CHAPTER

1

Deterministic Problems

1-1. Introduction

The use of high-speed digital computers not only allows more computations to be made than ever before, it makes practicable methods of solution too repetitious for hand calculation. In the past much effort was expended to analytically manipulate solutions into forms which minimized the computational effort. It is now often more convenient to use computer time to reduce the analytical effort. Approximation techniques, once considered a last resort, can be carried to such high orders on computers that they are for most purposes as good as exact answers. They also permit treatment of problems not solvable by exact methods.

This text has been written to provide a unified treatment of matrix methods for computing the solutions to field problems. The basic idea is to reduce a functional equation to a matrix equation, and then solve the matrix equation by known techniques. These concepts are best expressed in the language of linear spaces and operators. However, it is not necessary that the reader have prior knowledge of this theory, because we shall define and illustrate the concepts as they are introduced. A brief summary of linear spaces and operators is given in Appendix A. Detailed expositions may be found in many textbooks [1–3].[1]

In this chapter we consider equations of the inhomogeneous type

$$L(f) = g \tag{1-1}$$

[1] Bracketed numbers refer to the References at the end of each chapter.

where L is an *operator*, g is the *source* or *excitation* (known function), and f is the *field* or *response* (unknown function to be determined). By the term *deterministic* we mean that the solution to (1-1) is unique; that is, only one f is associated with a given g. A problem of *analysis* involves the determination of f when L and g are given. A problem of *synthesis* involves a determination of L when f and g are specified. In this text we consider only the analysis problem.

This chapter presents the basic mathematical techniques for reducing functional equations to matrix equations. A unifying principle for such techniques is found in the general *method of moments*, in terms of which most specific solutions can be interpreted. We shall consider a deterministic problem solved once it is reduced to a suitable matrix equation, since the solution is then given by matrix inversion. Most computers have subroutines available for matrix inversion, which is a relatively simple operation. For reference, the widely used Gauss-Jordan method is given in Appendix B.

The examples of this chapter are simple, chosen to illustrate the theory without clouding the picture with physical concepts or complicated mathematics. However, when these methods are applied to problems of practical interest the procedures are not so simple. The details vary according to the type of problem, and can be illustrated only by treating a variety of problems. For this reason we treat many specific problems in the subsequent chapters. It is hoped that these examples will not only allow the reader to solve similar problems, but will suggest extensions and modifications to treat other types. Although most of the examples are taken from electromagnetic theory, the procedures are general and apply to field problems of any kind.

1-2. Formulation of Problems

The general methods of solution will be discussed in the notation of linear spaces and operators, and hence specific problems should be put into this notation. Given a deterministic problem of the form $L(f) = g$, we must identify the operator L, its domain (the functions f on which it operates), and its range (the functions g resulting from the operation). Furthermore, we usually need an *inner product* $\langle f, g \rangle$, which is a scalar defined to satisfy[2]

$$\langle f, g \rangle = \langle g, f \rangle \tag{1-2}$$

$$\langle \alpha f + \beta g, h \rangle = \alpha \langle f, h \rangle + \beta \langle g, h \rangle \tag{1-3}$$

$$\begin{aligned} \langle f^*, f \rangle &> 0 \qquad \text{if } f \neq 0 \\ &= 0 \qquad \text{if } f = 0 \end{aligned} \tag{1-4}$$

[2] The usual definition of inner product in Hilbert space corresponds to $\langle f^*, g \rangle$ in our notation. For this text it is more convenient to show the conjugate operation explicity wherever it occurs, and to define the adjoint operator without conjugation.

where α and β are scalars and * denotes a complex conjugate. We sometimes need the *adjoint operator* L^a and its domain, defined by

$$\langle Lf, g\rangle = \langle f, L^a g\rangle \tag{1-5}$$

for all f in the domain of L. An operator is *self-adjoint* if $L^a = L$ and the domain of L^a is that of L.

Properties of the solution depend upon properties of the operator. An operator is *real* if Lf is real whenever f is real. An operator is *positive definite* if

$$\langle f^*, Lf\rangle > 0 \tag{1-6}$$

for all $f \neq 0$ in its domain. It is *positive semidefinite* if $>$ is replaced by $\geq$ in (1-6), *negative definite* if $>$ is replaced by $<$ in (1-6), etc. We shall identify other properties of operators as they are needed.

If the solution to $L(f) = g$ exists and is unique for all g, then the *inverse operator* L^{-1} exists such that

$$f = L^{-1}(g) \tag{1-7}$$

If g is known, then (1-7) represents the solution to the original problem. However, (1-7) is itself an inhomogeneous equation for g if f is known, and its solution is $L(f) = g$. Hence L and L^{-1} form a pair of operators, each of which is the inverse of the other.

Facility in formulating problems using the concepts of linear spaces comes only with practice, which will be provided by the many examples in the following chapters. For the present, let us consider a simple abstract example so that mathematical concepts may be illustrated without bringing physical concepts into the picture.

Example. Given $g(x)$, find $f(x)$ in the interval $0 \leq x \leq 1$ satisfying

$$-\frac{d^2 f}{dx^2} = g(x) \tag{1-8}$$

$$f(0) = f(1) = 0 \tag{1-9}$$

This is a boundary-value problem for which

$$L = -\frac{d^2}{dx^2} \tag{1-10}$$

The range of L is the space of all functions g in the interval $0 \leq x \leq 1$ that we wish to consider. The domain of L is the space of those functions f in the interval

$0 \leq x \leq 1$, satisfying the boundary conditions (1-9), and having second derivatives in the range of L. The solution to (1-8) is not unique unless appropriate boundary conditions are included. In other words, both the differential operator and its domain are required to define the operator.

