13 Numerical Analysis

13.1 ERRORS

All measurements are subject to errors. It is therefore essential in any scientific endeavor to analyze the results of experiments and to estimate the reliability of the data obtained. In general, experimental errors can be classified as

(i) systematic,
(ii) personal, and
(iii) random.

Systematic errors can arise in measurements made with a given scientific instrument. Often the acquired data do not represent directly the quantity of interest. Furthermore, the instrument used may not be correctly calibrated. These, and other sources of error that are inherent in an experiment, give rise to systematic errors. With care on the part of the experimentalist they can often be detected and, one hopes, corrected.

Personal errors sometimes occur due to inattention, or even prejudice, on the part of an observer. Sometimes an experimentalist would very much like to obtain data that agree with his hypothesis. His reaction – that is, to cheat slightly – may be entirely subjective. Furthermore, he may quite unconsciously make mistakes, either in his observation or in the subsequent presentation of his results. A well known example of the latter type of error was in an early report of the concentration of iron in spinach. According to that communication, spinach was found to be an incredibly rich source of iron. This result was propagated in the literature – including the image created by Popeye to encourage the young to eat more spinach. It was not a bad idea, of course, but a decimal-point error in the early experimental results was responsible for the exaggeration of the iron content of spinach.

If systematic errors can be traced, and perhaps eliminated, and personal errors can be minimized, the remaining random errors can be analyzed by statistical methods. This procedure will be summarized in the following sections.

13.1.1 The Gaussian distribution

Consider the probability

$$\mathcal{W}(x) = p^x q^{n-x} C(n, x) = \frac{p^x q^{n-x} n!}{(n-x)!x!}, \tag{1}$$

as given by Eq. (10-14) for the Bernoulli trials. Out of n trials, p is the number of successes and q the number of failures. When n is large, Eq. (1) can be approximated by a Gaussian distribution. This result is obtained by taking the logarithm and substituting Stirling's approximate expression [Eq. (10-21)] for each factorial. Then,

$$ln\, \mathcal{W}(x) = x\, ln \frac{np}{x} + (n-x)\, ln \frac{nq}{n-x} + \tfrac{1}{2}\, ln \frac{n}{2\pi x(n-x)}. \tag{2}$$

The logarithms in the first two terms in Eq. (2) can be developed in a power series, as shown in Section 2.9, *e.g.*

$$ln \frac{np}{x} = -\, ln\left(1 + \frac{x-np}{np}\right) = -\left[\frac{x-np}{np} - \tfrac{1}{2}\left(\frac{x-np}{np}\right)^2 + \cdots\right], \tag{3}$$

and similarly for the logarithm of $nq/(n-x)$. Then, as $p + q = 1$,

$$\mathcal{W}(x) \approx \frac{1}{\sqrt{2\pi npq}} e^{-(x-np)^2/2npq}, \tag{4}$$

if only the first term in each series is retained (problem 2).

The function $\mathcal{W}(x)$ is that of Gauss, which was discussed in Section 3.4.5. It is presented in Fig. 3-4, although the normalization condition is in this case somewhat different. As $\mathcal{W}(x)\,dx$ represents a probability, its integration over all of the sample space must yield the certainty. The function is thus normalized in the sense that

$$\int_{-\infty}^{\infty} \mathcal{W}(x)\, dx = 1. \tag{5}$$

The approximations introduced above are quite satisfactory close to the origin, although they become questionable further away. However, it is just in the latter regions that the exponential becomes weak. Thus, for most practical purposes Eq. (4) is a good approximation to the probability distribution, as the number of samples becomes large.

It is customary to define the dispersion of the distribution by

$$\sigma^2 = npq. \tag{6}$$

The quantity σ is known as the standard deviation. Furthermore, the mean value of the random variable x is given by $np = \bar{x}$. Then, Eq. (4) becomes

$$\mathcal{W}(x) \approx \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\bar{x})^2/2\sigma^2}. \tag{7}$$

The result obtained here has particular significance in the analysis of random errors of measurement. The substitution $t = (x - \bar{x})/\sigma$ in Eq. (7) leads to the expression

$$\mathcal{W}(t) \approx \frac{1}{\sqrt{2\pi}\sigma} e^{-t^2/2}, \tag{8}$$

which is Gauss's error function. From Eq. (7) it can be concluded that the probability that a given measurement yields a value of x in the interval $\pm x$ is given by

$$\mathcal{W}(x) \approx \frac{1}{\sqrt{2\pi}} \int_{-x}^{+x} e^{-t^2/2}\, dt = \frac{1}{\sqrt{\pi}} \int_{-x}^{+x} e^{-y^2}\, dy, \tag{9}$$

which can be written as

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{+x} e^{-y^2}\, dy. \tag{10}$$

The integral in Eq. (10) is the usual definition of the error function. A closely related function is the complementary error function

$$erfc(x) = 1 - erf(x) = \frac{2}{\sqrt{\pi}} \int_{x}^{\infty} e^{-y^2}\, dy. \tag{11}$$

The error function cannot be evaluated analytically, although it is readily available in the form of tables and evaluated in many computer programs.

13.1.2 The Poisson distribution*

In the previous section it was assumed that quantities of the order of $1/np$ and $1/npq$ were negligible. In that case the mean value of np is a large number. However, in many applications the quantity p is small and the product np remains finite. In this case the distribution is spread out, although the mean value remains small. The resulting distribution is no longer symmetrical. This behavior is illustrated in Fig. 1.

*Siméon Denis Poisson, French mathematician (1781–1840).

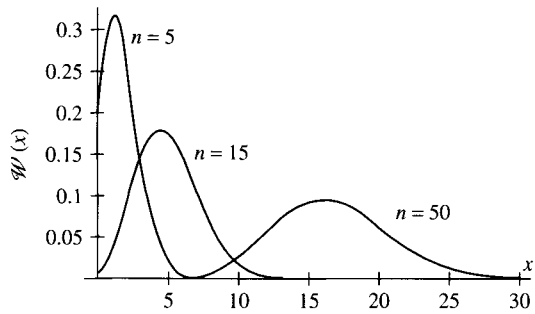

Fig. 1 The Poisson distribution [Eq. (13)] with $p = \frac{1}{3}$.

