) = y$ when f is known only for certain values of x, perhaps corresponding to entries in a table. One method is to construct an interpolating polynomial $p(x)$ and then solve

$$p(x) = y \tag{1.15}$$

This is known as *inverse interpolation.*

Inverse linear interpolation occurs when $p(x)$ is chosen to be linear. In this case, the table is first inspected and two consecutive entries x_1 and x_2 are determined between which x must lie. Then define

$$p(x) = \{(x_2 - x)f(x_1) + (x - x_1)f(x_2)\}/(x_2 - x_1)$$

and the solution of (1.15) is

$$x = [\{f(x_2) - y\}x_1 + \{y - f(x_1)\}x_2]/\{f(x_2) - f(x_1)\}.$$

If $p(x)$ is not chosen to be linear then more complicated methods must be used to solve (1.15). Examples are Muller's method, the secant method, the method of false position and the method of bisection described in §1.8.

An alternative way, if the function inverse to f is known, is to carry out interpolation on the inverse function. In general, this will be less reliable than inverse interpolation on f because, although a polynomial may well be a good approximation to f, there is no guarantee that the inverse function can be represented equally well by a polynomial. For example, if $f(x) = x^2$ the inverse function $x = \sqrt{y}$ does not have a good representation as a polynomial near the origin $x = 0$, $y = 0$.

1.3 Interpolation in two dimensions

The problem of interpolation in two or more dimensions is much more complicated than for one variable. In part, this is due to the fact that functions

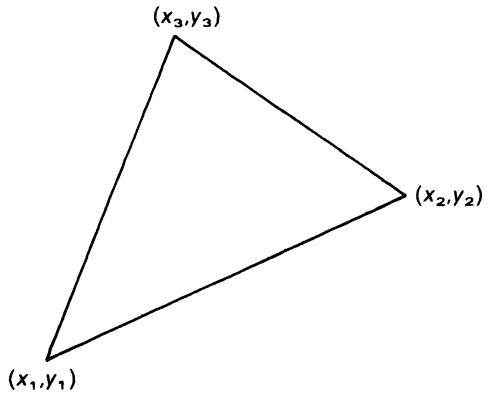

Fig. 1.4. Triangular interpolation.

may be specified on domains of highly irregular shape. It is usually assumed that any shape likely to arise in practice can be approximated to as high a degree of accuracy as required by a network of standard shapes, e.g. triangles or rectangles, provided that they are made sufficiently small. Therefore we restrict our attention to such shapes.

Suppose that we want an approximation F to $f(x, y)$ over the triangle shown in Fig. 1.4 and suppose that F has the form

$$F(x, y) = \alpha + \beta x + \gamma y$$

i.e. we make a *linear* approximation. If we impose the condition that F and f are to agree at the three vertices we discover that

$$F(x, y) = \alpha_1 f(x_1, y_1) + \alpha_2 f(x_2, y_2) + \alpha_3 f(x_3, y_3)$$

where

$$A\alpha_1 = x_2y_3 - x_3y_2 + (y_2 - y_3)x - (x_2 - x_3)y,$$

$$A\alpha_2 = x_3y_1 - x_1y_3 + (y_3 - y_1)x - (x_3 - x_1)y,$$

$$A\alpha_3 = x_1y_2 - x_2y_1 + (y_1 - y_2)x - (x_1 - x_2)y$$

and A, twice the area of the triangle, is given by

$$A = (x_2 - x_1)(y_3 - y_1) - (x_3 - x_1)(y_2 - y_1).$$

Take another triangle with vertices (x_1, y_1), (x_2, y_2) and (x_4, y_4) which does not overlap that of Fig. 1.4 and find a similar linear approximation F_1 to f over this triangle. Then, since both F and F_1 vary linearly along the side joining (x_1, y_1) and (x_2, y_2), and have the same values at the two vertices, they must be equal at every point of the side. In other words, F and F_1 are continuous across the common side. In this way, by selecting non-overlapping triangles to cover the region of interest, we obtain a linear approximant which is continuous throughout the region.

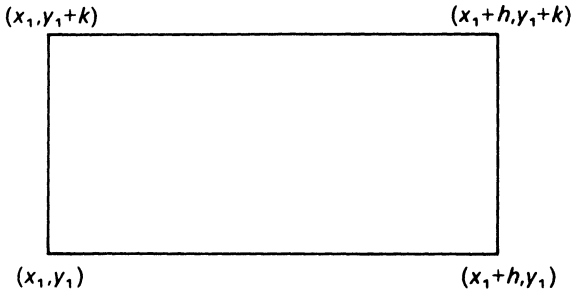

Fig. 1.5. Interpolation on a rectangle.

If rectangular elements are employed (Fig. 1.5) we can try the approximation $F(x, y) = \alpha + \beta x + \gamma y + \delta xy$. If we require that $F = f$ at the four vertices, we have

$$F(x, y) = \alpha_1 + \beta_1(x - x_1) + \gamma_1(y - y_1) + \delta_1(x - x_1)(y - y_1)$$

where

$$\alpha_1 = f(x_1, y_1), \quad \beta_1 = \{f(x_1 + h, y_1) - f(x_1, y_1)\}/h,$$

$$\gamma_1 = \{f(x_1, y_1 + k) - f(x_1, y_1)\}/k,$$

$$\delta_1 = \{f(x_1 + h, y_1 + k) - f(x_1 + h, y_1) - f(x_1, y_1 + k) + f(x_1, y_1)\}/hk.$$

For fixed y, F is a linear function of x and, for fixed x, a linear function of y. Consequently, F is known as a *bilinear* interpolant. On any side F depends only on the values at the two vertices so that, for two non-overlapping rectangles with a common side, the two bilinear interpolants take the same value on the common side. Thus bilinear interpolants yield a continuous approximant over the region covered by non-overlapping rectangles.

Exercises

1. The function $f(x)$ has the values shown

x	$f(x)$
0.1	1.10517
0.2	1.22140
0.3	1.34986
0.4	1.49182

Using linear interpolation determine an approximate value for $f(0.26)$.

2. *If* $f(x) = 3x^2 - 1$ find a piecewise linear interpolant which agrees with it at $x = 0$, 0.1, 0.2, 0.3, 0.4, 0.5. What approximation to $f(0.33)$ does it give?

3. If $f(x_i)$ and $f(x_{i+1})$ are increased by the small quantities ε_1 and ε_2 respectively, what is the change to the value of the linear interpolant for $f\{\frac{1}{2}(x_i + x_{i+1})\}$?
4. If the approximation F is linear on $[a, b]$ and agrees with f at the end-points, show that there is some c satisfying $a < c < b$ such that $f(x) - F(x) = \frac{1}{2}(x - a)(x - b)f''(c)$ if $f \in C^1[a, b]$ and f'' exists. What accuracy does this suggest for linear interpolation in a table of (i) $\sin x$, (ii) $\ln x$ when x is given at intervals of 0.01 between 1 and 2, while f is given to 5 decimal places?
5. Find a polynomial $P(x)$ of degree 2 or less such that $P(1) = 1, P(2) = 1, P'(1) = 1$.
6. Show that there is no polynomial $P(x)$ of degree 2 or less such that $P(x) = a$, $P(x + h) = b$, $P'(x + \frac{1}{2}h) \neq \frac{1}{2}(b - a)$.
7. For each of the functions (a) $\sin \frac{1}{2}\pi x$, (b) $\tan^{-1} x$, (c) $(1 + x^2)^{-1}$ determine a single polynomial and a cubic spline approximation which agrees over $-1 \leqslant x \leqslant 1$ at points separated by (i) 0.5. (ii) 0.25, (iii) 0.1, (iv) 0.01. Draw graphs of the original functions and their interpolants.
8. For the function of $Q.1$ find x_0 such that $f(x_0) = 1.3$.
9. Show that, for linear interpolation on a triangle, $\alpha_1 + \alpha_2 + \alpha_3 = 1$.
10. Prove that, in bilinear interpolation on a unit square, the basis function at an internal node is given by

$$B_{jk}(x, y) = \alpha_j(mx)\alpha_k(my) \qquad (1 \leqslant j, k \leqslant m - 1)$$

where

$$\begin{aligned} \alpha_j(x) &= x - j + 1 \qquad (j - 1 \leqslant x \leqslant j) \\ &= j + 1 - x \qquad (j \leqslant x \leqslant j + 1) \end{aligned}$$

and is zero elsewhere.

11. If

$$F(x, y) = \sum_{r=0}^{3} \sum_{s=0}^{3} \alpha_{rs} x^r y^s$$

express the coefficients α_{rs} in terms of the values of F, $\partial F/\partial x$, $\partial F/\partial y$, $\partial^2 F/\partial x\, \partial y$ at (0, 0), (0, 1), (1, 0), and (1, 1).

12. If $f(1, 11) = 1$, $f(3, 1) = 4$, $f(1, 2) = 5$, $f(3, 2) = 7$ find the approximate value of $f(\frac{3}{2}, \frac{5}{4})$ by (a) triangular interpolation over (1, 1), (3, 1), (1, 2), (b) bilinear interpolation, (c) interpolation over a triangle formed from a side and two diagonals.

1.4 Approximation

How do we know when an interpolant is a good approximation to a function? In a sense this question has no answer because what is regarded as good by one person will be deemed unsatisfactory by another. Nevertheless, certain measures of error have been introduced and once a particular measure has been adopted we have decided on a criterion which determines whether some errors are better than others.

One measure of the difference between two functions f and F over an interval $[a, b]$ is provided by

$$\sup_{a \leqslant x \leqslant b} |f(x) - F(x)|.$$

This is known as the *maximum* or *uniform norm* and measures the maximum

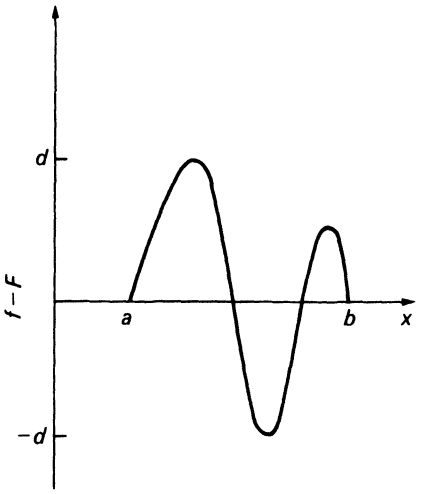

Fig. 1.6. A possible deviation in approximation.

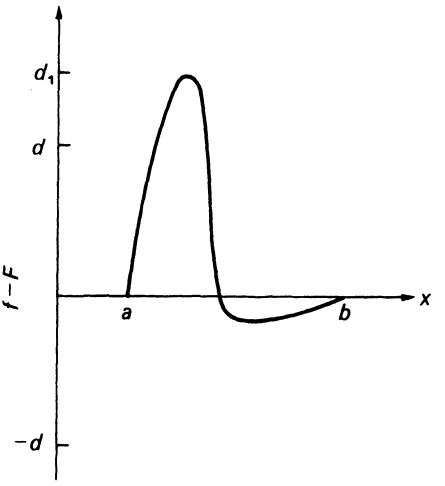

Fig. 1.7 Comparison of norms.

deviation that occurs between the two functions. Another measure which is often used is

$$\left[\int_a^b \{f(x) - F(x)\}^2 \, dx\right]^{1/2}.$$

It is known as the L_2 or *least squares norm*. The L_2-norm estimates the total deviation of f from F over the whole interval. In Fig. 1.6 the maximum norm has value d whereas, in Fig. 1.7, it has the greater value d_1. Therefore, if these figures represent different approximants F to the same f, Fig. 1.6 will be considered to be better than Fig. 1.7 as far as the maximum norm is concerned. On the other hand, the L_2-norm is larger in Fig. 1.6 than in Fig. 1.7 so that Fig. 1.7 will be preferred on the basis of the L_2-norm.

The maximum norm is the natural one if one wishes to be within an assigned accuracy at every point of the interval. In general, there is little virtue in arranging high accuracy throughout most of the interval with only moderate accuracy elsewhere. It is better to have the difference $f - F$ small over the whole interval and making small oscillations through positive and negative values.

For the maximum norm there are two theorems related to approximation and which will be quoted without proof.

THEOREM 1.4 (WEIERSTRASS). *If $f \in C[a, b]$ then, given any $\varepsilon > 0$, there is a polynomial $p_n(x)$ such that*

$$|p_n(x) - f(x)| \leqslant \varepsilon$$

for $x \in [a, b]$.

THEOREM 1.4*a*. *If $f \in C[a, b]$ and n is a given integer, there is a unique polynomial p_n of degree n or less such that*

$$\sup_{a \leqslant x \leqslant b} |p_n(x) - f(x)| \leqslant \sup_{a \leqslant x \leqslant b} |Q_n(x) - f(x)|$$

for every polynomial Q_n of degree n or less. The sup on the left is attained at $n + 2$ points at least.

There is no algorithm for calculating p_n in Theorem 1.4*a* in a finite number of stages. If, however, we only impose the condition at a finite number of points then we can construct an algorithm often known as the *first algorithm of Remes.* Let us denote the set of points by S and select from them $n + 2$ points $x_0, x_1, \ldots, x_{n+1}$ such that $x_0 < x_1 < \cdots < x_{n+1}$.

Define

$$\lambda_i = \prod_{j=0}^{n+1}{}' (x_i - x_j)^{-1} \tag{1.16}$$

where the prime means omit $j = i$ from the product, and then put

$$\eta \sum_{i=0}^{n+1} (-)^i \lambda_i = -\sum_{i=0}^{n+1} \lambda_i f(x_i). \tag{1.17}$$

Construct

$$p_n(x) = \sum_{i=0}^{n} \left\{ \prod_{j=0}^{n}{}' \frac{x - x_j}{x_i - x_j} \right\} \{ f(x_i) + (-)^i \eta \}.$$

Then

$$p_n(x_i) = f(x_i) + (-)^i \eta \tag{1.18}$$

for $i = 0, \ldots, n$. Also

$$\begin{aligned} p_n(x_{n+1}) &= -\sum_{i=0}^{n} \lambda_i \{ f(x_i) + (-)^i \eta \} / \lambda_{n+1} \\ &= f(x_{n+1}) + (-)^{n+1} \eta \end{aligned}$$

from (1.17). Thus (1.18) holds for $i = n + 1$ as well and we have ensured that, at $n + 2$ points, p_n does not differ from f by more than $|\eta|$.

Now check the other points of S. If the difference at them does not exceed $|\eta|$, then p_n is the required polynomial. Otherwise find the point x' of S when $|p_n - f|$ is a maximum. If $x_i \leqslant x' \leqslant x_{i+1}$ $(i = 0, \ldots, n)$ replace x_i by x' if $\{p_n(x') - f(x')\}\{p_n(x_i) - f(x_i)\} > 0$; otherwise replace x_{i+1} by x'. If $x' < x_0$ put x' for x_0 if $\{p_n(x') - f(x')\}\{p_n(x_0) - f(x_0)\} > 0$; otherwise replace x_{n+1} by x'. Operate similarly if $x' > x_{n+1}$. Return now to (1.16) and repeat the calculation with the new set of points. Proceeding in this way we shall, after a finite number of steps (since there is a finite number of selections of $n + 2$ points), reach a polynomial p_n for which the inequality of Theorem 1.4*a* is valid at all points of the set S.

Acceleration of the convergence may sometimes be achieved by the *second*

algorithm of Remes. Since $p_n - f$ changes sign in each of the intervals $[x_0, x_1], [x_1, x_2], \ldots, [x_n, x_{n+1}]$ it has at least one zero in each interval. Let y_i be a typical zero in $[x_i, x_{i+1}]$. In each of the intervals $[a, y_0], [y_0, y_1], \ldots, [y_n, b]$ find a value of x, say z_i, where $p_n(z_i) - f(z_i)$ is an extremum and has the same sign as $f(x_i)$. If, for some z_i,

$$|p_n(z_i) - f(z_i)| = \max_{x \in S} |p_n(x) - f(x)|$$

work with the set $z_0, \ldots, z_{n+1}$, otherwise find x' so that

$$|p_n(x') - f(x')| = \max_{x \in S} |p_n(x) - f(x)|$$

and replace one z_i by x' as in the preceding paragraph.

Exercise

13. Construct a computer program to carry out the first algorithm of Remes and use it to determine some best approximation over a finite set of points.

1.5 L_2-norm approximation

The determination of the best polynomial in the L_2 or least squares norm involves considerations which are more conveniently handled in a rather more general setting. If $\int_a^b |f|^2 \, dx$ exists we write $f \in L_2(a, b)$ or, more briefly, $f \in L_2$ when no confusion can arise.

When $f \in L_2$ and $g \in L_2$ we can introduce the *inner product* (f, g) by

$$(f, g) = \int_a^b f g^* \, dx \tag{1.19}$$

where g^* is the complex conjugate of g. Although we are only concerned with real functions at the moment, complex-valued ones will occur later and it makes little difference to the analysis to cover both cases at once.

We may verify that the right-hand side of (1.19) exists by deriving the *Schwarz inequality.* Clearly

$$\int_a^b (\lambda |f| + \mu |g|)^2 \, dx \geqslant 0$$

or

$$\lambda^2 \int_a^b |f|^2 \, dx + 2\lambda\mu \int_a^b |fg| \, dx + \mu^2 \int_a^b |g|^2 \, dx \geqslant 0$$

for any real λ and μ. The inequality on the quadratic form can hold only if

$$\left(\int_a^b |fg| \, dx \right)^2 \leqslant \int_a^b |f|^2 \, dx \int_a^b |g|^2 \, dx$$

whence

$$\left|\int_a^b fg^* \, dx\right|^2 \leqslant \int_a^b |f|^2 \, dx \int_a^b |g|^2 \, dx$$

which constitutes the Schwarz inequality.

The *norm* $\|f\|$ of f is defined by

$$\|f\| = (f, f)^{1/2}. \tag{1.20}$$

(When other norms are considered, a suffix will be added to this norm to distinguish it from the others.) The norm is always positive unless $f = 0$ almost everywhere. Further consideration of norms will be found in §1.11.

It will be remarked that, if c is a complex constant,

$$(cf, g) = c(f, g); \quad (f, cg) = c^*(f, g);$$
$$\|cf\| = |c| \|f\|; \quad (f, g) = (g, f)^*. \tag{1.21}$$

From the Schwarz inequality

$$|(f, g)| \leqslant \|f\| \, \|g\|. \tag{1.22}$$

Also

$$\int_a^b |f + g|^2 \, dx = \int_a^b |f|^2 \, dx + \int_a^b (fg^* + f^*g) \, dx + \int_a^b |g|^2 \, dx$$
$$\leqslant \left\{\left(\int_a^b |f|^2 \, dx\right)^{1/2} + \left(\int_a^b |g|^2 \, dx\right)^{1/2}\right\}^2 \tag{1.23}$$

by the Schwarz inequality. This may be expressed as

$$\|f + g\| \leqslant \|f\| + \|g\|. \tag{1.24}$$

On replacing f by $f_1 - f_2$ and g by $f_2 - f_3$,

$$\|f_1 - f_3\| \leqslant \|f_1 - f_2\| + \|f_2 - f_3\|. \tag{1.25}$$

If the norm of f is regarded as the length of f, (1.22) states that the modulus of the inner product of f and g is never greater than the product of their lengths. There is an obvious analogy with the scalar product of vectors and, if $(f, g) = 0$, we often say that f and g are *orthogonal*. Similarly, (1.25), expressed in terms of lengths, is the same as the triangle inequality of vectors. The distance between two functions f_1 and f_2 is $\|f_1 - f_2\|$ and is zero only when $f_1 = f_2$ almost everywhere. Approximation in the L_2-norm is an attempt to reduce the distance between two functions to a minimum, distance being understood in the sense above.

An important role is played by orthogonal elements. Suppose there is a finite or infinite set of functions $\phi_1, \phi_2, \ldots$, of L_2 such that

$$(\phi_m, \phi_n) = 0 \qquad (m \neq n), \tag{1.26}$$

$$(\phi_n, \phi_n) = \|\phi_n\|^2 = 1. \tag{1.27}$$

Such a set is said to be an *orthonormal set* and (1.26) and (1.27) are often abbreviated to $(\phi_m, \phi_n) = \delta_{mn}$.

