5

The Discretization Process: Basis/Testing Functions and Convergence

Chapters 2 and 3 presented a variety of examples involving the discretization of continuous equations into matrix form. The specific discretization places a limit on the accuracy of a numerical result for a fixed number of basis and testing functions and determines whether or not the numerical result will converge to the exact solution as the number of basis and testing functions is increased. These are fundamental issues, and they require an examination from a perspective that is more theoretical than operational. To accomplish this, we employ some of the tools of functional analysis, the branch of applied mathematics said to be concerned with providing "solutions to equations, often by means of convergent sequences of approximation" [1]. This chapter also introduces a variety of subsectional basis functions and explores the role that the basis and testing functions play in the solution accuracy.

5.1 INNER PRODUCT SPACES

We desire to solve an equation of the form $Lf = g$, where L is a continuous linear operator such as the integral operators of Chapter 2 or the differential operators of Chapter 3. The function f is the unknown to be determined, and g represents a known excitation. If a unique solution exists, it is given by $f = L^{-1}g$, where L^{-1} is the inverse operator. In practice, we are usually not able to determine L^{-1} and resort to numerical solutions.

The linear operator L maps functions in its domain (such as the unknown f) to functions in its range (such as the excitation g). As a general rule, the domain and range are different linear spaces. For instance, in the case where L is a differential operator, the domain of L will generally include boundary conditions not imposed on functions in the range.

It is convenient to introduce the notion of an inner product, which is a scalar quantity denoted $\langle a, b\rangle$ satisfying the following properties:

$$\langle a, b\rangle = \langle b, a\rangle^{\dagger} \tag{5.1}$$

$$\langle \alpha a, \beta b + c\rangle = \alpha^{\dagger}\beta\langle a, b\rangle + \alpha^{\dagger}\langle a, c\rangle \tag{5.2}$$

$$\begin{aligned} \langle a, a\rangle &> 0 \quad \text{if } a \neq 0 \\ \langle a, a\rangle &= 0 \quad \text{if } a = 0 \end{aligned} \tag{5.3}$$

where a, b, and c are functions and α and β are scalars. Complex conjugation is denoted using a dagger (†). Any inner product satisfying these properties can be used to define a natural norm

$$\|a\| = \sqrt{\langle a, a\rangle} \tag{5.4}$$

and the associated metric

$$d(a, b) = \|a - b\| \tag{5.5}$$

The metric provides us with the notion of "distance" between two functions.

Two functions a and b in an inner product space are said to be *orthogonal* if

$$\langle a, b\rangle = 0 \tag{5.6}$$

In a similar fashion, functions $\{B_n\}$ in an inner product space form an orthogonal set if

$$\langle B_m, B_n\rangle = 0 \qquad m \neq n \tag{5.7}$$

The set $\{B_n\}$ is said to be *complete* if the zero function is the only function in the inner product space orthogonal to each member of the set. A set $\{B_n\}$ that is both complete and orthogonal is said to be a basis and can be used to represent any function f in the inner product space in the sense that

$$\|f - \sum_n \alpha_n B_n\| = 0 \tag{5.8}$$

where the $\{a_n\}$ are scalar coefficients uniquely determined by [1, 2]

$$\alpha_n = \frac{\langle B_n, f\rangle}{\langle B_n, B_n\rangle} \tag{5.9}$$

In practice, we are forced to project the functions of interest onto a finite-dimensional subspace of the original inner product space. In the subspace, the basis is truncated to the form $\{B_1, B_2, \ldots, B_N\}$, and the representation is given by

$$f \cong f^N = \sum_{n=1}^{N} \alpha_n B_n \tag{5.10}$$

The scalar coefficients $\{\alpha_1, \alpha_2, \ldots, \alpha_N\}$ are selected to minimize the distance between the function f and the representation f^N. The error

$$d(f, f^N) = \|f - f^N\| \tag{5.11}$$

is minimized when the coefficients are chosen to make the error orthogonal to the N-dimensional basis, that is,

$$\langle B_n, f - f^N\rangle = 0 \qquad n = 1, 2, \ldots, N \tag{5.12}$$

This is known as an orthogonal projection. Because of the orthogonality of the basis functions, the coefficients are the same in the subspace as in the original inner product space. Therefore, the orthogonal projection (the "best" representation as measured by the metric) is realized using coefficients from Equation (5.9).

Now, consider the equation $Lf = g$. We seek a representation of the solution f in an N-dimensional subspace of the original domain of L, and Equation (5.10) provides the general form. The best approximation is obtained when the coefficients from Equation (5.9) are employed. Unfortunately, since f is not known, the coefficients $\{\alpha_n\}$ cannot be determined directly from (5.9).

On the other hand, quantities defined on the range of the linear operator L are known and might be more convenient to work with. If the set $\{T_n\}$ forms a basis (complete and orthogonal) for the range space of the operator L, any function in the range may be represented in the N-dimensional subspace spanned by $\{T_1, T_2, \ldots, T_N\}$ according to

$$g \cong g^N = \sum_{m=1}^{N} \beta_m T_m \tag{5.13}$$

The projection that minimizes the error $d(g, g^N)$ employs coefficients

$$\beta_m = \frac{\langle T_m, g \rangle}{\langle T_m, T_m \rangle} \tag{5.14}$$

If g is a known function, the coefficients $\{\beta_m\}$ are readily determined. In a similar fashion, the function LB_n can be represented by

$$LB_n \cong \sum_{m=1}^{N} l_{mn} T_m \tag{5.15}$$

where the coefficients $\{l_{mn}\}$ that minimize the error

$$\left\| LB_n - \sum_{m=1}^{N} l_{mn} T_m \right\| \tag{5.16}$$

are given by

$$l_{mn} = \frac{\langle T_m, LB_n \rangle}{\langle T_m, T_m \rangle} \tag{5.17}$$

The coefficients of Equation (5.17) achieve an orthogonal projection in the range of the operator and therefore provide the best approximation as measured by the metric.

Returning to the approximate solution of the equation $Lf = g$, we represent the unknown solution f in the form of Equation (5.10), where $\{\alpha_n\}$ are unknowns to be determined. This representation produces a function on the range space having the form

$$Lf^N = \sum_{n=1}^{N} \alpha_n LB_n \tag{5.18}$$

Projecting this function on the N-dimensional subspace spanned by the set $\{T_1, T_2, \ldots, T_N\}$ yields

$$Lf^N \cong \sum_{m=1}^{N} \sum_{n=1}^{N} l_{mn} \alpha_n T_m \tag{5.19}$$

where the coefficients $\{l_{mn}\}$ are obtained from Equation (5.17). Equating this representation for Lf^N with the representation from (5.13) for g^N produces the discrete system of equations

$$\sum_{n=1}^{N} l_{mn}\alpha_n = \beta_m \qquad m = 1, 2, \ldots, N \tag{5.20}$$

This system is an $N \times N$ matrix equation that can be solved for the coefficients $\{\alpha_n\}$.

Although the above procedure provides a way of obtaining f^N, the coefficients $\{\alpha_1, \alpha_2, \ldots, \alpha_N\}$ obtained from the solution of Equation (5.20) are generally *not* the same as those specified in Equation (5.9). In other words, despite the fact that the projections in the range space are orthogonal, (5.20) does not ensure an orthogonal projection in the domain space and will usually not produce the best approximation as measured by the metric $d(f, f^N)$ (see Prob. P5.3). We momentarily defer a discussion of the consequences of this fact.

5.2 THE METHOD OF MOMENTS [3]

The preceding discussion indicates that an approximate solution of the linear equation $Lf = g$ may be obtained in the form

$$f \cong \sum_{n=1}^{N} \alpha_n B_n \tag{5.21}$$

where the functions $\{B_n\}$ are known basis functions defined on the domain of L and the scalars $\{\alpha_n\}$ are unknown coefficients to be determined. From an operational standpoint, Equation (5.21) is substituted into $Lf = g$, and a system of linear equations is obtained by forcing the residual

$$L\left(\sum_{n=1}^{N} \alpha_n B_n\right) - g = \sum_{n=1}^{N} \alpha_n L B_n - g \tag{5.22}$$

to be orthogonal to a set of testing functions $\{T_1, T_2, \ldots, T_N\}$. This produces the matrix equation $\mathbf{L}\alpha = \beta$ having entries

$$l_{mn} = \langle T_m, LB_n \rangle \tag{5.23}$$

and

$$\beta_m = \langle T_m, g \rangle \tag{5.24}$$

The matrix equation $\mathbf{L}\alpha = \beta$ is formally identical to Equation (5.20), except for the normalization. Provided the matrix $\mathbf{L}$ is nonsingular, the unknown coefficients can be found using standard matrix solution algorithms (Chapter 4).

Since the system of equations is obtained by forcing the residuals to be orthogonal to the testing functions, this procedure is often given the name *weighted-residual method* [4, 5]. In electromagnetics, it is also known as the *method of moments* [3, 6–8]. The roots of this procedure originate in the methods of Rayleigh, Ritz, and Galerkin developed near the turn of the century [4, 9]. All discretization procedures can be placed on a direct correspondence with this approach, at least approximately. The *finite-element method* is

equivalent, although it is often presented in the context of minimizing a quadratic functional [10, 11]. The classical *finite-difference method* can also be interpreted in the context of basis and testing functions (Prob. P5.5).

The discretization of a continuous equation by the method of moments necessarily involves the projection of the continuous linear operator onto finite-dimensional subspaces defined by the basis and testing functions. Although the method-of-moments system $\mathbf{L}\alpha = \beta$ is formally equivalent to Equation (5.20), the process outlined in (5.21)–(5.24) can be applied to produce an approximate solution regardless of whether or not the functions $\{B_n\}$ and $\{T_m\}$ form complete, orthogonal sets. As illustrated in the following section, the basis and testing functions used in practice are often not orthogonal sets. (We continue to denote these as "basis functions" despite the fact that they do not satisfy the definition of a basis in the strict sense.) If the $\{T_m\}$ form an orthogonal set, the projection of the range space onto the testing functions is orthogonal and therefore a best approximation. Unfortunately, even if the basis functions are orthogonal, the projection of the domain space onto the basis functions is not guaranteed to be orthogonal. This makes it difficult to make firm statements about the convergence of the numerical approximation in Equation (5.21) to the exact solution as $N \to \infty$. In any case, since N is necessarily finite for numerical calculations, the result obtained from (5.21) is always approximate.

The choice of basis and testing functions is the principal issue arising within a method-of-moments implementation. As discussed by Harrington [3], practical factors affecting the selection of basis functions include the desired accuracy of the approximate solution, the relative complexity of the resulting matrix entries, and computational constraints that place an upper limit on the matrix size. While one would expect that the desired goal would always be to obtain the best accuracy with the fewest basis functions, the need to adapt the approach (i.e., the finished computer program) to a wide variety of different problems may motivate a less than optimal formulation. The basis and testing functions should be linearly independent and able to accurately approximate the f and Lf, respectively. (Actually, the basis functions do not have to be linearly independent, as long as the set $\{LB_n\}$ is linearly independent [12].) It cannot be overemphasized that for good results the basis and testing functions must be chosen with the particular operator L in mind. Although the domain of L may be restricted to functions that satisfy certain differentiability requirements, in many cases the differentiability, or "smoothness," requirements of the basis functions can be shifted to the testing functions. We have already seen this property illustrated in specific examples throughout Chapters 2 and 3 and consider it in more detail in Section 5.6.

The applications of interest range from those involving rather simple geometries to those involving structures of arbitrary shape and composition and include unknown quantities that may be scalar or vector functions of two or three variables. Because of this generality, functions that are commonly employed as a basis in other applications (such as the exponential or trigonometric functions) may not be well suited for many of the electromagnetic problems we encounter. From the examples presented in Chapters 2 and 3, we have seen that it is often more convenient to employ subsectional functions on irregular domains. Subsectional functions differ from the classical basis introduced in Section 5.1 in two respects. First, subsectional functions usually do not form orthogonal sets. Second, increasing the order N of the approximation usually alters every element of the set $\{B_1, B_2, \ldots, B_N\}$, rather than just adding an additional function. However, subsectional functions are quite flexible in that they can be easily adapted to arbitrary domains. The following section illustrates some of the common subsectional basis and testing functions in widespread use for discretizing electromagnetics equations.

5.3 EXAMPLES OF SUBSECTIONAL BASIS FUNCTIONS

The examples considered in Chapters 2 and 3 illustrate the use of the simplest subsectional basis and testing functions. In this section, several families of subsectional functions are introduced for one-dimensional scalar quantities. The generalization to the multidimensional, vector case is deferred until Chapter 9.