Return now to the binomial distribution [Eq. (1)] and let n approach infinity. The result is then

$$\mathscr{W}_n(x) = \frac{[1 - (np/n)]^n (np)^x}{x!} \frac{1[1 - (1/n)] \cdots [1 - (x - 1)/n]}{[1 - (np/n)]^x}; \tag{12}$$

thus,

$$\mathscr{W}(x) = \lim_{n \to \infty} \mathscr{W}_n(x) = \frac{(np)^x e^{-np}}{x!}. \tag{13}$$

Note that because the product np remains finite, the second factor in Eq. (12) approaches unity in the limit. Similarly, from the definition of the exponential (Section 1.4), $lim_{n \to \infty}[1 - (np/n)]^n = e^{-np}$. Equation (13) is an expression of the Poisson distribution.

The Poisson distribution is usually applied in the case of small values of np. For large values it is well approximated by the normal, or Gaussian, distribution. For a given value of p the distribution becomes more nearly symmetric with increasing values of n. It becomes wider and approaches a Gaussian form, as shown in Fig. 1. This distribution, and others, are often approximated by the normal (Gaussian) distribution in the region near the maximum. Although there are many applications of the Poisson distribution, the best known is in the area of atomic physics. The result of counting particles emitted by a radioactive substance is usually described by the Poisson distribution.

13.2 THE METHOD OF LEAST SQUARES

The normal distribution, as expressed by Eq. (7), can be employed in the analysis of random errors. If the error in a given measurement i is represented

by x_i, the probability that it lies between x_i and $x_i + \mathrm{d}x_i$ can be written as

$$P_i = \frac{1}{\sigma\sqrt{\pi}} e^{-x_i^2/\sigma^2}\,\mathrm{d}x_i \ . \tag{14}$$

Hence, for n independent measurements the combined probability is given by

$$P = \prod_{i=1}^{n} P_i = \left[\frac{1}{\sigma\sqrt{\pi}}\right]^n exp\left[-(1/\sigma^2)\sum_{i=1}^{n} x_i^2\right] \mathrm{d}x_1\,\mathrm{d}x_2 \ldots \mathrm{d}x_n \ . \tag{15}$$

For a given value of σ the probability is maximum when the sum in the exponent is minimum. Thus, the minimization of $\sum_{i=1}^{n} x_i^2$ becomes the criterion for the most probable value obtainable from n equally reliable measurements. This result is the basis for the various curve-fitting procedures that are commonly used in the analysis of experimental data.

It is very often of interest to fit a set of data points to a straight line. While it is possible to draw a line on a graph by eye, it is clearly preferable to have an objective method to establish the line with respect to the experimental points. Suppose that the straight line is specified by

$$Y_i = mx_i + b, \tag{16}$$

where m is its slope and b its intercept on the ordinate axis. The deviation of each point from the line is equal to $y_i - Y_i = \varepsilon_i$. The sum of squares, which is then given by

$$S = \sum_{i=1}^{n}(y_i - Y_i)^2 = \sum_{i=1}^{n}(y_i - mx_i - b)^2, \tag{17}$$

is the quantity to be minimized with respect to the two parameters m and b. Thus,

$$\frac{\partial S}{\partial b} = -2\sum_{i=1}^{n}(y_i - mx_i - b) = 0 \tag{18}$$

and

$$\frac{\partial S}{\partial m} = -2\sum_{i=1}^{n}(y_i - mx_i - b)x_i = 0. \tag{19}$$

The resulting expressions for the two parameters can be expressed as a function of the averages $\bar{x} = 1/n\sum_{i=1}^{n} x_i$, $\bar{y} = 1/n\sum_{i=1}^{n} y_i$, $\overline{x^2} = 1/n\sum_{i=1}^{n} x_i^2$ and $\overline{xy} = 1/n\sum_{i=1}^{n} x_i y_i$ in the form

$$m = \frac{\bar{x}\,\bar{y} - \overline{xy}}{\bar{x}^2 - \overline{x^2}} \tag{20}$$

and

$$b = \bar{y} - m\bar{x}. \tag{21}$$

Thus, the straight line corresponding to the best fit is established without ambiguity.

Many computer programs exist to achieve the linear, least-squares fitting (linear regression) to a given set of data. It is, however, worthwhile to apply the method to a simple problem in order to understand the basis. The data presented in Fig. 1-1 represent the weight of Miss X as a function of the date. While there is no reason to suppose that there is a linear relationship implied, the straight line provides her with an indication of her rate of weight loss, namely, the slope of the line. The least-squares fit to the data yields the relation $Y = -0.12x + 69.6$, as shown in the figure.

13.3 POLYNOMIAL INTERPOLATION AND SMOOTHING

Consider the simplest method of interpolating between two successive data points. It is linear, midpoint interpolation. This procedure is illustrated in Fig. 2. The ordinate value of the interpolated point is given by

$$Y = \tfrac{1}{2}y_{-1/2} + \tfrac{1}{2}y_{+1/2}, \tag{22}$$

the average of the values at the two points. The slope of the line segment connecting the two points is easily found as

$$\frac{dY}{dx} = -y_{-1/2} + y_{+1/2}, \tag{23}$$

where the interval Δx has been taken equal to one. The coefficients appearing in Eqs. (22) and (23), when arranged in matrix form yield

$$\boldsymbol{A} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ -1 & 1 \end{pmatrix}, \tag{24}$$

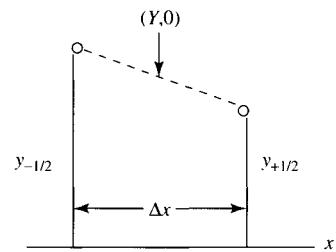

Fig. 2 Linear, midpoint interpolation.

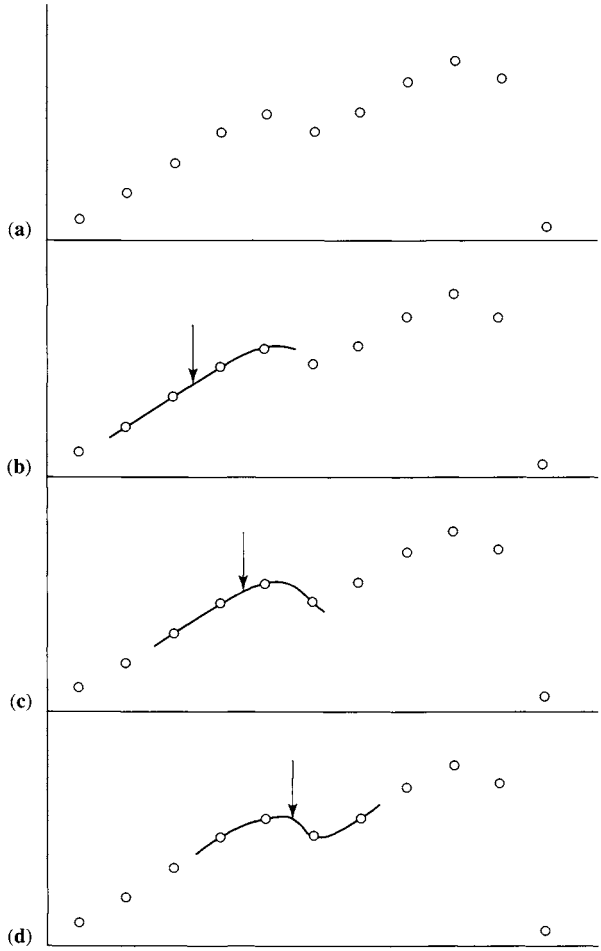

Fig. 3 Convolution with a four-point polynomial spline.

a matrix that will be defined in the following general development of the interpolation method.