Suppose we want to approximate a function $f \in L_2$ by means of an orthonormal set $\phi_1, \phi_2, \ldots, \phi_N$ using the L_2-norm. Then we wish to choose the coefficients c_n so that

$$\left\| f - \sum_{n=1}^{N} c_n \phi_n \right\|$$

is a minimum. Now, on account of (1.26) and (1.27)

$$\left\| f - \sum_{n=1}^{N} c_n \phi_n \right\|^2 = \|f\|^2 - \sum_{n=1}^{N} \{c_n^*(f, \phi_n) + c_n(\phi_n, f) - c_n c_n^*\}$$

$$= \|f\|^2 - \sum_{n=1}^{N} |(f, \phi_n)|^2 + \sum_{n=1}^{N} |(f, \phi_n) - c_n|^2.$$

Only the third term contains the coefficients c_n and, since no member of the series can be negative, it attains its smallest value of zero when

$$c_n = (f, \phi_n) \qquad (n = 1, 2, \ldots, N). \tag{1.28}$$

Thus (1.28) gives the rule for selecting the coefficients so that the norm is a minimum. When this choice is made

$$\left\| f - \sum_{n=1}^{N} c_n \phi_n \right\|^2 = \|f\|^2 - \sum_{n=1}^{N} |(f, \phi_n)|^2. \tag{1.29}$$

The left-hand side cannot be negative and so

$$\|f\|^2 \geqslant \sum_{n=1}^{N} |(f, \phi_n)|^2 \geqslant \sum_{n=1}^{N} |c_n|^2 \tag{1.30}$$

which is known as *Bessel's inequality*.

An orthonormal set is said to be *complete*, if for every $f \in L_2$, there is a linear combination such that the L_2-norm of the difference is arbitrarily small. If $(f, \phi_m) = 0$ for every ϕ_m of a complete orthonormal set all the coefficients c_m are zero so that the norm of the difference cannot be made arbitrarily small unless $f = 0$.

There is no loss of generality in assuming that the number of elements in a complete orthonormal set is infinite. Letting $N \to \infty$ in Bessel's inequality (1.30), we obtain

$$\sum_{n=1}^{\infty} |(f, \phi_n)|^2 \leqslant \|f\|^2 \tag{1.31}$$

which shows that the series on the left-hand side is convergent. Therefore

$$\left\| \sum_{k=m}^{n} (f, \phi_k)\phi_k \right\|^2 = \sum_{k=m}^{n} |(f, \phi_k)|^2$$

must tend to zero as m and n tend to infinity. It follows (from the Riesz–Fischer theorem) that there is a $g \in L_2$ such that

$$\lim_{n\to\infty} \left\| g - \sum_{k=1}^{n} (f, \phi_k)\phi_k \right\| = 0.$$

From the Schwarz inequality (1.22)

$$\left| \left(g - \sum_{k=1}^{n} (f, \phi_k)\phi_k, \phi_m \right) \right| \leqslant \left\| g - \sum_{k=1}^{n} (f, \phi_k)\phi_k \right\|.$$

Hence

$$(g, \phi_m) = \lim_{n\to\infty} \sum_{k=1}^{n} (f, \phi_k)(\phi_k, \phi_m) = (f, \phi_m).$$

Consequently, $(g - f, \phi_m) = 0$ for $m = 1, \ldots$ and since the orthonormal set is complete our earlier remarks entail $f = g$. We may summarize this by saying: *if $\phi_1, \phi_2, \ldots$ is a complete orthonormal set every $f \in L_2$ can be expressed as*

$$f = \sum_{k=1}^{\infty} (f, \phi_k)\phi_k,$$

the equality being understood to mean that

$$\lim_{n\to\infty} \left\| f - \sum_{k=1}^{n} (f, \phi_k)\phi_k \right\| = 0.$$

It follows from (1.29) that, for a complete orthonormal set, $f = \sum_{k=1}^{\infty} c_k\phi_k$ implies that

$$\|f\|^2 = \sum_{k=1}^{\infty} |c_k|^2. \tag{1.32}$$

If $g = \sum_{k=1}^{\infty} b_k\phi_k$ and we apply (1.32) to $f + g, f - g, f + ig, f - ig$ then, from the identity

$$\|f + g\|^2 - \|f - g\|^2 + i\|f + ig\|^2 - i\|f - ig\|^2 = 4(f, g),$$

is derived *Parseval's formula*

$$(f, g) = \sum_{k=1}^{\infty} c_k b_k^*.$$

Given a set of linearly independent elements $\psi_1, \psi_2, \ldots$ which will approximate any $f \in L_2$ arbitrarily close in L_2-norm we can always manufacture a complete orthonormal set by a method known as the *Schmidt process*. First define ϕ_1 by

$$\phi_1 = \psi_1/\|\psi_1\|.$$

Then pick $\phi_2 = g_2/\|g_2\|$ where $g_2 = \psi_2 - (\psi_2, \phi_1)\phi_1$; g_2 cannot be zero because ψ_1 and ψ_2 are linearly independent. Clearly $(\phi_2, \phi_1) = 0$. In general, $\phi_n = g_n/\|g_n\|$

where

$$g_n = \psi_n - (\psi_n, \phi_{n-1})\phi_{n-1} - (\psi_n, \phi_{n-2})\phi_{n-2} - \cdots - (\psi_n, \phi_1)\phi_1.$$

It is important to observe that in the whole of the preceding discussion concerning the minimization of the norm we have not used the specific form (1.19) but only properties of the inner product such as (1.20), (1.21), (1.22), and (1.24). Therefore we can draw the same conclusions if the inner product is defined in another way so long as it has the properties (1.20), (1.21), (1.22), and (1.24). For instance, if we choose

$$(f, g) = \sum_{i=1}^{M} f(x_i)g^*(x_i)$$

for some fixed x_i we can easily verify that the properties are valid and so we may deduce that $\| f - \sum_{n=1}^{N} c_n\phi_n \|$ or $\sum_{i=1}^{M} |f(x_i) - \sum_{n=1}^{N} c_n\phi_n(x_i)|^2$ is a minimum when

$$c_n = (f, \phi_n) = \sum_{i=1}^{M} f(x_i)\phi_n^*(x_i).$$

It is this kind of problem which arises in fitting data at a discrete number of points by the *method of least squares.* Note that it is frequently a computational advantage to employ orthonormal polynomials for least squares rather than expansions in non-orthogonal functions because the matrices tend to be diagonally dominant even when round-off error is present.

Another possibility is to take

$$(f, g) = \int_a^b w(x)f(x)g^*(x)\,dx$$

where w is a real non-negative function. This corresponds to varying the contribution from the various parts of the interval according to the *weight function* w. In this connection there is the following interesting result:

THEOREM 1.5. *If $\phi_1, \phi_2, \ldots$ is an infinite orthonormal set of polynomials on the finite interval $[a, b]$ with weight function w, i.e.*

$$\int_a^b w(x)\phi_m(x)\phi_n^*(x)\,dx = \delta_{mn},$$

then the orthonormal set is complete.

Proof. Theorem 1.4 ensures that, for continuous f, there is a polynomial $p(x)$ such that

$$|f(x) - p(x)| < \varepsilon.$$

The choice (1.28) guarantees a minimum of the L_2-norm so that

$$\int_a^b w(x)\left|f(x) - \sum_{n=1}^{N} c_n\phi_n(x)\right|^2 dx \leqslant \int_a^b w(x)|f(x) - p(x)|^2\, dx$$

provided that N is made larger than the degree of p. Since the right-hand side does not exceed $\varepsilon^2 \int_a^b w(x)\, dx$ and can be made arbitrarily small we have the desired result. Since any $f \in L_2$ can be approximated as close as one wishes by continuous functions the proof is terminated.

As an example let $a = -1$, $b = 1$, and $w = 1$; first construct an orthonormal set (which must be complete by Theorem 1.5) from the powers of x, i.e. with $\psi_j = x^{j-1}$. The Schmidt process gives

$$\phi_1 = 1/2^{1/2}, \quad \phi_2 = (3/2)^{1/2}x, \quad \phi_3 = (5/2)^{1/2}\tfrac{1}{2}(3x^2 - 1), \ldots,$$

which are multiples of the *Legendre polynomials* $P_n(x)$ which are defined by *Rodrigue's formula*

$$P_n(x) = \frac{1}{n!2^n}\frac{d^n}{dx^n}(x^2 - 1)^n.$$

The first few are

$$P_0(x) = 1, \quad P_1(x) = x, \quad P_2(x) = \tfrac{1}{2}(3x^2 - 1), \quad P_3(x) = \tfrac{1}{2}(5x^3 - 3x)$$

and they satisfy the recurrence relation

$$(n + 1)P_{n+1}(x) = (2n + 1)xP_n(x) - nP_{n-1}(x)$$

and have the orthogonal property

$$\int_{-1}^{1} P_m(x)P_n(x)\, dx = 2\delta_{mn}/(2n + 1).$$

In practical calculation it may be more convenient to compute the ϕ_k via the recurrence relations directly instead of deriving the analytical expressions first.

A second example is supplied by $a = -1$, $b = 1$, $w = (1 - x^2)^{-1/2}$. Again we start from the powers of x and find for our orthonormal set

$$\phi_1 = 1/\pi^{1/2}, \quad \phi_2 = (2/\pi)^{1/2}x, \quad \phi_3 = (2/\pi)^{1/2}(2x^2 - 1), \ldots$$

which are multiples of the *Chebyshev polynomials*. The Chebyshev polynomial T_n is defined by

$$\begin{aligned} T_n(x) &= \cos(n \cos^{-1} x) \\ &= \frac{n!(-2)^n}{(2n)!}(1 - x^2)^{1/2}\frac{d^n}{dx^n}[(1 - x^2)^{n-1/2}]. \end{aligned}$$

Some examples are

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

The term of the highest power in T_n is $2^{n-1}x^n$. The Chebyshev polynomial has a celebrated property concerning the maximum norm, namely

THEOREM 1.5a (CHEBYSHEV). *Of all polynomials of degree n in which the coefficient of the highest power is unity the one with the smallest maximum norm on* $[-1, 1]$ *is* $T_n(x)/2^{n-1}$ *and*

$$\|T_n(x)/2^{n-1}\|_\infty = 1/2^{n-1}.$$

Here the notation $\|f\|_\infty$ is employed to signify the maximum norm, i.e. $\sup|f|$ over the appropriate interval which, in this case, is $[-1, 1]$.

Proof. Assume that there is a polynomial $p_n(x)$ of degree n and with leading coefficient unity which is of smaller maximum norm than $T_n(x)/2^{n-1}$. Let

$$q(x) = p_n(x) - T_n(x)/2^{n-1}.$$

Then q is a polynomial of degree at most $n-1$. Since p_n has a smaller norm than $T_n/2^{n-1}$, q must be negative at the maxima of $T_m/2^{n-1}$ and positive at the minima of $T_n/2^{n-1}$. Now putting $x = \cos\theta$, $T_n(\cos\theta) = \cos n\theta$ so that $T_n(x)$ has zeros at $x = \cos\{(2k-1)\pi/2n\}$ for $k = 1, 2, \ldots, n$ and therefore possesses $n+1$ maxima and minima on $[-1, 1]$. Hence q must vanish at least n times which is contrary to its being a polynomial of degree $n-1$. Thus the first part of the theorem is proved and the second part follows from the form of T_n when $x = \cos\theta$.

Another way of expressing Theorem 1.5*a* is to say that of all polynomials of degree n with maximum norm unity on $[-1, 1]$, $T_n(x)$ has the largest leading coefficient, namely 2^{n-1}.

Series of Chebyshev polynomials can be readily summed on the computer by taking advantage of the recurrence formula

$$T_{n+1}(x) - 2xT_n(x) + T_{n-1}(x) = 0.$$

For instance, if

$$f(x) = \sum_{n=0}^{N} a_n T_n(x),$$

define $b_{N+1} = 0$, $b_N = a_N$ and then calculate $b_{N-1}, \ldots, b_1$ from

$$b_n = a_n + 2xb_{n+1} - b_{n+2}.$$

It follows from the recurrence formula for T_n that

$$f(x) = a_0 - b_2 + b_1 x.$$

The round-off characteristics of this method are no worse than those of ordinary polynomial evaluation and the same number of multiplications is used. In fact, the method can be used for any system of polynomials $p_n(x)$ which

satisfies a recurrent relation of the form

$$p_{n+1}(x) - p(x)p_n(x) + p_{n+1}(x) = 0$$

by putting

$$b_n = a_n + pb_{n+1} - b_{n+2}$$

and then

$$\sum_{n=0}^{N} a_n p_n(x) = (a_0 - b_2)p_0(x) + b_1 p_1(x).$$

Any power series can be expressed as an expansion in Chebyshev polynomials by employing formulae such as

$$1 = T_0(x), \quad x = T_1(x), \quad x^2 = \tfrac{1}{2}\{T_0(x) + T_2(x)\},$$
$$x^3 = \tfrac{1}{4}\{3T_1(x) + T_3(x)\}.$$

It is often possible to reduce the degree of an approximating polynomial and thereby economize in computation by implementing the properties of Chebyshev polynomials. For example, if the function f is approximated by the polynomial p_{n+1} where

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

consider the polynomial $\bar{p}_n$ defined by

$$\bar{p}_n(x) = p_{n+1}(x) - a_{n+1}T_{n+1}(x)/2^n.$$

Then $\bar{p}_n$ is of degree n and

$$\bar{p}_n - f = p_{n+1} - f - a_{n+1}T_{n+1}(x)/2^n.$$

Thus the error in $\bar{p}_n$ does not exceed that in p_{n+1} by more than $a_{n+1}T_{n+1}(x)/2^n$. Since $|T_{n+1}(x)| \leqslant 1$ on $[-1, 1]$, this error can be quite small when $a_{n+1}/2^n$ is small enough. In other words, truncation of the power series by removal of the higher powers by subtracting appropriate multiples of Chebyshev polynomials can lead to an effective measure of economization.

Although the properties of Chebyshev polynomials have been described for the interval $[-1, 1]$ they can be extended to other finite intervals such as $[x_1, x_2]$ by first making the substitution

$$y = -1 + 2\frac{x - x_1}{x_2 - x_1}.$$

Exercises

14. Express $1, x, \ldots, x^5$ in terms of Legendre polynomials.
15. Find the polynomial of degree 2 which gives the best L_2-norm approximation to e^x on $[0, 1]$.

16. The function $f(x)$ was determined experimentally and found to have the following values

x:	1.00	1.04	1.08	1.12	1.16	1.20
$f(x)$:	8.41	8.63	8.82	9.00	9.17	9.32

Find the polynomial of degree 2 which gives the best approximation in L_2-norm.

17. By making the substitution

$$x = \frac{(\sqrt{2}+1)y - 1}{(\sqrt{2}-1)y + 1}$$

express $\tan^{-1} y$ in terms of Chebyshev polynomials of x. If only those T_n are retained for which $n \leqslant 7$ show that the recurrence relation method gives

$$\tan^{-1}(1/\sqrt{3}) = 0.5235986.$$

18. By starting from the Taylor series for e^x up to powers of x^5 show that Chebyshev truncation leads to

$$e^x = (382 + 383x + 208x^2 + 68x^3)/384$$

with an error of not more than one unit in the second decimal place on $[-1, 1]$.

1.6 Rational approximation

Although Weierstrass's theorem tells us that any continuous function can be approximated as closely as we like on a finite interval, the degree of the polynomial may be unduly high for a specified level of accuracy. Again, the presence of a singularity in the complex plane near the real axis may render polynomial approximation awkward. For these reasons it is worth considering whether a rational function will give better accuracy as an approximant than a polynomial. It has been suggested (see, for example, Hart *et al.* (1968)) that for a given amount of computational effort rational functions give greater accuracy than polynomials.

Consider the possibility of constructing a rational approximation to f in a neighbourhood of the origin—there is no loss of generality in selecting the origin since any other point can be converted to it by a simple change of variable. We try $p_m(x)/q_n(x)$ where p_m and q_n are polynomials of degree m and n respectively, and are supposed to have no common zero since, otherwise, it could be cancelled. One method of specifying p_m and q_n is to require that p_m/q_n and its first $m + n$ derivatives agree with f and its first $m + n$ derivatives at $x = 0$; it is then called a *Padé approximant.*

For example, for a Padé approximant to $\ln(1 + x)$ with $m = 2$ and $n = 2$ we would want the coefficients in

$$(a_0 + a_1x + a_2x^2)/(b_0 + b_1x + b_2x^2)$$

chosen so that the expansion of the rational function near $x = 0$ was the same as $x - x^2/2 + \cdots$. To put it another way we wish to make as many powers of

x disappear from

$$a_0 + a_1x + a_2x^2 - (b_0 + b_1x + b_2x^2)(x - \tfrac{1}{2}x^2 + \cdots)$$

as possible. Therefore, select

$$a_0 = 0, \quad a_1 = b_0, \quad a_2 = b_1 - \tfrac{1}{2}b_0,$$
$$b_2 - \tfrac{1}{2}b_1 + \tfrac{1}{3}b_0 = 0, \quad -\tfrac{1}{2}b_2 + \tfrac{1}{3}b_1 - \tfrac{1}{4}b_0 = 0$$

so as to eliminate powers up to and including x^4; if we tried to remove x^5 we should find $b_0 = b_1 = b_2 = 0$ which is obviously unacceptable. Since we have one more coefficient than equations we normalize by putting $b_0 = 1$. Then $a_1 = 1$, $b_1 = 1$, $a_2 = \frac{1}{2}$, $b_2 = \frac{1}{6}$ and the Padé approximant to $\ln(1 + x)$ is

$$\frac{x + \frac{1}{2}x^2}{1 + x + \frac{1}{6}x^2}$$

agreeing to powers of up to x^4 in $\ln(1 + x)$.

Other Padé approximants can, of course, be constructed by choosing different values of m and n but, as a matter of practice, it is usually found that the best approximations are obtained by taking $m = n$ or possibly $m = n + 1$ provided that f has a Taylor expansion at the origin.

An alternative form of rational approximation may be derived from *Obreschkoff's formula*

$$\sum_{k=0}^{n} (-)^k \frac{n!(m+n-k)!}{(n-k)!(m+n)!} \frac{(x-x_1)^k}{k!} f^{(k)}(x)$$
$$= \sum_{k=0}^{m} \frac{n!(m+n-k)!}{(n-k)!(m+n)!} \frac{(x-x_1)^k}{k!} f^{(k)}(x_1)$$
$$+ \frac{1}{(m+n)!} \int_{x_1}^{x} (x-t)^m (x_1-t)^n f^{(m+n+1)}(t)\, dt$$

which may be verified by integrating the integral by parts $m + n + 1$ times. The integral is effectively of order $(x - x_1)^{m+n+1}$ and so can be ignored to a first approximation; its explicit form can be used to provide an estimate of the error made in such neglect.

As an example let $f(x) = x^\mu$ and $x_1 = 1$. Then, dropping the integral, we have with $m = n = 1$

$$x^\mu - \tfrac{1}{2}(x - 1)\mu x^{\mu-1} = 1 + \tfrac{1}{2}(x - 1)\mu$$

or

$$x^\mu = \frac{2 - \mu + \mu x}{\mu + (2 - \mu)x} x$$

as a rational approximation valid near $x = 1$ for any real μ.

Padé approximants usually become increasingly inaccurate as $|x|$ increases. So attempts have been made to minimize $|p_m/q_n - f|$ over an interval. Something like the second algorithm of Remes (§1.4) can be constructed but the algorithm may not converge if the initial approximation is not sufficiently good and, in any case, the solution of non-linear equations is involved at each stage.

A convenient method for evaluating rational functions is by *continued fractions*, which may also arise in other contents in numerical work. (Expansions for numerous functions in polynomials, Chebyshev polynomials, rational functions, and continued fractions can be found in Abramowitz and Stegun (1965).) To fabricate a continued fraction suppose we are given m/n. Divide m by n; let a_1 be the quotient and p the remainder so that

$$\frac{m}{n} = a_1 + \frac{p}{n} = a_1 + \frac{1}{n/p}.$$

Divide n by p; let a_2 be the quotient and q the remainder; then

$$\frac{n}{p} = a_2 + \frac{q}{p} = a_2 + \frac{1}{p/q}.$$

Proceeding in this way we obtain

$$\frac{m}{n} = a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cdots}} = a_1 + \frac{1}{a_2+}\,\frac{1}{a_3+}\cdots$$

More generally we can consider expressions of the form

$$b_0 + \frac{a_1}{b_1+}\,\frac{a_2}{b_2+}\cdots.$$

If the number of terms is finite it is called a *terminating continued fraction.* Otherwise, it is called an *infinite continued fraction* and the terminating fraction

$$f_n = b_0 + \frac{a_1}{b_1+}\,\frac{a_2}{b_2+}\cdots\frac{a_n}{b_n}$$

is called the n*th convergent.* If $\lim_{n\to\infty} f_n$ exists, an infinite continued fraction is said to be convergent. It can be proved that, if $a_i = 1$ and the b_i are integers, convergence is always secured.