For single-dimension scalar quantities, the simplest basis functions in use are illustrated in Figure 5.1. These include the Dirac delta function

$$B_0(x) = \delta(x - x_0) \tag{5.25}$$

the pulse, or piecewise-constant, function

$$B_1(x) = p(x; x_1, x_2) = \begin{cases} 1 & x_1 < x < x_2 \\ 0 & \text{otherwise} \end{cases} \tag{5.26}$$

and the subsectional triangle function

$$B_2(x) = t(x; x_3, x_4, x_5) = \begin{cases} \dfrac{x - x_3}{x_4 - x_3} & x_3 < x < x_4 \\ \dfrac{x_5 - x}{x_5 - x_4} & x_4 < x < x_5 \end{cases} \tag{5.27}$$

On uniform intervals, the subsectional pulse and triangle functions are actually the simplest members of a family of spline functions generated by the convolution

$$\begin{aligned} B_n(x) &= B_{n-1}(x) * \frac{1}{\Delta} p\left(x; -\frac{\Delta}{2}, \frac{\Delta}{2}\right) \\ &= \frac{1}{\Delta} \int_{-\Delta/2}^{\Delta/2} B_{n-1}(x - x')\, dx' \end{aligned} \tag{5.28}$$

The next member of this family is the quadratic spline

$$B_3(x) = q\left(x; -\frac{3\Delta}{2}, -\frac{\Delta}{2}, \frac{\Delta}{2}, \frac{3\Delta}{2}\right) = \begin{cases} 0 & x < -\dfrac{3\Delta}{2} \\ \dfrac{9}{8} + \dfrac{3x}{2\Delta} + \dfrac{x^2}{2\Delta^2} & -\dfrac{3\Delta}{2} < x < -\dfrac{\Delta}{2} \\ \dfrac{3}{4} - \dfrac{x^2}{\Delta^2} & -\dfrac{\Delta}{2} < x < \dfrac{\Delta}{2} \\ \dfrac{9}{8} - \dfrac{3x}{2\Delta} + \dfrac{x^2}{2\Delta^2} & \dfrac{\Delta}{2} < x < \dfrac{3\Delta}{2} \\ 0 & x > \dfrac{3\Delta}{2} \end{cases} \tag{5.29}$$

which is illustrated in Figure 5.2. Higher order functions can be constructed from the recursive formula in (5.28). The next such spline, denoted B_4, is a cubic function.

Suppose the domain of interest is divided into intervals, or *cells*, along the x-axis. The pulse function has support confined to a single cell and is orthogonal to pulse functions located at other cells. An expansion in pulse functions produces a piecewise-constant representation. The triangle function overlaps two adjacent cells and shares each cell with

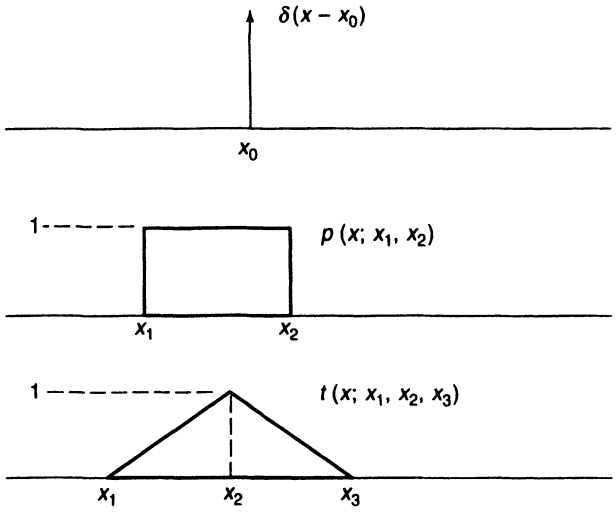

Figure 5.1 Definition of basis and testing functions.

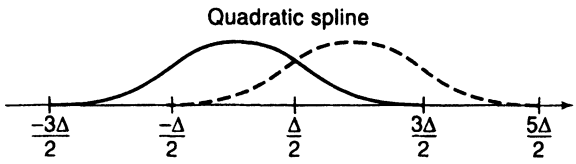

Figure 5.2 Quadratic spline functions.

adjacent triangle functions. By continuously superimposing triangle functions along the entire domain of interest, a global piecewise-linear approximation is achieved as depicted in Figure 5.3. The quadratic spline overlaps three cells and shares each cell with two other quadratic splines. A piecewise-quadratic representation throughout the domain can be obtained by a superposition of shifted quadratic splines (Figure 5.2). Triangle functions ensure continuity of the function they represent; quadratic splines provide continuity of the function and its first derivative. It is noteworthy that neither the triangle functions nor the quadratic splines form orthogonal sets.

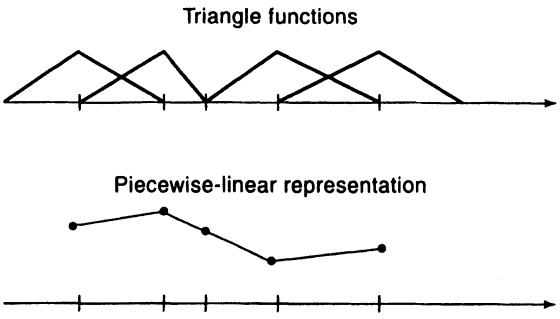

Figure 5.3 Linear interpolation using triangle functions.

Another important family of subsectional functions is obtained from the Lagrangian interpolation polynomials. These polynomials interpolate between function values at a finite number of locations throughout the domain of interest. The first-order Lagrangian functions

are piecewise linear and interpolate between the function at endpoints of the domain. These functions are identical to the subsectional triangle functions defined in Equation (5.27). The second-order Lagrangian functions consist of three quadratic polynomials that interpolate between two endpoints and one interior point in the interval of interest. On the interval spanning $[-1, 1]$ and employing the origin as the interior point, the three quadratic functions have the form

$$\phi_1(x) = \tfrac{1}{2}x(x - 1) \tag{5.30}$$

$$\phi_2(x) = 1 - x^2 \tag{5.31}$$

and

$$\phi_3(x) = \tfrac{1}{2}x(x + 1) \tag{5.32}$$

These are illustrated in Figure 5.4. All three functions span the domain $[-1, 1]$; the first equals one at $x = -1$ and vanishes at $x = 0$ and $x = 1$, the second has unity value at $x = 0$ and equals zero at the endpoints, and the third is zero at $x = -1$ and $x = 0$ and equals one at $x = 1$. The superposition of these functions, appropriately weighted, provides a quadratic representation of the function over the interval.

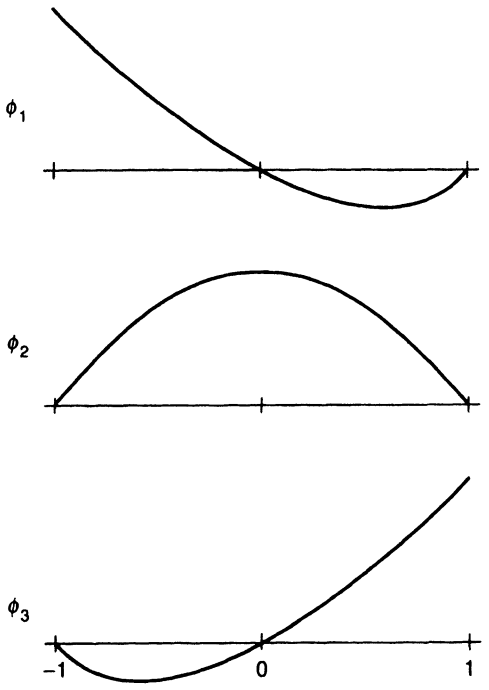

Figure 5.4 Quadratic Lagrangian functions.

In the general case, higher order Lagrangian functions can be defined so that the jth function of order $N - 1$ on a general point set within the interval $[x_1, x_N]$ is given by the formula

$$\phi_j(x) = \frac{(x - x_1)(x - x_2)\cdots(x - x_{j-1})(x - x_{j+1})\cdots(x - x_N)}{(x_j - x_1)(x_j - x_2)\cdots(x_j - x_{j-1})(x_j - x_{j+1})\cdots(x_j - x_N)} \tag{5.33}$$

where $j = 1, 2, \ldots, N$ and

$$x_1 < x_2 < \cdots < x_N \tag{5.34}$$

There are N polynomials of order $N - 1$ spanning the interval.

The Lagrangian polynomials could be used as an entire-domain basis spanning the global region of interest. In practice, however, higher order polynomial interpolation is often unstable [13]. Consequently, the traditional approach is to divide the global domain of interest into cells and employ a local expansion in Lagrangian polynomials of some fixed order within each cell. To improve the representation, polynomials of the original degree are employed on a finer subdivision of cells. In other words, the functions are used as a subsectional basis.

The two Lagrangian functions that interpolate (from each side) to a point located at a boundary between two cells can be thought of as one continuous function spanning two cells (Figure 5.5). Thus, the actual basis set consists of (a) functions spanning two cells that interpolate at cell boundaries and (b) functions spanning only one cell that interpolate to an interior point. It is also obvious from Figure 5.5 that the quadratic Lagrangian functions maintain continuity across cell boundaries but do not ensure continuity of any derivatives. This illustrates a significant difference between the quadratic Lagrangian functions and the quadratic spline functions defined in Equation (5.29). Another difference is that the splines are not interpolatory for orders higher than linear.

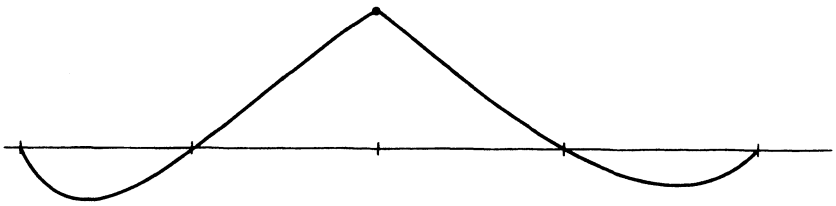

Figure 5.5 Global quadratic Lagrangian function interpolatory at a node between adjacent cells.

If it is necessary to ensure continuity of derivatives across cell boundaries, a set of functions known as Hermitian interpolates can be used. The lowest order Hermitian set consists of four cubic polynomials that interpolate between the function values at the endpoints of the interval. On the domain $[-1, 1]$, these four functions can be expressed as

$$\Psi_1(x) = \tfrac{1}{4}(1 - x)^2(2 + x) \tag{5.35}$$

$$\Psi_2(x) = \tfrac{1}{4}(1 + x)^2(2 - x) \tag{5.36}$$

$$\Psi_3(x) = \tfrac{1}{4}(1 - x^2)(1 - x) \tag{5.37}$$

$$\Psi_4(x) = \tfrac{1}{4}(-1 + x^2)(1 + x) \tag{5.38}$$

The four Hermitian functions are used in the same manner as the Lagrangian polynomials. Functions Ψ_1 and Ψ_2 interpolate the values at $x = -1$ and $x = 1$, respectively; these functions have vanishing first derivative at both endpoints. The functions Ψ_3 and Ψ_4 are zero at both endpoints but have unity first derivative at $x = -1$ and $x = 1$, respectively. Thus Ψ_1 and Ψ_2 maintain continuity of the function while Ψ_3 and Ψ_4 maintain continuity of

the first derivative across cell boundaries. The representation requires the superposition of all four functions. Each basis function can be thought of as spanning the two cells adjacent to their interpolation point.

A variety of other ad hoc subsectional basis functions are in use for applications such as wire antennas or scatterers [14, 15]. A common expansion function used in electromagnetics is the sinusoidal triangle function defined as

$$S(x) = \begin{cases} \dfrac{\sin(kx - kx_1)}{\sin(kx_2 - kx_1)} & x_1 < x < x_2 \\ \dfrac{\sin(kx_3 - kx)}{\sin(kx_3 - kx_2)} & x_2 < x < x_3 \end{cases} \tag{5.39}$$

In common with the piecewise-linear triangle function defined in Equation (5.27), this function overlaps two adjacent cells. In fact, if the range of kx is small, $S(x)$ is almost identical to the triangle function in shape. If the parameter k in the function is taken equal to the electromagnetic wavenumber in the surrounding medium, the Helmholtz operator $(\nabla\nabla\cdot + k^2)$ applied to the basis function produces the simple result

$$\left(\frac{\partial^2}{\partial x^2} + k^2\right) S(x) = \delta(x - x_1)\frac{k}{\sin(kx_2 - kx_1)} + \delta(x - x_3)\frac{k}{\sin(kx_3 - kx_2)} \tag{5.40}$$
$$-\delta(x - x_2)k[\cot(kx_3 - kx_2) + \cot(kx_2 - kx_1)]$$

Because of the delta functions, the expression for the electric field produced by an electric current density $S(x)$ is easily obtained in closed form using Equation (1.52). The closed-form evaluation of the field may be an advantageous property in practice.

An alternate expansion function can be constructed from the three-term sinusoid

$$I(x) = A + B\sin(kx - kx_0) + C\cos(kx - kx_0) \tag{5.41}$$

defined locally over a cell. Two of the three coefficients in Equation (5.41) are evaluated in order to ensure the proper continuity or discontinuity of the function at the two cell boundaries, leaving one unknown coefficient that represents the amplitude of the basis function. For instance, a sinusoidal type of spline function can be obtained from this general form by employing this expansion on three adjacent cells in order to obtain a continuous function that vanishes at the outer endpoints and has first derivatives that vanish at the outer endpoints. The three-term sinusoid is well suited for representing the current on thin wires and is used in the Numerical Electromagnetics Code (NEC) [16], where it provides the flexibility to build a derivative discontinuity into the current at junctions between wires of different radii.

A variety of other basis functions are presented in the literature [10, 11, 14, 15]. In recent years, the success of *wavelets* for signal-processing applications has motivated their use as basis functions for representing currents within integral equation formulations [17–19]. Because of their oscillatory nature, wavelets produce matrices that are effectively sparse and may offer computational advantages.

The extension of the scalar Lagrangian functions to the multidimensional case is discussed in Chapter 9. Chapter 9 also considers several *vector* expansion functions, which are often useful for multidimensional electromagnetic applications. These expansion functions share the characteristics that (1) they do not usually form orthogonal sets and (2) their "completeness" property is obtained by shrinking the domain of support rather than adding functions of greater polynomial order.