A suitable inner product for this problem is

$$\langle f, g \rangle = \int_0^1 f(x)g(x)\,dx \tag{1-11}$$

It is easily shown that (1-11) satisfies the postulates (1-2) to (1-4), as required. Note that the definition (1-11) is not unique. For example,

$$\int_0^1 w(x)f(x)g(x)\,dx \tag{1-12}$$

where $w(x) > 0$ is an arbitrary weighting function, is also an acceptable inner product. However, the adjoint operator depends on the inner product, which can often be chosen to make the operator self-adjoint.

To find the adjoint of a differential operator, we form the left side of (1-5), and integrate by parts to obtain the right side. For the present problem

$$\begin{aligned}\langle Lf, g \rangle &= \int_0^1 \left(-\frac{d^2f}{dx^2}\right) g\,dx \\ &= \int_0^1 \frac{df}{dx}\frac{dg}{dx}\,dx - \left[\frac{df}{dx}\,g\right]_0^1 \\ &= \int_0^1 f\left(-\frac{d^2g}{dx^2}\right) dx + \left[f\frac{dg}{dx} - g\frac{df}{dx}\right]_0^1\end{aligned} \tag{1-13}$$

The last terms are boundary terms, and the domain of L^a may be chosen so that these vanish. The first boundary terms vanish by (1-9), and the second vanish if

$$g(0) = g(1) = 0 \tag{1-14}$$

It is then evident that the adjoint operator to (1-10) for the inner product (1-11) is

$$L^a = L = -\frac{d^2}{dx^2} \tag{1-15}$$

Since $L^a = L$ and the domain of L^a is the same as that of L, the operator is self-adjoint.

It is also evident that L is a real operator, since Lf is real when f is real. That L is a positive definite operator is shown from (1-6) as follows:

$$\langle f^*, Lf\rangle = \int_0^1 f^*\left(-\frac{d^2f}{dx^2}\right) dx$$

$$= \int_0^1 \frac{df^*}{dx}\frac{df}{dx}\,dx - \left[f^*\frac{df}{dx}\right]_0^1$$

$$= \int_0^1 \left|\frac{df}{dx}\right|^2 dx \tag{1-16}$$

Note that L is a positive definite operator even if f is complex.

The inverse operator to L can be obtained by standard Green's function techniques.[3] It is

$$L^{-1}(g) = \int_0^1 G(x, x')g(x')\,dx' \tag{1-17}$$

where G is the Green's function

$$G(x, x') = \begin{cases} x(1-x') & x < x' \\ (1-x)x' & x > x' \end{cases} \tag{1-18}$$

We can verify that (1-17) is the inverse operator by forming $f = L^{-1}(g)$, differentiating twice, and obtaining (1-8). Note that no boundary conditions are needed on the domain of L^{-1}, which is characteristic of most integral operators. That L^{-1} is self-adjoint follows from the proof that L is self-adjoint, since

$$\langle Lf_1, f_2\rangle = \langle g_1, L^{-1}g_2\rangle \tag{1-19}$$

Of course, the self-adjointness of L^{-1} can also be proved directly. It similarly follows that L^{-1} is positive definite whenever L is positive definite, and vice versa.

1-3. Method of Moments

We now discuss a general procedure for solving linear equations, called the *method of moments* [4,5]. Consider the inhomogeneous equation

$$L(f) = g \tag{1-20}$$

---

[3] See, for example, reference [2], Chapter 3.

where L is a linear operator, g is known, and f is to be determined. Let f be expanded in a series of functions $f_1, f_2, f_3, \ldots$ in the domain of L, as

$$f = \sum_n \alpha_n f_n \tag{1-21}$$

where the α_n are constants. We shall call the f_n *expansion functions* or *basis functions*. For exact solutions, (1-21) is usually an infinite summation and the f_n form a complete set of basis functions. For approximate solutions, (1-21) is usually a finite summation. Substituting (1-21) in (1-20), and using the linearity of L, we have

$$\sum_n \alpha_n L(f_n) = g \tag{1-22}$$

It is assumed that a suitable inner product $\langle f, g \rangle$ has been determined for the problem. Now define a set of *weighting functions*, or *testing functions*, $w_1, w_2, w_3, \ldots$ in the range of L, and take the inner product of (1-22) with each w_m. The result is

$$\sum_n \alpha_n \langle w_m, Lf_n \rangle = \langle w_m, g \rangle \tag{1-23}$$

$m = 1, 2, 3, \ldots$. This set of equations can be written in matrix form as

$$[l_{mn}][\alpha_n] = [g_m] \tag{1-24}$$

where

$$[l_{mn}] = \begin{bmatrix} \langle w_1, Lf_1 \rangle & \langle w_1, Lf_2 \rangle & \cdots \\ \langle w_2, Lf_1 \rangle & \langle w_2, Lf_2 \rangle & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix} \tag{1-25}$$

$$[\alpha_n] = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \end{bmatrix} \qquad [g_m] = \begin{bmatrix} \langle w_1, g \rangle \\ \langle w_2, g \rangle \\ \vdots \end{bmatrix} \tag{1-26}$$

If the matrix $[l]$ is nonsingular its inverse $[l^{-1}]$ exists. The α_n are then given by

$$[\alpha_n] = [l_{nm}^{-1}][g_m] \tag{1-27}$$

and the solution for f is given by (1-21). For concise expression of this result, define the matrix of functions

$$[f_n] = [f_1 \quad f_2 \quad f_3 \quad \ldots] \tag{1-28}$$

and write

$$f = [f_n][\alpha_n] = [f_n][l_{mn}^{-1}][g_m] \tag{1-29}$$

This solution may be exact or approximate, depending upon the choice of the f_n and w_n. The particular choice $w_n = f_n$ is known as *Galerkin's method* [6,7].