If $f_n = A_n/B_n$ it may easily be verified that

$$A_n = b_nA_{n-1} + a_nA_{n-2}, \tag{1.33}$$

$$B_n = b_nB_{n-1} + a_nB_{n-2}, \tag{1.34}$$

$$A_nB_{n-1} - A_{n-1}B_n = (-1)^n$$

subject to $A_{-1} = 1$, $A_0 = b_0$, $B_{-1} = 0$, $B_0 = 1$. Hence

$$f_{n+1} - f_n = -a_{n+1}B_{n-1}(f_n - f_{n-1})/B_{n+1}.$$

If a_i and b_i are all positive, (1.34) indicates that $0 < a_{n+1}B_{n-1}/B_{n+1} < 1$. Thus $f_{n+1} - f_n$ is numerically less than, and of opposite sign to, $f_n - f_{n-1}$. Now, in this case, b_0 is less than the continued fraction since part is omitted while the convergent $b_0 + a_1/b_1$ is greater than the continued fraction because the denominator is too small. Following this route we conclude that, when the a_i and b_i are positive, every convergent of odd order is greater than the continued fraction and every convergent of even order is less than the continued fraction; moreover

$$f_{2n+1} < f_{2n-1}, \quad f_{2n} > f_{2n-2}$$

so that the convergents of odd order steadily decrease while those of even order steadily increase.

These properties make continued fractions very convenient for computation. Since, for any rational function an equivalent terminating continued fraction can be manufactured (clearly, a terminating continued fraction in which a_i and b_i are polynomials is equivalent to a rational function), the continued fraction may be evaluated more economically, as far as the number of arithmetical operations is concerned, than calculating the numerator and denominator of the rational function separately and then dividing.

For the conversion of series the following terminating continued fractions may be noted:

$$1 + b_2 + b_2b_3 + \cdots + b_2b_3 \cdots b_n = \frac{1}{1-}\,\frac{b_2}{b_2+1-}\,\frac{b_3}{b_3+1-}\cdots\frac{b_n}{-b_n+1}, \tag{1.35}$$

$$\frac{1}{u_1} + \frac{1}{u_2} + \cdots + \frac{1}{u_n} = \frac{1}{u_1-}\,\frac{u_1^2}{u_1+u_2-}\cdots\frac{u_{n-1}^2}{-u_{n-1}+u_n}, \tag{1.36}$$

$$\frac{1}{a_0} - \frac{x}{a_0a_1} + \frac{x^2}{a_0a_1a_2} \cdots + \frac{(-)^n x^n}{a_0a_1 \ldots a_n} = \frac{1}{a_0+}\,\frac{a_0x}{a_1-x+}\,\frac{a_1x}{a_2-x+}\cdots\frac{a_{n-1}x}{a_n-x}. \tag{1.37}$$

Infinite series may be handled via

$$\sum_{n=0}^{\infty} a_n x^n = \frac{\alpha_0}{1-}\,\frac{\alpha_1 x}{1+\alpha_1 x-}\,\frac{\alpha_2 x}{1+\alpha_2 x-}\cdots \tag{1.38}$$

where $\alpha_0 = a_0$, $\alpha_n = a_n/a_{n-1}$ $(n \geqslant 1)$. Alternate expressions can be derived by

using the fact that the nth convergent can be written as

$$f_n = b_0 + \frac{c_1a_1}{c_1b_1+} \frac{c_1c_2a_2}{c_2b_2+} \frac{c_2c_3a_3}{c_3b_3+} \cdots \frac{c_{n-1}c_na_n}{c_nb_n}$$

for arbitrary non-zero c_i.

Exercises

19. (i) Construct the Padé approximant with $m = n = 2$ for e^x in the neighbourhood of the origin.

    (ii) Find the maximum norm of the difference between the Padé approximant and e^x on [0, 1]. Compare your result with the polynomial of degree 5 obtained by the first algorithm of Remes with S the set 0(0.1)1.

20. Find the Padé approximants with (i) $m = 2$, $n = 2$, (ii) $m = 3$, $n = 2$ for $\sin x$ near the origin.

21. Use Obreschkoff's formula to obtain the approximations

$$\text{(i) } \ln(1 + x) = -\frac{x(x + 2)}{6(x + 1)^2}(2x^2 - 3x - 3),$$

$$\text{(ii) } e^x = \frac{1 + \frac{1}{2}x + \frac{x^2}{12}}{1 - \frac{1}{2}x + \frac{x^2}{12}}$$

    near the origin. How does (ii) compare with 19(i)?

22. Find a, b, and c so that

$$\max_{0 \le x \le 1} \left| e^x - \frac{a + bx}{1 + cx} \right|$$

    is a minimum. Compare the corresponding Padé approximant with $m = n = 1$.

23. Calculate successive convergents to

$$\text{(i) } 2 + \frac{1}{6+} \frac{1}{1+} \frac{1}{1+} \frac{1}{11+} \frac{1}{2}.$$

$$\text{(ii) } \frac{1}{2+} \frac{1}{2+} \frac{1}{3+} \frac{1}{1+} \frac{1}{4+} \frac{1}{2+} \frac{1}{6}.$$

24. A metre equals 1.0936 yards. Find limits to the error in taking 222/203 yards as equivalent to a metre.

25. Show that

$$\text{(i) } \tan x = \frac{x}{1-} \frac{x^2}{3-} \frac{x^3}{5-} \cdots,$$

$$\text{(ii) } \ln \frac{1 + x}{1 - x} = \frac{2x}{1-} \frac{x^2}{3-} \frac{(2x)^2}{5-} \frac{(3x)^2}{7-} \cdots$$

26. The numerator and denominator of a rational function, both of degree n, are expressed in terms of Chebyshev polynomials. Obtain the formulae converting it to a continued fraction of the form

$$a_0 + \frac{b_1}{a_1 + x +} \frac{b_2}{a_2 + x +} \cdots.$$

1.7 Trigonometric interpolation

The approximation of a function f on $[0, 2\pi]$ by a series of the form

$$\tfrac{1}{2}a_0 + \sum_{n=0}^{N} (a_n \cos nx + b_n \sin nx)$$

is a particular case of the general theory developed in §1.5. Nevertheless some of the formulae are of interest and will be needed subsequently. By the general theory the best L_2-norm approximation to f is obtained when $a_n = \alpha_n$ and $b_n = \beta_n$ where

$$\alpha_n = (1/\pi) \int_0^{2\pi} f(x) \cos nx \, dx,$$

$$\beta_n = (1/\pi) \int_0^{2\pi} f(x) \sin nx \, dx.$$

The coefficients α_n and β_n are, of course, those which would occur in the infinite Fourier series representation of f. This infinite series may not converge to f but, if f has only a finite number of discontinuities which are finite jumps, the series converges to $\frac{1}{2}\{f(x + 0) + f(x - 0)\}$ at interior points and $\frac{1}{2}\{f(0 + 0) + f(2\pi - 0)\}$ at $x = 0, 2\pi$ (when f is piecewise smooth). However, since at the moment we are concerned with finite trigonometric series the problem of convergence does not arise.

Suppose now that we ask that the trigonometric expansion be specified not by the L_2-norm but by being required to agree with f at certain points. Let the points be chosen as kh $(k = 0, 1, \ldots, M)$ where M is a positive integer and $h = 2\pi/M$. Then we try to find a_n and b_n so that

$$\tfrac{1}{2}a_0 + \sum_{n=1}^{N} (a_n \cos nkh + b_n \sin nkh) = f(kh) \qquad (k = 1, \ldots, M - 1)$$

$$= \tfrac{1}{2}\{f(0) + f(2\pi)\} \qquad (k = 0, M) \qquad (1.39)$$

Now

$$\sum_{k=1}^{M} e^{inkh} = \frac{e^{inh} - e^{i(M+1)nh}}{1 - e^{inh}}$$

unless $e^{inh} = 1$. But $e^{iMnh} = 1$, since n is an integer and so the series is zero if

$e^{inh} \neq 1$. If, however, $e^{inh} = 1$, each term in the series is 1 and so

$$\sum_{k=1}^{M} e^{inkh} = M \qquad \text{(if } n/M \text{ is an integer)}$$
$$= 0 \qquad \text{(otherwise)} \tag{1.40}$$

since n/M being an integer is the condition for $e^{inh} = 1$. If m and n are integers we see from (1.40) that

$$\sum_{k=1}^{M} e^{i(m+n)kh} = M \qquad \text{(if } (m+n)/M \text{ is an integer),}$$

$$\sum_{k=1}^{M} e^{i(n-m)kh} = M \qquad \text{(if } (n-m)/M \text{ is an integer)}$$

and otherwise the sum of each series is zero. With $\mathscr{R}$ denoting the real part

$$\cos nkh \cos mkh = \tfrac{1}{2}\mathscr{R}\{e^{i(m+n)kh} + e^{i(n-m)kh}\}$$

and hence

$$\sum_{k=1}^{M} \cos nkh \cos mkh = 0 \quad \text{or} \quad \tfrac{1}{2}M \quad \text{or} \quad M \tag{1.41}$$

according as (a) neither $(n+m)/M$ nor $(n-m)/M$ is an integer, (b) one but not both of $(n+m)/M$ and $(n-m)/M$ is an integer, (c) both $(n+m)/M$ and $(n-m)/M$ are integers.

Similarly, from

$$\sum_{k=1}^{M} \sin nkh \sin mkh = \mathscr{R}\frac{1}{2}\sum_{k=1}^{M}\{e^{i(n-m)kh} - e^{i(n+m)kh}\},$$

$$\sum_{k=1}^{M} \cos nkh \sin mkh = \mathscr{I}\frac{1}{2}\sum_{k=1}^{M}\{e^{i(n+m)kh} - e^{i(n-m)kh}\}$$

we deduce that

$$\sum_{k=1}^{M} \sin nkh \sin mkh = 0 \quad \text{or} \quad -\tfrac{1}{2}M \quad \text{or} \quad \tfrac{1}{2}M \tag{1.42}$$

according as (a) both $(n+m)/M$ amd $(n-m)/M$ are integers or neither is, (b) $(n+m)/M$ is an integer but $(n-m)/M$ is not, (c) $(n-m)/M$ is an integer but $(n+m)/M$ is not, and that

$$\sum_{k=1}^{M} \cos nkh \sin mkh = 0. \tag{1.43}$$

Multiply the kth equation of (1.39) by $\cos mkh$, where m is one of the integers

$0, \ldots, N$, and add. Then

$$\sum_{k=1}^{M-1} f(kh) \cos mkh + \tfrac{1}{2}\{f(0) + f(2\pi)\} \cos 2\pi m$$

$$= \sum_{k=1}^{M} \left\{ \tfrac{1}{2}a_0 + \sum_{n=1}^{N} (a_n \cos nkh + b_n \sin nkh) \right\} \cos mkh. \quad (1.44)$$

Suppose how that M is even; select $N = \frac{1}{2}M$. Then, from (1.40), (1.41), and (1.43) the right-hand side of (1.44) is $\frac{1}{2}Ma_m$ if $m \neq \frac{1}{2}M$ and Ma_N if $m = \frac{1}{2}M = N$. In a similar way the right-hand side of

$$\sum_{k=1}^{M-1} f(kh) \sin mkh + \tfrac{1}{2}\{f(0) + f(2\pi)\} \sin 2\pi m$$

$$= \sum_{k=1}^{M} \left\{ \tfrac{1}{2}a_0 + \sum_{n=1}^{N} (a_n \cos nkh + b_n \sin nkh) \right\} \sin mkh$$

is $\frac{1}{2}Mb_m$ when $m \neq 0, \frac{1}{2}M$.

Thus the solution to our problem when M is even is

$$\tfrac{1}{2}a_0 + \tfrac{1}{2}a_N \cos Nx + \sum_{n=1}^{N-1} (a_n \cos nx + b_n \sin nx)$$

where $N = \frac{1}{2}M$ and

$$a_m = \frac{2}{M} \sum_{k=1}^{M} f(kh) \cos mkh, \quad (1.45)$$

$$b_n = \frac{2}{M} \sum_{k=1}^{M} f(kh) \sin mkh \quad (1.46)$$

with the understanding that $f(Mh)$ means $\frac{1}{2}\{f(0) + f(2\pi)\}$. It will be observed that there is no other solution since the coefficients a_m and b_m vanish when f is zero in (1.45) and (1.46).

If M is odd, an analogous procedure gives the expansion

$$\tfrac{1}{2}a_0 + \sum_{n=1}^{N} (a_n \cos nx + b_n \sin nx)$$

where $N = \frac{1}{2}(M - 1)$ and the coefficients a_m, b_m are still given by (1.45) and (1.46).

The analysis of the inner product $\sum f(x_i)g^*(x_i)$ in §1.5 demonstrates that, not only does the trigonometric polynomial agree with the function at the specified points, but also it is the same as would be obtained by the method of least squares in fitting the data by a trigonometric polynomial of degree N.

Exercises

27*a*. Find the trigonometric interpolant on $[0, 2\pi]$ for $f(x) = x$ with $M = 4$ and show that it is badly in error at the end-points.

27*b*. If $f(x) = x(0 \leqslant x \leqslant \pi)$, $= 2\pi - x(\pi \leqslant x \leqslant 2\pi)$ obtain the trigonometric interpolant when $M = 3$ and when $M = 10$. Compare the graphs of the interpolants with the original function.

SOLUTION OF EQUATIONS

1.8 Solution of an equation

Often one is faced with the problem of finding the values of x which satisfy an equation of the form

$$f(x) = 0. \tag{1.47}$$

Such a value of x is called a *root* of (1.47) or a *zero* of f. Since the number of equations which can be solved analytically is very limited, the devising of numerical techniques is of paramount importance.

It is necessary to be aware right from the start that it will rarely be possible to find the roots of (1.47) exactly by numerical methods. There are several reasons for this. In the first place, unless f is a very elementary function, it will usually have to be replaced by some approximant—perhaps one of the types discussed in preceding sections. Such replacement is bound to introduce some error. Secondly, any computation will usually involve round-off error. Thirdly, any computer can carry only a certain set of rational numbers so that if the root of (1.47) is not a rational number or is a rational number outside the computer set its representation in the computer must inevitably be in error.

Given that these sources of error are virtually inescapable it is vital to arrange that techniques produce answers which can be related to the roots of (1.47) and, in particular, do not supply more or less zeros of f than were originally present.

Suppose that f is continuous for $a \leqslant x \leqslant b$ and that $f(a)$ and $f(b)$ have opposite signs, i.e. $f(a)f(b) < 0$. Then we know that $f(x) = 0$ has at least one root in $[a, b]$. In the *bisection method* we aim to locate a root by taking a sequence of intervals, each half the size of the previous one and each containing a root. The actual algorithm is:

Define $a_0 = a$, $b_0 = b$ *and then form the numbers* $a_1, b_1, a_2, b_2, \ldots$ *successively by the following procedure. Put*

$$c_r = \tfrac{1}{2}(a_{r-1} + b_{r-1})$$

and calculate $f(c_r)$. *If* $f(c_r) = 0$ *then* $x = c_r$ *is the root sought. If* $f(c_r) \neq 0$ *then either* (*i*) $f(c_r)f(a_{r-1}) > 0$ *and then we define* $a_r = c_r$, $b_r = b_{r-1}$, *or* (*ii*) $f(c_r)f(a_{r-1}) < 0$ *and then we define* $a_r = a_{r-1}$, $b_r = c_r$. *Stop the process when* $|a_r - b_r| \leqslant \varepsilon$, *where* ε *is some pre-assigned number.*

In general ε is selected so that desired accuracy is attained or so as to keep the number of iterations down to a specified level. The convergence of the process is governed by Theorem 1.8.

THEOREM 1.8. *Under the conditions of the algorithm*

(i) $b_r - a_r = (b - a)/2^r$

and, if x_0 *is the root of* $f(x) = 0$,

(ii) $|x_0 - \frac{1}{2}(a_r + b_r)| < \frac{1}{2}(b_r - a_r) < (b - a)/2^{r+1}$.

Proof. If (i) of the algorithm applies

$$b_r - a_r = b_{r-1} - c_r = \tfrac{1}{2}(b_{r-1} - a_{r-1}).$$

If (ii) applies

$$b_r - a_r = c_r - a_{r-1} = \tfrac{1}{2}(b_{r-1} - a_{r-1})$$

so that there is the same connection between the lengths of successive intervals in both cases. Part (i) of the theorem is an immediate consequence.

For part (ii) remark that

$$x_0 - \tfrac{1}{2}(a_r + b_r) = \tfrac{1}{2}(x_0 - a_r) + \tfrac{1}{2}(x_0 - b_r).$$

Now $x_0 - a_r$ is positive and $x_0 - b_r$ is negative so that the right-hand side must be less than $\frac{1}{2}(x_0 - a_r)$ and greater than $\frac{1}{2}(x_0 - b_r)$. However, $x_0 < b_r$ so that $x_0 - a_r < b_r - a_r$, and $x_0 > a_r$ so that $x_0 - b_r > a_r - b_r$. Thus the right-hand side is smaller than $\frac{1}{2}(b_r - a_r)$ and larger than $\frac{1}{2}(a_r - b_r)$, i.e.

$$|x_0 - \tfrac{1}{2}(a_r + b_r)| < \tfrac{1}{2}(b_r - a_r).$$

The final statement in part (ii) follows from part (i) and the proof is complete.

Theorem 1.8 (i) tells us that successive intervals containing the root become smaller and smaller so that the root can be placed to any desired degree of accuracy. From (ii) we see that if the iteration is stopped when $b_r - a_r \leqslant \varepsilon$ the error in $\frac{1}{2}(a_r + b_r)$ as an approximation to x_0 does not exceed $\frac{1}{2}\varepsilon$. Furthermore, the number of iterations to achieve this accuracy satisfies $2^r \geqslant (b - a)/\varepsilon$.

These conclusions and Theorem 1.8 assume that f is calculated exactly. As we have already remarked this is not true in general. However, reasonable results can be expected provided that ε is not chosen too small, e.g. it must be greater than the minimum distance between two consecutive numbers of the computer set.

A variant of the bisection method is the *method of false position.* In this the approximation c_{r+1} to the root, instead of being taken as $\frac{1}{2}(a_r + b_r)$, is chosen as the point where the straight line joining $(a_r, f(a_r))$ and $(b_r, f(b_r))$ cuts the x-axis in the $(x, f(x))$-plane. Consequently,

$$c_{r+1} = \frac{b_r f(a_r) - a_r f(b_r)}{f(a_r) - f(b_r)}. \tag{1.48}$$

Apart from this change the method of false position has the same procedure as the bisection method. It can be proved that the method of false position converges to a root under the same conditions as Theorem 1.8 but the convergence is generally much slower than that for the bisection method.

A relation of the method of false position is the secant method, in which a sequence of points $x_1, x_2, \ldots$ is generated via (1.48) so that

$$x_{r+1} = \frac{x_{r-1}f(x_r) - x_r f(x_{r-1})}{f(x_r) - f(x_{r-1})} \tag{1.49}$$

with $x_1 = a$, $x_2 = b$. Here there is no requirement that $f(a)f(b) < 0$ but now we have no guarantee of convergence. Indeed, if there is convergence, the denominator of (1.49) must approach zero which can make for numerical difficulty. There is, of course, complete failure if $f(x_r) = f(x_{r-1})$. On the other hand, the secant method will, when it converges, usually do so faster than the bisection method or the method of false position.

The iterative methods that have been discussed so far and those to be mentioned subsequently are all of the type

$$x_{r+1} = F(x_r). \tag{1.50}$$

If $\lim_{r\to\infty} x_r = x_0$ and F is continuous in a neighbourhood of x_0, $\lim_{r\to\infty} F(x_r) = F(x_0)$. Hence, a convergent iteration with continuous F leads to a root of

$$x = F(x). \tag{1.51}$$

Thus the main question is whether the sequence converges and the answer to this may depend not only on the form of F but also the starting value x_1.

A somewhat stronger condition than continuity is to require

$$|F(x) - F(y)| \leqslant M|x - y| \tag{1.52}$$

which is a *Lipschitz condition.* If F is differentiable the mean value theorem asserts that

$$F(x) - F(y) = F'(\xi)(x - y)$$

for some ξ in (x, y). Thus, if $|F'(\xi)| \leqslant M$, F satisfies the Lipschitz condition (1.52)

We now prove

THEOREM 1.8*a. If F satisfies* (1.52) *for all x, y with $M < 1$ then* (1.51) *has a unique root x_0 and the iteration* (1.50) *converges to it for any x_1.*

Proof. From (1.50) and (1.52)

$$|x_{r+1} - x_r| = |F(x_r) - F(x_{r-1})| \leqslant M|x_r - x_{r-1}| \leqslant M^{r-1}|x_2 - x_1|$$

by repeated application. Hence, for any integer $s \geqslant 1$,

$$\begin{aligned} |x_{r+s} - x_r| &\leqslant |x_{r+s} - x_{r+s-1}| + |x_{r+s-1} - x_{r+s-2}| + \cdots + |x_{r+1} - x_r| \\ &\leqslant (M^{r+s-2} + M^{r+s-1} + \cdots + M^{r-1})|x_2 - x_1| \\ &\leqslant M^{r-1}|x_2 - x_1|/(1 - M). \end{aligned}$$

Since $M < 1$, the right-hand side tends to zero as $r \to \infty$ and therefore so does

the left-hand side. But this is the standard Cauchy condition for the convergence of the sequence $\{x_r\}$ to a limit x_0. Because (1.52) implies that F is continuous, x_0 is a solution of (1.51).