5.4 INTERPOLATION ERROR

If a linear combination of the basis functions used in a discretization can exactly represent the solution, the method-of-moments process should adjust the coefficients to produce the exact solution. If the basis functions cannot exactly represent the solution, even the best possible choice of coefficients leaves some residual error, known as *interpolation error*. The interpolation error associated with a piecewise-polynomial basis function is relatively easy to characterize.

Consider the function

$$f(x) = a + bx + cx^2 \tag{5.42}$$

Suppose $f(x)$ is approximated on the interval $(-\frac{1}{2}\Delta < x < \frac{1}{2}\Delta)$ using two of the subsectional triangle functions defined in (5.27), that is,

$$f(x) \cong f_{ap}(x) = f(-\tfrac{1}{2}\Delta)B_1(x) + f(\tfrac{1}{2}\Delta)B_2(x) \tag{5.43}$$

where B_1 and B_2 are defined throughout the interval by

$$B_1(x) = \frac{1}{2} - \frac{x}{\Delta} \tag{5.44}$$

$$B_2(x) = \frac{1}{2} + \frac{x}{\Delta} \tag{5.45}$$

By direct substitution, the function f_{ap} can be expressed in the form

$$f_{ap}(x) = a + bx + c(\tfrac{1}{2}\Delta)^2 \tag{5.46}$$

The approximation captures the correct constant and linear dependence but obviously cannot represent the quadratic term. In this case, the error between f and f_{ap} reaches a maximum at $x = 0$ and has the peak value

$$|\text{Error}| = \tfrac{1}{4}c\Delta^2 \tag{5.47}$$

As the interval size shrinks to zero, the peak error decreases as $O(\Delta^2)$. In other words, a 50% decrease in Δ causes a 75% decrease in the error. The same result holds if $f(x)$ is an arbitrary polynomial and a piecewise-linear representation is employed.

This result is easily generalized to other polynomial orders (Prob. P5.6), with the result that a representation of polynomial degree p results in an interpolation error of order Δ^{p+1} as $\Delta \to 0$. (Strictly speaking, this estimate requires continuity of the derivative of the function being approximated. With Lagrangian functions, however, the estimate holds even in that case if a cell boundary coincides with the point of derivative discontinuity.) In a situation where variable-sized cells are employed, the error estimate is valid provided that Δ corresponds to the largest cell in the discretization.

Using some of the sample results presented in Chapters 2 and 3, we can investigate the solution accuracy in actual method-of-moments calculations. For example, Table 2.2 shows the error in the surface current density as the number of basis functions is increased for an integral equation formulation. As expected for piecewise-constant basis functions, this error closely follows an $O(\Delta)$ behavior as $\Delta \to 0$. From Table 3.2 and the result of Prob. P3.9, we observe that the error in the field produced by a finite-element discretization of the Helmholtz equation behaves as $O(\Delta^2)$ when piecewise-linear basis functions are employed. This also agrees with the theoretical prediction $O(\Delta^{p+1})$. Although we generally expect

additional error in method-of-moments results due to the fact that the coefficients are not optimal, the primary error in these examples appears to be interpolation error.

Occasionally, however, the coefficient values at specific interpolation points are more accurate than $O(\Delta^{p+1})$ as $\Delta \to 0$. This phenomenon, known as superconvergence, occurs when errors combine in such a way that their leading order terms happen to cancel. A similar phenomenon sometimes occurs with secondary quantities derived from the numerical results of integral equation formulations and will be considered in Section 5.12.

5.5 DISPERSION ANALYSIS [20, 21]

An alternate way of investigating the error associated with a basis representation is to consider the distortion that a plane-wave solution undergoes as it propagates across the computational domain. Consider the one-dimensional scalar Helmholtz equation

$$\frac{d^2E_z}{dx^2} + k^2E_z(x) = 0 \tag{5.48}$$

and the use of piecewise-linear basis and testing functions. In addition, we assume that the mesh is uniform and large in extent and ignore boundaries. Under these conditions, the mth finite-element equation can be written (Prob. P5.5)

$$2E_m - E_{m-1} - E_{m+1} - k^2\Delta^2\left(\tfrac{2}{3}E_m + \tfrac{1}{6}E_{m-1} + \tfrac{1}{6}E_{m+1}\right) = 0 \tag{5.49}$$

where Δ is the cell size and E_{m-1}, E_m, and E_{m+1} are the basis function coefficients for $E_z(x)$ at $x_m - \Delta$, x_m, and $x_m + \Delta$, respectively.

Equation (5.48) has a traveling-wave solution $E_z(x) = E_0e^{\pm jkx}$ and motivates an investigation to determine whether a similar solution exists for (5.49). In fact, the discrete solution

$$E_z(x_m) = E_0e^{\pm j\beta x_m} \tag{5.50}$$

satisfies (5.49) exactly provided that

$$\beta = \frac{1}{\Delta}\cos^{-1}\left(\frac{1 - (k\Delta)^2/3}{1 + (k\Delta)^2/6}\right) \tag{5.51}$$

Consequently, Equations (5.50) and (5.51) constitute the numerical solutions that would be obtained for waves on an infinite, uniform mesh. In a lossless medium, β is real valued for

$$k\Delta \leq \sqrt{12} \tag{5.52}$$

and complex valued for larger $k\Delta$. Therefore, for cell sizes smaller than $\Delta \cong 0.55\lambda$, *the error in the numerical result is entirely phase error.* For larger cell sizes, a complex-valued β indicates error in both the magnitude and phase of the wave.

The phase error across a single cell is given by

$$k\Delta - \beta\Delta \tag{5.53}$$

and is easily tabulated as a function of Δ from Equation (5.51). For example, the error across a cell of width $\Delta = 0.1\lambda$ is about 0.57°. This phase error builds progressively across a mesh, so cells of width 0.1λ spanning a 10λ region would produce 57° of total error.

For this cell size, an error of 180° would be reached in about 32 wavelengths. Table 5.1 summarizes the predicted phase error per wavelength as a function of Δ.

TABLE 5.1 Predicted Phase Error per Wavelength and the Percent Error in the Field as a Function of Cell Size Δ

Δ	Phase error per λ (deg)	Percent error per λ
0.2λ	20.103	34.9
0.1λ	5.670	9.9
0.05λ	1.464	2.6
0.025λ	0.369	0.64
0.0125λ	0.0925	0.16
0.00625λ	0.0231	0.04

Note: From Equations (5.51) and (5.53).

These theoretical estimates correlate with observed numerical error. Figure 5.6 depicts the error in a finite-element solution of the Helmholtz equation for a plane-wave incident on a dielectric slab of thickness 1.0λ and relative permittivity $\varepsilon_r = 2$ [22]. A region of 20λ is included in the finite-element mesh on either side of the slab. Results are shown as a function of the uniform cell size Δ used throughout the free-space region of the computational domain. The fact that the error is primarily phase error explains the oscillatory behavior of the $\Delta = 0.2\lambda$ curve (the error increasing to a maximum, subsequently decreasing, and increasing again throughout the region). For $\Delta = 0.1\lambda$, a phase error of 180° is reached at a distance of about 32λ from the side of the region where the excitation is coupled into the equation, as predicted. Despite the fact that the region is not entirely homogeneous and is terminated by radiation boundary conditions, the observed discretization error exhibits excellent agreement with the theoretical predictions from (5.53).

From an examination of Table 5.1, it is easily discerned that the error decreases as $O(\Delta^2)$ as $\Delta \to 0$, in agreement with the theoretical interpolation error associated with piecewise-linear basis functions. In fact, although we have obtained the preceding results for the one-dimensional case, the phase error estimates arising from this dispersion analysis remain essentially unchanged for waves on triangular-cell models in two dimensions. Irregular cell sizes are typical for the two-dimensional case, and a worst-case approach involving a general triangular-cell model would assign the length of the largest cell edge to Δ, although this may overestimate the error.

Given the size of the computational domain and the phase constant of the medium, the cell density required to produce an acceptable error can be estimated in advance. For example, a cell density of 30 cells/λ would limit the maximum phase error across a free-space region spanning 10λ to about 10°. In order to model a one-dimensional region spanning 100λ with a 10° maximum phase error, however, the minimum required cell density must be increased to about 80 cells/λ. Because the phase error builds on itself, the peak error increases with the size of the region being modeled. *This suggests that to limit the growth in phase error, the cell density must increase as the size of the computational domain increases.* These findings directly impact finite-element discretizations of the Helmholtz equation, where the entire region of interest must often be contained within the computational domain.

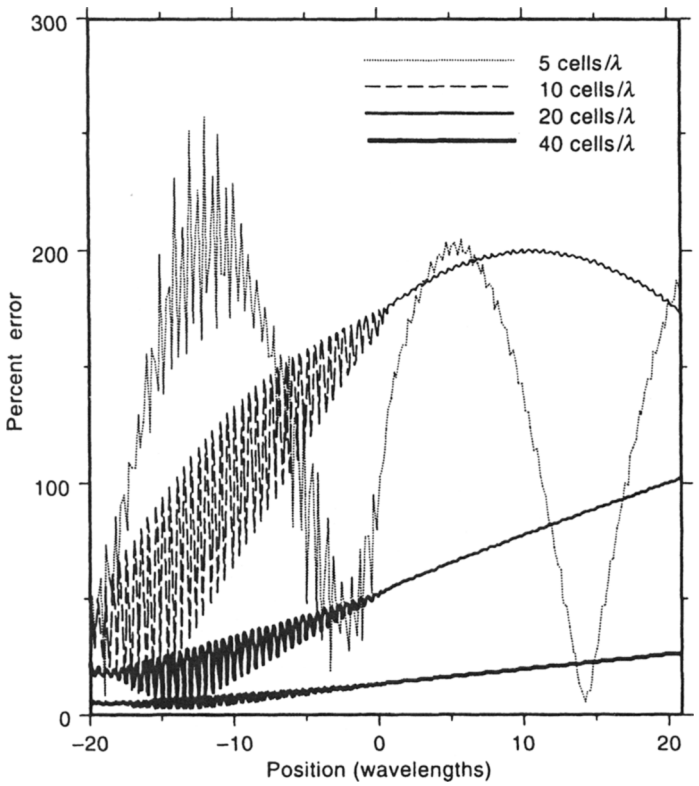

Figure 5.6 Percent error in the finite-element result for the one-dimensional scalar Helmholtz equation. The curves suggest that the error is primarily associated with the phase of the result. After [22]. ©1991 IEEE.

Integral equation formulations using an exact Green's function to span a large region will not incur this cumulative error.

The growth of error with domain size poses an obvious difficulty when attempting to model electrically large structures. However, the error can be reduced by the use of higher order polynomial interpolation functions [21], such as those introduced in Section 5.3. A dispersion analysis for quadratic basis functions is suggested in Prob. P5.10 and indicates a substantial reduction in error.

5.6 DIFFERENTIABILITY CONSTRAINTS ON BASIS AND TESTING FUNCTIONS

We now turn our attention to the role of the testing functions in the numerical solution process. Previous sections have considered the intrinsic error introduced by the basis functions, under the assumption that functions of any polynomial order can be employed. In fact, the specific linear operator to be discretized dictates a minimum polynomial degree for the basis and testing functions. In practice, this degree must increase in proportion to

the number of derivatives operating on the unknown function. We will explore this issue in the context of several typical equations.

Consider an integral equation of the form

$$E^i(x) = \frac{k\eta}{4}\int_a^b J(x')H_0^{(2)}(k|x-x'|)\,dx' \qquad a < x < b \tag{5.54}$$

where $J(x)$ is an unknown to be determined and $E^i(x)$ is a known excitation. For instance, $J(x)$ could represent the TM current density induced on a conducting strip. The integrand in this case involves the Hankel function $H_0^{(2)}$, which is a weakly singular function behaving as

$$H_0^{(2)}(kx) \approx 1 - j\frac{2}{\pi}\ln\left(\frac{\gamma kx}{2}\right) \tag{5.55}$$

as $x \to 0$, where γ is defined in Equation (2.13). We wish to discretize Equation (5.54) into matrix form using the method of moments with subsectional basis and testing functions. We first define an inner product

$$\langle a, b\rangle = \int_a^b a^\dagger(x)b(x)\,dx \tag{5.56}$$

Using this inner product, a generic combination of basis and testing functions produces a matrix equation $\mathbf{L}\alpha = \beta$ having entries of the form

$$l_{mn} = \frac{k\eta}{4}\int_a^b T_m^\dagger(x)\int_a^b B_n(x')H_0^{(2)}(k|x-x'|)\,dx'\,dx \tag{5.57}$$

We are now in a position to consider what constraints, if any, must be imposed on the basis and testing functions to ensure a meaningful numerical solution.

Throughout this text, we will adhere to the heuristic assumption that to ensure a meaningful numerical solution *it is necessary to employ a combination of basis and testing functions that keep the coefficients in Equation (5.57) finite and well defined for any location of the basis or testing functions throughout the domain.* From an examination of the integral operator in Equation (5.54), we observe that if piecewise-constant or pulse basis functions are used to represent $J(x)$, the integral produces a function that is continuous and bounded. Thus, the use of Dirac delta functions as testing functions would produce a finite result for each entry of the matrix (regardless of where the testing functions were placed throughout the range of the operator). We also observe that if Dirac delta functions were used as both basis and testing functions, the entries would be infinite if the testing location coincided with a basis function location. Therefore, we conclude that pulse basis functions and Dirac delta testing functions provide the minimum degree of smoothness necessary to discretize an integral operator having a weakly singular kernel.