If the matrix $[l]$ is of infinite order, it can be inverted only in special cases, for example, if it is diagonal. The classical eigenfunction method leads to a diagonal matrix, and can be thought of as a special case of the method of moments. If the sets f_n and w_n are finite, the matrix is of finite order, and can be inverted by known methods (Appendix B).

One of the main tasks in any particular problem is the choice of the f_n and w_n. The f_n should be linearly independent and chosen so that some superposition (1-21) can approximate f reasonably well. The w_n should also be linearly independent and chosen so that the products $\langle w_n, g \rangle$ depend on relatively independent properties of g. Some additional factors which affect the choice of f_n and w_n are (1) the accuracy of solution desired, (2) the ease of evaluation of the matrix elements, (3) the size of the matrix that can be inverted, and (4) the realization of a well-conditioned matrix $[l]$.

Example. Consider the same equation as in the example of Section 1-2, but with the specific source $g = 1 + 4x^2$. Hence our problem is

$$-\frac{d^2f}{dx^2} = 1 + 4x^2 \tag{1-30}$$

$$f(0) = f(1) = 0 \tag{1-31}$$

This is, of course, a simple boundary-value problem with solution

$$f(x) = \frac{5x}{6} - \frac{x^2}{2} - \frac{x^4}{3} \tag{1-32}$$

To illustrate the procedure, the problem will be reconsidered by the method of moments.

For a power-series solution, let us choose

$$f_n = x - x^{n+1} \tag{1-33}$$

$n = 1, 2, 3, \ldots, N$, so that the series (1-21) is

$$f = \sum_{n=1}^{N} \alpha_n (x - x^{n+1}) \tag{1-34}$$

Note that the term x is needed in (1-33), else the f_n will not be in the domain of L; that is, the boundary conditions will not be satisfied. For testing functions, choose

$$w_n = f_n = x - x^{n+1} \tag{1-35}$$

in which case the method is that of Galerkin. In Section 1-8 it is shown that the w_n should be in the domain of the adjoint operator. Since L is self-adjoint for this problem, the w_n should be in the domain of L, as are those of (1-35).

Evaluation of the matrices (1-25) and (1-26) for the inner product (1-11) and $L = -d^2/dx^2$ is straightforward, and results in

$$l_{mn} = \langle w_m, Lf_n \rangle = \frac{mn}{m+n+1} \tag{1-36}$$

$$g_m = \langle w_m, g \rangle = \frac{m(3m+8)}{2(m+2)(m+4)} \tag{1-37}$$

For any fixed N (number of expansion functions), the α_n are given by (1-27) and the approximation to f by (1-34).

To illustrate convergence, let us consider successive approximations as N is increased. For $N = 1$, we have $l_{11} = 1/3$, $g_1 = 11/30$, and hence from (1-24) $\alpha_1 = 11/10$. For $N = 2$, the matrix equation (1-24) becomes

$$\begin{bmatrix} \frac{1}{3} & \frac{1}{2} \\ \frac{1}{2} & \frac{4}{5} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} \frac{11}{30} \\ \frac{7}{12} \end{bmatrix} \tag{1-38}$$

from which the α's are found as

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{10} \\ \frac{2}{3} \end{bmatrix} \tag{1-39}$$

For $N = 3$, the matrix equation (1-24) becomes

$$\begin{bmatrix} \frac{1}{3} & \frac{1}{2} & \frac{3}{5} \\ \frac{1}{2} & \frac{4}{5} & 1 \\ \frac{3}{5} & 1 & \frac{9}{7} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} = \begin{bmatrix} \frac{11}{30} \\ \frac{7}{12} \\ \frac{51}{70} \end{bmatrix} \tag{1-40}$$

from which the α's are found as

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ 0 \\ \frac{1}{3} \end{bmatrix} \tag{1-41}$$

Note that this third-order solution is the exact solution, (1-32). For $N = 4$ we

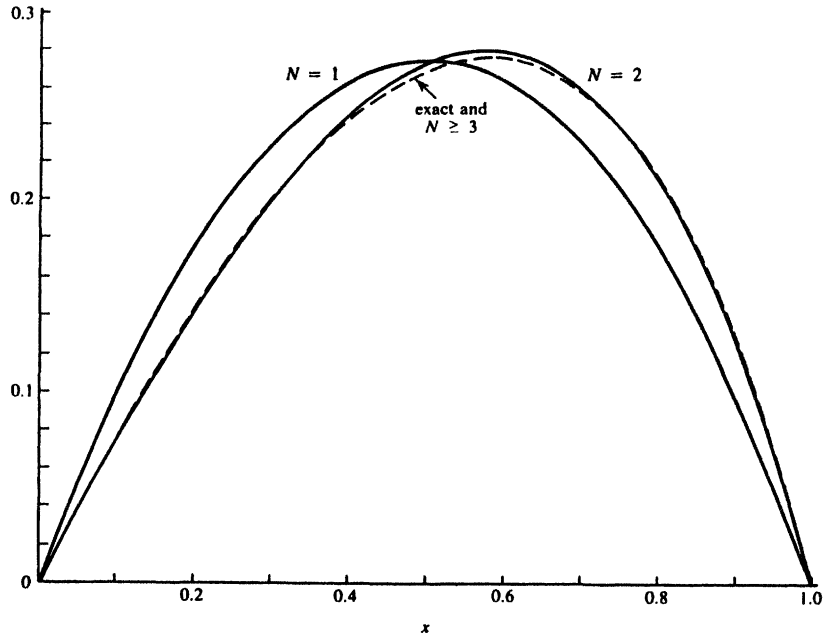

Figure 1-1. Solutions using $f_n = x - x^{n+1}$ and Galerkin's method.

again obtain the exact solution, and so on for higher N. Plots of the various solutions are shown in Fig. 1-1.