To complete the proof it remains to show that there is no other root. Suppose there were another root y_0. Then

$$|y_0 - x_0| = |F(y_0) - F(x_0)| \leqslant M|y_0 - x_0|$$

from (1.52). On account of $M < 1$, the only possibility is $y_0 = x_0$ and the proof is terminated.

The disadvantage of Theorem 1.8*a* is that it needs the Lipschitz condition (1.52) to hold for all x and y. If we are prepared to assume that x_0 exists in some interval we can lighten this restriction.

THEOREM 1.8*b*. *Let* $x_0 = F(x_0)$ *and assume that* (1.52) *holds with* $M < 1$ *for all* x, y *in the interval* $[x_0 - a, x_0 + a]$ *for some* $a > 0$. *If* $x_0 - a < x_1 < x_0 + a$ *the iteration* (1.50) *has the properties*

(i) $x_0 - a < x_r < x_0 + a$,
(ii) $\lim_{r\to\infty} x_r = x_0$
(iii) $|x_{r+1} - x_0| \leqslant M^r|x_2 - x_1|/(1 - M)$.

The result (i) ensures that all iterates stay within the given interval while (ii) shows that the iteration converges to the root. An estimate of the distance of an iterate from the root is supplied by (iii).

Proof. Assume firstly that, for some r, $x_0 - a < x_r < x_0 + a$. Then

$$|x_{r+1} - x_0| = |F(x_r) - F(x_0)| \leqslant M|x_r - x_0| \tag{1.53}$$

from (1.52). Hence $|x_{r+1} - x_0| < a$. Therefore, if the result is true for r it is true for $r + 1$. Since $|x_1 - x_0| < a$, the validity of (i) follows by induction.

Inequality (1.53) implies that

$$|x_{r+1} - x_0| \leqslant M^r|x_1 - x_0|$$

whence $\lim_{r\to\infty} |x_{r+1} - x_0| = 0$ and (ii) is proved.

Further

$$|x_2 - x_0| = |F(x_1) - F(x_2) + F(x_2) - F(x_0)| \leqslant M|x_1 - x_2| + M|x_2 - x_0|$$

so that $|x_2 - x_0| \leqslant M|x_1 - x_2|/(1 - M)$. From (1.53) $|x_{r+1} - x_0| \leqslant M^{r-1}|x_2 - x_0|$ and the proof of the theorem is finished.

THEOREM 1.8*c*. *If* F *is continuous and differentiable on* $[x_0 - a, x_0 + a]$ *where* $x_0 = F(x_0)$, *and* $|F'(x)| \leqslant M < 1$ *then Theorem* 1.8*b holds and*

$$\lim_{r\to\infty} \frac{x_{r+1} - x_0}{x_r - x_0} = F'(x_0).$$

Proof. We have already seen that the differentiability of F entails the conditions of Theorem 1.8*b* so only the last part needs proof. Now

$$\lim_{r\to\infty}\frac{x_{r+1}-x_0}{x_r-x_0}=\lim_{r\to\infty}\frac{F(x_r)-F(x_0)}{x_r-x_0}=F'(x_0)$$

from the definition of a derivative and Theorem 1.8*b* (ii).

It should be remarked that Theorem 1.8*c* states that the iteration converges if $|F'|<1$ but this does not imply that the iteration diverges if $|F'|\geqslant 1$. In fact, we could permit $F'(x_0)=1$ without invalidating the theorem. More generally, if $x-F(x)>0$ and $F'(x)>0$ for $a+x_0\geqslant x>x_0$ then $a+x_0\geqslant x_r>x_0$ has the consequence

$$x_{r+1}=F(x_r)<x_r$$

while the mean value theorem

$$x_{r+1}-x_0=(x_r-x_0)F'(c_r),$$

with c_r between x_r and x_0, shows that $x_{r+1}>x_0$. Therefore, if $a+x_0\geqslant x_1>x_0$, induction demonstrates that $x_0<x_{r+1}<x_r$ for all r. Thus the sequence converges to a limit $L\geqslant x_0$. By continuity, $L=F(L)$ and so $L=x_0$. Thus the sequence converges to x_0.

Similarly, the conditions $F(x)-x>0$, $F'(x)>0$ for $x_0-a\leqslant x<x_0$ give a sequence converging to x_0 if $x_0-a\leqslant x_1<x_0$.

Newton's method for finding x_0 so that $f(x_0)=0$ may be derived in the following manner. Let x_r be an approximation to x_0. Then

$$f(x_0)=f(x_r)+(x_0-x_r)f'(x_r)+\tfrac{1}{2}(x_0-x_r)^2f''\{x_r+\theta(x_0-x_r)\} \quad (1.54)$$

where $0<\theta<1$. *If* x_r is a good approximation to x_0, x_0-x_r can be expected to be small and then, if f'' is not too large, the last term can be neglected, i.e.

$$f(x_0)\approx f(x_r)+(x_0-x_r)f'(x_r).$$

This will make $f(x_0)$ zero if

$$x_0-x_r=-f(x_r)/f'(x_r).$$

In other words, if x_r is an approximation to x_0, $x_r-f(x_r)/f'(x_r)$ should be a better one. Calling this new approximation x_{r+1} we have the iteration formula

$$x_{r+1}=x_r-\frac{f(x_r)}{f'(x_r)}. \quad (1.55)$$

Note that if x_r converges we expect its limit to be a zero of f if f' does not vanish there. In fact, the iteration will converge to a multiple zero as will be seen later.

Sometimes to simplify the computation $f'(x_r)$ is replaced by $f'(x_1)$ but we shall consider only the form (1.55).

The eqn (1.55) has the structure of (1.50) if

$$F(x) = x - f(x)/f'(x).$$

Hence

$$F'(x) = f(x)f''(x)/\{f'(x)\}^2$$

and Theorem 1.8*c* tells us that Newton's method converges to a simple zero of f if $|ff''/f'^2| < 1$ in a neighbourhood of the zero. Since f is small near zero, the basic assertion is that the method will converge if x_1 is close enough to the zero.

However, it must not be concluded that, if x_1 is closer to one zero than another, the iteration will necessarily converge to the nearby zero. For example, the iteration for

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

will converge to -1 if $x_1 = \frac{1}{4}$ even though x_1 is closer to 1 than -1.

A modification of Newton's method is *Cauchy's method* in which θ is placed equal to zero in (1.54). Then x_{r+1} is chosen as the root of

$$\tfrac{1}{2}(x_{r+1} - x_r)^2 f''(x_r) + (x_{r+1} - x_r)f'(x_r) + f(x_r) = 0$$

for which $x_{r+1} - x_r$ has the smallest modulus. The obvious disadvantage of Cauchy's method is that it requires the calculation of two derivatives as well as the solution of a quadratic equation.

An iteration scheme which is a generalization of the secant method is *Muller's method.* For this, three starting values, say x_1, x_2, and x_3, are necessary. Then one constructs a polynomial of degree 2 which has the values $f(x_1)$, $f(x_2)$, and $f(x_3)$ at x_1, x_2, and x_3 respectively. The polynomial has two zeros; choose the one x_4 for which $|x_4 - x_3|$ is smallest. Then repeat the process starting with x_2, x_3, and x_4. The polynomial always possesses a root unless $f(x_r) = f(x_{r+1}) = f(x_{r+2})$ when it represents a straight line parallel to the x-axis. Hence, provided that this situation is never met, the iteration can proceed.

The advantage of Muller's method over Newton's is that no computation of a derivative has to be undertaken. Also Muller's method offers the possibility of finding complex roots, which are excluded by Newton's method when f is real.

To discuss the convergence of an iterative process we say that, if

$$\lim_{r\to\infty} \frac{|x_{r+1} - x_0|}{|x_r - x_0|^p} = b$$

where b is finite and non-zero, *the iterative method is of order p.* If

$$\sup_{r\geq s} \frac{|x_{r+1} - x_0|}{|x_r - x_0|} = B$$

we have

$$|x_{r+s+1} - x_0| \leqslant B|x_{r+s} - x_0|^p$$
$$\leqslant B^{1+p}|x_{r+s-1} - x_0|^{p^2}$$

and, continuing in this, we obtain

$$|x_{r+s+1} - x_0| \leqslant B^c|x_{s+1} - x_0|^{p^r}$$

where

$$c = 1 + p + p^2 + \cdots + p^{r-1}.$$

If $p = 1$, $c = r$ and

$$|x_{r+s+1} - x_0| \leqslant B^r|x_{s+1} - x_0| \tag{1.56}$$

whereas, if $p \neq 1$, $c = (p^r - 1)/(p - 1)$ and

$$|x_{r+s+1} - x_0| \leqslant \frac{1}{B^{1/(p-1)}}\{B^{1/(p-1)}|x_{s+1} - x_0|\}^{p^r}. \tag{1.57}$$

It is evident that, if $p = 1$, convergence is relatively slow and only certain if $B < 1$. On the other hand, if $p > 1$ and

$$|x_{s+1} - x_0|B^{1/(p-1)} < 1$$

convergence will be very fast. Therefore iterative methods of higher order are to be preferred from the point of view of speed of convergence.

A theorem on the order of an iterative procedure is

THEOREM 1.8*d*. *Let* $\lim_{r\to\infty} x_r = x_0$ *where* $x_{r+1} = F(x_r)$ *and* F *is continuous on* $x_0 - a \leqslant x \leqslant x_0 + a$ $(a > 0)$.

(i) *If* $F'(x)$ *exists on* $x_0 - a < x < x_0 + a$ *and* $F'(x_0) \neq 0$, *the iterative method is of order* 1.

(ii) *If* $F'(x_0) = 0$ *and* $F''(x)$ *is continuous on* $x_0 - a < x < x_0 + a$, *then the iterative method is of order* 2 *if* $F''(x_0) \neq 0$.

Proof. As in Theorem 1.8*c*

$$\lim_{r\to\infty}\left|\frac{x_{r+1} - x_0}{x_r - x_0}\right| = |F'(x_0)|$$

so that, when $F'(x_0) \neq 0$, the method is of order 1.

In case (ii) $\lim_{r\to\infty} x_r = x_0$ implies that all x_r from some r onwards will certainly lie between $x_0 - a$ and $x_0 + a$. For such r Taylor's theorem gives

$$F(x_r) = F(x_0) + (x_r - x_0)F'(x_0) + \tfrac{1}{2}(x_r - x_0)^2F''(c_r)$$

where c_r is between x_r and x_0. Since $F'(x_0) = 0$,

$$\lim_{r\to\infty}\left|\frac{x_{r+1}-x_0}{(x_r-x_0)^2}\right| = \lim_{r\to\infty}\left|\frac{F(x_r)-F(x_0)}{(x_r-x_0)^2}\right|$$

$$= \lim_{r\to\infty}|\tfrac{1}{2}F''(c_r)|$$

$$= |\tfrac{1}{2}F''(x_0)|$$

because F'' is continuous and $c_r \to x_0$ since $x_r \to x_0$. Since $F''(x_0) \neq 0$ the proof of the theorem is complete.

In Newton's method $F'(x_0) = 0$ and $F''(x_0) = f''(x_0)/f'(x_0)$. Therefore, if $f''(x_0) \neq 0$, Newton's method is of order 2 for a simple root provided that f''' is continuous on an interval including x_0.

If x_0 is a q-fold root where $f(x_0)=f'(x_0)=\cdots=f^{(q-1)}(x_0)=0$ but $f^{(q)}(x_0)\neq 0$, Newton's method may still be shown to converge when $f^{(q)}$ is continuous in a neighbourhood of x_0. First, observe that

$$x_{r+1} - x_0 = \frac{(x_r - x_0)f'(x_r) - f(x_r)}{f'(x_r)}.$$

By Taylor's theorem

$$f(x_r) = (x_r - x_0)^q f^{(q)}(\xi_1)/q!,$$
$$f'(x_r) = (x_r - x_0)^{q-1} f^{(q)}(\xi_2)/(q-1)!$$

where both ξ_1 and ξ_2 lie between x_r and x_0. Hence

$$x_{r+1} - x_0 = (x_r - x_0)\left\{1 - \frac{f^{(q)}(\xi_1)}{q f^{(q)}(\xi_2)}\right\}.$$

As $x_r \to x_0$, $\xi_1 \to x_0$ and $\xi_2 \to x_0$ so that, from the continuity of $f^{(q)}$,

$$x_{r+1} - x_0 \approx (x_r - x_0)(1 - 1/q).$$

This demonstrates that the convergence is much slower than in the case of a simple root and can be very slow indeed if q is large.

For a multiple root the convergence of Newton's method can be improved by adopting the formula

$$x_{r+1} = x_r - qf(x_r)/f'(x_r). \tag{1.58}$$

Using the same technique as just above but taking one extra term in the Taylor expansions we obtain

$$x_{r+1} - x_0 = -\frac{(x_r - x_0)^2}{q+1}\frac{f^{(q+1)}(x_0)}{f^{(q)}(x_0)}$$

so that the method is of order 2. However, one should be warned that if (1.58) is employed near a simple root convergence may fail.

It can be demonstrated that the secant method is of order 1.62 approximately and Muller's method of order 1.84 approximately.

A standard scheme for accelerating the convergence of an iteration procedure is *Aitken's δ^2-method.* In this method, starting from x_r, we generate $y_{r+1} = F(x_r)$, $y_{r+2} = F(y_{r+1})$ and then define

$$x_{r+1} = y_{r+2} - \frac{(y_{r+2} - y_{r+1})^2}{y_{r+2} + x_r - 2y_{r+1}}.$$

Analysis reveals that this scheme is of order 2 if $F'(x_0) \neq 1$ and neither $F'(x_0)$ nor $F''(x_0)$ is zero. If $F'(x_0) = 1$ the scheme is of order 1.

Exercises

28. Use the bisection method to solve
    (i) $8x^3 - 4x - 5 = 0$,
    (ii) $2x = \tan x$,
    correct to two decimal places.
29. On $0 \leqslant x \leqslant \frac{1}{2}$, $f(x) = \frac{1}{2}$ and on $\frac{1}{2} \leqslant x \leqslant 1$,

$$f(x) = 6x - 1 - 6x^2.$$

    Obtain the value of c_{r+1} in the method of false position.
30. Solve $3 \sin x = 2$ correct to three decimal places by the secant method.
31. Use Newton's method to find $\sqrt{7}$ correct to 2 decimal places, starting from $x_1 = 3$.
32. Obtain by Newton's method a root of
    (i) $x^3 - 2x^2 - 5x + 10 = 0$, starting from $x_1 = 3$,
    (ii) $x^3 - 6x^2 + 13x - 9 = 0$, starting from $x_1 = 2$.
33. Find the root of $27x^3 + 18x - 25 = 0$ between 0 and 1 using the iteration

$$x_{r+1} = \tfrac{1}{3}(25 - 18x_r)^{1/3},$$

    checking whether Theorem 1.8*c* is satisfied. Is the iteration

$$x_{r+1} = (15 - 27x_r^3)/18$$

    better?
34. Examine the iterations
    (i) $x_{r+1} = (x_r^2 + c)/b$,
    (ii) $x_{r+1} = b - (c/x_r)$
    as possible schemes for determining the larger root of $x^2 - bx + c = 0$ when $b > 0$, $\frac{1}{4}b^2 > c > 0$.
35. What happens when Newton's method is applied to $x^2 - 2x + 2 = 0$?
36. Solve $x^3 = 3$ by Cauchy's method starting from $x_1 = 3$.
37. Find a root of $\sin x + 2 = x$ by Muller's method starting with $x_1 = -1$, $x_2 = 0$, $x_3 = 1$.
38. If $F(x) = x + h(x)f(x)$ find h so that the iteration method is of order 2.

39. To calculate $\sqrt{a}$ when $a > 0$ the following iteration is suggested:

$$x_{r+1} = \frac{x_r^3 + 3ax_r}{3x_r^2 + a}.$$

Show that it is of order 3.

1.9 Systems of non-linear equations

The solution of simultaneous non-linear equations is complicated and we shall be content to describe how Newton's method can be generalized. Suppose the values of x and y are required which simultaneously satisfy

$$f(x, y) = 0, \quad g(x, y) = 0.$$

By Taylor's theorem, if we neglect second orders,

$$f(x_{r+1}, y_{r+1}) = f(x_r, y_r) + (x_{r+1} - x_r)f_x + (y_{r+1} - y_r)f_y,$$

$$g(x_{r+1}, y_{r+1}) = g(x_r, y_r) + (x_{r+1} - x_r)g_x + (y_{r+1} - y_r)g_y$$

where f_x denotes the partial derivative $\partial f/\partial x$ and all the partial derivatives are evaluated at (x_r, y_r). If we hope that (x_{r+1}, y_{r+1}) is close to a zero we want the left-hand sides to be zero. This can be arranged by putting

$$x_{r+1} = x_r + (gf_y - fg_y)/J, \tag{1.59}$$

$$y_{r+1} = y_r + (fg_x - gf_x)/J \tag{1.60}$$

where J is the *Jacobian* defined by

$$J = f_x g_y - f_y g_x.$$

Eqns (1.59) and (1.60) constitute the generalization of Newton's method to two equations, all quantities on the right-hand side being calculated at (x_r, y_r).

MATRICES

1.10 Matrices

It is assumed that the reader has some acquaintance with the theory of matrices so that the treatment here will be somewhat cursory (see, for example, Liebeck (1969)). A general matrix consists of mn entries arranged in m rows and n columns, giving an $m \times n$ array, to be denoted by a capital letter such as A:

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ & & \cdots & \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}.$$

The symbol a_{ij} denotes the element in the ith row and jth column and often we shall abbreviate the notation by writing $A = (a_{ij})$.

The matrix is called *square* and of *order* n if $m = n$. If $n = 1$ so that the matrix consists of a single column we shall call the matrix a *column vector* and signify its special nature by using bold type, e.g.

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix}.$$

The elements a_{ii} for $i = 1, 2, \ldots, n$ in a square matrix are said to be the *diagonal elements.*

The elementary rules of combination are:

$A = B$ if and only if $a_{ij} = b_{ij}$ all i, j
$A + B = (a_{ij} + b_{ij})$,
$\alpha A = (\alpha a_{ij})$.

Multiplication of A and B is possible only if A has the same number of columns as B has rows. If A is $m \times n$ and B is $n \times p$ then

$$AB = \left(\sum_{k=1}^{n} a_{ik} b_{kj} \right)$$

the result being an $m \times p$ matrix. In general, two matrices do not commute, i.e. $AB \neq BA$ even if both are square.

The *unit matrix* I of order n is a square matrix all of whose elements are zero except the diagonal ones which are unity. Thus $AI = A$.

The *transpose* of a $m \times n$ matrix $A = (a_{ij})$ is the $n \times m$ matrix whose ijth element is a_{ji}. The symbol A^{T} will be used to indicate the transpose. Note that the transpose $\mathbf{a}^{\mathrm{T}}$ of a column matrix will be a *row matrix*, i.e. a matrix whose elements lie in a single row. There is no difficulty in verifying that

$$(A + B)^{\mathrm{T}} = A^{\mathrm{T}} + B^{\mathrm{T}}, \quad (A^{\mathrm{T}})^{\mathrm{T}} = A, \quad (AB)^{\mathrm{T}} = B^{\mathrm{T}} A^{\mathrm{T}}.$$

If A is $m \times n$ and $\mathbf{x}$ is a column matrix with n elements $A\mathbf{x}$ is a column matrix whose ith element is

$$\sum_{j=1}^{n} a_{ij} x_j.$$

Observe that $x^{\mathrm{T}} A^{\mathrm{T}}$ is a row matrix. If B is a $n \times m$ matrix such that $BA = I$ then B is called a *left-inverse* of A. Similarly, if C is $n \times m$ and $AC = I$ then C is called a *right-inverse* of A. Suppose A is square and has both a left-inverse and a right-inverse then

$$B = BI = (B(AC) = (BA)C = IC = C.$$

Thus there is only one left-inverse and only one right-inverse and both are equal. This unique matrix is called the *inverse* of A and denoted by A^{-1}. Clearly,

$$(A^{-1})^{-1} = A, \quad (AB)^{-1} = B^{-1}A^{-1}$$

but, in general, $(A + B)^{-1} \neq A^{-1} + B^{-1}$.