We will now demonstrate that, for many basis and testing functions, the entries in Equation (5.57) remain the same if the basis and testing functions are exchanged. To show this, we employ a Fourier transformation in conjunction with the convolution theorem. The Fourier transform can be defined as

$$F\{A(x)\} = \tilde{A}(k_x) = \int_{-\infty}^{\infty} A(x)e^{-jk_x x}\,dx \tag{5.58}$$

The transformation converts functions of x to functions of the variable k_x. The inverse

Fourier transform, which converts functions of k_x to functions of x, is defined as

$$F^{-1}\{\tilde{A}(k_x)\} = A(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \tilde{A}(k_x)e^{jk_x x}\,dk_x \tag{5.59}$$

The convolution of two functions is given by

$$A(x) * B(x) = \int_{-\infty}^{\infty} A(x')B(x-x')\,dx' \tag{5.60}$$

The convolution theorem associated with the Fourier transform states that

$$A(x) * B(x) = F^{-1}\{\tilde{A}(k_x)\tilde{B}(k_x)\} \tag{5.61}$$

The integral operator of (5.54) is obviously a convolution, but it is not as obvious that the double integral appearing in (5.57) can be thought of as a double convolution. However, since subsectional functions B_n and T_m are usually shifted to the location of cells n and m, respectively, we can redefine the functions as

$$B_n(x) = B(x - x_n) \tag{5.62}$$

and

$$T_m(x) = T(x - x_m) \tag{5.63}$$

It follows that the entries of the method-of-moments matrix $\mathbf{L}$ can be written as

$$l_{mn} = \frac{k\eta}{4} T^R(x) * \left[B(x) * H_0^{(2)}(k|x|)\right]\Big|_{x=x_m-x_n} \tag{5.64}$$

where T^R is the space reversal of the complex conjugate of the testing function, that is,

$$T^R(x) = T^\dagger(-x) \tag{5.65}$$

For many of the real-valued subsectional functions used in practice (e.g., splines), $T^R = T$. The convolution theorem in (5.61) shows that the convolutions can be performed in any order. It follows that interchanging the basis and testing functions does not change the entries of $\mathbf{L}$.

Since the entries of $\mathbf{L}$ are the same if the basis and testing functions are interchanged, the new arrangement also satisfies our criteria for minimum differentiability. This means that it is appropriate to employ Dirac delta functions as basis functions to discretize Equation (5.54), provided that pulse functions (or functions smoother than pulse functions) are used as testing functions. In principle, we are modifying the definitions of the domain and range of the linear operator when we consider the use of delta functions as a basis for this problem. This is possible because, as a consequence of the testing function introduced during discretization, the original operator is replaced by a "weak" operator that imposes less stringent mathematical properties.

Suppose we now consider the integro-differential equation

$$E^i(x) = \frac{\eta}{4k}\left(\frac{\partial^2}{\partial x^2} + k^2\right)\int_a^b J(x')H_0^{(2)}(k|x-x'|)\,dx' \qquad a < x < b \tag{5.66}$$

where E^i and J play the same role as they did in the preceding example. This equation could represent TE scattering from conducting strips. The presence of two additional derivatives in Equation (5.66) indicates that the minimum differentiability requirements have increased

by two polynomial orders. Employing the convolution notation, the matrix entries have the form

$$l_{mn} = \left[\frac{\eta}{4k}\left(\frac{\partial^2}{\partial x^2} + k^2\right) T^R(x) * \left[B(x) * H_0^{(2)}(k|x|)\right]\right]_{x=x_m - x_n} \tag{5.67}$$

It follows from the properties associated with the Fourier transform of derivatives and Equation (5.61) that the convolution and differentiation operations in (5.67) commute. Thus, the differentiation may be transferred to either of the basis or testing functions in whole or part. For example, to satisfy our heuristic assumption of keeping the matrix elements finite for any location of basis or testing function, we might employ (a) quadratic spline basis functions with Dirac delta testing functions, (b) triangle basis functions with pulse testing functions, (c) pulse basis functions with triangle testing functions, or (d) Dirac delta basis functions with quadratic spline testing functions. Of course, smoother combinations would also satisfy the minimum criterion.

The differential operators considered in Chapter 3 require similar constraints on basis and testing functions. For example, a one-dimensional weak equation similar in form to (3.5) would require entries of the **L** matrix

$$l_{mn} = \int_a^b \left(\frac{1}{\mu_r}\frac{dT_m}{dx}\frac{dB_n}{dx} - \varepsilon_r k^2 T_m(x) B_n(x)\right) dx \tag{5.68}$$

The minimum differentiability condition suggests the use of piecewise-linear basis functions (triangles) and piecewise-constant testing functions (pulses). An equivalent combination such as quadratic spline basis functions and Dirac delta testing functions also meet the minimum smoothness condition.

Whether the operator is of the integral or differential type, the minimum smoothness condition should be satisfied if reasonable accuracy is expected in the numerical solution. Our heuristic condition places a lower limit on the net differentiability of the basis and testing functions, but not a specific constraint on either function. Since derivatives appearing in the expression for l_{mn} can be moved to the basis or testing function using the convolution idea or straightforward integration by parts, other factors can dictate the specific choice of basis and testing functions. Violation of the minimum smoothness condition will likely degrade the accuracy of the result and prevent numerical solutions from converging to the exact as the number of expansion functions is increased.

Although the exchange of basis and testing functions does not alter the entries of **L** for convolutional operator equations, it does change the entries of the excitation vector β. The need to maintain a well-defined β will also constrain the process of selecting basis and testing functions. In addition, if the representation of the current density is to be used within secondary calculations, that fact may motivate the use of smoother basis functions than the minimum identified above. (For example, it may be necessary to evaluate fields at specific points near the scatterer, for which a delta function, pulse function, or even triangle function representation of the current is likely to produce erratic near fields.) A third consideration is the combination of different geometrical structures generally requiring different types of basis and testing functions. As an example, the NEC [16] employs Dirac delta testing functions for discretizing both the EFIE for conducting wires and the MFIE for conducting surfaces, which simplifies the calculation of mutual interaction terms (the same subroutines can be used regardless of whether the testing location is on a wire or on a surface).

Other factors enter into the choice of basis and testing functions. In practice, it may not be desired to represent the unknown quantity by continuous functions. For instance,

the true fields at a dielectric boundary may exhibit a discontinuity or a derivative discontinuity. Often, approximations employed to simplify the scatterer geometry may overwhelm improvements in the basis functions. In several of the examples discussed in Chapter 2, the smooth surface of a scatterer was approximated by a flat-strip model. Once that type of approximation is employed, it makes little sense to try to use basis functions having continuous derivatives at the strip edge. In the past, the numerical integration effort associated with the matrix entries of integral equation formulations sometimes dictated the use of simple integrands. Recent trends seem to favor the use of sophisticated quadrature libraries [23] to efficiently evaluate the necessary integrals to a rather arbitrary degree of accuracy, making the use of more complicated basis functions a fairly convenient task.

We have not yet considered the impact of testing function choice on the relative accuracy of the numerical solution. To illustrate the accuracy for a specific example, consider the EFIE for TM scattering from perfectly conducting cylinders (Section 2.1)

$$E_z^{\text{inc}}(t) = jk\eta \int J_z(t') \frac{1}{4j} H_0^{(2)}(kR)\, dt' \tag{5.69}$$

where

$$R = \sqrt{[x(t) - x(t')]^2 + [y(t) - y(t')]^2} \tag{5.70}$$

and t is a parametric variable around the contour of the cylinder. We wish to test the accuracy of the method-of-moments discretization of this equation using the spline basis and testing functions defined in (5.25)–(5.28).

For circular cylinders excited by a uniform plane wave, exact and numerical solutions can be systematically compared using a normalized error

$$\text{Percent error} = \frac{\left\| J_z^{\text{exact}} - J_z^{\text{numerical}} \right\|}{\left\| J_z^{\text{exact}} \right\|} \times 100 \tag{5.71}$$

based on the norm

$$\left\| J_z^{\text{exact}} - J_z^{\text{numerical}} \right\| = \sqrt{\int \left| J_z^{\text{exact}}(t) - J_z^{\text{numerical}}(t) \right|^2 dt} \tag{5.72}$$

The integration required in (5.72) will be performed over the actual basis function set used in each case to represent the current density, with no additional smoothing or interpolation after the coefficients are determined.

Table 5.2 summarizes the results for a cylinder of size $ka = 6$ as a function of the order of the splines employed as basis and testing functions [24, 25]. Results are presented for 20 basis and testing functions, corresponding to a density of only 3.3 basis functions per wavelength, and 60 basis and testing functions, corresponding to a density of 10 basis functions per wavelength. From Table 5.2, we see that the solution accuracy is primarily determined by the order of the basis functions, or equivalently by the polynomial interpolation error associated with the approximation of J_z. Although the testing functions are expected to play a role in the accuracy of the basis function coefficients, for this example the coefficients appear to be almost independent of the testing function order.

TABLE 5.2 Error in Surface Current Density

	Percent Error	
Order of Testing	20 Unknowns	60 Unknowns
	Order of Basis 1	
0	34.9	11.3
1	35.0	11.3
2	35.6	11.3
3	35.8	11.3
4	36.0	11.3
5	36.2	11.3
	Order of Basis 2	
0	11.7	0.89
1	11.7	0.89
2	11.8	0.89
3	11.9	0.89
4	11.9	0.89
	Order of Basis 3	
0	6.56	0.10
1	6.51	0.10
2	6.53	0.10
3	6.55	0.10

Note: As measured by Equation (5.71) for a TM circular cylinder with circumference of 6λ as a function of the order of the splines employed as basis and testing functions [24, 25]. Basis and testing functions of the indicated order were employed with equal-size cells to construct the matrix equation. A spline of order n has polynomial order $n - 1$ with order zero denoting a Dirac delta function.

5.7 EIGENVALUE PROJECTION THEORY

If the domain and range of the continuous operator L coincide, there may be solutions to the eigenvalue equation

$$Le = \lambda e \tag{5.73}$$

where λ is an eigenvalue and e an eigenfunction of L. There are several reasons for considering Equation (5.73). Applications involving cavities or waveguides lead naturally to eigenvalue equations, and it may be necessary to construct numerical solutions in that context. On the other hand, the behavior of the continuous equation $Lf = g$ will also depend on the eigenfunctions and eigenvalues of L, and these properties are projected in some sense onto the method-of-moments matrix $\mathbf{L}$. Since the numerical stability of matrix solution algorithms (Chapter 4) depends on the condition number and eigenvalues of the system matrix, it is natural to inquire into the relationship between the eigenvalues of the

continuous operator and those of the matrix. (Although a general linear operator L may not have eigenvalues, the matrix $\mathbf{L}$ always has N eigenvalues.)

To study the relationship, consider the discretization of the eigenvalue equation (5.73) using basis functions $\{B_n\}$ and testing functions $\{T_m\}$. In other words, we employ the expansion

$$e \cong \sum_{n=1}^{N} e_n B_n \tag{5.74}$$

and weigh the residual equations to zero with testing functions $\{T_m\}$ in order to construct the discrete equation

$$\sum_{n=1}^{N} \langle T_m, LB_n \rangle e_n = \lambda \sum_{n=1}^{N} \langle T_m, B_n \rangle e_n \qquad m = 1, 2, \ldots, N \tag{5.75}$$

This is a generalized matrix eigenvalue equation of the form $\mathbf{Le} = \lambda \mathbf{Se}$, where the entries of $\mathbf{L}$ are

$$l_{mn} = \langle T_m, LB_n \rangle \tag{5.76}$$

and the entries of $\mathbf{S}$ are

$$s_{mn} = \langle T_m, B_n \rangle \tag{5.77}$$

As long as the basis and testing functions are each linearly independent sets, $\mathbf{S}$ is nonsingular and Equation (5.75) can be written as

$$\mathbf{S}^{-1}\mathbf{Le} = \lambda \mathbf{e} \tag{5.78}$$

This is an ordinary eigenvalue equation and can be thought of as a discretization of Equation (5.73). It follows that the eigenvalues of the product matrix $\mathbf{S}^{-1}\mathbf{L}$ should approximate those of the original operator L. Furthermore, the corresponding eigenvectors of $\mathbf{S}^{-1}\mathbf{L}$ provide coefficients for Equation (5.74) to approximate the eigenfunctions of L. The accuracy of a particular matrix eigenvalue should depend on the ability of the basis functions to represent the associated eigenfunction.

Conceptually, the eigenvalues of L are projected from the continuous operator onto the method-of-moments matrix $\mathbf{L}$ by the discretization. However, the matrix $\mathbf{S}$ provides a scaling that alters the direct projection and complicates the relationship between the original and matrix eigenvalues. In the special case where the basis functions and testing functions are orthonormal, so that $\mathbf{S}$ is an identity matrix, the eigenvalues of $\mathbf{L}$ are a direct approximation to those of the continuous operator. We will study the numerical accuracy of the eigenvalue projection process in the following section using several canonical examples.

To summarize, if it is necessary to discretize the continuous eigenvalue equation, Equation (5.78) provides the matrix analog. In addition, the relationship between the eigenvalue spectrum of the continuous operator L and the matrix operator $\mathbf{L}$ can be explored in order to learn more about the deterministic equation $Lf = g$ and its numerical treatment. The knowledge gained from this relationship will be the focus for the following sections of this chapter and parts of Chapter 6.