The reason an exact solution is obtained for this problem is that some combination of the f_n can exactly represent the solution, and any N linearly independent tests must correctly determine the coefficients. If the solution cannot be expressed as a finite series of the f_n, then we continue to obtain approximate solutions converging to the exact solution in the sense of projections, as discussed in Section 1-8.

More important than solving any particular equation, the inverse matrix $[l^{-1}]$ gives a representation of the inverse operator L^{-1}. Hence we have a solution (usually approximate) to $Lf = g$ for *any* g. In physical problems, L represents the system, g the excitation, and f the response. A determination of the $[l^{-1}]$ matrix therefore gives us a general solution for the system, that is, the response f for arbitrary excitation g, assuming that g is reasonably well behaved.

1-4. Point Matching

The integration involved in evaluating the $l_{mn} = \langle w_m, Lf_n \rangle$ of (1-25) is often difficult to perform in problems of practical interest. A simple way to obtain approximate solutions is to require that equation (1-22) be satisfied at discrete

points in the region of interest. This procedure is called a *point-matching method.* In terms of the method of moments, it is equivalent to using Dirac delta functions as testing functions. The following example illustrates this in the one-dimensional case.

Example. Reconsider the problem of Section 1-3, stated by (1-30) and (1-31). Again we choose expansion functions (1-33), so that (1-22) becomes

$$\sum_{n=1}^{N} \alpha_n \left[-\frac{d^2}{dx^2}(x - x^{n+1}) \right] = 1 + 4x^2 \tag{1-42}$$

For a point-matching solution, let us take the points

$$x_m = \frac{m}{N+1} \qquad m = 1, 2, \ldots, N \tag{1-43}$$

which are equispaced in the interval $0 \leq x \leq 1$. Requiring (1-42) to be satisfied at each x_m gives us the matrix equation (1-24), with elements

$$l_{mn} = n(n+1)\left(\frac{m}{N+1}\right)^{n-1} \tag{1-44}$$

$$g_m = 1 + 4\left(\frac{m}{N+1}\right)^2 \tag{1-45}$$

Note that this result is identical to choosing weighting functions

$$w_m = \delta(x - x_m) \tag{1-46}$$

where $\delta(x)$ is the Dirac delta function, and applying the method of moments with inner product (1-11).

To illustrate some numerical results, consider the solution as N is increased. For $N = 1$, we have $l_{11} = 2$, $g_1 = 2$, and from (1-27) $\alpha_1 = 1$. For $N = 2$, the matrix equation is

$$\begin{bmatrix} 2 & 2 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} \frac{13}{9} \\ \frac{25}{9} \end{bmatrix} \tag{1-47}$$

from which the α's are found as

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{18} \\ \frac{2}{3} \end{bmatrix} \tag{1-48}$$

For $N = 3$, the exact solution (1-41) must again be obtained, since the exact solution is a linear combination of the f_n's and we are applying N independent tests. Similarly, for $N > 3$ we continue to obtain the exact answer for the same reason. Plots of these solutions differ to some extent from those of Fig. 1-1 but are qualitatively similar. The point-matching solutions in this case are actually less accurate than the corresponding Galerkin approximations, but for low orders of solution they are usually sensitive to the particular points of match. For high-order solutions the use of equispaced points normally gives excellent results.

Note that even though the $[l]$ matrices of (1-36) and (1-44) are quite different in form, they give similar results. There are infinitely many possible sets of basis functions and of testing functions. Some sets may give faster convergence than others, or give matrices easier to evaluate, or give acceptable results with smaller matrices, etc. For any particular problem one of our tasks is to choose sets well suited to the problem.

1-5. Subsectional Bases

Another approximation useful for practical problems is the *method of subsections*. This involves the use of basis functions f_n each of which exists only over subsections of the domain of f. Then each α_n of the expansion (1-21) affects the approximation of f only over a subsection of the region of interest. This procedure often simplifies the evaluation and/or the form of the matrix $[l]$. Sometimes it is convenient to use the point-matching method of Section 1-4 in conjunction with the subsectional method.

Example. Again consider the problem of Section 1-3, stated by (1-30) and (1-31). N equispaced points on the interval $0 \le x \le 1$ are defined by the x_m of (1-43). A subinterval is defined to be of width $1/(N+1)$ centered on the x_m. This is shown for case $N = 5$ in Fig. 1-2(a). A function which exists over only one subinterval is the *pulse function*

$$P(x) = \begin{cases} 1 & |x| < \dfrac{1}{2(N+1)} \\ 0 & |x| > \dfrac{1}{2(N+1)} \end{cases} \tag{1-49}$$

For $N = 5$, the function $P(x - x_2)$ is shown in Fig. 1-2(b). A linear combination of $f_n = P(x - x_n)$ according to (1-21) gives a *step approximation* to f, as represented by Fig. 1-2(c). However, for $L = -d^2/dx^2$, the operation LP does not yield a function in the range of L. Hence the pulse functions cannot be used as basis functions unless we extend the operator (Section 1-7) or use an approximate operator (Section 1-6).