A matrix is called *symmetric* if $A = A^{\mathrm{T}}$ and *anti-symmetric* if $A = -A^{\mathrm{T}}$. A matrix such that $A^{-1} = A^{\mathrm{T}}$ is known as *orthogonal.*

From now on we shall be concerned primarily with square matrices A. It will therefore be assumed that A is square and of order n unless otherwise is specifically stated.

It is known that the equations

$$A\mathbf{x} = \mathbf{0}$$

possess a solution with $\mathbf{x} \neq \mathbf{0}$ if and only if $\det A = 0$, where det signifies the determinant of the matrix.

The quantities λ_i such that

$$A\mathbf{x}_i = \lambda_i \mathbf{x}_i \tag{1.61}$$

has solutions $x_i \neq 0$ are called the *eigenvalues* of A. The λ_i are solutions of

$$\det(A - \lambda I) = 0$$

and are therefore n in number, though some of them may be multiple roots. Since the determinant of the transpose of a matrix is the same as the determinant of the original matrix

$$\det(A^{\mathrm{T}} - \lambda I) = 0.$$

Consequently, there are $\mathbf{y}_j \neq \mathbf{0}$ such that

$$A^{\mathrm{T}}\mathbf{y}_j = \lambda_j \mathbf{y}_j \tag{1.62}$$

Hence A and A^{T} have the same eigenvalues.

Multiply (1.61) by $\mathbf{y}_j^{\mathrm{T}}$ and (1.62) by $\mathbf{x}_i^{\mathrm{T}}$ and subtract. Then

$$\mathbf{y}_j^{\mathrm{T}} A\mathbf{x}_i - \mathbf{x}_i^{\mathrm{T}} A^{\mathrm{T}}\mathbf{y}_j = \lambda_i \mathbf{y}_j^{\mathrm{T}}\mathbf{x}_i - \lambda_j \mathbf{x}_i^{\mathrm{T}}\mathbf{y}_j.$$

The left-hand side vanishes and so

$$(\lambda_i - \lambda_j)\mathbf{y}_j^{\mathrm{T}}\mathbf{x}_i = 0.$$

If $\lambda_i \neq \lambda_j$ then $\mathbf{y}_j^{\mathrm{T}}\mathbf{x}_i = 0$, i.e. the eigenvectors of A and A^{T} corresponding to distinct eigenvalues are orthogonal.

Moreover, the eigenvectors corresponding to *distinct* eigenvalues of A are linearly independent. Suppose, to the contrary, that s are linearly dependent and that any smaller number are linearly independent. Then

$$\alpha_1\mathbf{x}_1 + \cdots + \alpha_s\mathbf{x}_s = \mathbf{0} \tag{1.63}$$

where all the α_i are non-zero. On multiplying by A we obtain

$$\alpha_1\lambda_1\mathbf{x}_1 + \cdots + \alpha_s\lambda_s\mathbf{x}_s = \mathbf{0}.$$

If $\lambda_1 = 0$, $s-1$ vectors would be linearly dependent contrary to our hypothesis. If $\lambda_1 \neq 0$ multiply (1.63) by λ_1 and subtract; then

$$\alpha_2(\lambda_2 - \lambda_1)\mathbf{x}_2 + \cdots + \alpha_s(\lambda_s - \lambda_1)\mathbf{x}_s = \mathbf{0}.$$

Since $\lambda_i - \lambda_1 \neq 0$ for $i = 2, \ldots, s$ this gives a linear relation between $s-1$ vectors. Again, a contradiction occurs and the statement is proved.

One consequence is that, if A has n distinct eigenvalues, $\mathbf{y}_i^{\mathrm{T}}\mathbf{x}_i \neq 0$. For, if this were not true, $\mathbf{y}_i$ would be orthogonal to the n independent vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ which is impossible because $\mathbf{y}_i \neq \mathbf{0}$. It is therefore always possible to select $\mathbf{y}_i$ so that $\mathbf{y}_i^{\mathrm{T}}\mathbf{x}_i = 1$.

Moreover, if A has n distinct eigenvalues, define X as the matrix with columns $\mathbf{x}_i$, i.e.

$$X = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n).$$

Then, with $\mathbf{y}_i$ picked so that $\mathbf{y}_i^{\mathrm{T}}\mathbf{x}_i = 1$,

$$X^{-1} = \begin{pmatrix} \mathbf{y}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{y}_n^{\mathrm{T}} \end{pmatrix}$$

because of the orthogonal relations. Hence

$$\begin{aligned} X^{-1}AX &= X^{-1}(\lambda_1\mathbf{x}_1, \lambda_2\mathbf{x}_2, \ldots, \lambda_n\mathbf{x}_n) \\ &= \mathrm{diag}(\lambda_i) \end{aligned} \tag{1.64}$$

where diag is used to denote a *diagonal matrix*, i.e. a matrix whose non-diagonal elements are all zero.

Two matrices A and B are said to be *similar* if there is a *non-singular* matrix R (i.e. $\det R \neq 0$) such that $B = R^{-1}AR$. Sometimes, A is said to have undergone a *similarity transformation*. The eigenvaues of similar matrices are the same because $A\mathbf{x} = \lambda\mathbf{x}$ can be written as

$$(R^{-1}AR)R^{-1}\mathbf{x} = \lambda R^{-1}\mathbf{x}$$

showing that $R^{-1}\mathbf{x}$ is an eigenvector of $R^{-1}AR$.

What has been demonstrated above is, if A has n distinct eigenvalues, that A is similar to a diagonal matrix whose entries are the eigenvalues of A. If A is also symmetric then $\mathbf{y}_i = \mathbf{x}_i$ and $X^{-1} = X^{\mathrm{T}}$ so that, in this case, there is an orthogonal similarity transformation converting A to diagonal form.

If A is symmetric with multiple eigenvalues it can be shown that there is still an orthogonal similarity transformation which changes A to $\mathrm{diag}(\lambda_i)$. If, however, A has multiple eigenvalues but is not symmetric the situation is more complicated. What can be demonstrated is that there is a non-singular R such that

$$R^{-1}AR = J \tag{1.65}$$

where J is the *Jordan canonical form* of A and has the following structure: J is a *block-diagonal* matrix

$$J = \begin{pmatrix} J_1 & & & 0 \\ & J_2 & & \\ & & \ddots & \\ 0 & & & J_k \end{pmatrix}$$

where each J_i is either the number λ_i or a matrix of the form

$$J_i = \begin{pmatrix} \lambda_i & 1 & & & 0 \\ & \lambda_i & 1 & & \\ & & \ddots & 1 & \\ & & & \ddots & 1 \\ 0 & & & & \lambda_i \end{pmatrix} \tag{1.66}$$

which is an *upper triangular matrix* since all the elements below the diagonal are zero. The Jordan canonical form is the most compact to which a general matrix can be reduced by a similarity transformation. The same eigenvalue may occur in different J_i, but the total number of times that a given eigenvalue occurs in the diagonal of J is the same as the multiplicity of the eigenvalue. The number of linearly independent eigenvectors of A is k, i.e. the number of Jordan blocks in the canonical form. In particular, if J_i is $m_i \times m_i$ and $\mathbf{r}_i$ is the ith column of R then $\mathbf{r}_1, \mathbf{r}_{m_1+1}, \ldots, \mathbf{r}_{m_1+\cdots m_{k-1}+1}$ are the eigenvectors of A.

If the elements of A are changed continuously, then $\det(A - \lambda I)$ varies continuously and so the eigenvalues of A change continuously. In general, however, the eigenvectors do not alter continuously.

If $p(t) = a_0 + a_1 t + \cdots + a_m t^m$ is a polynomial in t, a corresponding *matrix polynomial* $p(A)$ can be defined by

$$p(A) = a_0 + a_1 A + \cdots + a_m A^m$$

where, of course, $A^r = AA^{r-1}$. It is immediate that any eigenvector of A is an eigenvector of $p(A)$ with eigenvalue $p(\lambda_i)$. If A has an inverse the eigenvalues of A^{-1} are $1/\lambda_i$.

If the elements of A are complex, the matrix A^* is obtained by replacing each element of A by its complex conjugate. Write $A^{\mathrm{H}} = A^{*\mathrm{T}}$. Then a matrix is said to be *unitary* if $A^{\mathrm{H}}A = I$. It is called *Hermitian* if $A^{\mathrm{H}} = A$. Note that the Hermitian matrices include the real symmetric matrices.

If λ is the eigenvalue of a Hermitian matrix A so that $A\mathbf{x} = \lambda\mathbf{x}$ then

$$\mathbf{x}^{\mathrm{H}}A\mathbf{x} = \lambda\mathbf{x}^{\mathrm{H}}\mathbf{x}. \tag{1.67}$$

Now $(\mathbf{x}^H A\mathbf{x})^H = (\mathbf{x}^T A^* \mathbf{x}^*)^T = \mathbf{x}^H A^H \mathbf{x} = \mathbf{x}^H A\mathbf{x}$ so that $\mathbf{x}^H A\mathbf{x}$ is real. Since $\mathbf{x}^H\mathbf{x}$ is real and non-zero it follows that λ is real, i.e. *the eigenvalues of a Hermitian matrix are real.* Furthermore, if $\mathbf{x}_i$ and $\mathbf{x}_j$ are two eigenvectors,

$$\lambda_i \mathbf{x}_j^H \mathbf{x}_i = \mathbf{x}_j^H A\mathbf{x}_i = (\mathbf{x}_i^H A\mathbf{x}_j)^H = \lambda_j (\mathbf{x}_i^H x_j)^H = \lambda_j \mathbf{x}_j^H \mathbf{x}_i$$

from which we deduce that $\mathbf{x}_j^H \mathbf{x}_i = 0$, i.e. the vectors are orthogonal, if $\lambda_i \neq \lambda_j$.

It can be shown that if A is Hermitian there is a unitary matrix U such that

$$U^H A U = \text{diag}(\lambda_i). \tag{1.68}$$

Moreover, the eigenvectors can be arranged to be mutually orthogonal. Consequently, any vector $\mathbf{y}$ can be expressed in the form

$$\mathbf{y} = \sum_{i=1}^{n} a_i \mathbf{x}_i.$$

Hence

$$\mathbf{y}^H A\mathbf{y} = \mathbf{y}^H \sum_{i=1}^{n} a_i \lambda_i \mathbf{x} = \sum_{i=1}^{n} \lambda_i |a_i|^2$$

since $\mathbf{x}_j^H \mathbf{x}_i = 0$ $(i \neq j)$ and the magnitude may be made to satisfy $\mathbf{x}_i^H \mathbf{x}_i = 1$. Put the eigenvalues in order so that $\lambda_1 \leqslant \lambda_2 \leqslant \cdots \leqslant \lambda_n$. Then

$$\lambda_n \sum_{i=1}^{n} |a_i|^2 \geqslant \mathbf{y}^H A\mathbf{y} \geqslant \lambda_1 \sum_{i=1}^{n} |a_i|^2.$$

In other words, for arbitrary $\mathbf{y}$,

$$\lambda_n \mathbf{y}^H \mathbf{y} \geqslant \mathbf{y}^H A\mathbf{y} \geqslant \lambda_1 \mathbf{y}^H \mathbf{y} \tag{1.69}$$

when A is Hermitian and λ_1, λ_n are the least and largest eigenvalues respectively of A.

A Hermitian matrix is said to be *positive definite* if

$$\mathbf{x}^H A\mathbf{x} > 0$$

for every $\mathbf{x} \neq \mathbf{0}$ and *positive semi-definite* if

$$\mathbf{x}^H A\mathbf{x} \geqslant 0$$

for every $\mathbf{x} \neq \mathbf{0}$. A deduction from (1.67) is that a Hermitian matrix is positive definite if, and only if, all its eigenvalues are positive. It is positive semi-definite if and only if all its eigenvalues are non-negative.

A measure of the eigenvalues of a matrix is provided by the *trace* Tr A defined by

$$\text{Tr } A = a_{11} + a_{22} + \cdots + a_{nn}.$$

Obviously

$$\text{Tr}(kA) = k \text{ Tr } A, \tag{1.70}$$

$$\text{Tr}(A + B) = \text{Tr } A + \text{Tr } B. \tag{1.71}$$

Also

$$\text{Tr}(AB) = \sum_{i=1}^{n} \sum_{k=1}^{n} a_{ik} b_{ki} = \sum_{k=1}^{n} \sum_{i=1}^{n} b_{ki} a_{ik} = \text{Tr}(BA). \tag{1.72}$$

A deduction from (1.72) is that

$$\mathrm{Tr}(R^{-1}AR) = \mathrm{Tr}(ARR^{-1}) = \mathrm{Tr}\, A.$$

It therefore follows from the Jordan canonical form (1.65) and (1.66) that

$$\mathrm{Tr}\, A = \lambda_1 + \lambda_2 + \cdots + \lambda_n. \tag{1.73}$$

Exercises

40. Find the eigenvalues, eigenvectors, and Jordan canonical form of

(i) $\begin{pmatrix} 1 & 2 \\ 4 & 3 \end{pmatrix}$ (ii) $\begin{pmatrix} -1 & 0 & 0 \\ 1 & 0 & -1 \\ 1 & 1 & -2 \end{pmatrix}$

41. If A and B are symmetric prove that AB is symmetric if and only if $AB = BA$.
42. Show that A and A^{T} are similar.
43. If $\det A \neq 0$ prove that $A^{\mathrm{H}}A$ is positive definite.
44. Prove that the eigenvalues of $A^m(A + \mu I)^{-1}$ are $\lambda_i^m(\lambda_i + \mu)^{-1}$ given $\mu \neq -\lambda_i$ for any i.
45. Prove that the eigenvalues of

$$\begin{pmatrix} 1 + \dfrac{1}{2^r}\cos 2^{r+1} & -2^r \sin 2^{r+1} \\ -2^r \sin 2^{r+1} & 1 - \dfrac{1}{2^r}\cos 2^{r+1} \end{pmatrix}$$

are 1 ± 2^{-r}. Deduce that the eigenvalues tend to 1 as $r \to \infty$ but that the eigenvectors do not have a limit.

46. A is real positive semi-definite and R is an orthogonal matrix such that $R^{\mathrm{T}}AR = \mathrm{diag}(\lambda_i)$. If $B = \mathrm{diag}(\sqrt{\lambda_i})$ and $C = RBR^{\mathrm{T}}$ prove that $C^2 = A$ so that a *square root* of A may be defined as $A^{1/2} = C$.
47. Show that the Hermitian matrix A is positive semi-definite if and only if there is a matrix B such that $A = BB^{\mathrm{H}}$.
48. Show that (i) $\mathrm{Tr}(AA^{\mathrm{H}}) > 0$, (ii) if A is anti-symmetric $\mathrm{Tr}\, A = 0$.

1.11 Matrix norms

The modulus of a complex number gives an idea of its size and it is desirable to have a single number which plays a similar role for matrices and vectors. This quantity will be known as a *norm* (see also §1.5). We define the norm in terms of its properties, and not by means of a specific formula. In this way it is possible to define many different kinds of norm associated with a vector. In fact, any formula for the norm $\|\mathbf{x}\|$ of a vector $\mathbf{x}$ will be acceptable if it has the properties

(a) $\|\mathbf{x}\| > 0$ if $\mathbf{x} \neq \mathbf{0}$; $\|\mathbf{x}\| = 0$ only if $\mathbf{x} = \mathbf{0}$;
(b) $\|\alpha\mathbf{x}\| = |\alpha|\,\|\mathbf{x}\|$ for any complex number α;
(c) $\|\mathbf{x} + \mathbf{y}\| \leqslant \|\mathbf{x}\| + \|\mathbf{y}\|$.

If $\mathbf{x}$ has elements $x_1, \ldots, x_n$ standard norms are the l_p-*norms* defined by

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} \qquad (1 \leqslant p < \infty).$$

The l_∞-*norm* is defined by

$$\|\mathbf{x}\|_\infty = \max_{1 \leqslant i \leqslant n} |x_i|$$

and corresponds to the uniform norm we have already considered.

Norms can always be obtained from inner products as we have seen in §1.5 and we now take this opportunity to define an inner product formally. An inner product (x, y) is required to satisfy

(a) $(x, x) > 0$ if $x \neq 0$; $(x, x) = 0$ only if $x = 0$;
(b) $(x, y) = (y, x)^*$;
(c) $(x + y, z) = (x, z) + (y, z)$, $(\alpha x, y) = \alpha(x, y)$.

An inner product supplies a norm via $\|x\| = (x, x)^{1/2}$ and the Schwarz inequality

$$|(x, y)| \leqslant \|x\| \, \|y\|$$

always holds.

The l_2-norm is often known as the *Euclidean norm* since it stems from the inner product $\mathbf{x}^H\mathbf{y}$. Note that in inner product notation

$$\mathbf{x}^H A\mathbf{y} = (\mathbf{x}, A\mathbf{y}) = (A^H\mathbf{x}, \mathbf{y}).$$

A matrix norm can also be introduced by asking that it has the properties

(a) $\|A\| > 0$ if $A \neq 0$; $\|A\| = 0$ only if $A = 0$;
(b) $\|\alpha A\| = |\alpha| \, \|A\|$ for any complex number α;
(c) $\|A + B\| \leqslant \|A\| + \|B\|$;
(d) $\|AB\| \leqslant \|A\| \, \|B\|$.

If $\| \ \|'$ is a matrix norm and $\| \ \|$ is a vector form, the matrix and vector norms are said to be *compatible* if

$$\|A\mathbf{x}\| \leqslant \|A\|' \|\mathbf{x}\| \tag{1.74}$$

A matrix norm can be constructed from a vector norm by defining

$$\|A\| = \sup_{\|\mathbf{x}\| = 1} \|A\mathbf{x}\|; \tag{1.75}$$

such a matrix norm is said to be *subordinate* to the given vector norm. It is obvious that the subordinate norm is compatible. From (1.75) can be seen by putting $A = I$ that any subordinate norm has the property $\|I\| = 1$.

From now on the only matrix norm which will be considered is the one

specified by (1.75). Corresponding to the vector l_p-norms we have

$$\|A\|_1 = \max_{1 \leqslant j \leqslant n} \sum_{i=1}^{n} |a_{ij}|, \tag{1.76}$$

$$\|A\|_\infty = \max_{1 \leqslant i \leqslant n} \sum_{j=1}^{n} |a_{ij}|, \tag{1.77}$$

$$\|A\|_2 = \sqrt{\mu} \tag{1.78}$$

where μ is the largest eigenvalue of $A^{\mathrm{H}}A$. Sometimes $\|A\|_2$ is known as the *spectral norm* of A. From (1.76) and (1.77) $\|A\|_1 = \|A^{\mathrm{T}}\|_\infty$.

To prove these results we remark that

$$\|A\mathbf{x}\|_1 = \sum_{i=1}^{n} \left| \sum_{j=1}^{n} a_{ij}x_j \right| \leqslant \sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}||x_j| \leqslant \sum_{j=1}^{n} |x_j| \sum_{i=1}^{n} |a_{ij}|$$

$$\leqslant \left(\max_{1 \leqslant j \leqslant n} \sum_{i=1}^{n} |a_{ij}| \right) \|\mathbf{x}\|_1.$$

This inequality shows that the norm certainly does not exceed the value given in (1.76). Moreover, if

$$\sum_{i=1}^{n} |a_{ij}|$$

is largest when $j = k$ choose $x_i = 0$ $(i \neq k)$, $= 1$ $(i = k)$ and then the value in (1.76) is actually attained and so (1.76) is proved.

The proof for $\|A\|_\infty$ is similar except that in the last stage, if

$$\sum_{j=1}^{n} |a_{ij}|$$

is greatest for $i = k$, we choose $x_i = a_{ki}/|a_{ki}|$ $(a_{ki} \neq 0)$, $= 1$ $(a_{ki} = 0)$ to achieve the supremum.

For $\|A\|_2$ we remark that $\|A\mathbf{x}\|_2 = (\mathbf{x}^{\mathrm{H}}A^{\mathrm{H}}A\mathbf{x})^{1/2}$ and the result follows from (1.69).