5.8 CLASSIFICATION OF OPERATORS FOR SEVERAL CANONICAL EQUATIONS

We have seen that electromagnetic scattering problems can be posed in terms of equations involving integral or differential operators. To characterize the typical behavior of the specific linear operators of interest, we first classify them into one of several types. These classifications follow the conventions of functional analysis [1, 2, 7, 26–29].

Section 2.1 presented the EFIE for TM scattering from perfectly conducting strips or cylinders. The TM EFIE operator involves an integration over the weakly singular Hankel function $H_0^{(2)}(kR)$. Under mild assumptions concerning the smoothness of the scatterer surface, this integral is an example of a *compact operator.*

Consider the special case of a circular cylinder of radius a. The TM EFIE operator is given by

$$L_{\text{EFIE}}^{\text{TM}}(J_z) = \frac{k\eta}{4}\int_{\phi'=0}^{2\pi} J_z(\phi')H_0^{(2)}(kR)a\,d\phi' \tag{5.79}$$

where

$$R = 2a\left|\sin\left[\tfrac{1}{2}(\phi - \phi')\right]\right| \tag{5.80}$$

For this simple geometry, solutions to the eigenvalue equation $Le_n = \lambda_n e_n$ are easily found in closed form (Prob. P5.11). The eigenfunctions of this operator are the exponential functions

$$e_n(\phi) = e^{jn\phi} \tag{5.81}$$

and the corresponding eigenvalues are

$$\lambda_n^{\text{TM, EFIE}} = \tfrac{1}{2}(\eta\pi ka)J_n(ka)H_n^{(2)}(ka) \tag{5.82}$$

where J_n and H_n are the nth order Bessel and Hankel functions, respectively.

The eigenvalues in Equation (5.82) are complex valued and lie in the right-half complex plane. A plot of λ_0, λ_1, and λ_2 as a function of ka is shown in Figure 5.7. There are a number of noteworthy characteristics that can be gleaned from the behavior of the eigenvalues. For electrically small cylinders, the eigenvalues can be simplified using asymptotic formulas for the Bessel and Hankel functions as $ka \to 0$, yielding

$$\lambda_0^{\text{TM}} \approx \frac{\eta\pi ka}{2}\left[1 - j\frac{2}{\pi}\ln\left(\frac{\gamma ka}{2}\right)\right] \to 0 \tag{5.83}$$

$$\lambda_n^{\text{TM}} \approx j\frac{\eta ka}{2|n|} \to 0 \qquad n \neq 0 \tag{5.84}$$

The asymptotic behavior for large n is also given by Equation (5.84) and indicates that the eigenvalues cluster at the origin as $n \to \infty$. This is a typical characteristic of a compact operator and leads directly to the conclusion that the inverse of a compact operator is unbounded (since the reciprocal eigenvalues "cluster" at infinity).

As $ka \to \infty$, straightforward analysis yields asymptotic formulas such as

$$\lambda_0^{\text{TM}} \approx \eta\cos^2\left(ka - \tfrac{1}{4}\pi\right) - j\eta\cos\left(ka - \tfrac{1}{4}\pi\right)\sin\left(ka - \tfrac{1}{4}\pi\right) \tag{5.85}$$

which is the equation of a circle in the complex plane. For large ka, the eigenvalues tend to follow a common circular trajectory. It is interesting that they pass through the

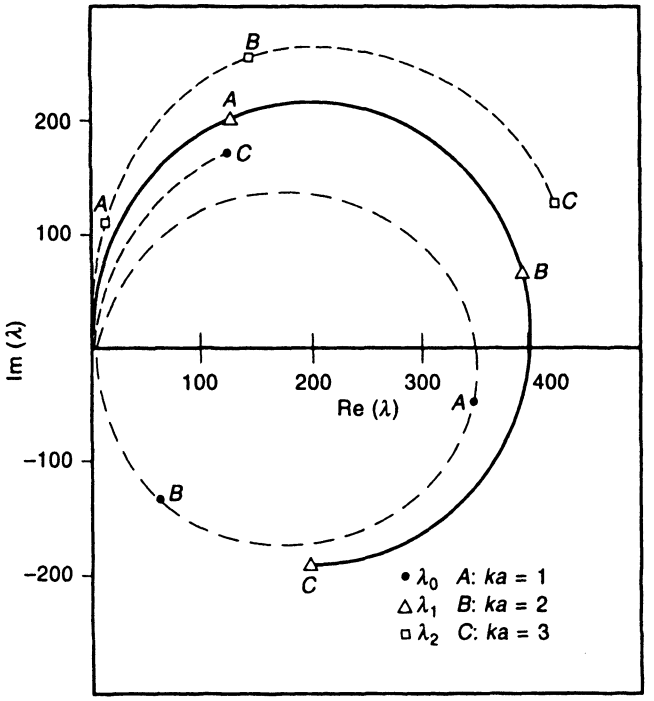

Figure 5.7 Plot of the three dominant eigenvalues of the TM EFIE as a function of ka. After [30]. ©1990 Hemisphere Publishing Corporation.

origin whenever $J_n(ka) = 0$. At these values of ka, the EFIE has homogeneous solutions associated with interior resonant cavity modes. Surface integral equations applied to closed geometries often exhibit homogeneous solutions. (A discussion of the interior solutions is deferred until Chapter 6.)

The behavior of the EFIE eigenvalues is directly indicative of the behavior of the eigenvalues associated with the method-of-moments matrix. The approach of Section 2.1 employed pulse basis functions and Dirac delta testing functions. Under these conditions, the scaling matrix **S** from Equation (5.78) is an identity matrix, and the eigenvalues of the method-of-moments matrix **L** should be a direct approximation to those of the original operator. To test this theory, Table 5.3 compares eigenvalues of **L** with the analytical eigenvalues from Equation (5.82) for a circular cylinder with $ka = 1$. Good agreement is observed between the numerical data and the exact eigenvalues.

The TE EFIE operator for perfectly conducting strips or cylinders was presented in Section 2.4. For a circular cylinder of radius a, the TE EFIE operator has the form

$$L_{\text{EFIE}}^{\text{TE}}(J_\phi) = \frac{\eta}{4k}\hat{\phi} \cdot (\nabla\nabla \cdot + k^2) \int_{\phi'=0}^{2\pi} \hat{\phi}(\phi') J_\phi(\phi') H_0^{(2)}(kR) a \, d\phi' \tag{5.86}$$

where R is defined in Equation (5.80). The eigenfunctions of this operator are also given by (5.81), with associated eigenvalues

$$\lambda_n^{\text{TE, EFIE}} = \tfrac{1}{2}(\eta\pi ka) J_n'(ka) H_n^{(2)\prime}(ka) \tag{5.87}$$

There are many similarities in form between the TE and TM eigenvalues. The TE eigen-

TABLE 5.3 First 10 Distinct Eigenvalues of TM EFIE Compared to Those of Moment-Method Matrix for Circular p.e.c. Cylinder with Circumference 1λ

TM EFIE (5.82)	30×30 Matrix
$346.50 - j\,39.96$	$346.43 - j\,41.95$
$114.59 + j\,203.43$	$114.55 + j\,201.56$
$7.81 + j\,112.24$	$7.82 + j\,110.35$
$0.23 + j\,67.40$	$0.23 + j\,65.62$
$0.00 + j\,48.77$	$0.00 + j\,47.08$
$0.00 + j\,38.49$	$0.00 + j\,36.96$
$0.00 + j\,31.85$	$0.00 + j\,30.51$
$0.00 + j\,27.19$	$0.00 + j\,26.13$
$0.00 + j\,23.74$	$0.00 + j\,23.02$
$0.00 + j\,21.06$	$0.00 + j\,20.82$

values also lie in the right-half complex plane, as illustrated by a plot of λ_0, λ_1, and λ_2 as a function of ka provided in Figure 5.8. As $ka \to \infty$, they follow a circular trajectory similar to that of the TM EFIE eigenvalues. The TE eigenvalues pass through the origin whenever $J_n'(ka) = 0$, indicating homogeneous solutions of the EFIE at those values of ka (see Chapter 6). However, their behavior differs from the TM EFIE eigenvalues in the limiting cases of small ka and large n. As $ka \to 0$,

$$\lambda_0^{\text{TE}} \approx j\frac{\eta ka}{2} \to 0 \tag{5.88}$$

$$\lambda_n^{\text{TE}} \approx -j\frac{\eta |n|}{2ka} \to -j\infty \qquad n \neq 0 \tag{5.89}$$

The large spread between the eigenvalues in the low-frequency case suggests that the TE EFIE is unstable for electrically small scatterers, an issue that will be considered in Section 10.6. The asymptotic form of the eigenvalues as $n \to \infty$ is also given by (5.89) and indicates that the TE eigenvalues cluster at infinity. This is a consequence of the derivatives appearing in Equation (5.86), and in common with differential operators this EFIE is known as an *unbounded operator.*

The method-of-moments procedure described in Section 2.4 for discretizing the TE EFIE involved the use of subsectional triangle basis functions and pulse testing functions. Because of the overlap between adjacent basis and testing functions, the scaling matrix $\mathbf{S}$ from Equation (5.78) is not an identity matrix. Table 5.4 compares the eigenvalues of $\mathbf{L}$ and $\mathbf{S}^{-1}\mathbf{L}$ with the analytical eigenvalues from Equation (5.87) for a circular cylinder with $ka = 1$. The eigenvalues of $\mathbf{S}^{-1}\mathbf{L}$ exhibit reasonable agreement with the eigenvalues of the continuous operator.

Section 2.2 presented an alternate approach for TE scattering from closed conducting cylinders using the MFIE. The MFIE operator is characteristic of a third type that we will call an *"identity-plus-compact" operator.* This type of operator is generally considered to possess the nicest mathematical properties of the three categories and is considered in detail in several texts [27–29].

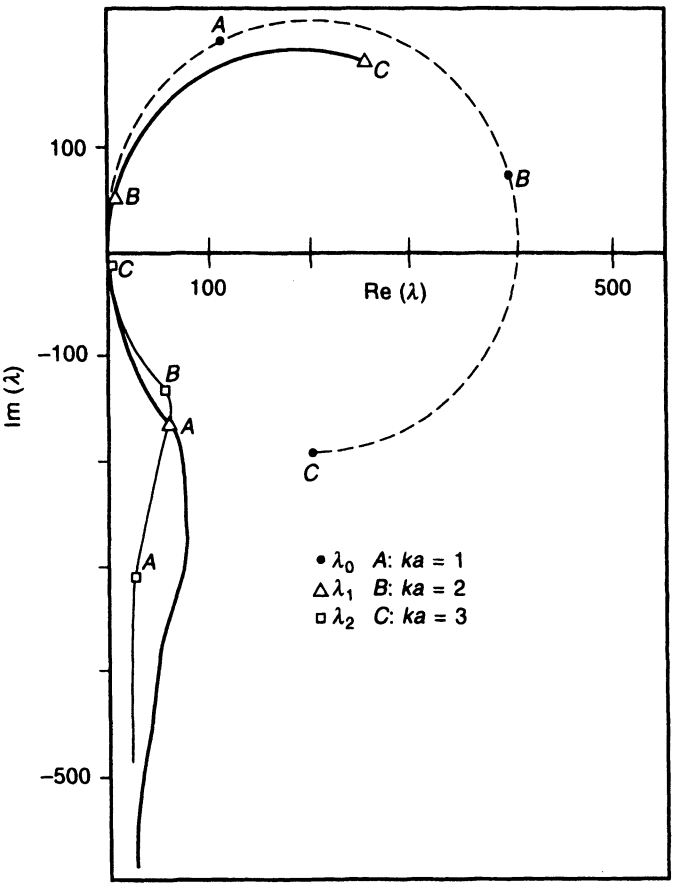

Figure 5.8 Plot of the three dominant eigenvalues of the TE EFIE as a function of ka. After [30]. ©1990 Hemisphere Publishing Corporation.

TABLE 5.4 First 10 Distinct Eigenvalues of TE EFIE Compared to Those of Moment-Method L-Matrix and Product Matrix $\mathbf{S}^{-1}\mathbf{L}$ for Circular p.e.c. Cylinder with Circumference 1λ

TE EFIE (5.87)	L-Matrix	$\mathbf{S}^{-1}\mathbf{L}$-Matrix
114.6 + j 203.4	111.6 + j 200.1	112.2 + j 201.1
62.6 – j 167.3	61.7 – j 166.6	62.4 – j 168.4
26.2 – j 313.5	25.6 – j 306.4	26.3 – j 314.9
1.9 – j 526.1	1.8 – j 500.5	1.9 – j 528.5
0.1 – j 727.4	0.1 – j 668.9	0.1 – j 733.2
0.0 – j 921.9	0.0 – j 813.0	0.0 – j 934.3
0.0 – j 1114	0.0 – j 935.6	0.0 – j 1137
0.0 – j 1305	0.0 – j 1037	0.0 – j 1344
0.0 – j 1495	0.0 – j 1120	0.0 – j 1555
0.0 – j 1685	0.0 – j 1185	0.0 – j 1770

Note: The order of S and L is 30.