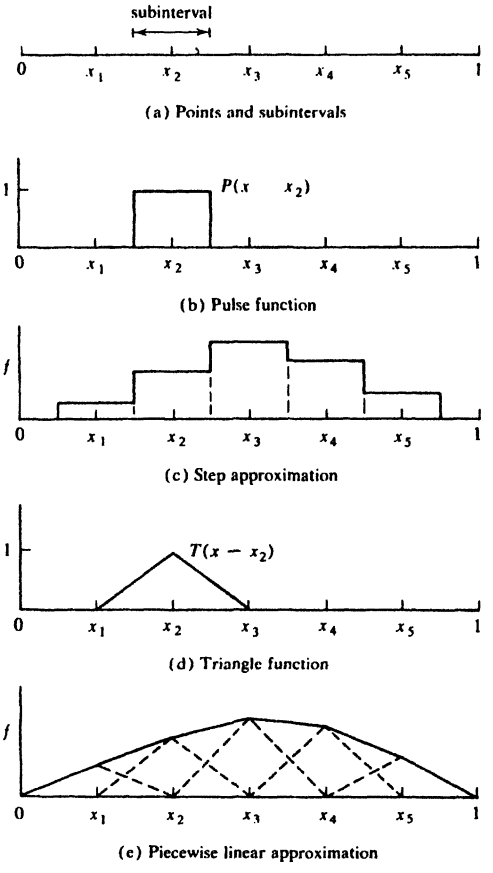

***Figure 1-2.* Subsectional bases and functional approximations.**

A better-behaved function is the *triangle function*, defined as

$$T(x) = \begin{cases} 1 - |x|(N+1) & |x| < \dfrac{1}{N+1} \\ 0 & |x| > \dfrac{1}{N+1} \end{cases} \tag{1-50}$$

For the case $N = 5$ the function $T(x - x_2)$ is shown in Fig. 1-2(d). A linear

combination of triangle functions of the form

$$f = \sum_{n=1}^{N} \alpha_n T(x - x_n) \tag{1-51}$$

gives a *piecewise linear approximation* to f, as represented by Fig. 1-2(e). For $L = -d^2/dx^2$, the operation LT gives the symbolic function

$$LT(x - x_n) = (N + 1)[-\delta(x - x_{n-1}) + 2\delta(x - x_n) - \delta(x - x_{n+1})] \tag{1-52}$$

where $\delta(x)$ is the Dirac delta function. We can use this result in the method of moments as long as the w_n are not also symbolic functions. We cannot use a point-matching procedure in this case.

To follow through the method of moments, let $f_n = T(x - x_n)$, that is, use the expansion (1-51). As testing functions, choose $w_m = P(x - x_m)$. For inner product (1-11), the matrix elements of (1-25) and (1-26) are easily evaluated as

$$l_{mn} = \begin{cases} 2(N+1) & m = n \\ -(N+1) & |m-n| = 1 \\ 0 & |m-n| > 1 \end{cases} \tag{1-53}$$

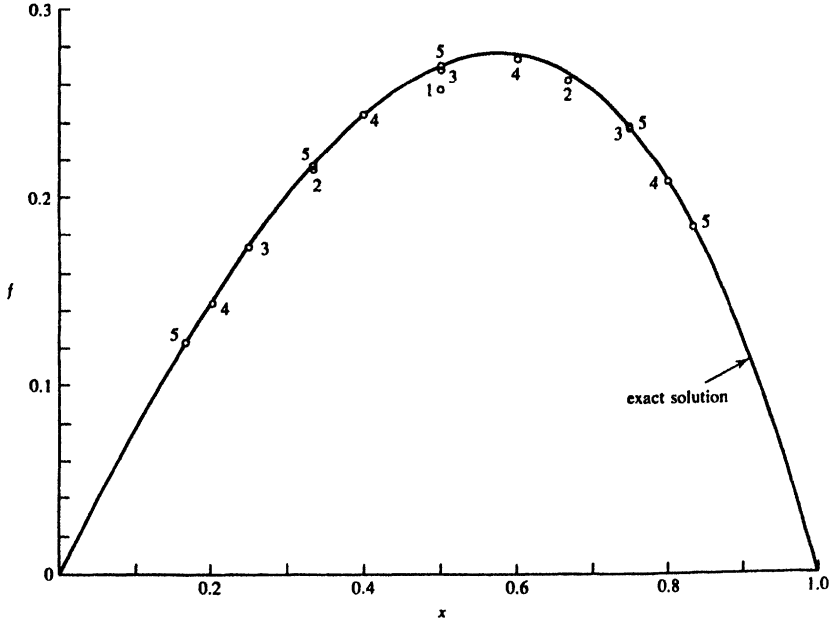

Figure 1-3. **Moment solutions using triangles for expansion and pulses for testing. Numbers adjacent to points denote order of solution.**

$$g_m = \frac{1}{N+1}\left[1 + \frac{4m^2 + (1/3)}{(N+1)^2}\right] \tag{1-54}$$

Note the particularly simple form of $[l]$. We shall encounter this form again in connection with difference equations (Section 1-6).

Figure 1-3 illustrates the convergence of the above solution as N (number of subsections) is increased. Only the break points of the piecewise linear solution are shown; the functional approximation is given by straight lines joining these points. The break points are, of course, also the α_n, since they are the peaks of the triangle-function components.

1-6. Approximate Operators

In complex problems it is sometimes convenient to approximate the operator to obtain approximate solutions. For differential operators, the finite-difference approximation has been widely used [8]. For integral operators, an approximate operator can be obtained by approximating the kernel of the integral operator [6]. Any method whereby a functional equation is reduced to a matrix equation can be interpreted in terms of the method of moments. Hence for any matrix solution using approximation of the operator there will be a corresponding moment solution using approximation of the function.