A very simple example of interpolation was provided above with the use of a linear function. However, interpolation involving polynomials of higher degree, with more points on either side of the interpolated one is relatively complicated. In effect, the matrix $\boldsymbol{A}$ is then not easily found by inspection.

In the precomputer era a series of experimental points on a graph such as shown in Fig. 3a, was "fitted" with a spline – a sort of flexible ruler that could be adjusted to fit approximately a certain limited number of points. This

procedure is illustrated in Fig. 3b–d. The spline forms a smooth curve which can be used by the draftsman to interpolate between successive data points. Furthermore, if there is a certain amount of scatter in the data, a smoothing operation can be carried out, albeit with certain artistic licence.

In computer programs that are now devoted to these problems the interpolation and smoothing of data are special cases of convolution of the data with a set of numerical coefficients, represented for the present by the vector $\boldsymbol{A}_1(x - x_i)$. These coefficients can be determined in advance and placed in memory to be used as needed. If the data points are entered as a vector $\boldsymbol{y}(x_i)$, the convolution can be written in the form

$$\boldsymbol{Y}(x) = \sum_i \boldsymbol{A}_1(x - x_i) \cdot \boldsymbol{y}(x_i). \tag{25}$$

This expression is the discrete form of the convolution integral defined in Eq. (11-13).

In practice the experimental values $\boldsymbol{y}(x_i)$ are usually measured at equally spaced abscissa values and the convolution is applied in succession to limited portions of the experimental data. In principle the equal spacing of data points along the x axis is not necessary, although it is essential in most numerical applications. It is useful to define the difference $\boldsymbol{y} - \boldsymbol{Y} = \boldsymbol{\varepsilon}$, the vector of "errors" at each point. The chosen function $\boldsymbol{Y}(x)$ will be assumed here to be a polynomial of degree $k - 1$, although it can be a more general function. Then, if $\boldsymbol{\theta}$ is a vector composed of the k coefficients in the polynomial

$$\boldsymbol{Y}(x) = \boldsymbol{X}\boldsymbol{\theta}, \tag{26}$$

where $\boldsymbol{X}$ is a $2m \times k$ matrix (with $2m \geq k$) whose elements are powers of x. Specifically this matrix is of the general form

$$\boldsymbol{X} = \begin{pmatrix} (-m + \frac{1}{2})^0 & (-m + \frac{1}{2})^1 & \cdots & (-m + \frac{1}{2})^{k-1} \\ (-m + \frac{3}{2})^0 & & & \\ \vdots & & & \\ (-\frac{1}{2})^0 & & & \vdots \\ (+\frac{1}{2})^0 & & & \\ \vdots & & & \\ (m - \frac{1}{2})^0 & (m - \frac{1}{2})^1 & \cdots & (m - \frac{1}{2})^{k-1} \end{pmatrix}. \tag{27}$$

Then

$$\boldsymbol{\varepsilon} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta} \tag{28}$$

and it is the quantity $S \equiv \tilde{\boldsymbol{\varepsilon}}\boldsymbol{\varepsilon}$ that is minimized in the application of the least-squares criterion used in the previous section. Thus,

$$\frac{\partial S}{\partial \theta_j} = 0, \qquad j = 1, 2, \ldots, k \tag{29}$$

which leads to the matrix relation

$$2\tilde{\boldsymbol{X}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) = \boldsymbol{0}. \tag{30}$$

Its solution for $\boldsymbol{\theta}$ is in the form (problem 8).

$$\boldsymbol{\theta} = (\tilde{\boldsymbol{X}}\boldsymbol{X})^{-1}\tilde{\boldsymbol{X}}\boldsymbol{y} = \boldsymbol{A}\boldsymbol{y} \tag{31}$$

and thus the matrix $\boldsymbol{A}$ [see Eq. (25)] can be constructed from the matrix $\boldsymbol{X}$. It should be noted that the derivatives of order n can also be evaluated, as

$$\left(\frac{\partial^n \boldsymbol{Y}}{\partial \boldsymbol{X}^n}\right)_{X=0} = n!\,\theta_{n+1}\ . \tag{32}$$

The first row of the matrix $\boldsymbol{A}$ consists of the coefficients for the interpolation of the values of y_i, while subsequent rows provide the values of the corresponding derivative coefficients.

The application of the general method can be illustrated by the example shown in Fig. 3. The series of data points is fitted by a polynomial of second degree. Two points will be employed on either side of the point to be interpolated. Thus, $m = 2$ and the matrix $\boldsymbol{X}$ is of the form

$$\boldsymbol{X} = \begin{pmatrix} (-\frac{3}{2})^0 & (-\frac{3}{2})^1 & (-\frac{3}{2})^2 \\ (-\frac{1}{2})^0 & (-\frac{1}{2})^1 & (-\frac{1}{2})^2 \\ (+\frac{1}{2})^0 & (+\frac{1}{2})^1 & (+\frac{1}{2})^2 \\ (+\frac{3}{2})^0 & (+\frac{3}{2})^1 & (+\frac{3}{2})^2 \end{pmatrix} = \begin{pmatrix} 1 & -\frac{3}{2} & \frac{9}{4} \\ 1 & -\frac{1}{2} & \frac{1}{4} \\ 1 & +\frac{1}{2} & \frac{1}{4} \\ 1 & +\frac{3}{2} & \frac{9}{4} \end{pmatrix} \tag{33}$$

and

$$\boldsymbol{A} = (\tilde{\boldsymbol{X}}\boldsymbol{X})^{-1}\tilde{\boldsymbol{X}} = \begin{pmatrix} -\frac{1}{16} & \frac{9}{16} & \frac{9}{16} & -\frac{1}{16} \\ -\frac{3}{10} & -\frac{1}{10} & \frac{1}{10} & \frac{3}{10} \\ \frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & \frac{1}{4} \end{pmatrix} \tag{34}$$

(problem 9).