If applied to a circular cylinder of radius a, the TE MFIE operator has the form

$$L_{\text{MFIE}}^{\text{TE}}(J_\phi) = J_\phi + \frac{1}{4j}\hat{\phi} \cdot \nabla \times \int_{\phi'=0}^{2\pi} \hat{\phi}(\phi') J_\phi(\phi') H_0^{(2)}(kR) a \, d\phi' \tag{5.90}$$

where R is defined in Equation (5.80). For the circular geometry, the eigenfunctions are the set $\{e^{jn\phi}\}$ and the corresponding eigenvalues are given by

$$\lambda_n^{\text{TE, MFIE}} = \tfrac{1}{2}(j\pi ka) J_n(ka) H_n^{(2)\prime}(ka) \tag{5.91}$$

These eigenvalues also lie in the right-half complex plane, as illustrated by a plot of λ_0, λ_1, and λ_2 as a function of ka provided in Figure 5.9. However, the limiting cases differ from the EFIE eigenvalues. For instance, as $ka \to 0$,

$$\lambda_0^{\text{TE, MFIE}} \approx 1 \tag{5.92}$$

$$\lambda_n^{\text{TE, MFIE}} \approx \tfrac{1}{2} \qquad n \neq 0 \tag{5.93}$$

As $n \to \infty$, the MFIE eigenvalues cluster at $0.5 + j0.0$ in the complex plane, a finite location bounded away from the origin.

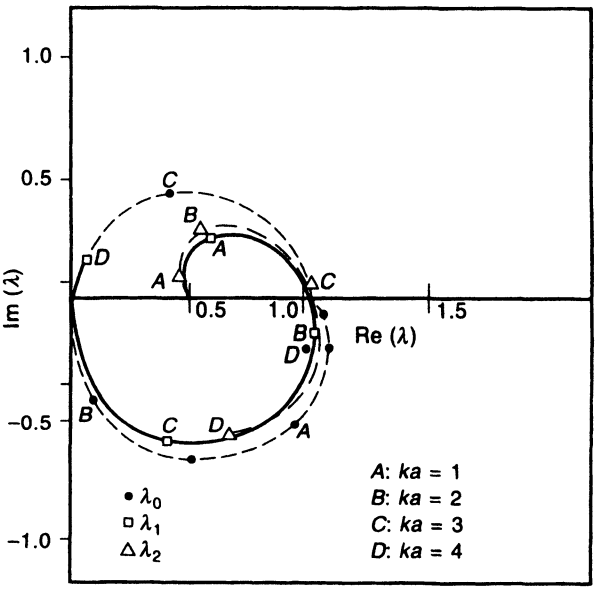

Figure 5.9 Plot of the three dominant eigenvalues of the TE MFIE as a function of ka. After [30]. ©1990 Hemisphere Publishing Corporation.

We have identified the three typical types of operators arising in electromagnetic scattering problems: the compact, unbounded, and "identity-plus-compact" operators. The distinction between these operators is easily illustrated by the eigenvalue behavior identified for the special case of circular cylinders, although the classification holds for similar operators even if no eigenvalue interpretation is possible. The characterization of the operator is important when discussing the convergence of numerical solutions, since convergence proofs assume some particular form. We now turn our attention to some of the arguments that have been proposed to prove convergence.

5.9 CONVERGENCE ARGUMENTS BASED ON GALERKIN'S METHOD [24, 31]

If the identical set of functions is used for basis and testing within a method-of-moments discretization, the procedure is denoted *Galerkin's method.* This approach is advantageous in the special case when (1) the operator L is self-adjoint with respect to the inner product, so that

$$\langle La, b\rangle = \langle a, Lb\rangle \tag{5.94}$$

and (2) the operator is positive definite, that is,

$$\langle a, La\rangle > 0 \text{ for all nonzero } a \tag{5.95}$$

In this situation, a second inner product space can be defined in terms of the new inner product

$$\langle a, b\rangle_2 = \langle a, Lb\rangle \tag{5.96}$$

If $T_m = B_m$, the orthogonal projection of the residuals onto the testing functions described by Equations (5.13)–(5.20) is equivalent to enforcing

$$\langle B_m, L(f^N - f)\rangle = 0 \tag{5.97}$$

in the original inner product space, which can be expressed as

$$\langle B_m, f^N - f\rangle_2 = 0 \tag{5.98}$$

in the new inner product space. This is a statement that the projection of the error in f^N is orthogonal to the basis in the new space.

Consequently, in the special case when L is a positive-definite, self-adjoint operator and the basis functions form a complete, orthogonal set in the *new* inner product space, the method-of-moments projection of the unknown f onto the basis functions is an orthogonal projection. It immediately follows that the numerical approximation f^N will converge to f in the limit as $N \to \infty$.

The constraint on orthogonality can be relaxed. In fact, it may be difficult to identify a practical basis satisfying the orthogonality

$$\langle B_m, LB_n\rangle = 0 \qquad m \neq n \tag{5.99}$$

However, a Gram–Schmidt procedure can be used to show the equivalence of any linearly independent set $\{B_n\}$ to an orthogonal set $\{\Psi_n\}$ satisfying (5.99). Since $\{B_1, \ldots, B_N\}$ and $\{\Psi_1, \ldots, \Psi_N\}$ span the same N-dimensional subspace, it is sufficient to force the residual error to be orthogonal to the set $\{B_n\}$. Thus, the orthogonality of the basis set is not necessary to establish the convergence of $f^N \to f$.

The preceding idea is often used to demonstrate the convergence of solutions to a finite-element discretization of Laplace's equation. Consider an inner product defined according to (5.56), for instance, with the second inner product defined by (5.96). Subsectional Lagrangian expansion functions satisfy the necessary completeness property in the new inner product space, and the operator is positive definite and self-adjoint. Unfortunately, the same argument is not directly applicable to the scalar Helmholtz equation when the

operator is indefinite, since then the second inner product is not valid. For the indefinite Helmholtz equation, convergence must be established by other means [32, 33].

The integral operators arising in electromagnetics are usually not positive definite or self-adjoint except in the static limit, and consequently the above convergence proof does not apply to the EFIE or MFIE. There may, however, be persuasive reasons for choosing $T_m = B_m$ in practice. For instance, using Galerkin's method with the EFIE produces matrices with diagonal symmetry, which permits a 50% reduction in computation during matrix construction and solution (recall Prob. P2.1 and the discussion in Section 4.4.)

5.10 CONVERGENCE ARGUMENTS BASED ON DEGENERATE KERNEL ANALOGS

We now turn our attention to an integral equation of the form $f + Lf = g$, where f is the unknown to be determined and L is a compact integral operator

$$Lf = \int K(x, x') f(x')\, dx' \tag{5.100}$$

defined in terms of a weakly singular kernel K. (This equation might represent the MFIE of Section 2.2 applied to a closed, smooth, perfectly conducting surface.) We define an inner product

$$\langle a, b \rangle = \int a^\dagger(x) b(x)\, dx \tag{5.101}$$

and postulate a (complete, orthogonal) basis $\{B_n\}$ for the domain space and a second basis $\{T_m\}$ for the range space. We therefore intend to seek a solution of the form

$$f(x) \cong \sum_{n=1}^{N} \alpha_n B_n(x) \tag{5.102}$$

Instead of substituting Equation (5.102) into the integral equation, however, we proceed in a different manner. Suppose that we project the kernel of the integral operator as a function of x' onto the basis $\{B_n^\dagger\}$ constructed from the complex conjugates of $\{B_n\}$. This yields

$$K(x, x') \cong \sum_{n=1}^{N} \gamma_n(x) B_n^\dagger(x') \tag{5.103}$$

where, to obtain an orthogonal projection,

$$\gamma_n(x) = \frac{\langle B_n^\dagger, K \rangle}{\langle B_n^\dagger, B_n^\dagger \rangle} = \frac{LB_n}{\langle B_n, B_n \rangle} \tag{5.104}$$

(The integration within the inner products is carried out over the primed coordinates.) We also project the kernel as a function of x onto the set $\{T_m\}$, to obtain

$$K(x, x') \cong K_N(x, x') = \sum_{m=1}^{N} \sum_{n=1}^{N} k_{mn} T_m(x) B_n^\dagger(x') \tag{5.105}$$

Here, K_N is an example of a *degenerate* kernel. The coefficients

$$k_{mn} = \frac{\langle T_m, \gamma_n \rangle}{\langle T_m, T_m \rangle} = \frac{\langle T_m, LB_n \rangle}{\langle T_m, T_m \rangle \langle B_n, B_n \rangle} \tag{5.106}$$

provide the best representation by minimizing the norm of the error

$$\| K - K_N \| = \left\| K(x, x') - \sum_{m=1}^{N} \sum_{n=1}^{N} k_{mn} T_m(x) B_n^{\dagger}(x') \right\| \tag{5.107}$$

This error norm converges to zero as $N \to \infty$ as long as the kernel K is either a continuous or a weakly singular function [34].

We now consider the equation

$$f^N(x) + L^N(f^N) = f^N(x) + \int K_N(x, x') f^N(x')\, dx' = g(x) \tag{5.108}$$

obtained by replacing the original kernel by the degenerate kernel K_N. We want to establish the relationship between f^N and the actual solution f. First, however, note that this system can be expressed as

$$f^N(x) + \sum_{m=1}^{N} \sum_{n=1}^{N} k_{mn} T_m(x) \langle B_n, f^N \rangle = g(x) \tag{5.109}$$

which illustrates that f^N can be obtained exactly in the form

$$f^N(x) = g(x) - \sum_{m=1}^{N} \delta_m T_m(x) \tag{5.110}$$

where

$$\delta_m = \sum_{n=1}^{N} k_{mn} \langle B_n, f^N \rangle = \sum_{n=1}^{N} \frac{\langle T_m, LB_n \rangle}{\langle T_m, T_m \rangle} \frac{\langle B_n, f^N \rangle}{\langle B_n, B_n \rangle} = \frac{\langle T_m, L^N f^N \rangle}{\langle T_m, T_m \rangle} \tag{5.111}$$

If we introduce

$$c_n = \langle B_n, f^N \rangle \tag{5.112}$$

the coefficients $\{c_n\}$ and $\{\delta_n\}$ can be found by constructing an inner product of Equation (5.93) with B_j to produce

$$c_j + \sum_{n=1}^{N} c_n \left(\sum_{m=1}^{N} \frac{\langle T_m, LB_n \rangle}{\langle T_m, T_m \rangle} \frac{\langle B_j, T_m \rangle}{\langle B_n, B_n \rangle} \right) = \langle B_j, g \rangle \tag{5.113}$$

Solving this $N \times N$ linear system completes the solution. In addition to obtaining f^N in the form of Equation (5.110), we have also determined the projection of f^N onto the basis $\{B_n\}$ as

$$f^N(x) \cong \sum_{n=1}^{N} \frac{c_n}{\langle B_n, B_n \rangle} B_n(x) \tag{5.114}$$

Although this orthogonal projection is the best representation of f^N in the subspace spanned by the basis $\{B_n\}$, it is not an exact equality, as is Equation (5.110).

It is a straightforward matter to show that f^N converges to f as $N \to \infty$ [27]. Using the identity

$$f - f^N + L^N f - L^N f^N = L^N f - Lf \tag{5.115}$$

we write

$$f - f^N = (I + L^N)^{-1}(L^N - L)f \tag{5.116}$$

where I denotes the identity operator. We then construct the inequality

$$\frac{\|f - f^N\|}{\|f\|} \leq \|(I + L^N)^{-1}\| \|L^N - L\| \tag{5.117}$$

Since L and L^N are compact operators, the first norm on the right-hand side is bounded as long as -1 is not an eigenvalue of L.[1] The second norm converges to zero as $N \to \infty$ as a consequence of (5.106). Therefore (as long as -1 is not an eigenvalue of the operator L), f^N converges to f.

We now investigate the relationship between f^N and the formal method-of-moments solution of the original equation $f + Lf = g$. To establish this relationship, first substitute (5.114) into (5.109) to obtain

$$\sum_{n=1}^{N} \frac{c_n}{\langle B_n, B_n \rangle} B_n(x) + \sum_{m=1}^{N} \sum_{n=1}^{N} k_{mn} c_n T_m(x) \cong g(x) \tag{5.118}$$

where the coefficients $\{c_n\}$ are defined in Equation (5.112) and produce an orthogonal projection of f^N onto the basis functions. Because (5.114) is only exact in the subspace spanned by $\{B_1, B_2, \ldots, B_N\}$, Equation (5.118) is not an equality except in that subspace. It follows that a projection onto the testing functions

$$\sum_{n=1}^{N} \frac{c_n}{\langle B_n, B_n \rangle} \langle T_m, B_n \rangle + \sum_{n=1}^{N} k_{mn} c_n \langle T_m, T_m \rangle \cong \langle T_m, g \rangle \qquad m = 1, 2, \ldots, N \tag{5.119}$$

is also only approximate. But, by rewriting the left side of this equation to obtain

$$\sum_{n=1}^{N} \frac{c_n}{\langle B_n, B_n \rangle} \langle T_m, (I + L)B_n \rangle \cong \langle T_m, g \rangle \qquad m = 1, 2, \ldots, N \tag{5.120}$$

we arrive at the method-of-moments system representing the original equation $f + Lf = g$. This demonstrates that the coefficients produced by solving the $N \times N$ method-of-moments matrix equation

$$\sum_{n=1}^{N} \langle T_m, (I + L)B_n \rangle \alpha_n = \langle T_m, g \rangle \qquad m = 1, 2, \ldots, N \tag{5.121}$$

are *not* the best representation of f^N in the subspace spanned by $\{B_1, B_2, \ldots, B_N\}$ [except in the special case $T_m = B_m$, in which case (5.120) is an exact equality]. However, in either case the approximate equality in (5.118) becomes exact in the limit as $N \to \infty$, suggesting that the method-of-moments solution does converge to the orthogonal projection.