Example. Let us consider the problem (1-30) and (1-31) by a finite-difference approximation. This involves replacing all derivatives by finite differences; that is, for a given Δx,

$$\begin{aligned}
\frac{df}{dx} &\approx \frac{1}{\Delta x}\left[f\left(x + \frac{\Delta x}{2}\right) - f\left(x - \frac{\Delta x}{2}\right)\right] \\
\frac{d^2f}{dx^2} &\approx \frac{1}{\Delta x}\left[f'\left(x + \frac{\Delta x}{2}\right) - f'\left(x - \frac{\Delta x}{2}\right)\right] \\
&\approx \frac{1}{(\Delta x)^2}\left[f(x - \Delta x) - 2f(x) + f(x + \Delta x)\right]
\end{aligned} \tag{1-55}$$

For our present problem, consider the interval $0 \le x \le 1$ divided into $N + 1$ segments, with end points x_n, as depicted in Fig. 1-2(a). For Δx equal to one segment, $\Delta x = 1/(N + 1)$, and a finite-difference approximation to $L = -d^2/dx^2$ is

$$L^d f = (N+1)^2\left[-f\left(x - \frac{1}{N+1}\right) + 2f(x) - f\left(x + \frac{1}{N+1}\right)\right] \tag{1-56}$$

Note that $L^d \to L$ as $N \to \infty$ for all f in the domain of L.

We can now apply the method of moments to the approximate equation

$$L^d f = 1 + 4x^2 \tag{1-57}$$

subject to boundary conditions $f(0) = f(1) = 0$. Most commonly this is done by a point-matching procedure at the x_m. The result is a matrix equation of the form (1-24), where the α_n correspond to $f(x_n)$,

$$l_{mn} = \begin{cases} 2(N+1)^2 & m = n \\ -(N+1)^2 & |m-n| = 1 \\ 0 & |m-n| > 1 \end{cases} \tag{1-58}$$

$$g_m = 1 + 4\left(\frac{m}{N+1}\right)^2 \tag{1-59}$$

Note that the $[l]$ matrix of (1-58) is the same form as that of (1-53) obtained from a subsectional basis. [The trivial difference in the position of $N + 1$ can be taken care of by choosing $w_m = (N + 1)P(x - x_m)$ in the solution of Section 1-5.] The g_m of (1-59) and (1-54) are slightly different, and hence the two solutions will be slightly different. However, as N becomes larger the two g_m approach one another, so the rates of convergence of the two solutions are about the same.

Numerical results for the above solution are similar to those of Fig. 1-3. Iterative procedures are sometimes used to solve the matrix equations obtained by difference approximations [9]. However, iterative procedures usually converge slowly, and with high-speed large-memory computers it is often simpler to invert the matrix. Because of the tridiagonal form of $[l]$, special techniques can be used to invert it [10].

1-7. Extended Operators

As noted earlier, an operator is defined by an operation (for example, $L = -d^2/dx^2$) plus a domain (space of functions to which the operation may be applied). We can *extend the domain* of an operator by redefining the operation to apply to new functions (not in the original domain) as long as this extended operation does not change the original operation in its domain. If the original operator is self-adjoint, it is desirable to make the extended operator self-adjoint also. By this procedure we can use a wider class of functions for solution by the method of moments. This becomes particularly important in multivariable problems (fields in multidimensional space), where it is not always easy to find simple functions in the domain of the original operator.

Example A. Suppose we wish to use pulse functions for an expansion of f in a moment solution for the operator $L = -d^2/dx^2$. As noted in Section 1-5, these are not in the original domain of L. However, for any functions w and f in the original domain,

$$\langle w, Lf \rangle = \int_0^1 \frac{dw}{dx}\frac{df}{dx}\,dx \tag{1-60}$$

obtained from (1-11) by integration by parts. If Lf does not exist, but df/dx does exist, (1-60) can be used to define an extended operator. This extends the domain of L to include functions f whose second derivatives do not exist, but whose first derivatives do exist. It is still assumed that $f(0) = f(1) = 0$. Actually, the type of extension represented here is precisely that which gives rise to the theory of symbolic functions. By using Dirac delta functions in earlier sections we anticipated this concept of extending the domain of a differential operator.

To apply the method of moments using pulse functions and the extended operator, let

$$f = \sum_{n=1}^{N} \alpha_n P(x - x_n) \tag{1-61}$$

where P are the pulse functions defined by (1-49). For testing functions, let $w_m = T(x - x_m)$, where T are the triangle functions defined by (1-50). The elements of the $[l]$ matrix are found using (1-60) as

$$l_{mn} = \langle w_m, Lf_n \rangle = \begin{cases} 2(N+1) & m = n \\ -(N+1) & |m-n| = 1 \\ 0 & |m-n| > 1 \end{cases} \tag{1-62}$$

Note that these are identical to the elements (1-53), which were for f_n and w_m reversed from those of the present solution. We could have anticipated this result because L is self-adjoint. The elements of the $[g]$ matrix are now given by

$$g_m = \int_0^1 T(x - x_m)(1 + 4x^2)\,dx \tag{1-63}$$

which yields a result slightly different from (1-54). However, the two g_m approach each other as N becomes large, and the convergence of the two solutions is about the same.

Numerical results for the above example are similar to those of Fig. 1-3 for various N. However, the functional approximation in this case is a step approximation; that is, the points are midpoints of steps, instead of break points of a piecewise linear approximation as in Fig. 1-3.

Example B. As a second example, let us extend the original domain of $L = -d^2/dx^2$ to apply to functions not satisfying the boundary conditions $f(0) = f(1) = 0$. Referring to (1-13), we note that boundary terms appear if the functions do not obey the given boundary conditions. However, if an extended operator L^e is defined by

$$\langle w, L^e f\rangle = \int_0^1 wLf\,dx - \left[f\frac{dw}{dx}\right]_0^1 \tag{1-64}$$

we have $\langle w, L^e f\rangle = \langle f, L^e w\rangle$ even if the original boundary conditions are not met. Hence the extended operator is self-adjoint regardless of boundary conditions. A method-of-moments solution therefore proceeds in this extended domain in the same manner as for the original domain, except that the expansion and testing functions need not satisfy boundary conditions.