If A is Hermitian, (1.78) implies that

$$\|A\|_2 = \max_{1 \leqslant i \leqslant n} |\lambda_i|.$$

The *spectral radius* $\rho(A)$ of a matrix is defined by

$$\rho(A) = \max_{1 \leqslant i \leqslant n} |\lambda_i|.$$

Thus $\|A\|_2 = \{\rho(A^{\mathrm{H}}A)\}^{1/2}$ which simplifies, if A is Hermitian, to $\|A\|_2 = \rho(A)$. In general, if $\mathbf{x}$ is an eigenvector of A,

$$\|A\mathbf{x}\| = \|\lambda\mathbf{x}\| = |\lambda|\|\mathbf{x}\|$$

which demonstrates via (1.75) that $|\lambda| \leqslant \|A\|$, i.e.

$$\rho(A) \leqslant \|A\| \tag{1.79}$$

for any norm of A. However, if the norms are badly chosen the norm and spectral radius need not be close; for example

$$\left\| \begin{pmatrix} 0 & 10 \\ 0 & 0 \end{pmatrix} \right\|_1 = 10$$

but the spectral radius is zero. In contrast, it can be shown that there is always some norm which is arbitrarily close to the spectral radius.

A useful theorem is:

THEOREM 1.11. $\lim_{r \to \infty} A^r = 0$ *if and only if* $\rho(A) < 1$.

Proof. It is evident from the Jordan canonical form that A^r can approach the zero matrix if and only if each of its eigenvalues tends to zero. But its eigenvalues are λ_i^r which can vanish as $r \to \infty$ if and only if $|\lambda_i| < 1$ and the theorem is proved.

THEOREM 1.11*a*. *If* $\rho(A) < 1$ *then* $(I - A)^{-1}$ *exists and*

$$(I - A)^{-1} = \lim_{m \to \infty} \sum_{i=0}^{m} A^i.$$

Proof. Since $\rho(A) < 1$, $I - A$ has no zero eigenvalues and so possesses an inverse. Also

$$(I - A)(I + A + \cdots + A^{m-1}) = I - A^m$$

and so

$$I + A + \cdots + A^{m-1} = (I - A)^{-1} - (I - A)^{-1} A^m.$$

The result now follows from Theorem 1.11 by letting $m \to \infty$.

It will be remarked that a sufficient condition for the validity of Theorems 1.11 and 1.11*a* is that $\|A\| < 1$, on account of (1.79).

A matrix A is called *strictly diagonal dominant* if

$$\sum_{j=1}^{n}{}' |a_{ij}| < |a_{ii}| \qquad (i = 1, \ldots, n)$$

where the prime on $\sum$ means omit the term $j = i$. The importance of this concept arises from:

THEOREM 1.11*b*. *If* A *is strictly diagonal dominant then* A^{-1} *exists.*

Proof. Suppose there is $\mathbf{x} \neq \mathbf{0}$ such that $A\mathbf{x} = \mathbf{0}$. Let $x_m = \max_{1 \leqslant i \leqslant n} |x_i|$.

Then, from

$$\sum_{j=1}^{n} a_{mj}x_j = 0,$$

we obtain

$$|a_{mm}||x_m| = \left|\sum_{j\neq m} a_{mj}x_j\right| \leqslant |x_m| \sum_{j\neq m} |a_{mj}|$$

which contradicts the condition of strict diagonal dominance.

We can now prove:

THEOREM 1.11*c* (GERSCHGORIN CIRCLE THEOREM). *Every eigenvalue of A lies in one of the complex domains*

$$|z - a_{ii}| \leqslant \sum_{j=1}^{n}{}' |a_{ij}| \quad (i = 1, 2, \ldots, n).$$

Proof. Let λ be any eigenvalue which does not lie in one of these domains. Then

$$|\lambda - a_{ii}| > \sum_{j=1}^{n}{}' |a_{ij}| \qquad (i = 1, \ldots, n).$$

Hence $A - \lambda I$ is strictly diagonal dominant and hence, by Theorem 1.11*b*, has an inverse. But this is impossible because λ is an eigenvalue and the theorem is proved.

One consequence of Gerschgorin's theorem is that

$$\rho(A) \leqslant \min\left(\max_i \sum_j |a_{ij}|; \max_j \sum_i |a_{ij}|\right).$$

Gerschgorin's theorem and (1.69) provide rules for locating the positions of eigenvalues. While they may not always be very precise they do at any rate limit the possibilities.

A type of matrix often encountered with difference equations is a *Stieltjes matrix* which is a real positive definite matrix with all its off-diagonal elements non-positive.

Exercises

50. Prove that $\|\mathbf{x}\|_1 \leqslant n\|\mathbf{x}\|_\infty$ and $\|\mathbf{x}\|_2 \leqslant \sqrt{n}\|\mathbf{x}\|_\infty$.
51. If U is unitary prove that (i) $\|U\mathbf{x}\|_2 = \|\mathbf{x}\|_2$, (ii) $\|UA\|_2 = \|A\|_2$, (iii) $\|UAU^{\mathrm{H}}\|_2 = \|A\|_2$.
52. If a is real and $A = \begin{pmatrix} a & 4 \\ 0 & a \end{pmatrix}$ show that

$$\|A^r\|_2 = a^r\left[1 + \frac{8r^2}{a^2}\left\{1 + \left(1 + \frac{a^2}{4r^2}\right)^{1/2}\right\}\right]^{1/2}.$$

53. Prove that $\max|a_{ij}| \leqslant \|A\|_2 \leqslant n \max|a_{ij}|$, the maximum being taken over all i and j.
54. Prove $\|A\|_2^2 \leqslant \|A\|_1 \|A\|_\infty$.
55. If a is real show that the spectral radius and spectral norm of

$$\begin{pmatrix} 1 & 0 & \alpha \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

are equal only if $a = 0$.

56. Show that (i) $[\rho\{(A^H A)^{-1}\}]^{-1} \leqslant |\lambda_i|^2 \leqslant \rho(A^H A)$, (ii) $\sum_{i=1}^n |\lambda_i|^2 \leqslant \{\sum_{i=1}^n \sum_{j=1}^n |a_{ij}|^2\}^{1/2}$.
57. For the *tridiagonal* matrix

$$\begin{matrix} a & b & 0 & 0 & \cdots & 0 \\ c & a & b & 0 & \cdots & 0 \\ 0 & c & a & b & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & c & a & b & \\ 0 & \cdots & 0 & c & a & \end{matrix}$$

show that

$$\lambda_j = a + 2(bc)^{1/2} \cos\{j\pi/(n+1)\}.$$

Check that this satisfies Gerschgorin's theorem.

58. In the tridiagonal matrix

$$\begin{pmatrix} a_1 & d_1 & \cdots & 0 \\ d_1 & a_2 & \cdots & 0 \\ & & \cdots & \\ & & \cdots & d_{n-1} \\ 0 & \cdots & d_{n-1} & a_n \end{pmatrix}$$

$d_i = (b_i c_i)^{1/2}$ where $b_i c_i > 0$, $a_i \geqslant |b_i| + |c_{i-1}|$ $(i = 2, \ldots, n-1)$, $a_1 > |b_1|$ and $a_n > |c_{n-1}|$. Prove that it is positive definite.

LINEAR EQUATIONS

1.12 Linear equations—direct methods

The solution of the system of linear equations $A\mathbf{x} = \mathbf{b}$ where A is non-singular is simple in principle. In fact, the solution can be written as $\mathbf{x} = A^{-1}\mathbf{b}$. When A^{-1} can be calculated easily by analytical means this is often satisfactory. However, for numerical work, it must be recognized that simple analytical formulae for A^{-1} usually involve the ratio of determinants and the computation of determinants of order 4 or higher is a complex task. Therefore, if we are to realize efficient numerical methods we must seek other ways of finding a solution.

In practice, the systems of equations which arise are frequently of two types:

(i) The matrix A may be of moderate order, say $n < 100$, and *dense*, i.e. nearly all its elements are non-zero;

(ii) A may be of large order, say $n > 1000$, and *sparse*, i.e. it contains a large number of zero elements.

The type of matrix is of prime importance in deciding on a method of solution. For dense matrices, direct methods are appropriate and will be described in this section. Sparse matrices should be treated by iterative techniques (see next section) and it may be possible to economize in computer storage by retaining only non-zero elements.

Perhaps the most popular direct method for the numerical solution of a system of linear equations is *Gaussian elimination.* It has two parts—an *elimination* or *triangularization* procedure and *back substitution.* Its principle is simple and can be illustrated by the problem of finding the unknowns x_1, x_2, and x_3 in

$$\begin{aligned} x_1 + x_2 + 4x_3 &= 7, \\ x_1 - 2x_2 + 6x_3 &= 15, \\ 2x_1 + x_2 - x_3 &= 7. \end{aligned}$$

Subtract the first equation from the second; then subtract twice the first from the third. There results

$$\begin{aligned} x_1 + x_2 + 4x_3 &= 7, \\ -3x_2 + 2x_3 &= 8, \\ -x_2 - 9x_3 &= -7. \end{aligned}$$

The first equation, which was used to remove x_1 from the other two equations, is known as the *pivot equation* and the coefficient of x_1 in it is called the *pivot.* Now, we make the new second equation the pivot equation and use it to eliminate x_2 from the third. Thus, by subtracting $\frac{1}{3}$ of the second from the last, we reach

$$\begin{aligned} x_1 + x_2 + 4x_3 &= 7, \\ -3x_2 + 2x_3 &= 8, \\ -29x_3/3 &= -29/3. \end{aligned}$$

If these were written in matrix form, the matrix on the left would be upper triangular which explains why the process is sometimes called triangularization. Clearly, if we reversed our steps we should recover the original equations so the two systems are equivalent.

The final step of back substitution is now undertaken. From the last equation $x_3 = 1$. Substituting this value in the second we obtain $x_2 = -2$. Then the first equation gives $x_1 = 5$.

The method can obviously be generalized to a system of n equations in n unknowns and is very easy to program. A simple count of the operations

involved reveals that about $\frac{1}{3}n^3$ multiplications and additions are required to solve a system. By taking $\mathbf{b}$ as a unit vector and by employing the special properties of unit vectors we find that the inverse of A can be found in n^3 (and not $\frac{1}{3}n^4$ as might be expected) multiplications and additions. If m_n multiplications are required for a determinant of order n and additions are ignored, expansion in co-factors gives $m_{n+1} = (n+1)m_n$ whence m_n is about $nn!$ Thus Gaussian elimination gives a dramatic improvement over Cramer's rule. Even if more sophisticated methods of evaluating determinants are adopted this statement remains true (see, for example, Kunz (1957)).

It may happen that during the elimination the normal pivot is zero. In that case two equations are interchanged so that the pivot is non-zero (there must be at least one equation with non-zero pivot so long as A is non-singular) but, if the pivot is non-zero but very small compared with other coefficients in its column, *numerical instability* can arise. This is caused by the fact that computers have a finite word length.

As an example suppose that the computer can store only three significant digits in *floating point* and is working in the base of 10. Let the equations be (Forsythe and Moler (1967))

$$1.00 \times 10^{-4}x_1 + 1.00x_2 = 1.00,$$

$$1.00x_1 + 1.00x_2 = 2.00.$$

Gaussian elimination, taking account of the limitation of word length, supplies

$$1.00 \times 10^{-4}x_1 + 1.00x_2 = 1.00,$$

$$-1.00 \times 10^{4}x_2 = -1.00 \times 10^{4}.$$

From back substitution, $x_2 = 1.00$ and $x_1 = 0.00$ which is obviously incorrect. By reversing the order of the original equations and performing Gaussian elimination we obtain $x_2 = 1.00$ and $x_1 = 1.00$ which is acceptable.

The general rule therefore, in eliminating x_r from some equations, is to select as the pivot the coefficient of x_r which has the largest magnitude; this is termed *partial pivoting*. If, however, the element of largest magnitude in both rows and columns is chosen as pivot the process is known as *complete pivoting*. According to Wilkinson (1965) and Ralston and Wilf (1967) it is doubtful whether complete pivoting warrants the additional complication and computer time. Wilkinson also shows that partial pivoting is not necessary for numerical stability even if it is sufficient, e.g. if A is a real symmetric positive definite matrix, or if A is strictly diagonal dominant.

Although pivotal strategy can control the difficulty of large multipliers in Gaussian elimination it still leaves open the possibility that the solution is very sensitive to small changes in the coefficients, i.e. the system is *ill-conditioned*. Suppose that A (which is non-singular) is perturbed to $A + B$ and $\mathbf{b}$ to $\mathbf{b} + \mathbf{c}$.

These perturbations cause a change in $\mathbf{x}$, altering it to $\mathbf{x} + \mathbf{y}$ (say). Then

$$(A + B)(\mathbf{x} + \mathbf{y}) = \mathbf{b} + \mathbf{c}.$$

A bound for $\mathbf{y}$ is provided by:

THEOREM 1.12. *If* $\|B\|\,\|A^{-1}\| < 1$ *then*

$$\|\mathbf{y}\| \leqslant C\|A^{-1}\|(\|\mathbf{c}\| + \|B\|\,\|\mathbf{x}\|) \quad \textit{and} \quad \frac{\|\mathbf{y}\|}{\|\mathbf{x}\|} \leqslant C\kappa\left(\frac{\|\mathbf{c}\|}{\|\mathbf{b}\|} + \frac{\|B\|}{\|A\|}\right)$$

where $\kappa = \|A\|\,\|A^{-1}\|$ *and* $C = \|I\| + \|A^{-1}B\|/(1 - \|A^{-1}B\|)$.

Proof. Since A is non-singular

$$(I + A^{-1}B)\mathbf{y} = A^{-1}(\mathbf{c} - B\mathbf{x}).$$

From (1.79) and Theorem 1.11*a*, $\|A^{-1}B\| < 1$ implies that $I + A^{-1}B$ has an inverse. Consequently,

$$\mathbf{y} = (I + A^{-1}B)^{-1}A^{-1}(\mathbf{c} - B\mathbf{x})$$

whence

$$\|\mathbf{y}\| \leqslant \|(I + A^{-1}B)^{-1}\|\,\|A^{-1}\|(\|\mathbf{c}\| + \|B\|\,\|\mathbf{x}\|).$$

Because of the expansion in Theorem 1.11*a*,

$$\|(I + A^{-1}B)^{-1}\| \leqslant \|I\| + \|A^{-1}B\|(1 - \|A^{-1}B\|)^{-1};$$

also $\|\mathbf{b}\| \leqslant \|A\|\,\|\mathbf{x}\|$ and the result stated in the theorem follows.

If κ is small then small changes in $\mathbf{b}$ and A will produce only small changes in $\|\mathbf{x}\|$ and the equations can be regarded as *well-conditioned.* However, large κ does not necessarily mean that the system is ill-conditioned because only upper bounds occur in Theorem 1.12. Nevertheless, we cannot improve those bounds, when $B = 0$ at any rate, since examples are known (see, for example, Forsythe and Moler (1967)) in which equality is achieved in Theorem 1.12.

We call κ the *condition number* of the system. Its precise value depends upon the choice of norm. If the spectral norm $\|\ \|_2$ is selected then, from (1.78), $\kappa = (\mu_1/\mu_n)^{1/2}$ where μ_1 and μ_n are the largest and smallest eigenvalues of $A^{\mathrm{H}}A$; this κ is sometimes described as the *spectral condition number.*

The condition number can be altered by *scaling*, i.e. by multiplying each equation in $A\mathbf{x} = \mathbf{b}$ by some integer power of 10, in the decimal system, though the same power need not be used for each row. If κ is large, whatever scale factors are employed, the equations are ill-conditioned. A small value of κ indicates a well-conditioned system. It is desirable to have available a systematic technique for scaling that ensures that κ is small as possible. Unfortunately, no method is known which applies to arbitrary matrices and arbitrary norms. One practical method is to attempt to arrange that n elements of A are of order unity, no two of these elements being in the same row or column, all other

elements of A being less than unity in magnitude. Round-off error may be reduced by using the power of the machine number base closest to the largest element.

It should be remarked that det A not being large does not necessarily signify ill-conditioning. Examples are available in which det A is small and $\|\mathbf{y}\|_2/\|\mathbf{x}\|_2 = \|\mathbf{c}\|_2/\|\mathbf{b}\|_2$. If, by scaling, it is arranged that $\|A\|_2 = 1$ then $\kappa = 1/\mu_n^{1/2}$ and the spectral condition number is large if and only if μ_n is small. Let det $A \to 0$; then $\mu_n \to 0$ provided that μ_1 is fixed. In other words, if A is normalized so as to keep μ_1 fixed then det A is closely related to the condition of A. But, in general, the largeness of the condition number is more significant than the smallness of det A as a criterion for determining ill-conditioning.

Gaussian elimination is applicable to *complex* equations if the computer has a facility for complex arithmetic. Otherwise the equations must be separated into their real and imaginary parts and the resulting real equations solved.

There are other inversion algorithms. A popular one is *triangular decomposition* in which the aim is to write A in the form

$$A = LU \tag{1.80}$$

where L is a *lower triangular matrix* (i.e. all elements above the diagonal are zero) and U is an upper triangular matrix. If this can be done in such a way that L and U are non-singular then, by putting $U\mathbf{x} = \mathbf{y}$, we have to solve the two systems

$$L\mathbf{y} = \mathbf{b},$$

$$U\mathbf{x} = \mathbf{y}.$$

Since both systems are triangular the first can be solved for $\mathbf{y}$ by back substitution and then $\mathbf{x}$ can be determined from the second by back substitution. The effort involved in the back substitution is substantially less than that in the triangular decomposition.

Conditions which permit (1.80) are contained in:

THEOREM 1.12*a*. *Let A_k be the matrix formed by the first k rows and columns of A. If $A_1, A_2, \ldots, A_{n-1}, A$ are all non-singular $A = LU$ and the decomposition is unique if the diagonal elements of either L or U are specified.*

Proof. Only the situation in which all the diagonal elements of L are chosen to be unity will be considered, the general case being left to the reader.

With

$$L = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ l_{21} & 1 & \cdots & 0 \\ & \cdots & & \\ l_{n1} & l_{n2} & \cdots & 1 \end{pmatrix}, \quad U = \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ 0 & u_{22} & \cdots & u_{2n} \\ & \cdots & & \\ 0 & 0 & \cdots & u_{nn} \end{pmatrix}$$

we require that

$$A = \begin{pmatrix} u_{11} & u_{12} & u_{13} & \cdots \\ l_{21}u_{11} & l_{21}u_{12} + u_{22} & l_{21}u_{13} + u_{23} & \cdots \\ l_{31}u_{11} & l_{31}u_{12} + l_{32}u_{22} & l_{31}u_{13} + l_{32}u_{23} + u_{33} & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{pmatrix},$$

To get the first row of A right we need

$$u_{1j} = a_{1j} (j = 1, \ldots, n).$$

The first column of A will be given correctly if

$$l_{i1} = a_{i1}/u_{11} \qquad (i = 1, \ldots, n)$$

which is possible since $u_{11} = a_{11} \neq 0$ by assumption. Now the second row of A is obtained by taking

$$u_{2j} = a_{2j} - l_{21}u_{1j} \quad (j = 2, \ldots, n)$$

and then the second column may be realized by

$$l_{i2} = (a_{i2} - l_{i1}u_{12})/u_{22} \qquad (i = 2, \ldots, n).$$

The last formula is legitimate provided that $u_{22} \neq 0$, i.e. $a_{22} - a_{21}a_{12}/a_{11} \neq 0$ which is true since $\det A_2 \neq 0$. Proceeding in this way, a row and a column at a time, we construct L and U and the construction obviously leads to unique elements.

Remark that, since A is non-singular, neither L nor U can be singular.

If A is real, symmetric, and positive definite we can show by the same procedure that there is a real lower triangular matrix L such that

$$A = LL^{\mathrm{T}}.$$

This is known as *Cholesky decomposition.* The algorithm, in this case, is highly stable. If A is Hermitian positive definite then $A = LL^{\mathrm{H}}$ where the diagonal elements of L are positive.

Finally, we observe that it is sometimes possible to improve the accuracy of a computed solution to a system of linear equations by iteration. If $\mathbf{x}^{(1)}$ is the first approximation, calculate the *residual*

$$\mathbf{r}^{(1)} = \mathbf{b} - A\mathbf{x}^{(1)}$$

as accurately as possible. Then solve the system $A\mathbf{y} = \mathbf{r}^{(1)}$ and take $\mathbf{x}^{(2)} = \mathbf{x}^{(1)} + \mathbf{y}$. Clearly this is the first stage of an iterative procedure which, under suitable circumstances, will lead to more accurate numerical values.

Exercises

59. Is the system with

$$A = \begin{pmatrix} 10^{-4} & 0.1 & -2 \times 10^{-4} \\ 0.2 & 1.0 & 0.1 \\ -10^{-4} & 0.2 & -10^{-4} \end{pmatrix}$$

badly scaled?