The coefficients appearing in the first row of the matrix $\boldsymbol{A}$ provide the weight attributed to each of the four data points that determine the interpolated point indicated by the arrow in Fig. 3b. The elements in the second and third

rows are similarly employed if there is an interest in calculating the first and second derivatives, respectively. The determination of the interpolated points at $i + 1$ and $i + 2$ is carried out in the same way, as indicated in Figs. 3c and 3d. This procedure is continued to complete the convolution represented by Eq. (25). The matrices $\boldsymbol{A}$ that have been calculated for polynomials of various degrees and number of points, have been published and are available in certain computer programs.

The interpolation method outlined above can be applied as well to the "smoothing" of experimental data. In this case a given experimental point is replaced by a point whose position is calculated from the values of m points on each side. The matrix $\boldsymbol{X}$ then contains an odd number of columns, namely $2m + 1$. The matrices $\boldsymbol{A}$ have also been tabulated for this application. This smoothing method has been used for a number of years by molecular spectroscopists, who generally refer to it as the method of Savitzky and Golay.*

13.4 THE FOURIER TRANSFORM

13.4.1 The discrete Fourier transform (DFT)

The Fourier transform was defined by Eq. (11-2) as

$$F(k) = \int_{-\infty}^{\infty} f(x)e^{2\pi ikx}\,\mathrm{d}x. \tag{35}$$

As experimental data represented by $f(x)$ are usually symmetrical (even), or can be made so, it is the cosine transform that is appropriate, *viz.*

$$F(k) = \int_{-\infty}^{\infty} f(x)\cos(2\pi kx)\,\mathrm{d}x. \tag{36}$$

In spectroscopy, for example, the Fourier transform of an interferogram, $f(x)$ is sampled at regular intervals, Δx. Equation (36) is then replaced by the summation

$$F(k) = \Delta x \sum_{m=-M/2}^{m=M/2-1} f(m\Delta x)\cos(2\pi km\Delta x), \tag{37}$$

where M is the number of points sampled. As $f(x)$ has been assumed to be an even function, Eq. (37) can be written as

*George Boris Savitzky, American physical chemist (1925–); Marcel J. E. Golay, Swiss-American physicist (1902–).

$$F(k) = \Delta x \left\{ F(0) + \sum_{m=1}^{m=M/2-1} f(m\Delta x)\cos(2\pi km\Delta x) + f[(M/2)\Delta x]\cos[2\pi k(M/2)\Delta x] \right\} \tag{38}$$

if an even number of points has been chosen. The total number of terms to be evaluated is then reduced from M to $(M/2) + 1$.

The computer evaluates the cosine functions appearing in Eq. (38) from their series expansions, as given by Eq. (1-34). As M is usually a large number, the time required for the evaluation of the sum can be considerable. However, the arguments of the cosines are simply related because the data points are separated by the constant interval Δx. Given the relation

$$\cos\alpha + \cos\beta = 2\cos\left[\tfrac{1}{2}(\alpha+\beta)\right]\cos\left[\tfrac{1}{2}(\alpha-\beta)\right], \tag{39}$$

(problem 10), its application to the present problem can be written as

$$\cos[(n+1)\eta] = 2\cos n\eta\cos\eta - \cos[(n-1)\eta] \tag{40}$$

where $\eta = \frac{1}{2}(\alpha - \beta)$. This recurrence relation is known as that of Chebyshev. If $\eta = 2\pi km\Delta x$ is the argument of the cosines in the summation [Eq. (38)], all of the other cosines can be calculated from this expression. The result is a considerable saving in the calculation time.

Furthermore, as the output of this calculation is normally in the form of regularly spaced data points, it can be expressed as

$$F(n\Delta k) = \Delta x \sum_{m=1}^{m=M/2-1} f(m\Delta x)\cos(2\pi mn\Delta k\Delta x). \tag{41}$$

In this case the cosines can be arranged in matrix form, *viz.*

$$\boldsymbol{C} = \begin{pmatrix} 1 & 1 & 1 & 1 & \cdots \\ 1 & \cos u & \cos 2u & \cos 3u & \\ 1 & \cos 2u & \cos 4u & \cos 6u & \\ 1 & \cos 3u & \cos 6u & \cos 9u & \\ \vdots & & & & \ddots \end{pmatrix}, \tag{42}$$

where $u = 2\pi\Delta k\Delta x$. If there are the same number of points in x-space as in k-space, as is usually the case, the matrix $\boldsymbol{C}$ is square and symmetrical with respect to the principal diagonal. With the aid of the Chebyshev recurrence relation the elements of this matrix can be rapidly calculated, once $\cos u$ has been evaluated. If the ordinate values of $f(x)$ are arranged in the form of a

column vector $\boldsymbol{f}$ and those of $(\Delta x)F(k)$ as a vector $\boldsymbol{F}$, the discrete Fourier transform is then calculated by the simple matrix multiplication,

$$\boldsymbol{F} = \boldsymbol{C}\boldsymbol{f}. \tag{43}$$

This numerical method of computing the DFT is sometimes referred to as the slow Fourier transform – by comparison with the fast Fourier transform (FFT) described in the following section.

13.4.2 The fast Fourier transform (FFT)

The fast Fourier transform can be carried out by rearranging the various terms in the summations involved in the discrete Fourier transform. It is, in effect, a special book-keeping scheme that results in a very important simplification of the numerical evaluation of a Fourier transform. It was introduced into the scientific community in the mid-sixties and has resulted in what is probably one of the few significant advances in numerical methods of analysis since the invention of the digital computer.

The basic argument in the FFT algorithm is determined by the initial requirement that

$$k_{max} = \frac{M}{2}\Delta k = \frac{1}{2\Delta x}, \tag{44}$$

where M is the number of measured points. Thus, the point-by-point accumulation of data in k space is made symmetrically with respect to the maximum at k_{max}. Equation (44) corresponds to $\Delta k \Delta x = M^{-1}$. Furthermore, the matrix $\boldsymbol{C}$ is always taken to be square, *viz.* $N = M$ and of rank 2^{ℓ}, where ℓ is an integer. Under these conditions the cosines appearing in the matrix $\boldsymbol{C}$ will all be of the form $cos(2\pi mn/M)$. Here, the (independent) indices n and m have been chosen to run from zero to $M - 1$. In this case the general expression for the discrete Fourier transform [Eq. (41)] can be written as

$$F(n) = \Delta x \sum_{m=0}^{M-1} f(m)\, cos\left(\frac{2\pi mn}{M}\right). \tag{45}$$

Furthermore, the choice of the cosine transform implies that $f(m)$ is symmetrical about its maximum value; thus, $f(m - M) = f(m)$.