To illustrate the procedure, consider the choice

$$f_n = w_n = x^n \qquad n = 1, 2, \ldots, N \tag{1-65}$$

For $N \geq 4$ these functions form a basis for the exact solution (1-32), and hence the exact solution should be obtained. Evaluating the matrices in the usual way, using the extended operator for $l_{mn} = \langle w_m, L^e f_n\rangle$, for $N = 4$ we obtain the matrix equation

$$\begin{bmatrix} -1 & -2 & -3 & -4 \\ -2 & -\frac{8}{3} & -\frac{7}{2} & -\frac{22}{5} \\ -3 & -\frac{7}{2} & -\frac{21}{5} & -5 \\ -4 & -\frac{22}{5} & -5 & -\frac{40}{7} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{bmatrix} = \begin{bmatrix} \frac{3}{2} \\ \frac{17}{15} \\ \frac{11}{12} \\ \frac{27}{35} \end{bmatrix} \tag{1-66}$$

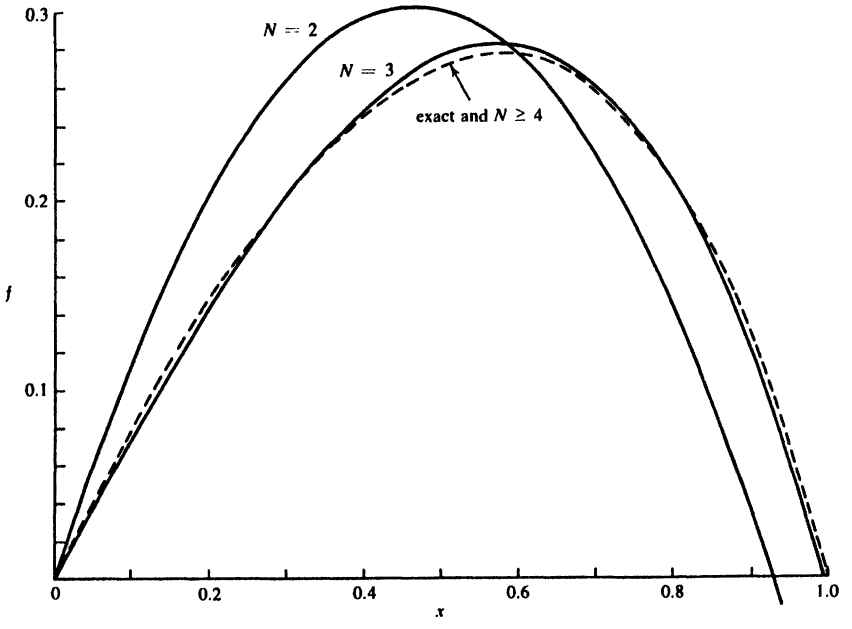

Figure 1-4. **Extended operator moment solutions using powers of x for expansion and testing.**

This may be solved for the α's to obtain

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{bmatrix} = \begin{bmatrix} \frac{5}{6} \\ -\frac{1}{2} \\ 0 \\ -\frac{1}{3} \end{bmatrix} \tag{1-67}$$

which is indeed the exact solution. Note that if (1-65) are used with the original operator $L = -d^2/dx^2$ a singular $[I]$ matrix results, and hence no solution is obtained. To illustrate convergence using the extended operator, Fig. 1-4 shows plots of the cases $N = 2$ and $N = 3$, plus the exact solution ($N \geq 4$).

1-8. Variational Interpretation

It is well known that Galerkin's method ($w_n = f_n$) is equivalent to the Rayleigh-Ritz variational method [6,7]. That the general method of moments is also a variational method is usually not noted, but the proof is essentially the same as for Galerkin's method [7].

Let us first interpret the method of moments according to the concepts of linear spaces. Let $\mathcal{S}(Lf)$ denote the range of L, $\mathcal{S}(Lf_n)$ denote the space spanned by the Lf_n, and $\mathcal{S}(w_n)$ denote the space spanned by the w_n. The method of moments (1-23) then equates the projection of Lf onto $\mathcal{S}(w_n)$ to the projection of the approximate Lf onto $\mathcal{S}(w_n)$. Figure 1-5 represents this pictorially. In the

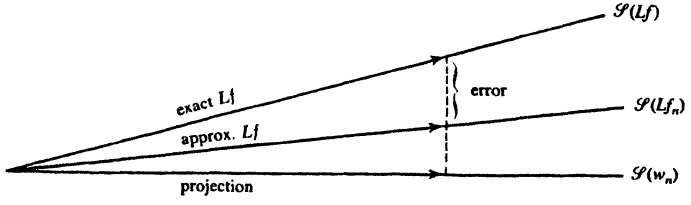

Figure 1-5. **Pictorial representation of the method of moments in function space.**

special case of Galerkin's method, $\mathcal{S}(w_n) = \mathcal{S}(f_n)$. Because the process of obtaining projections minimizes an error, the method of moments is an error-minimizing procedure. Because the error is orthogonal to the projections, it is of second order. This same conclusion is obtained from the calculus of variations [7]. The derivation of the variational results will not be given here, but we shall summarize the conclusions.