60. Find the triangular decomposition of

$$\begin{pmatrix} 2 & 2 & 4 \\ 4 & -2 & 2 \\ 2 & 4 & 0 \end{pmatrix}.$$

61. Suppose there are two triangular decompositions

$$A = L_1 U_1 = L_2 U_2$$

with L_1 and L_2 having units on the diagonal. If A is non-singular prove, without using Theorem 1.12*a*, that U_1 and U_2 are not singular. Deduce from $L_1^{-1} L_2 = U_1 U_2^{-1}$ that the decomposition is, in fact, unique.

1.13 Iterative methods

Iterative methods, which are appropriate for large sparse matrices, are based upon the idea of starting with an initial guess $\mathbf{x}^{(0)}$ to the solution of $A\mathbf{x} = \mathbf{b}$ and then deriving a sequence $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots$ which converges to the exact solution.

All of the methods are based on rewriting A in the form

$$A = L + D + U$$

where D is a diagonal matrix with diagonal elements the same as those of A, L is lower triangular with zeros on the diagonal, and U is upper triangular with zeros on the diagonal. For instance, if we express $A\mathbf{x} = \mathbf{b}$ as

$$D\mathbf{x} = \mathbf{b} - (L + U)\mathbf{x}$$

this suggests the iterative scheme

$$\mathbf{x}^{(r+1)} = D^{-1}\mathbf{b} - D^{-1}(L + U)\mathbf{x}^{(r)}$$

which is known as the *Jacobi method*. A necessary condition for the application of this method is that all the diagonal elements of A are non-zero. Alternatively, the form

$$(L + D)\mathbf{x} = \mathbf{b} - U\mathbf{x}$$

suggests the iteration

$$\mathbf{x}^{(r+1)} = (L + D)^{-1}\mathbf{b} - (L + D)^{-1}U\mathbf{x}^{(r)}$$

which is the *Gauss–Seidel* method. Again, by introducing the non-zero scalar parameter ω and writing

$$(\mathrm{D} + \omega L)\mathbf{x} = \{-\omega U + (1 - \omega)D\}\mathbf{x} + \omega\mathbf{b},$$

we derive the procedure

$$\mathbf{x}^{(r+1)} = (D + \omega L)^{-1}\{-\omega U + (1 - \omega)D\}\mathbf{x}^{(r)} + (D + \omega L)^{-1}\omega\mathbf{b}.$$

This is the method of *successive over-relaxation* (*SOR*). It reduces to the Gauss–Seidel method if $\omega = 1$.

Of course, one does not in practice calculate the inverse matrices on the right-hand sides in the iterative schemes but, instead, solves the linear system which arises before the application of the inverse matrix. One advantage is evident in that zero elements of the matrix need not be stored and successive vectors can be overwritten on their predecessors so that considerable economy of computer storage can be achieved.

The iterations are all examples of taking an equation

$$\mathbf{x} = B\mathbf{x} + \mathbf{c}$$

and replacing it by

$$\mathbf{x}^{(r+1)} = B\mathbf{x}^{(r)} + \mathbf{c}.$$

Let $\mathbf{e}^{(r)} = \mathbf{x}^{(r)} - \mathbf{x}$ be the error at a particular stage. Then, by subtraction, we see that $\mathbf{e}^{(r+1)} = B\mathbf{e}^{(r)}$ whence

$$\mathbf{e}^{(r)} = B^r\mathbf{e}^{(0)}.$$

Now $\mathbf{x}^{(r)}$ converges to $\mathbf{x}$ if and only if $\mathbf{e}^{(r)}$ approaches zero which can happen for every choice of $\mathbf{e}^{(0)}$ if and only if $B^r \to 0$. From Theorem 1.11 we deduce:

THEOREM 1.13. *The iterative scheme converges to the correct solution if and only if* $\rho(B) < 1$.

Thus the convergence of the three schemes turns upon the spectral radii of the relevant matrices, i.e. of $D^{-1}(L + U)$, $(L + D)^{-1}U$ and $\mathscr{L}_\omega = (D + \omega L)^{-1} \times \{-\omega U + (1 - \omega)D\}$. The smaller the spectral radius the more rapid the convergence. One aim of successive overrelaxation is to choose ω so that the spectral radius of $\mathscr{L}_\omega$ is as small as possible. However, we are limited in our choice by:

THEOREM 1.13*a*. *If* $\rho(\mathscr{L}_\omega) < 1$ *then* $0 < \omega < 2$.

Proof. If λ is an eigenvalue of $\mathscr{L}_\omega$, $\det(\mathscr{L}_\omega - \lambda I) = 0$. Now, in the polynomial equation for λ which results, the coefficient of λ^n is ± 1 and the constant term is $\det \mathscr{L}_\omega$ which, from the structure of D, L, and U, is $(\det D)^{-1}(1 - \omega)^n \det \mathrm{D}$. Thus the product of the roots of the characteristic equation is $\pm(1 - \omega)^n$. Hence, at least one of the eigenvalues must have a modulus as great as $|1 - \omega|$. Since $\rho(\mathscr{L}_\omega) < 1$ this implies that $|1 - \omega| < 1$ and the theorem is proved.

As a consequence of Theorem 1.13*a* there is no point in considering values of ω outside the interval (0, 2) and, in practice, it is normal to choose ω so that $1 < \omega < 2$. In general, the determination of the optimal value of ω is extremely complicated (see, for example, Mitchell (1969)).

It can be shown that the Gauss–Seidel method converges if A is a real positive definite matrix, though the Jacobi method may not, and other conditions for convergence are known (see, for example, Varga (1962)). In general the convergence of the Gauss–Seidel method is faster than that of the Jacobi method and SOR is usually appreciably better than either (see subsequent exercises).

Another iterative scheme which is often employed is the *Peaceman–Rachford method*. In this method we write $A = A_1 + A_2 + A_3$ where A_1, A_2, and A_3 have certain properties which will not be elaborated here. An intermediate iteration is inserted so that $\mathbf{x}^{(r)} \to \mathbf{y}^{(r)} \to \mathbf{x}^{(r+1)}$ by the process

$$(A_1 + \omega_1 A_2 + \omega_2 I)\mathbf{y}^{(r)} = \mathbf{b} - \{A_3 + (1 - \omega_1)A_2 - \omega_2 I\}\mathbf{x}^{(r)},$$

$$(A_3 + \omega_3 A_2 + \omega_4 I)\mathbf{x}^{(r+1)} = \mathbf{b} - \{A_1 + (1 - \omega_3)A_2 - \omega_4 I\}\mathbf{y}^{(r)}.$$

The analysis of this scheme is highly complex but it does seem to be a profitable method in connection with difference equations (see, for example, Mitchell (1969)).

Particular methods for sparse matrices have been the subject of considerable research in recent years (see, for example, Duff (1976)).

Exercises

62. Find the spectral radii of the Jacobi and Gauss–Seidel methods for

$$A = \begin{pmatrix} 1 & -\frac{1}{2} & 0 \\ -\frac{1}{2} & 1 & -\frac{1}{2} \\ 0 & -\frac{1}{2} & 1 \end{pmatrix}$$

and show that both methods converge.

63. If A is symmetric show that an eigenvalue λ and the associated eigenvector $\mathbf{u}$ (possibly complex) of the Gauss–Seidel method satisfy

$$\{(\mathbf{u}, D\mathbf{u}) + (\mathbf{u}, L\mathbf{u})\}\lambda = (\mathbf{u}, L^{\mathrm{T}}\mathbf{u}) = (\mathbf{u}, L\mathbf{u})^*.$$

By forming $|\lambda|^2$ deduce that, if A is a real positive definite matrix, $|\lambda| < 1$ and that the Gauss–Seidel method converges.

64. Let A be a tridiagonal matrix. Let λ, $\mathbf{u}$ be an eigenvalue and associated eigenvector of the matrix of the Jacobi method so $(L + U)\mathbf{u} = \lambda D\mathbf{u}$. By applying $(L + U)D^{-1}$ show that λ is a diagonal element of $(UD^{-1}L + LD^{-1}U)D^{-1}$.

If μ is an eigenvalue of the SOR method use this procedure to demonstrate that

$$(1 - \omega - \mu)^2 = \omega^2 \mu \lambda^2.$$

Deduce that the Jacobi and Gauss–Seidel methods either both converge or both diverge for a tridiagonal matrix and that the convergence of the Gauss–Seidel is faster.

If all λ are real and $\lambda_1(<1)$ is the largest show that the optimal choice for ω is $2\{1 + (1 - \lambda_1^2)^{1/2}\}^{-1}$ and determine the corresponding value of μ_1. If, say, $\lambda_1 = 0.995$ then $\mu_1 \approx 0.8$ and the convergence of SOR is much more rapid than that of Gauss–Seidel.

1.14 Matrix eigenvalues

The matrix eigenvalue problem occurs in many applications and is concerned with solving

$$(A - \lambda I)\mathbf{x} = \mathbf{0},$$

i.e. with determining the eigenvalues λ and associated eigenvectors $\mathbf{x}$ of A. The eigenvalues must satisfy $\det(A - \lambda I) = 0$ and it is tempting to try expanding this as a polynomial in λ, whose roots can be found by one of the methods of §1.8. Apart from the difficulty of calculating the coefficients there is a classic example due to Wilkinson (1965) which demonstrates why this should not be done. The polynomial

$$(x - 1)(x - 2)\ldots(x - 20) - 2^{-23}x^{19}$$

is a slight perturbation from a polynomial with zeros at $1, 2, \ldots, 20$. But only ten of its zeros are real and two of the complex ones have imaginary parts of about 2.8 in magnitude. Since round-off error can easily introduce perturbations in the coefficients of a polynomial the characteristic polynomial is never used for the computation of eigenvalues. Indeed, it may be wiser to calculate the zeros of a polynomial by solving an associated eigenvalue problem.

To begin with we discuss an iterative scheme known as the *power method.*

THEOREM 1.14. *Let A have n linearly independent eigenvectors $\boldsymbol{x}_i$ and the corresponding eigenvalues λ_i satisfy*

$$|\lambda_1| > |\lambda_2| \geqslant |\lambda_3| \geqslant \cdots \geqslant |\lambda_n|.$$

Then the sequence $\boldsymbol{u}_{r+1} = A\boldsymbol{u}_r$ is such that

$$\lim_{r\to\infty} (\mathbf{u}_{r+1})_j/(\mathbf{u}_r)_j = \lambda_1$$

where $(\mathbf{u}_r)_j$ denotes the jth element of $\boldsymbol{u}_r$.

Proof. Since the $\mathbf{x}_i$ are linearly independent, there are constants α_i such that

$$\mathbf{u}_0 = \alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \cdots + \alpha_n\mathbf{x}_n.$$

Hence

$$\mathbf{u}_r = A^r\mathbf{u}_0 = \alpha_1\lambda_1^r\mathbf{x}_1 + \cdots + \alpha_n\lambda_n^r\mathbf{x}_n$$

and

$$\frac{(\mathbf{u}_{r+1})_j}{(\mathbf{u}_r)_j} = \lambda_1 \frac{\alpha_1(\mathbf{x}_1)_j + \alpha_2(\lambda_2/\lambda_1)^{r+1}(\mathbf{x}_2)_j + \cdots}{\alpha_1(\mathbf{x}_1)_j + \alpha_2(\lambda_2/\lambda_1)^r(\mathbf{x}_2)_j + \cdots}.$$

By hypothesis, $(\lambda_i/\lambda_1)^r \to 0$ as $r \to \infty$ if $i \neq 1$ and the theorem is proved.

Strictly, the proof of the theorem requires that $\alpha_1 \neq 0$, otherwise λ_2 is obtained as the limit rather than λ_1. However, round-off error is almost certain to introduce a small component of $\mathbf{x}_1$ and the effect of this will be to direct the convergence towards λ_1.

It is also common to normalize the iteration by putting $\mathbf{v}_{r+1} = A\mathbf{u}_r$ and then defining $\mathbf{u}_{r+1} = \mathbf{v}_{r+1}/\|\mathbf{v}_{r+1}\|_\infty$. The effect of this is that $\|\mathbf{v}_r\| \to \lambda_1$ as $r \to \infty$. Furthermore, $\mathbf{u}_r \to \mathbf{x}_1/\|\mathbf{x}_1\|_\infty$ so that the associated eigenvector is supplied at the same time.

The iteration fails if there are a number of unequal eigenvalues of the same modulus. There are ways of overcoming this difficulty but details will not be given here. (See, for example, Gourlay and Watson (1973).)

Aitken's method (§1.8) can be employed to accelerate convergence. Another technique is to work with $A - qI$ instead of A. The eigenvalues of the new matrix are $\lambda_i - q$ and, provided that $\lambda_1 - q$ is still dominant, it may be possible to choose q so that $|(\lambda_2 - q)/(\lambda_1 - q)|$ is much smaller than $|\lambda_2/\lambda_1|$.

Another iterative plan is that of *inverse iteration*. In this procedure we form $\mathbf{w}_{r+1} = (A - qI)^{-1}\mathbf{w}_r$; actually we determine $\mathbf{w}_{r+1}$ by solving the linear system

$$(A - qI)\mathbf{w}_{r+1} = \mathbf{w}_r.$$

By Theorem 1.14 inverse iteration provides an eigenvalue of $(A - qI)^{-1}$, i.e. one of $1/(\lambda_i - q)$. With the normalization described above the associated eigenvector is also obtained.

Inverse iteration is capable, by judicious choice of q, of finding any eigenvalue or eigenvector of A. It also has a fast rate of convergence. It is one of the most powerful and accurate methods available.

While the above methods compute a single eigenvalue at a time there are others which aim for the complete eigensystem right from the beginning, usually at the price of requiring A to be symmetric or Hermitian.

Let e_{ij} denote the (i,j) element of the unit matrix. A *plane rotation matrix* R_{ij} is a matrix derived from the unit matrix by replacing four elements according to the following scheme

$$\begin{pmatrix} e_{ii} & e_{ij} \\ e_{ji} & e_{jj} \end{pmatrix} \to \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

so that

$$R_{ij} = \begin{pmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ & & & \cdots & & & \\ 0 & \cdots & \cos\theta & \cdots & -\sin\theta & \cdots & 0 \\ & & & \cdots & & & \\ 0 & \cdots & \sin\theta & \cdots & \cos\theta & \cdots & 0 \\ & & & \cdots & & & \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{pmatrix}$$

The nomenclature stems from the fact that the replacement represents a rotation of two-dimensional axes through an angle θ.

It is immediate that R_{ij} is orthogonal, i.e. $R_{ij}R_{ij}^T I$.

Let A be a real symmetric matrix and then put

$$B = R_{ij}^{\mathrm{T}} A R_{ij}.$$

The elements of B are the same as those of A except for

$$b_{ik} = b_{ki} = a_{ik}\cos\theta + a_{jk}\sin\theta, \qquad k \neq i, j$$

$$b_{jk} = b_{kj} = -a_{ik}\sin\theta + a_{jk}\cos\theta, \qquad k \neq i, j$$

$$b_{ii} = a_{ii}\cos^2\theta + 2a_{ij}\sin\theta\cos\theta + a_{jj}\sin^2\theta,$$

$$b_{ij} = b_{ji} = a_{ij}\cos 2\theta + \tfrac{1}{2}(a_{jj} - a_{ii})\sin 2\theta,$$

$$b_{jj} = a_{ii}\sin^2\theta - 2a_{ij}\sin\theta\cos\theta + a_{jj}\cos^2\theta.$$

We now choose θ so that $b_{ij} = 0$. If $a_{jj} \neq a_{ii}$ we take θ so that $-\frac{1}{4}\pi < \theta < \frac{1}{4}\pi$ and

$$\tan 2\theta = 2a_{ij}/(a_{jj} - a_{ii}).$$

If $a_{jj} = a_{ii}$ and $a_{ij} \neq 0$ we select $\theta = \frac{1}{4}\pi(a_{ij}/|a_{ij}|)$; if $a_{ij} = 0$ no choice is necessary.

So far i and j are at our disposal. They are determined by searching the elements of A *above* the diagonal and finding the element a_{ij} of maximum modulus. Having fixed θ the resulting matrix B is denoted by A_1. The largest off-diagonal element of A_1 is now reduced to zero by the same procedure of applying a plane rotation matrix. Denoting the new matrix by A_2 we note that the (i, j) element which was reduced to zero in A_1 is no longer zero in A_2. However, it can be shown that if the procedure is repeated indefinitely, the limit of the sequence A_r is a diagonal matrix with the eigenvalues of A on its diagonal.

This algorithm, which is known as the *Jacobi method*, is easy to program but it is not very efficient. For small matrices which can be held entirely in the fast store it may be appropriate because it is very reliable.

Elements annihilated by a plane rotation in the Jacobi method may be

recreated at later stages. An algorithm to overcome this is provided by the *Givens method.* Instead of choosing θ so that $b_{ij} = 0$, pick some $k \neq i$ and ask that $b_{kj} = 0$, i.e. select

$$\tan \theta = a_{jk}/a_{ik}.$$

Suppose, in fact, $A_1 = R_1^{\mathrm{T}} A R_1$ where R_1 is the plane rotation which annihilates a_{13} (say, $i = 2, j = 3, k = 1$). Next form $A_2 = R_2^{\mathrm{T}} A_1 R_2$ where R_2 reduces the (1, 4) element of A_1 to zero with $i = 2, j = 4, k = 1$; the (1, 3) element which is zero in A_1 remains zero in A_2. By repeating the process we can make $(n - 2)$ elements in the first row zero; by starting from (2, 4) we can operate on the second row without affecting the first. In this way a real symmetric can be transformed by tridiagonal form by plane rotations.

Denote the tridiagonal matrix by B and let $b_{ii} = b_i$, $b_{i,i+1} = b_{i+1,i} = c_i$. Let $p_r(\lambda)$ be the determinant formed by the first r rows and colums of $B - \lambda I$ and define $p_0(\lambda) = 1$. Then, it may easily be established that

$$p_0(\lambda) = 1, p_1(\lambda) = b_1 - \lambda,$$

$$p_r(\lambda) = (b_r - \lambda)p_{r-1}(\lambda) - c_{r-1}^2 p_{r-2}(\lambda) \qquad (r = 2, 3, \ldots, n).$$

The zeros of $p_{r-1}(\lambda)$ lie between those of $p_r(\lambda)$. Also, define the sequence $s_r(\lambda)$ by $s_r(\lambda) = s_{r-1}(\lambda) + 1$ if $p_r(\lambda)$ is zero or has the same sign as $p_{r-1}(\lambda)$ and by $s_r(\lambda) = s_{r-1}(\lambda)$ otherwise. Starting from $s_0(\lambda) = 0$, we generate $s_r(\lambda)$ as either zero or a positive integer. Define $s(\lambda) = s_n(\lambda)$, then the *Sturm sequence property* states that $s(\lambda)$ is equal to the number of eigenvalues of B which are strictly greater than λ; in fact, this property holds if B is symmetric instead of being tridiagonal.

The Sturm sequence property permits the location of an interval $(\lambda_1^{(0)}, \lambda_2^{(0)})$ in which only one eigenvalue λ lies, i.e. for some $m < n$, $s(\lambda_2^{(0)}) = m$, $s(\lambda_1^{(0)}) = m + 1$. Let $\mu^{(0)} = \frac{1}{2}(\lambda_1^{(0)} + \lambda_2^{(0)})$ then $s(\mu^{(0)})$ is either m or $m + 1$ and a smaller interval for λ has been determined. This is akin to the method of bisection (§1.8) and is very efficient for finding the eigenvalues in a particular interval.

Once the eigenvalues have been calculated it is tempting to find the eigenvectors by solving $(B - \lambda_i I)\mathbf{x} = \mathbf{0}$ by omitting the last equation of the system and solving the first $(n - 1)$ for $x_1, \ldots, x_{n-1}$ in terms of an arbitrary x_n. In general, this is catastrophically unstable and should never be undertaken. The preferred method is inverse iteration.

Another technique for the reduction of a real symmetric A to tridiagonal form is the *Givens–Householder* method. Here, one considers matrices of the type

$$P = I - 2\mathbf{w}\mathbf{w}^{\mathrm{T}}$$

where the vector $\mathbf{w}$ satisfies $\|\mathbf{w}\| = 1$ but is otherwise at our disposal. Observe firstly that P is symmetric. Also

$$PP^{\mathrm{T}} = I - 4\mathbf{w}\mathbf{w}^{\mathrm{T}} + 4\mathbf{w}\mathbf{w}^{\mathrm{T}}\mathbf{w}\mathbf{w}^{\mathrm{T}} = I$$

so that P is, in fact, orthogonal.