With the arguments of the preceding paragraph in mind it becomes possible to construct the functions $F(n)$ in k space. This procedure is best explained with the aid of an example. Consider the simple case in which $M = 8 = 2^3$. With $\Delta x = \frac{1}{2}$, Eq. (45) yields the expressions

$$\begin{aligned} F(0) &= \tfrac{1}{2}\{f(0) + f(4) + f(2) + f(6) + f(1) + f(5) + f(3) + f(7)\} \\ &= \tfrac{1}{2}\{f(0) + f(4) + 2f(2) + 2f(1) + 2f(3)\}, \end{aligned} \tag{46}$$

$$F(1) = \tfrac{1}{2}\{f(0) - f(4) + 0f(2) + 0f(6) + [f(1) - f(5) - f(3) + f(7)]\cos\tfrac{\pi}{4}\}$$
$$= \tfrac{1}{2}\{f(0) - f(4) + 2[f(1) + -f(5)]\cos\tfrac{\pi}{4}\}, \quad (47)$$

$$F(2) = \tfrac{1}{2}\{f(0) + f(4) - f(2) - f(6)\}, \quad (48)$$

$$F(3) = \tfrac{1}{2}\{f(0) - f(4) - [f(1) - f(5) - f(3) + f(7)]\cos\tfrac{\pi}{4}\} \quad (49)$$

and

$$F(4) = \tfrac{1}{2}\{f(0) + f(4) + f(2) + f(6) - f(1) - f(5) - f(3) - f(7)\} \quad (50)$$

(problems 11).

The order in which the functions $f(m)$ are presented in the above relations is specific. First, note that all of the functions of even values of m are specified before those of odd values. Moreover, the order employed here is referred to as reverse binary order,* which does not correspond to the order that might be intuitively established, namely, $m = 0, 1, 2, \ldots, 7$. Furthermore, each is multiplied by a value of $\cos(2\pi nm/8)$, as $M = 8$ in this case. Clearly, Eqs. (46–50) can be recast in matrix form. However, with the addition of the symmetry conditions $F(5) = F(3)$, $F(6) = F(2)$ and $F(6) = F(1)$ the appropriate 8×8 matrix $\boldsymbol{C}$ can be easily constructed. On the other hand, if the inverse binary order is also imposed on the elements of the vector $F(n)$, a considerable simplification results.

Continuing with the eight-point transform, Eq. (43) can be written in the form

$$\begin{pmatrix} F(0) \\ F(4) \\ F(2) \\ F(6) \\ F(1) \\ F(5) \\ F(3) \\ F(7) \end{pmatrix} = \frac{1}{2} \left(\begin{array}{cccc:cccc} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & 0 & 0 & 0 & 0 \\ 1 & 1 & -1 & -1 & 0 & 0 & 0 & 0 \\ \hdashline 1 & -1 & 0 & 0 & c & -c & -c & c \\ 1 & -1 & 0 & 0 & -c & c & c & -c \\ 1 & -1 & 0 & 0 & -c & c & c & -c \\ 1 & -1 & 0 & 0 & c & -c & -c & c \end{array}\right) \begin{pmatrix} f(0) \\ f(4) \\ f(2) \\ f(6) \\ f(1) \\ f(5) \\ f(3) \\ f(7) \end{pmatrix}, \quad (51)$$

*In binary algebra [Boolean, after George Boole, British mathematician (1815–1864)] the indices $0, 1, 2, \ldots, 7$ are represented by 000, 001, 010, 011, 100, 101, 110, and 111, respectively. The reversal of these binary numbers yields the values of m in the order indicated in Eqs. (46)–(50).

(problem 12). Note that the matrix $\boldsymbol{C}$ is now symmetric with respect to the principal diagonal. Furthermore, there is symmetry with respect to the parity of both m and n. In the 4×4 block that corresponds to both m and n even the columns occur in identical pairs, while in the block with both m and n odd the result is analogous, although the signs are reversed. It is apparent that only four matrix elements need be evaluated, *viz.* $c = cos(\pi/4)$, plus the trivial ones $cos\ 0 = 1, cos\ \pi/2 = 0$ and $cos\ \pi = -1$. The matrix $\boldsymbol{C}$ given in Eq. (51) should be compared with that obtained by application of Eq. (42). Note that the factor $\frac{1}{2}$ is just $\Delta x = 1/\ell$.

It is instructive to consider a specific example of the method outline above. The triangle function $(1/\ell) \wedge (x/\ell)$ was discussed in Section 11.1.2. It was pointed out there that it arises in dispersive spectroscopy as the slit function for a monochromator, while in Fourier-transform spectroscopy it is often used as an apodizing function.* Its Fourier transform is the function $sinc^2$, as shown in Fig. (11-2). The eight points employed to construct the normalized triangle function define the matrix

$$\boldsymbol{f} = \begin{pmatrix} \frac{1}{2} \\ 0 \\ \frac{1}{4} \\ \frac{1}{4} \\ \frac{3}{8} \\ \frac{1}{8} \\ \frac{1}{8} \\ \frac{3}{8} \end{pmatrix} \tag{52}$$

where it is essential to preserve the order of the elements as given in Eq. (51). Multiplication of the vector $\boldsymbol{f}$ by the matrix $\boldsymbol{C}$ of Eq. (51) yields

$$\boldsymbol{F} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ \frac{1}{4}(1+c) \\ \frac{1}{4}(1-c) \\ \frac{1}{4}(1-c) \\ \frac{1}{4}(1+c) \end{pmatrix}, \tag{53}$$

*An apodizing function is employed to reduce oscillations in an observed spectrum due to discontinuities at the ends of an interferogram.

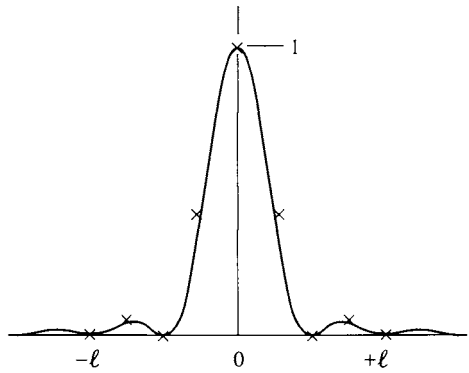

Fig. 4 The Fourier transform of the triangle, $(1/\ell)\,\Lambda(x/\ell)$. The points calculated with the use of Eq. (53) are indicated by $\times$.

which is the desired Fourier transform. This result is compared with the function $sinc^2$ in Fig. 4. Considering the limited number of points employed, this agreement is excellent. Clearly, it would be necessary to carry out the transform with a larger number of points to obtain a more convincing description of the function $sinc^2$.