To summarize, assuming that the basis and testing functions are complete, orthogonal sets and that L in the equation $f + Lf = g$ is a compact operator, the method-of-moments solution converges to f^N in the limit as $N \to \infty$, and f^N in turn converges to f. Whether

[1] In practice, -1 may be an eigenvalue of L for certain geometries; refer to Chapter 6.

or not the method-of-moments solution is an orthogonal projection of f^N onto the basis functions is not particularly relevant, since there is no reason to believe that f^N is an orthogonal projection of f onto the basis functions for finite N in the first place.

We next consider the equation $Lf = g$, where L is a compact operator. (The TM EFIE representing scattering from smooth perfectly conducting cylinders is an example of such an equation.) Unfortunately, the convergence arguments outlined above cannot be applied to this equation. Replacing the original kernel by the degenerate kernel K_N leads to the equation

$$\sum_{m=1}^{N}\sum_{n=1}^{N} k_{mn} T_m(x) \langle B_n, \tilde{f} \rangle \cong g(x) \tag{5.122}$$

The approximate equality is a consequence of the fact that a finite combination of the $\{T_m\}$ may not be able to represent g exactly. We can obtain an exact equality by considering the modified equation

$$\sum_{m=1}^{N}\sum_{n=1}^{N} k_{mn} T_m(x) \langle B_n, f^N \rangle = \sum_{m=1}^{N} \frac{\langle T_m, g \rangle}{\langle T_m, T_m \rangle} T_m(x) \tag{5.123}$$

obtained by projecting the excitation $g(x)$ onto the basis $\{T_m\}$. Since this is an orthogonal projection, the representation for g converges as $N \to \infty$ as long as g is well behaved. The modified equation is just the method-of-moments system

$$\sum_{n=1}^{N} \langle T_m, LB_n \rangle \alpha_n = \langle T_m, g \rangle \qquad m = 1, 2, \ldots, N \tag{5.124}$$

In this case, we have defined f^N so that its projection into the subspace spanned by $\{B_1, B_2, \ldots, B_N\}$ is the method-of-moments solution. By definition, the $\{\alpha_n\}$ produce an orthogonal projection of f^N onto $\{B_1, B_2, \ldots, B_N\}$.

However, we are not able to show that f^N converges to f as $N \to \infty$. For an equation of the form $Lf = g$, (5.115) is modified to

$$L^N f - L^N f^N = L^N f - Lf + g - g^N \tag{5.125}$$

from which we construct

$$f - f^N = (L^N)^{-1}[(L^N - L)f + (g - g^N)] \tag{5.126}$$

The appropriate inequality obtained from this result is

$$\|f - f^N\| \leq \|(L^N)^{-1}\| \|(L^N - L)f + (g - g^N)\| \tag{5.127}$$

Since L^N is compact, its inverse increases without bound as $N \to \infty$, and the inequality does not bound the error in f^N. Consequently, we cannot use this argument to prove general convergence for an equation such as the TM EFIE with the form $Lf = g$, where L is a compact operator. (In fact, the difficulty here is of a fundamental nature, and solutions to this type of equation may not exist in the absence of additional constraints on g. See Prob. P5.15.)

The convergence argument also fails to apply in the case where L represents an unbounded operator. In that situation, the kernel K is strongly singular and the projection onto the basis and testing functions in Equation (5.105) may not converge as $N \to \infty$. Therefore the degenerate kernel analog cannot be used to show convergence for unbounded integro-differential equations, such as the TE EFIE.

5.11 CONVERGENCE ARGUMENTS BASED ON PROJECTION OPERATORS

The preceding arguments require the basis and testing functions to form complete, orthogonal sets. However, the expansion functions used in practice seldom satisfy the orthogonality criterion. Furthermore, the testing functions used in practice may include Dirac delta functions, which cannot be used in the preceding convergence arguments. Fortunately, some of the assumptions needed to show convergence can be relaxed in a more general framework based on the concept of a *projection operator.* The following discussion has been adapted from Atkinson [27].

A bounded projection operator P_N can be defined to map functions in some space S to a subspace S^N in such a manner that

$$P_N f = f \quad \text{for all } f \text{ in } S^N \tag{5.128}$$

A simple example of a projection operator is that obtained from linear interpolation on a point set $\{x_1, x_2, \ldots, x_N\}$. Suppose we employ as basis functions the subsectional triangle functions defined in Equation (5.27), that is,

$$B_n = \begin{cases} \dfrac{x - x_{n-1}}{x_n - x_{n-1}} & x_{n-1} < x < x_n \\ \dfrac{x_{n+1} - x}{x_{n+1} - x_n} & x_n < x < x_{n+1} \end{cases} \tag{5.129}$$

These functions provide a linear interpolation, and the representation

$$f(x) \cong \sum_{n=1}^{N} f(x_n) B_n(x) \tag{5.130}$$

is clearly a projection onto a subspace spanned by $\{B_1, B_2, \ldots, B_N\}$. The projection also satisfies Equation (5.128). Furthermore, if f has a continuous first derivative, and if the points are spaced at equal intervals throughout the domain, the error associated with linear interpolation is known to be $O(\Delta^2)$, where $\Delta = x_n - x_{n-1}$ is the interval size. Under these conditions, (5.130) converges to $f(x)$ pointwise as $N \to \infty$, which is a stronger statement of convergence than the convergence-in-norm used in previous sections.

To place the solution of an equation $f + Lf = g$ in the context of projection operators, consider

$$f^N(x) + P_N L f^N = P_N g \tag{5.131}$$

where f^N lies in the subspace S^N. The relationship between f and f^N can be studied by combining Equation (5.131) with

$$P_N(f + Lf) = P_N f + P_N L f = P_N g \tag{5.132}$$

to obtain

$$f - f^N + P_N L f - P_N L f^N = f - P_N f \tag{5.133}$$

From Equation (5.133), we construct the inequality

$$\|f - f^N\| \leq \|(I + P_N L)^{-1}\| \|f - P_N f\| \tag{5.134}$$

Therefore, convergence of f to f^N is assured only if $(I + P_N L)^{-1}$ is bounded and if $P_N f$ converges to f as $N \to \infty$.

Atkinson has shown [27, pp. 51–54] that if L is compact, the pointwise convergence of $P_N f$ to f (for all f in the space) is sufficient to establish a bound on $(I + P_N L)^{-1}$. The complete development of this result is lengthy and will be omitted. The pointwise convergence is also sufficient to show that

$$\|f - P_N f\| \to 0 \quad \text{as } N \to \infty \tag{5.135}$$

It follows that f^N converges to f.

We now demonstrate the connection between Equation (5.131) and the method-of-moments system for two examples involving compact integral operators of the form

$$Lf = \int K(x, x') f(x')\, dx' \tag{5.136}$$

In the first example, we employ subsectional triangle functions from Equation (5.27) in connection with point matching on the set of equally spaced points $\{x_1, x_2, \ldots, x_N\}$ in order to obtain the discrete system

$$f_m + \sum_{n=1}^{N} f_n \int K(x_m, x') B_n(x')\, dx' = g_m \qquad m = 1, 2, \ldots, N \tag{5.137}$$

This system is identical in form to (5.131), with $P_N f$ defined as the projection operator that produces the element of the subspace spanned by $\{B_1, B_2, \ldots, B_N\}$ that interpolates to $f(x)$ at $(x_1, x_2, \ldots, x_N)$. The convergence of f^N to f is readily established for this example from the preceding results.

Therefore, in the case of an identity-plus-compact operator, a bounded projection operator is obtained when employing the method of moments with subsectional triangle basis functions and point matching. Furthermore, the method-of-moments solution converges to f as $N \to \infty$. This shows that a convergence proof is possible even if the testing functions are Dirac delta functions.

As a second example, consider the use of a linearly independent set of functions $\{B_1, B_2, \ldots, B_N\}$, defined in some subspace of an inner product space where

$$\langle a, b \rangle = \int a^{\dagger}(x) b(x)\, dx \tag{5.138}$$

Specifically, we want to consider the use of subsectional spline functions of some order (Section 5.3) for $\{B_n\}$. We seek a solution in the form

$$f^N(x) = \sum_{n=1}^{N} \alpha_n B_n(x) \tag{5.139}$$

The method-of-moments system obtained by employing the same functions for basis and testing (i.e., Galerkin's method) is

$$\sum_{n=1}^{N} \langle B_m, (I + L) B_n \rangle \alpha_n = \langle B_m, g \rangle \qquad m = 1, 2, \ldots, N \tag{5.140}$$

Let $\{\Psi_n\}$ denote the orthonormal basis constructed in the subspace S^N from $\{B_n\}$ by a Gram–Schmidt procedure. The projection operator associated with this discretization can

be defined as providing the orthogonal projection onto $\{\Psi_n\}$, that is,

$$P_N f = \sum_{n=1}^{N} (\Psi_n, f)\Psi_n(x) \tag{5.141}$$

Since Equation (5.140) forces the residual

$$f^N(x) + Lf^N - g(x) \tag{5.142}$$

to be orthogonal to each entry of $\{B_1, \ldots, B_N\}$, the residual is also orthogonal to the set $\{\Psi_1, \ldots, \Psi_N\}$, and (5.140) is equivalent to

$$P_N f^N + P_N L f^N = P_N g \tag{5.143}$$

Since $P_N f^N = f^N$, this method-of-moments system is also equivalent to that obtained from the projection operator defined in (5.131).

To show convergence of f^N to f, it is again necessary to demonstrate the pointwise convergence of $P_N f$ to f for all f in the space. Atkinson outlines a proof valid for piecewise-polynomial representations, which encompass the spline functions considered here, and we refer the readers to his text for the details [27, p. 68].

The preceding discussion introduced the "projection operator" interpretation of the method-of-moments procedure for solving $f + Lf = g$, where L is a compact operator. The analysis is applicable to a variety of practical discretizations involving subsectional basis and testing functions (which, as we have seen, may not always form orthogonal sets), even if Dirac delta testing functions are involved. Consequently, this approach is somewhat more general than the degenerate kernel analysis considered in Section 5.10.

As with the degenerate kernel approach, however, the convergence of f^N to f cannot be established by the preceding arguments for equations of the form $Lf = g$, where L is either compact or unbounded. For these systems, Equation (5.134) is replaced by

$$\|f - f^N\| \leq \|(P_N L)^{-1}\| \|(P_N L - L)f + (g - P_N g)\| \tag{5.144}$$

To establish convergence, $(P_N L)^{-1}$ must be bounded as $N \to \infty$, which is not the case for L compact. For unbounded L, the problem arises in showing that

$$\|P_N L - L\| \to 0 \quad \text{as } N \to \infty \tag{5.145}$$

Although this condition is certainly true in particular situations, it appears to be difficult to demonstrate for the general case.

5.12 THE STATIONARY CHARACTER OF FUNCTIONALS EVALUATED USING NUMERICAL SOLUTIONS [25]

In electromagnetic radiation and scattering problems, we are often primarily concerned with the far-zone fields needed to characterize the scattering cross section of a target or the radiation pattern of an antenna. These quantities can be expressed as quadratic functionals of the surface current density. Such functionals can sometimes be defined in a way that ensures that they have a stationary point at the true solution and consequently exhibit an error of only $O(\varepsilon^2)$ as $\varepsilon \to 0$ when the surface current used in their evaluation has an error of $O(\varepsilon)$ [35]. In this section, we show that when the surface current density is computed

via the method of moments, the error in the far fields or any other continuous functional depends equally on the basis and testing functions used in the method-of-moments process. Furthermore, if the basis and testing functions have similar smoothness characteristics, any continuous functional exhibits error of the same order as a stationary expression. Therefore, there appears to be no advantage in using a strictly stationary functional.

In general, suppose we want to compute an expression

$$Q = \langle f, h \rangle \tag{5.146}$$

where f is a solution of the linear equation $Lf = g$, h is a given function, and the inner product is defined in accordance with (5.1)–(5.3). We can seek a method-of-moments approximation

$$f \cong f^N = \sum_{n=1}^{N} \alpha_n B_n \tag{5.147}$$

obtained by forcing the residual $Lf^N - g$ to be orthogonal to testing functions $\{T_m\}$. In other words, the coefficients $\{\alpha_n\}$ are obtained from the system

$$\sum_{n=1}^{N} \alpha_n^{\dagger} \langle LB_n, T_m \rangle = \langle g, T_m \rangle \qquad m = 1, 2, \ldots, N \tag{5.148}$$

Let us assume that the error in the result is

$$f - f^N = \varepsilon_f \tag{5.149}$$

and investigate the error in the approximation $\langle f^N, h \rangle$.