Given an operator equation $Lf = g$, it is desired to determine a functional of f (number depending on f)

$$\rho(f) = \langle f, h \rangle \tag{1-68}$$

where h is a given function. If h is a continuous function, then $\rho(f)$ is a *continuous linear functional.* The functional ρ may be f itself if h is an impulse function, but then ρ is no longer a continuous functional. Now let L^a be the adjoint operator to L, and define an adjoint function f^a (adjoint field) by

$$L^a f^a = h \tag{1-69}$$

By the calculus of variations, it can then be shown that [7]

$$\rho = \frac{\langle f, h \rangle \langle f^a, g \rangle}{\langle Lf, f^a \rangle} \tag{1-70}$$

is a variational formula for ρ with stationary point (1-68) when f is the solution to $Lf = g$ and f^a the solution to (1-69). For an approximate evaluation of ρ, let

$$f = \sum_n \alpha_n f_n \qquad f^a = \sum_m \beta_m w_m \tag{1-71}$$

Substitute these in (1-70), and apply the Rayleigh-Ritz conditions $\partial\rho/\partial\alpha_i = \partial\rho/\partial\beta_i = 0$ for all i. The result is that the necessary and sufficient conditions for ρ to be a stationary point are equations (1-23). Hence the method of moments is equivalent to the Rayleigh-Ritz variational method [7]. The method of moments is closely related to the direct methods of the calculus of variations, so called because they yield a solution to the variational problem without recourse to the associated differential equation.

The above variational interpretation can be used to give additional insight into how to choose the testing functions. It is evident from (1-69) and (1-71) that the w_n should be chosen so that some linear combination of them can closely represent the adjoint field f^a. When we calculate f itself by the method of moments, h of (1-68) is a Dirac delta function and f^a of (1-69) is a Green's function. This implies that some combination of the w_n should be able to approximate the Green's function. Since a Green's function is usually poorly behaved, we should expect computation of a field by the method of moments to converge less slowly than computation of a continuous linear functional. This is found actually to be the case.

1-9. Perturbation Solutions

Sometimes the problem under consideration is only slightly different (perturbed) from a problem which can be solved exactly (the unperturbed problem). A first-order solution to the perturbed problem can then be obtained by using the solution to the unperturbed problem as a basis for the method of moments. This procedure is called a *perturbation method.* Higher-order perturbation solutions

can be obtained by using the unperturbed solution plus correction terms in the method of moments. Sometimes this is done as successive approximations by including one correction term at a time, but for machine computations it is usually easier to include all correction terms at once.

To express these concepts in equation form, let

$$L_0(f_0) = g \tag{1-72}$$

represent the unperturbed problem for which the solution f_0 is known. Let $M = L - L_0$ be the difference operator, and hence

$$L(f) = (L_0 + M)(f) = g \tag{1-73}$$

represents the perturbed problem for which the solution f is desired. For a first-order perturbation solution, let

$$f = \alpha f_0 \tag{1-74}$$

and apply the method of moments. If L is self-adjoint, the testing function $w = f_0$ may be chosen; otherwise we should choose $w = f_0^a$, the solution to the unperturbed adjoint problem. An application of the method of moments to this one-term expansion yields

$$(\langle f_0, L_0 f_0 \rangle + \langle f_0, M f_0 \rangle)\alpha = \langle f_0, g \rangle \tag{1-75}$$

Now, by (1-72), $\langle f_0, L_0 f_0 \rangle = \langle f_0, g \rangle$, and the above equation can be written

$$\alpha = 1 - \frac{\langle f_0, M f_0 \rangle}{\langle f_0, g \rangle + \langle f_0, M f_0 \rangle} \tag{1-76}$$

If the perturbation is truly small, the second term in the denominator of (1-76) will be small compared to the first term, and from (1-74) and (1-76)

$$f \approx \left(1 - \frac{\langle f_0, M f_0 \rangle}{\langle f_0, g \rangle}\right) f_0 \tag{1-77}$$

This is the first-order perturbation solution.

For higher-order solutions, we merely choose $f_1 = f_0$ in the general method of moments (Section 1-3) and $f_2, f_3, \ldots$ serve as correction terms. For self-adjoint operators, choose $w_1 = f_0$; otherwise choose $w_1 = f_0^a$. The advantage of a perturbation approach over other moment solutions rests primarily in the faster convergence of the perturbation solution.

References

[1] B. Z. Vulikh, *Introduction to Functional Analysis for Scientists and Technologists*, translated by I. N. Sneddon, Pergamon Press, Oxford, 1963.

[2] B. Friedman, *Principles and Techniques of Applied Mathematics*, John Wiley & Sons, Inc., New York, 1956.

[3] J. W. Dettman, *Mathematical Methods in Physics and Engineering*, McGraw-Hill Book Co., New York, 1962.

[4] L. V. Kantorovich and G. P. Akilov, *Functional Analysis in Normed Spaces*, translated by D. E. Brown, Pergamon Press, Oxford, 1964, pp. 586–587.

[5] R. F. Harrington, "Matrix Methods for Field Problems," *Proc. IEEE*, Vol. **55**, No. 2, Feb. 1967, pp. 136–149.

[6] L. V. Kantorovich and V. I. Krylov, *Approximate Methods of Higher Analysis*, 4th ed., translated by C. D. Benster, John Wiley & Sons, Inc., New York, 1959, Chap. IV.

[7] D. S. Jones, "A Critique of the Variational Method in Scattering Problems," *IRE Trans.*, Vol. **AP-4**, No. 3, 1956, pp. 297–301.

[8] G. E. Forsythe and W. R. Wasow, *Finite-Difference Methods for Partial Differential Equations*, John Wiley & Sons, Inc., New York, 1960.

[9] R. V. Southwell, *Relaxation Methods in Theoretical Physics*, Vol. **1**, Oxford University Press, London, 1946.

[10] P. Henrici, *Discrete Variable Methods in Ordinary Differential Equations*, John Wiley & Sons, Inc., New York, 1962, pp. 350–355.