The symmetry and simplicity of the matrix $\boldsymbol{C}$ (and hence the extreme rapidity of the FFT) is determined by the particular order employed in both the input vector $\boldsymbol{f}$ and the output $\boldsymbol{F}$. Thus, both sets of data must be rearranged from what would be normally expected. While this problem represents an inconvenience for a programmer, it is carried out automatically in available programs. Although it would probably go un-noticed by the user, it is important for him or her to understand the fundamental algorithm of the FFT, which is based on the inverse binary order explained here.

13.4.3 An application: interpolation and smoothing

Both interpolation and smoothing of experimental data are of particular importance in all branches of spectroscopy. One approach to this problem was considered in Section 13.3. However, with the development of the FFT another, often more convenient, method has become feasible. The basic argument is illustrated in Fig. 5. Given a particular problem whose solution may appear to be difficult, it is sometimes possible to resolve it *via* recourse to the Fourier transform.

Consider the problem of smoothing an experimental curve, such as represented in Fig. 6a. It might well correspond to a spectrum, as observed in absorption, emission or, say, Raman scattering. The noise is, usually at least,

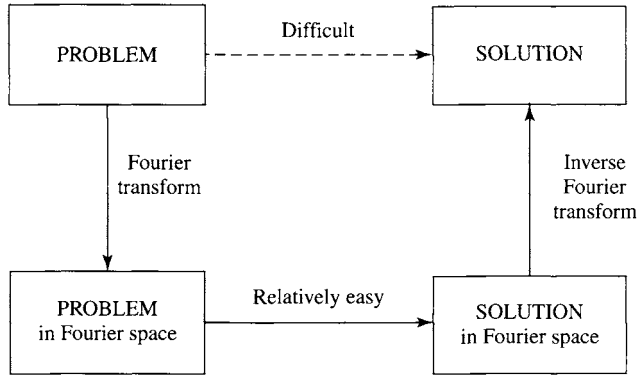

Fig. 5 The solution of a problem with the aid of the Fourier transform.

largely random. As it can be reasonably represented by a normal distribution, it would seem appropriate to smooth the observed data by convolution with a Gaussian function. This operation is conveniently carried out by first transforming the data into Fourier space, where they can simply be multiplied by the appropriate Gaussian, as the latter is of course the Fourier transform of the original Gaussian (see Fig. 11-3). The product is subsequently transformed back as a smoothed spectrum. An example is shown in Fig. 6. If convoluted by a Gaussian of width 5 points, the experimental spectrum of Fig. 6a is smoothed as in Fig. 6b. Further smoothing, for example by a Gaussian of width 30 points, results in a substantial loss of information. Thus, as in the case of polynomial smoothing, this method must be used with discretion.

The principle presented above can also be applied to interpolate points in an experimental profile. If the original function $\boldsymbol{F}$ (a spectrum, for example) is transformed with the use of the FFT algorithm, the result is a function of the same number of points in Fourier space. It might be, for example, the original interferogram $\boldsymbol{f}$ that was used to generate the spectrum. The number of points can be augmented by simply adding zeros to the vector $\boldsymbol{f}$. If the number of points is doubled, the result of carrying out the inverse transform is to produce the vector $\boldsymbol{F}$ with twice as many points as before. This procedure corresponds exactly to midpoint interpolation of the original spectrum by the function *sinc*. This result should become evident if it is recalled that the *sinc* function is the Fourier transform of the boxcar (Fig. 11-1), whose width has been doubled by the operation of "zero filling". Obviously, no new information is obtained by this procedure, but the result may be of esthetic value in the presentation of the spectrum. This method yields better results than the more usual polynomial interpolation method presented in Section 13.3.

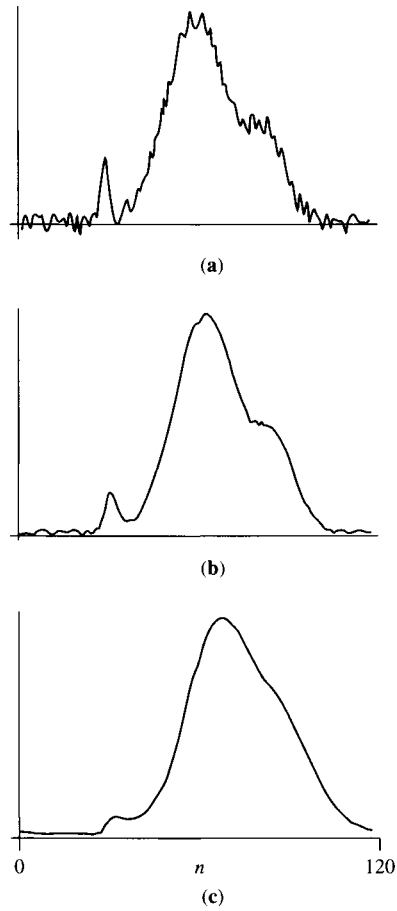

Fig. 6 Gaussian convolution of an experimental profile: (**a**) raw spectrum; (**b**) after convolution by a Gaussian of width 5 points; (**c**) after convolution by a Gaussian of 30 points. The ordinate scale is arbitrary.

13.5 NUMERICAL INTEGRATION

The numerical evaluation of definite integrals can be carried out in several ways. However, in all cases it must be assumed that the function, as represented by a table of numerical values, or perhaps a known function, is well behaved. While this criterion is not specific, it suggests that the functions having pathological problems, *e.g.* singularities, discontinuities, ..., may not survive under the treatment in question.

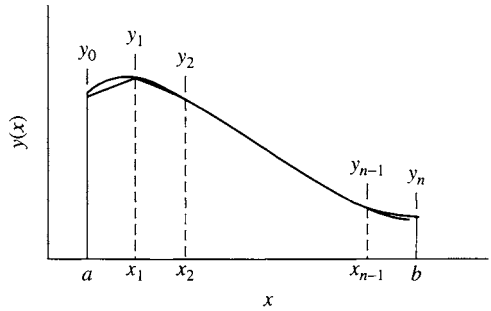

Fig. 7 The trapezoids.