If $\mathbf{u}$ and $\mathbf{v}$ are real vectors such that $\|\mathbf{u}\| = \|\mathbf{v}\|$ put

$$\mathbf{w} = (\mathbf{v} - \mathbf{u})/\|\mathbf{v} - \mathbf{u}\|.$$

Then

$$\|\mathbf{v} - \mathbf{u}\|^2 = (\mathbf{v}, \mathbf{v}) - (\mathbf{v}, \mathbf{u}) - (\mathbf{u}, \mathbf{v}) + (\mathbf{u}, \mathbf{u}) = 2(\mathbf{u} - \mathbf{v}, \mathbf{u})$$

and

$$P\mathbf{u} = \mathbf{v}.$$

Thus, with this choice of w, P converts to a real vector $\mathbf{u}$ into another one v of the same length.

In particular, let $\mathbf{a}_1$ be the first column of A. Let $\mathbf{b}_1$ be a vector with element b_{11} in the first row and zero elsewhere. Take

$$b_{11} = -a_{11}\|\mathbf{a}_1\|/|a_{11}|$$

and then $\|\mathbf{b}_1\| = \|\mathbf{a}_1\|$ so that we may place $\mathbf{u} = \mathbf{a}_1$, $\mathbf{v} = \mathbf{b}_1$ above. Consequently, PA is a matrix whose first column is zeros except the diagonal element which is b_{11}. An *arbitrary* $m \times n$ matrix can be transformed by this approach to the form QU where Q is orthogonal and U upper triangular.

For our particular purposes the transformation is not quite suitable since we want $P^{\mathrm{T}}AP$, rather than PA, to be simpler than A. However, it indicates the direction in which to go. When A is real and symmetric partition it according to

$$A = \begin{pmatrix} a_{11} & \boldsymbol{\alpha}^{\mathrm{T}} \\ \boldsymbol{\alpha} & A_1 \end{pmatrix}$$

and introduce

$$M = \begin{pmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & P \end{pmatrix}$$

where P is an $(n-1) \times (n-1)$ matrix of Givens–Householder type. Then

$$M^{\mathrm{T}}AM = \begin{pmatrix} a_{11} & \boldsymbol{\alpha}^{\mathrm{T}}P \\ P^{\mathrm{T}}\boldsymbol{\alpha} & P^{\mathrm{T}}A_1P \end{pmatrix}$$

so that, if we choose P so that the last $n-2$ elements of $P^{\mathrm{T}}\boldsymbol{\alpha}$ are zero, the matrix $M^{\mathrm{T}}AM$ will have $n-2$ zeros on its first row and column. We remark that

$$M = I - 2\boldsymbol{\omega}\boldsymbol{\omega}^{\mathrm{T}}$$

where $\boldsymbol{\omega} = \begin{pmatrix} 0 \\ \mathbf{w} \end{pmatrix}$ and $\mathbf{w}$ is an $(n-1)$ vector. This suggests that for the next step we try $M_1 = I - 2\boldsymbol{\omega}_1\boldsymbol{\omega}_1^{\mathrm{T}}$ where

$$\boldsymbol{\omega}_1 = \begin{pmatrix} 0 \\ 0 \\ \mathbf{w}_1 \end{pmatrix}$$

and $\mathbf{w}_1$ is an $(n-2)$ vector. There is no difficulty in checking that $M_1^T M^T A M M_1$ has the same zeros as $M^T A M$ in the first row and column and we can choose $\mathbf{w}_1$ so that there are $n-3$ additional zeros in the second row and column. It is now obvious how we may create a tridiagonal matrix from a real symmetric A by the Givens–Householder process.

The Givens–Householder method requires $\frac{2}{3}n^3$ multiplications as compared with $\frac{4}{3}n^3$ for the Givens. It is therefore more efficient and generally regarded as one of the best methods for finding the eigenvalues of a Hermitian matrix though the Givens method is sometimes valuable for sparse matrices. For non-symmetric matrices where the two methods reduce A to the more complicated *Hessenberg form* (which is the same as a tridiagonal matrix below the diagonal but may have non-zero elements anywhere above the diagonal) the situation is less clear.

In the symmetric case both methods lead to a tridiagonal matrix and we have already described one way of dealing with eigenvalue problems by Sturm sequences. Another technique is based on triangular decomposition. Suppose

$$A_1 = L_1 U_1$$

where L_1 is lower triangular with units on the diagonal and U_1 is upper triangular (cf. Theorem 1.12*a*). Let $A_2 = U_1 L_1$. Then $A_2 = L_1^{-1} A_1 L_1$ and is similar to A_1. This suggests the iteration: given A_s, write it as $A_s = L_s U_s$ and then form $A_{s+1} = U_s L_s$. This is known as the *LR algorithm* (LR because Rutishauser, who introduced it, called the decomposition left, right instead of lower, upper). It is subject to many shortcomings but its introduction led to the *QR algorithm*, one of the most powerful devices for the matrix eigenvalue problem.

The QR algorithm is based on the fact that, for any non-singular A, there is a decomposition

$$A = QU$$

in which Q is unitary and U is upper triangular. Moreover the decomposition is unique if we impose the condition that the diagonal elements of U are positive. The iteration is now performed as: write $A_s = Q_s U_s$ and then define $A_{s+1} = U_s Q_s$. Since $A_{s+1} = Q_s^H A_s Q_s$, A_{s+1} is unitarily similar to A_s and therefore to A.

In the QR algorithm the tridiagonal property is preserved, i.e. if A is tridiagonal so are its iterates. In spite of the power of the algorithm it has been suggested that a matrix should be reduced to Hessenberg or tridiagonal form before the algorithm is applied. (See, for example, Ralston and Wilf (1967).)

Exercises

65. Use the power method to find the largest eigenvalue of

(a) $\begin{pmatrix} 6 & 2 & 1 \\ 2 & 3 & 1 \\ 1 & 1 & 1 \end{pmatrix}$, (b) $\begin{pmatrix} 0 & 1 & 1 \\ -1 & 0 & 1 \\ -1 & -1 & 0 \end{pmatrix}$.

66. Use inverse iteration to find the eigenvalues of the matrix in 65(a).

67. Find the eigenvectors of $\begin{pmatrix} 4 & 0 & 0 \\ 0 & 3 & 1 \\ 0 & 1 & 3 \end{pmatrix}$.

68. Use the Givens and Givens–Householder methods to reduce to tridiagonal form

(a) $\begin{pmatrix} 4 & 4 & 2 \\ 4 & 4 & 1 \\ 2 & 1 & 8 \end{pmatrix}$, (b) $\begin{pmatrix} 1 & 2 & 1 & 2 \\ 2 & 2 & -1 & 1 \\ 1 & -1 & 1 & 1 \\ 2 & 1 & 1 & 1 \end{pmatrix}$.

69. Show how $P = I - 2\mathbf{w}\mathbf{w}^{\mathrm{H}}$ may be used to transform a Hermitian matrix to tridiagonal form.

70. Show that $\begin{pmatrix} -2 & 1 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \\ 1 & 0 & 1 & -2 \end{pmatrix}$ has two eigenvalues in $(-2, 0)$ and is, in fact, negative definite.

71. Find the eigenvalues of $\begin{pmatrix} 1 & 2 & 0 \\ 2 & -1 & 1 \\ 0 & 1 & 3 \end{pmatrix}$ by Sturm sequences.

72. In the QR algorithm prove that

$$A_1^s = Q_1 Q_2 \dots Q_s U_s \dots U_1.$$

GENERALIZED INVERSE

1.15 The generalized inverse

It is not uncommon in applications to encounter the problem of solving the system

$$A\mathbf{x} = \mathbf{b} \tag{1.81}$$

where A is not square but $m \times n$. For example, in making observations it may be that we have data from less points than we have unknowns so that $m < n$, or we may have more data points than unknowns in which case $m > n$. In the former case the linear system possesses an infinite number of solutions while,

in the latter, the system may strictly have no solution because the equations are inconsistent with one another. Yet it may be important for the application to identify a single entity which one is prepared to accept as 'the solution' of the system.

One method which suggests itself is that of least squares. There are at least two ways in which this could be applied. We could consider minimizing the sum of the squares of the residual $\mathbf{r}^T\mathbf{r}$ where $\mathbf{r} = \mathbf{b} - A\mathbf{x}$ or we might try minimizing $\mathbf{x}^T\mathbf{x}$. Let us deal with residuals first.

The *rank* of a matrix is the order of the largest non-singular submatrix in the matrix. It is not difficult to confirm that, when A is real, A and A^TA have the same rank. With that notion we can formulate

THEOREM 1.15. *If real A is $m \times n$ with $m > n$ and of rank n, the solution of* (1.81) *which minimizes $\mathbf{r}^T\mathbf{r}$ is given by*

$$\mathbf{x} = (A^TA)^{-1}A^T\mathbf{b}.$$

Proof. Since A^TA is $n \times n$ and of rank n, it is non-singular and so the formula makes sense. Also

$$\mathbf{r}^T\mathbf{r} = \mathbf{b}^T\mathbf{b} - \mathbf{x}^TA^T\mathbf{b} - \mathbf{b}^TA\mathbf{x} + \mathbf{x}^TA^TA\mathbf{x}$$

so that $\partial(\mathbf{r}^T\mathbf{r})/\partial x_i = 0$ for $i = 1, \ldots, n$ leads to

$$A^TA\mathbf{x} = A^T\mathbf{b} \tag{1.82}$$

and the theorem is proved.

For the case $m < n$ we have

THEOREM 1.15*a*. *If real A is $m \times n$ with $m < n$ and of rank m, the solution of* (1.81) *which minimizes $\mathbf{x}^T\mathbf{x}$ is given by*

$$\mathbf{x} = A^T(AA^T)^{-1}\mathbf{b}.$$

Proof. The formula makes sense because AA^T is $m \times m$ of rank m and therefore non-singular. The minimum of $\mathbf{x}^T\mathbf{x}$ subject to $A\mathbf{x} = \mathbf{b}$ is found by Lagrange multipliers, i.e. by minimizing

$$S = \mathbf{x}^T\mathbf{x} + \boldsymbol{\lambda}^T(\mathbf{b} - A\mathbf{x})$$

where $\boldsymbol{\lambda}$ is a column vector with m elements. From $\partial S/\partial x_i = 0$, $i = 1, \ldots, n$ we obtain $\mathbf{x} = A^T\boldsymbol{\lambda}$ and from $\partial S/\partial \lambda_j = 0, j = 1, \ldots, n$ we have $A\mathbf{x} = \mathbf{b}$. Hence

$$AA^T\boldsymbol{\lambda} = A\mathbf{x} = \mathbf{b}$$

which can be solved for $\boldsymbol{\lambda}$ and the theorem follows.

It is desirable to relax the conditions on rank in Theorems 1.15 and 1.15*a* and find a single formula which encompasses all possibilities. To this end we

examine whether there is a matrix A^+ such that $\mathbf{x} = A^+\mathbf{b}$. In the case of Theorem 1.15 we have

$$A^+ = (A^TA)^{-1}A^T$$

and we remark that

$$A^+AA^+ = (A^TA)^{-1}A^TA(A^TA)^{-1}A^T = (A^TA)^{-1}A^T = A^+.$$

Similarly $AA^+A = A$. Also AA^+ and A^+A are symmetric. Now, for Theorem 1.15*a*, $A^+ = A^T(AA^T)^{-1}$ and a check reveals that it has the same three properties. This prompts:

DEFINITION. *A matrix A^+ with the properties (i) $A^+AA^+ = A^+$, (ii) $AA^+A = A$, (iii) AA^+ amd A^+A are symmetric, is called the generalized inverse of A. If A is complex replace symmetric in (iii) by Hermitian.*

It has already been verified that the inverses of Theorems 1.15 and 1.15*a* comply with this definition. If A is square and non-singular, the inverse A^{-1} obviously satisfies it. Moreover, we can show that A possesses only one generalized inverse so that $A^+ = A^{-1}$ when A^{-1} exists.

Suppose, in fact, that real A had a second generalized inverse B^+. Then, from property (ii)

$$A^+A = A^+AB^+A,$$

and then property (iii) implies that

$$A^+A = B^+AA^+A = B^+A$$

whence $B^+ = A^+AB^+$. Similarly, $AA^+ = AB^+$ from which $A^+ = A^+AB^+$ and hence $A^+ = B^+$ so that the generalized inverse is unique.

By taking the transpose of the quantities in the definition we deduce that $(A^+)^T = (A^T)^+$ so that, if A is symmetric so is A^+. Obviously, $(A^+)^+ = A$ and A^+ has the same rank as A.

To obtain an explicit formula for A^+ it is convenient to derive first some properties of *complex* $m \times n$ matrices. If μ is a *non-zero* number and there are vectors $\mathbf{u}$, $\mathbf{v}$ such that

$$A\mathbf{u} = \mu\mathbf{v}, \quad A^H\mathbf{v} = \mu\mathbf{u}$$

then μ is known as a *singular value* of A and $\mathbf{u}$, $\mathbf{v}$ as the corresponding pair of *singular vectors.*

Now

$$A^HA\mathbf{u} = \mu A^H\mathbf{v} = \mu^2\mathbf{u}.$$

Since A^HA is positive semi-definite the values of μ^2 are real and non-negative. Also A^HA, which is of order $n \times n$, possesses n linearly independent eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_n$ which can be arranged to be orthonormal. If the rank of A is k so is the rank of A^HA and precisely k of the values of μ^2 are non-zero. Pick the order of the eigenvectors so that $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k$ correspond to the non-zero

eigenvalues $\mu_1^2, \ldots, \mu_k^2$. Note that $\mathbf{u}_{k+1}, \ldots, \mathbf{u}_n$ can be chosen to satisfy $A\mathbf{u}_i = \mathbf{0}$.

Define $\mathbf{v}_i$ for $i = 1, \ldots, k$ by $\mathbf{v}_i = A\mathbf{u}_i/\mu_i$ with μ_i the positive square root of μ_i^2. Then

$$AA^{\mathrm{H}}\mathbf{v}_i = AA^{\mathrm{H}}A\mathbf{u}_i/\mu_i = \mu_i A\mathbf{u}_i = \mu_i^2\mathbf{v}_i$$

so that $\mathbf{v}_i$ is an eigenvector of AA^{H} and, since $A^{\mathrm{H}}\mathbf{v}_i = \mu_i\mathbf{u}_i$, $\mu_1, \ldots, \mu_k$ are the positive singular values of A and $\mathbf{u}_i, \mathbf{v}_i$ the corresponding singular vectors.

The vectors $\mathbf{v}_i$ are orthonormal because

$$\mu_i\mu_j(\mathbf{v}_i, \mathbf{v}_j) = (A\mathbf{u}_i, A\mathbf{u}_j) = (\mathbf{u}_i, A^{\mathrm{H}}A\mathbf{u}_j) = \mu_j^2(\mathbf{u}_i, \mathbf{u}_j).$$

The set may be completed by adding on $m - k$ orthogonal vectors satisfying $A^{\mathrm{H}}\mathbf{v} = \mathbf{0}$; they are automatically eigenvectors of AA^{H} corresponding to the eigenvalue zero.

Define the $n \times n$ matrix U and the $m \times m$ matrix V by

$$U = (\mathbf{u}_1, \ldots, \mathbf{u}_n), \quad V = (\mathbf{v}_1, \ldots, \mathbf{v}_m).$$

Then we have

THEOREM 1.15*b*. *If A is of rank k there are unitary matrices U and V such that*

$$V^{\mathrm{H}}AU = \begin{pmatrix} \Lambda & 0 \\ 0 & 0 \end{pmatrix}$$

where Λ is a diagonal matrix of order k whose diagonal elements are the singular values of A.

Proof. By construction $U^{\mathrm{H}}U$ is the $n \times n$ unit matrix so that U is unitary. Similarly V is unitary. Also

$$AU = (\mu_1\mathbf{v}_1, \ldots, \mu_k\mathbf{v}_k, 0, \ldots, 0)$$

and the theorem follows from the orthonormal property of the $\mathbf{v}_i$.

Since the diagonal elements of Λ are non-zero, Λ^{-1} exists. This fact enables us to state

THEOREM 1.15*c*. *If A is of rank k, its generalized inverse is given by*

$$A^+ = U\begin{pmatrix} \Lambda^{-1} & 0 \\ 0 & 0 \end{pmatrix}V^{\mathrm{H}}.$$

Proof.

$$A^+AA^+ = U\begin{pmatrix} \Lambda^{-1} & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} \Lambda & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} \Lambda^{-1} & 0 \\ 0 & 0 \end{pmatrix}V^{\mathrm{H}}$$

from Theorem 1.15*b*. Property (i) of the Definition follows at once. Further,

since U and V are unitary

$$A = V\begin{pmatrix}\Lambda & 0\\ 0 & 0\end{pmatrix}U^{\mathrm{H}}$$

so that

$$AA^{+} = V\begin{pmatrix}I & 0\\ 0 & 0\end{pmatrix}V^{\mathrm{H}}$$

$$A^{+}A = U\begin{pmatrix}I & 0\\ 0 & 0\end{pmatrix}U^{\mathrm{H}}$$

which show that property (iii) is satisfied. Finally,

$$AA^{+}A = V\begin{pmatrix}\Lambda & 0\\ 0 & 0\end{pmatrix}U^{\mathrm{H}} = A$$

and the proof is complete.

Remark now that

$$A^{\mathrm{H}}AA^{+} = A^{\mathrm{H}}$$

so that $\mathbf{x} = A^{+}\mathbf{b}$ satisfies (1.82) with the affix T replaced by H. Moreover, in the analogue of Theorem 1.15*a*, we need a solution of $AA^{\mathrm{H}}\boldsymbol{\lambda} = \mathbf{b}$. But AA^{H} is Hermitian and so has the structure of Theorem 1.15*b* with $V = U$ whence $(AA^{\mathrm{H}})^{+} = (A^{\mathrm{H}})^{+}A^{+}$. Therefore $\mathbf{x} = A^{\mathrm{H}}(A^{\mathrm{H}})^{+}A^{+}\mathbf{b} = A^{+}\mathbf{b}$. Therefore, we have proved

THEOREM 1.15*d*. *If A is a complex $m \times n$ matrix of rank k the vector $\mathbf{x}$ which minimizes* (*a*) $(\mathbf{b} - A\mathbf{x}, \mathbf{b} - A\mathbf{x})$ *and* (*b*) $(\mathbf{x}, \mathbf{x})$ *subject to $A\mathbf{x} = \mathbf{b}$ is given by $\mathbf{x} = A^{+}\mathbf{b}$ where A^{+} is specified in Theorem* 1.15*c*.

It is sometimes possible to derive formulae for A^{+} which do not involve finding U and V (see exercises). In practice, the system (1.82) may be ill-conditioned and, indeed, worse than the original system. For consider the case when A is square. Then, if $A\mathbf{x} = \mathbf{b}$ is ill-conditioned, $\det A$ is likely to be small and $\det(A^{\mathrm{T}}A) = (\det A)^2$ will be much smaller again. There are similar arguments if A is not square. For this reason, when A is real, advantage is sometimes taken of the result derived in the previous section that $A = QU_1$ where Q is orthogonal and U_1 upper triangular to solve instead

$$U_1\mathbf{x} = Q^{\mathrm{T}}\mathbf{b}.$$

The condition of the system may then be considerably improved.

Exercises

73. If $\mathbf{u}$ and $\mathbf{v}$ are non-zero real vectors prove that (i)$\mathbf{u}^+ = (\mathbf{u}^T\mathbf{u})^{-1}\mathbf{u}^T$, (ii) $A^+ = A^T/(\mathbf{v}^T\mathbf{v})(\mathbf{u}^T\mathbf{u})$ where $A = \mathbf{u}\mathbf{v}^T$.
74. If $A = BC$ where B is $m \times k$, C is $k \times n$ and all three matrices are of rank k prove that
$$A^+ = C^T(CC^T)^{-1}(B^TB)^{-1}B^T.$$
75. Give an example in which $(AB)^+ \neq B^+A^+$.
76. Calculate the singular values of $\begin{pmatrix} -1 & 1 \\ 1 & -1 \\ -2 & 2 \end{pmatrix}$.
77. If $A\mathbf{x} = \mathbf{b}$ is a consistent system prove that
$$\mathbf{x} = \sum_{i=1}^{k} (\mathbf{v}_i, \mathbf{b})\mathbf{u}_i/\mu_i + \sum_{i=k+1}^{n} \alpha_i\mathbf{u}_i$$
in the notation of this section, the α_i being arbitrary constants.
78. If A is Hermitian with eigenvalues λ_i and orthonormal eigenvectors $\mathbf{x}_i$ show that
$$A = \sum_{i=1}^{n} \lambda_i\mathbf{x}_i\mathbf{x}_i^H.$$
79. Prove from (1.83) that $A^+ = (A^HA)^{-1}A^H$ if A^HA is non-singular.
80. If A^HA is non-singular prove that the vector $\mathbf{x}$ which minimizes r^HBr, where B is positive definite, is $\mathbf{x} = (A^HBA)^{-1}A^HB\mathbf{b}$.