13.5.1 The trapezoid rule

Consider a function $y(x)$ as shown in Fig. 7. If the interval of integration, say from a to b, is divided in n equal intervals,

$$x_k = a + k\Delta x, \tag{54}$$

where $k = 0, 1, 2, \ldots, n$, $\Delta x = (b - a)/n$ and $y_k = y(x_k)$. If $y(x)$ is expanded in a Taylor series (Section 2.9),

$$y(x) = y(x_0) + (x - x_0)y'(x_0) + \tfrac{1}{2!}(x - x_0)^2 y''(x_0) + \cdots. \tag{55}$$

The integral of this expression from x_0 to x_1 is given by

$$\begin{aligned}\int_{x_0}^{x_1} y(x)\,dx &= (x - x_0)y(x_0) + \left[\frac{(x - x_0)^2}{2!}y'(x_0)\right]_{x_0}^{x_1} \\ &\quad + \left[\frac{(x - x_0)^3}{3!}y''(x_0)\right]_{x_0}^{x_1} + \cdots \\ &= \Delta x y(x_0) + \left[\frac{\overline{\Delta x}^2}{2!}y'(x_0)\right] + \left[\frac{\overline{\Delta x}^3}{3!}y''(x_0)\right] + \cdots. \end{aligned} \tag{56}$$

From Eq. (55)

$$y_1 = y(x_1) = y_0 + \Delta x y'(x_0) + \frac{\overline{\Delta x}^2}{2!}y''(x_0) + \cdots, \tag{57}$$

which when multiplied by $\overline{\Delta x}^2/2$ and substituted in Eq. (56) yields

$$\int_{x_0}^{x_1} y(x)\,dx = \frac{\Delta x}{2}[y(x_0) + y(x_1)] - \frac{\overline{\Delta x}^3}{12}y''(x_0) + \cdots. \tag{58}$$

The integral over the entire region can then be written as

$$\int_a^b y(x)\,dx = \int_{x_0}^{x_n} y(x)\,dx \approx \frac{\Delta x}{2}[y_0 + 2y_1 + 2y_2 + \cdots + 2y_{n-1} + y_n] + \mathcal{O}(\overline{\Delta x}^2), \tag{59}$$

where the term $\mathcal{O}(\overline{\Delta x}^2) = -(\overline{\Delta x}^2/12)[y'(b) - y'(a)]$ represents the error in the simple trapezoid method of numerical integration.

13.5.2 Simpson's rule*

In the method presented in the previous section each vertical "slice" was defined by two successive points, $x_0, x_1; x_1, x_2$; *etc.* If now the successive points are selected three-by-three, they can be connected by a parabola. The approximate integral over the first two slices can then be written as

$$\int_{x_0}^{x_2} y(x)\,dx \approx \frac{\Delta x}{3}[y(x_0) + 4y(x_1) + y(x_2)]. \tag{60}$$

The correction to this expression involves multiple derivatives, although the basic equation, Eq. (60), does not. The development of this result, as above for the trapezoid rule, leads to the relation for the integral over the range a to b in the form

$$\int_a^b y(x)\,dx \approx \frac{\Delta x}{3}[y(x_0) + 4y(x_1) + 2y(x_2) + 4y(x_3) + \cdots + 2y(x_{n-2}) + 4y(x_{n-1}) + y(x_n)]. \tag{61}$$

It should be noted that n has been assumed here to be even. Equation (61), without the inclusion of the correction term in $\overline{\Delta x}^4$, is the one usually used in the numerical evaluation of integrals. When higher precision is required, and a suitable computer is available, the algorithm described in the following section can be employed.

13.5.3 The method of Romberg†

The two well-known methods of numerical integration described in the previous sections can be generalized. Represent the sum on the right-hand side of Eq. (59) as $S_0(n)$. This function converges but very slowly towards

*Thomas Simpson, English mathematician (1710–1761).

†Werner Romberg, German mathematician (1909–).

the exact value of the integral as $n \to \infty$. However, the following method is much more efficient.

Reconsider Eq. (59) in the form

$$S_0(n) = \int_a^b f(x)\,dx + C(\overline{\Delta x}^2) + \mathcal{O}(\overline{\Delta x}^4), \tag{62}$$

where C is a constant. If the number of intervals n is now doubled, this expression becomes

$$S_0(2n) = \int_a^b f(x)\,dx + \tfrac{1}{4}C(\overline{\Delta x}^2) + \mathcal{O}(\overline{\Delta x}^4). \tag{63}$$

By eliminating the constant C between Eqs. (62) and (63) the relation

$$\begin{aligned}\int_a^b f(x)\,dx &= \frac{4S_0(2n) - S_0(n)}{3} + \mathcal{O}(\overline{\Delta x}^4) \\ &= S_1(2n) + \mathcal{O}(\overline{\Delta x}^4)\end{aligned} \tag{64}$$

can easily be established. If the correction terms are neglected, this result is equivalent to Simpson's rule for the division of the interval a,b in $2n$ equal slices. If the process of halving the intervals is continued, the expression

$$S_0(n) = \frac{\overline{\Delta x}}{2}[y(x_0) + 2y_1 + 2y_2 + \cdots + 2y_{n-1} + y_n] \qquad n = 1, 2, 4, 8, \ldots \tag{65}$$

can be obtained for the application of the simple trapezoid rule for each value of n. This result is the starting point for the application of Romberg's method. It is continued by application of the recursion relation that is obtained by generalizing of Eq. (64). It is given by

$$S_m(2n) = \frac{4^m S_{m-1}(2n) - S_{m-1}(n)}{4^m - 1}, \tag{66}$$

with $m = 1, 2, 3, \ldots$ and $n = 2^{m-1}, 2^m, 2^{m+1}, \ldots$.

As an example of the application of Romberg's method, consider the integral

$$I(T/\theta_D) = \int_0^{T/\theta_D} \frac{x^3}{e^x - 1}\,dx, \tag{67}$$

that arises in Debye's theory of the heat capacity of solids.* In Eq. (67), T is the absolute temperature and θ_D is referred to as the Debye temperature. In the low-temperature limit the integral in Eq. (67) is given approximately by

*Petrus Debye, Dutch-American physicist and chemist (1884–1966).

$\frac{1}{3}(T/\theta_D)^3$. This limiting expression is known as Debye's third-power law for the heat capacity (problem 15). It is employed in thermodynamics to evaluate the low-temperature contribution to the absolute entropy.

The integral in Eq. (67) cannot be evaluated analytically. However, for a given upper limit T/θ_D, it can be calculated, in principle to any desired precision, with the application of the methods outlined above. The results for $T/\theta_D = 1.6$ are summarized in Table 1.

Table 1 Evaluation of Eq. (67) with $T/\theta_D = 1.6$.

n	(Trapezoid) $S_0(n)$	(Simpson) $S_1(n)$	(Milne) $S_2(n)$	$S_3(n)$
1	0.8289332462284			
2	0.74868638720142	0.72193743419243		
4	0.724310280204	0.71618491120487	0.7158014096724	
8	0.71795245734086	0.71583318305313	0.71580973450967	0.71580986664995
16	0.71634660087305	0.71581131538377	0.71580985753913	0.71580985949197
32	0.71594411304265	0.7158099504325	0.71580985943573	0.71580985946584
64	0.71584342712365	0.71580986515063	0.71580985946513	0.71580985946559

The slowest part of the construction of this table is the evaluation of the entries in the first column. The simple trapezoid rule, as given by Eq. (65), is applied with successive sectioning of the slices. It can be seen that by descending the column a limiting value can, in principle, be obtained. However, the convergence is very slow. With the use of the recursion relation given by Eq. (66), each successive pair of entries in the first column can be employed to calculate the values presented in the second column. The results shown for this example are equal to those obtained by Simpson's method [Eq. (61)].

The third column of Table 1 is calculated by applying the recursion relation to the values shown in the second column, *etc.* It corresponds to the method of Milne.* It is apparent that the convergence becomes much more rapid with each successive column. For this particular example the same limiting values is obtained as either n or m becomes very large.

13.6 ZEROS OF FUNCTIONS

13.6.1 Newton's method

Given a function $f(x)$, if its derivatives can be evaluated numerically, Newton's method can often serve as an algorithm for the determination of

*William. E. Milne, American mathematician (1890–1971).

the zeros of the equation $f(x) = 0$. Assume that x_0 is an estimated value of one of the roots. Then, at least in principle, an improved value of the root is given by

$$x = x_0 + \Delta x, \tag{68}$$

where $\Delta x = -f(x_0)/f'(x_0)$. The next approximation is found by replacing x_0 by x in Eq. (68) to get a new value of Δx. This procedure is continued as long as is necessary to obtain the desired accuracy. Usually, after a few successive approximations, the value of the derivative will change little; hence, $f'(x)$ need not be recalculated each time. It should be obvious that the solution will be found more rapidly if the initial value x_0 is wisely chosen.

13.6.2 The bisection method

In the application of the bisection method it is assumed only that the function $f(x)$ is continuous. It requires that two initial values of x, say x_a and x_b, be chosen so that they straddle the desired zero. Thus, $f(x_a)$ and $f(x_b)$ will have opposite signs and their product will be negative. Now, take the midpoint $x_m = (x_a + x_b)/2$ and calculate $f(x_m)$. If, for example, the product $f(x_a)f(x_m) < 0$, the desired root lies between x_a and x_m. The midpoint between these two limits is then calculated and the process is repeated to the desired degree of accuracy. Here again, the better the choice of the initial limits, the fewer the number of bisections that will be required.

13.6.3 The roots: an example

The function $f(x) = (5-x)e^x - 5$ arises in the theory of black-body radiation. Obviously, it has a zero at $x = 0$. A plot of this function (Fig. 8) shows that it has a second zero near $x = 5$. As this function appears to be well behaved in this region, Newton's method might be expected to yield a value for the second root.

If, as a guess, the initial value of x is chosen to be $x_0 = 4.5$, convergence to the value $x = 4.96511$ will occur within a few iterations. On the other side, where $x > 5$, even wilder guesses will yield the same, correct answer. However, if $x_0 = 4$ is taken as a starting point, disaster will result. Reference to the plot of this function in Fig. 8 indicates that this point is at the maximum. As the slope is then equal to zero, the computer will yield a "division by zero" message for the calculation of Δx and the method fails. Of course if $x_0 = 3$ were chosen as the initial value, the procedure will converge to the root at $x = 0$. Clearly, the function must be plotted if such pitfalls are to be avoided.

As the bisection method does not depend on the derivatives of the function in question, it can be applied with confidence, even if there are stationary points within the chosen limits, x_a and x_b. However, convergence is often

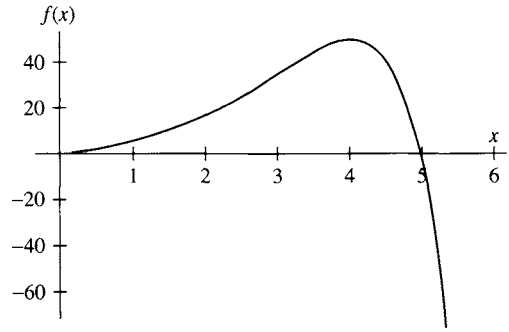

Fig. 8 The function $f(x) = (5-x)e^x - 5$ as a function of x.

somewhat slower. It is to be emphasized that it is assumed in this method that the function is continuous between the chosen limits. Here again, it is essential to plot the function before undertaking the evaluation of its roots.

A final remark should be added that applies to both of the methods outlined above. As both are iterative, any computer program must specify either the number of iterations or the precision of the desired result. Or better, both should be included and employed – whichever comes first.

PROBLEMS

1. Make the indicated substitution to yield Eq. (2).

2. Develop the logarithms in Eqs. (2) and (3) to obtain Eq. (4).

3. Show that the Gaussian function given by Eq. (4) is correctly normalized.

4. Verify Eqs. (12) and (13).

5. Derive the expressions for m and b in Eq. (16). Ans. Eqs. (20) and (21)

6. Verify the least-squares fit to the data given in Fig. 1-1.

7. Show that in the application of linear, midpoint interpolation

$$X = \begin{pmatrix} (-\frac{1}{2})^0 & (-\frac{1}{2})^1 \\ (+\frac{1}{2})^0 & (+\frac{1}{2})^1 \end{pmatrix} = \begin{pmatrix} 1 & -\frac{1}{2} \\ 1 & +\frac{1}{2} \end{pmatrix}$$

and thus $\boldsymbol{A}$ is given by Eq. (24).

8. Derive the general expression for θ [Eq. (31)].

9. Verify the matrix $\boldsymbol{A}$ given by Eq. (34).

10. Derive Eq. (39).

11. Check the relations given by Eqs. (46)–(50).

12. Construct the matrix given in Eq. (51).

13. Carry out the matrix multiplication indicated to obtain Eq. (53).

14. Prove Eq. (64).

15. Derive Debye's third-power law.