To study the error in $\langle f^N, h \rangle$, we must first consider the equation $L^A e = h$, where L^A is the adjoint of L with respect to the inner product and h is the function appearing in (5.146). The adjoint operator is defined according to the relationship $\langle La, b \rangle = \langle a, L^A b \rangle$. Suppose we construct a method-of-moments approximation for the adjoint equation using basis functions $\{T_n\}$ to represent e and testing functions $\{B_m\}$ (the opposite of what we used above). In other words, e is replaced by

$$e \cong e^N = \sum_{n=1}^{N} \beta_n T_n \tag{5.150}$$

and the associated system of equations is

$$\sum_{n=1}^{N} \beta_n \langle B_m, L^A T_n \rangle = \langle B_m, h \rangle \qquad m = 1, 2, \ldots, N \tag{5.151}$$

Let us assume that the error in e^N will be

$$e - e^N = \varepsilon_e \tag{5.152}$$

The role of the adjoint equation becomes apparent if we rewrite Q as

$$Q = \langle f, h \rangle = \langle f, L^A e \rangle = \langle Lf, e \rangle \tag{5.153}$$

and consider the approximation

$$Q \cong Q^N = \langle Lf^N, e^N \rangle \tag{5.154}$$

Because of the way that f^N and e^N were constructed, we can show that

$$Q^N = \langle Lf^N, e^N \rangle = \langle Lf^N, e \rangle = \langle Lf, e^N \rangle \tag{5.155}$$

without approximation. This follows from Equations (5.148) and (5.151), for instance,

$$\begin{aligned}\langle Lf^N, e^N\rangle &= \sum_{m=1}^{N} \beta_m \left\langle L\left(\sum_{n=1}^{N} \alpha_n B_n\right), T_m \right\rangle \\ &= \sum_{m=1}^{N} \beta_m \langle g, T_m\rangle \\ &= \sum_{m=1}^{N} \beta_m \langle Lf, T_m\rangle \\ &= \langle Lf, e^N\rangle\end{aligned} \tag{5.156}$$

where the second equality is established by (5.148). Using (5.155), the error in Q can be expressed as

$$\begin{aligned}Q - Q^N &= \langle Lf, e\rangle - \langle Lf^N, e^N\rangle \\ &= \langle Lf, e\rangle - \langle Lf^N, e\rangle \\ &= \langle L(f - f^N), e\rangle \\ &= \langle L(f - f^N), e^N\rangle + \langle L\varepsilon_f, \varepsilon_e\rangle\end{aligned} \tag{5.157}$$

The first inner product is identically zero, by (5.156). Therefore, we obtain [36]

$$Q - Q^N = \langle L\varepsilon_f, \varepsilon_e\rangle \tag{5.158}$$

Although the expression $\langle Lf^N, e^N\rangle$ was used to obtain this error estimate, (5.155) shows that the identical error estimate applies to $\langle f^N, h\rangle$ or $\langle g, e^N\rangle$. Thus, there is no need to solve two method-of-moments problems to obtain $O(\varepsilon_f\varepsilon_e)$ error.

To summarize, if $f^N \cong f$ is obtained from a method-of-moments solution of $Lf = g$ using basis functions $\{B_n\}$ and testing functions $\{T_n\}$, the approximation $\langle f^N, h\rangle$ contains error of $O(\varepsilon_f\varepsilon_e)$, where ε_f and ε_e are defined in (5.149) and (5.152). We realize an error proportional to ε_e even though we do not actually solve the adjoint equation $L^A e = h$. From the previous discussion of solution error in Sections 5.4–5.6, it is reasonable to assume that ε_f is primarily determined by the ability of the set $\{B_n\}$ to represent f, while ε_e is primarily determined by the ability of the set $\{T_n\}$ to represent e. These observations suggest that the accuracy of $\langle f, h\rangle$ is enhanced if the set $\{B_n\}$ is a good basis for the domain space of L and the set $\{T_n\}$ is a good basis for the range space of L.

In the special case when L is self-adjoint, f satisfies $Lf = g$, and the functional to be computed is $\langle f, g\rangle$, we have

$$Q = \langle f, g\rangle = \langle Lf, f\rangle \cong \langle Lf^N, f^N\rangle \tag{5.159}$$

If the same functions are used for $\{B_n\}$ and $\{T_n\}$ (Galerkin's method), the approximation in (5.159) exhibits an error of $O(\varepsilon_f^2)$. This estimate follows directly from equation (5.158), since under these assumptions $Lf = g$ and its adjoint equation are identical, and e^N is replaced by f^N throughout. The $O(\varepsilon_f^2)$ error is often reported in the literature [37, 38] and sometimes used as an argument in favor of Galerkin's method. Although Equation (5.159) satisfies the strict definition of a stationary functional, it is important to note that the error in (5.159) is not necessarily any better than the error in the more general case described by (5.158). As long as ε_f and ε_e are comparable, the general functional $\langle f^N, h\rangle$ is also a stationary quantity. Thus, there is no definitive advantage to this special case and no

reason to use testing functions that are identical to the basis functions, in preference to an alternative choice that provides similar error levels.

To illustrate these calculations in practice, consider the scattered electric field produced by a point source illuminating a perfectly conducting target. A specific functional can be obtained from a reciprocity relationship. Consider two sets of sources, $(\bar{J}_1, \bar{K}_1)$ and $(\bar{J}_2, \bar{K}_2)$, and the fields produced by these in a common region, $(\bar{E}_1, \bar{H}_1)$ and $(\bar{E}_2, \bar{H}_2)$. In a manner similar to that developed in Section 1.6, Maxwell's curl equations can be written for each source–field pair and combined to produce

$$\begin{aligned} & E_2 \cdot J_1 - H_2 \cdot K_1 - E_1 \cdot J_2 + H_1 \cdot K_2 \\ & \quad = \bar{H}_2 \cdot \nabla \times \bar{E}_1 - \bar{E}_1 \cdot \nabla \times \bar{H}_2 + \bar{E}_2 \cdot \nabla \times \bar{H}_1 - \bar{H}_1 \cdot \nabla \times \bar{E}_2 \\ & \quad = \nabla \cdot (\bar{E}_1 \times \bar{H}_2 - \bar{E}_2 \times \bar{H}_1) \end{aligned} \tag{5.160}$$

After integrating throughout the region and applying the divergence theorem, we obtain

$$\begin{aligned} & \iiint_V \bar{E}_2 \cdot \bar{J}_1 - \bar{H}_2 \cdot \bar{K}_1 - \bar{E}_1 \cdot \bar{J}_2 + \bar{H}_1 \cdot \bar{K}_2 \\ & \quad = \iiint_V \nabla \cdot (\bar{E}_1 \times \bar{H}_2 - \bar{E}_2 \times \bar{H}_1) \\ & \quad = \iint_S (\bar{E}_1 \times \bar{H}_2 - \bar{E}_2 \times \bar{H}_1) \cdot \hat{n}\, dS \\ & \quad = 0 \end{aligned} \tag{5.161}$$

as the closed surface S containing the region recedes to infinity, provided that all the sources are contained within S. Therefore, we obtain the reciprocity relationship

$$\iiint_V \bar{E}_2 \cdot \bar{J}_1 - \bar{H}_2 \cdot \bar{K}_1 = \iiint_V \bar{E}_1 \cdot \bar{J}_2 - \bar{H}_1 \cdot \bar{K}_2 \tag{5.162}$$

where the volume V contains all sources. Equation (5.162) can be specialized to the case where the sources are entirely electric currents, as might describe a p.e.c. scatterer illuminated by an electric point source

$$\bar{J}_0 = \hat{u} J_0 \delta(\bar{r} - \bar{r}_0) \tag{5.163}$$

where $\bar{r}_0$ represents a location outside the scatterer. This source produces an incident electric field $\bar{E}^{\text{inc}}$, and in turn an induced surface current density $\bar{J}_s$ on the scatterer. The $\hat{u}$-component of the scattered electric field at the point $\bar{r}_0$ is given by

$$\hat{u} \cdot \bar{E}^s|_{\bar{r}=\bar{r}_0} = \frac{1}{J_0} \iiint_V \bar{E}^s \cdot \bar{J}_0 \tag{5.164}$$

However, using (5.162), this expression can be rewritten as

$$\hat{u} \cdot \bar{E}^s|_{\bar{r}=\bar{r}_0} = \frac{1}{J_0} \iiint_V \bar{E}^{\text{inc}} \cdot \bar{J}_s \tag{5.165}$$

Equation (5.165) is a functional for the scattered field at the source point.

We next introduce a method-of-moments formulation for $\bar{J}_s$, in terms of the EFIE. Since the EFIE operator is not self-adjoint with respect to an inner product defined according

to (5.1)–(5.3), instead we introduce a symmetric product defined as

$$(\bar{A}, \bar{B}) = \iint_S \bar{A}_{\tan} \cdot \bar{B}_{\tan}\, dS \tag{5.166}$$

These symmetric and inner products exhibit similar properties, except that the symmetric product does not produce a norm. Since $\bar{J}_s$ is confined to the surface of the scatterer, the functional in (5.165) can be written as

$$\hat{u} \cdot \bar{E}^s|_{\bar{r}=\bar{r}_0} = \frac{1}{J_0}(\bar{E}^{\text{inc}}, \bar{J}_s) = \frac{1}{J_0}(L\bar{J}_s, \bar{J}_s) \tag{5.167}$$

where L denotes the EFIE operator

$$L\bar{J}_s = \frac{\nabla\nabla\cdot + k^2}{j\omega\varepsilon} \iint \bar{J}_s(\bar{r}') \frac{e^{-jk|\bar{r}-\bar{r}'|}}{4\pi|\bar{r}-\bar{r}'|}\, dS \tag{5.168}$$

and $\bar{J}_s$ is the solution to $L\bar{J}_s = \bar{E}^{\text{inc}}$. A method-of-moments approximation $\bar{J}_s^N$ obtained using the identical functions for basis and testing (Galerkin's method) satisfies the conditions necessary for Equation (5.167) to be a stationary functional of the form of (5.159). Thus, if $\bar{\varepsilon}_J = \bar{J}_s - \bar{J}_s^N$, it follows that

$$\frac{1}{J_0}(\bar{E}^{\text{inc}}, \bar{J}_s) - \frac{1}{J_0}(\bar{E}^{\text{inc}}, \bar{J}_s^N) = \frac{1}{J_0}(L\bar{\varepsilon}_J, \bar{\varepsilon}_J) \tag{5.169}$$

The error in this approximation for $\bar{E}^s$ at $\bar{r}_0$ is $O(|\bar{\varepsilon}_J|^2)$. If the problem was discretized using a non-Galerkin selection of basis and testing functions, the error estimate would revert to $O(|\bar{\varepsilon}_J||\bar{\varepsilon}_E|)$, where $\bar{\varepsilon}_E$ denotes the error in the solution of the adjoint equation using the opposite functions for basis and testing.

To illustrate the nature of these estimates, Table 5.5 shows the error in the monostatic scattering cross section for the circular cylinder example used to generate Table 5.2, for various order splines employed as basis and testing functions. These data support the notion that the error is actually a function of the combined order of the basis and testing functions ($P + Q$ in the table). In other words, ε_E is comparable to ε_J for a given degree spline function, and the Galerkin solution ($P = Q$) is no more accurate than a non-Galerkin solution obtained with the same total $P + Q$.

TABLE 5.5 Percentage Error in Backscattered Far Field for TM Circular Cylinder with Circumference of 6λ as Function of Order of Splines Employed as Basis (P) and Testing (Q) Functions

	Percent Error				
Q	$P=1$	$P=2$	$P=3$	$P=4$	$P=5$
1	4.1×10^{-2}	9.2×10^{-5}	2.1×10^{-5}	2.4×10^{-6}	4.5×10^{-7}
2	9.2×10^{-5}	2.1×10^{-5}	2.4×10^{-6}	4.5×10^{-7}	
3	2.1×10^{-5}	2.4×10^{-6}	4.5×10^{-7}		
4	2.4×10^{-6}	4.5×10^{-7}			
5	4.5×10^{-7}				

Note: A total of 60 basis and testing functions of the indicated order were employed with equal-size cells to construct the matrix equation. A spline of order P has polynomial order $P-1$. After [25].

5.13 SUMMARY

This chapter has introduced several types of one-dimensional subsectional expansion sets, including polynomial spline functions and interpolative Lagrangian functions. The extension of these to the multidimensional and vector case will be considered in Chapter 9. In addition, the role of the basis and testing functions was discussed, and their impact on the accuracy of numerical solutions was explored. While the accuracy of a numerical result f^N appears to depend primarily on the ability of the basis functions to represent f, the testing functions play an equal role in the error associated with a secondary quantity $\langle f, h \rangle$.

The question of whether a particular discretization produces a numerical solution f^N that converges to the true solution f as $N \to \infty$ is fundamental. In order to introduce the reader to mathematical convergence proofs, a number of concepts from functional analysis have been reviewed. Sections 5.9–5.11 consider several approaches for establishing the convergence of f^N to f for integral equations with operators of the identity-plus-compact type. The specific arguments presented in these sections cannot easily be extended to more general operators. Our previous experience with integral equation formulations (Chapter 2) supports the notion that, if constructed with sufficient care, numerical solutions appear to converge under much more general conditions. Despite this observation, the authors are not aware of more general convergence proofs applicable to the specific integral operators arising in electromagnetic scattering.

A principal assumption required in order to relate the operators arising in electromagnetics with the "compact" or identity-plus-compact operators used within the convergence proofs is the boundedness of the inverse operators. Unfortunately, when applied to closed scatterers, surface integral equations such as the EFIE and MFIE may not always have bounded inverses. In fact, there are certain discrete frequencies where these equations do not produce unique solutions. Chapter 6 investigates this topic in detail and presents several alternate formulations that circumvent the difficulty.

REFERENCES

[1] D. H. Griffel, *Applied Functional Analysis*, New York: Wiley, 1981.

[2] I. Stakgold, *Green's Functions and Boundary Value Problems*, New York: Wiley, 1979.

[3] R. F. Harrington, *Field Computation by Moment Methods*, Malabar, FL: Krieger, 1982 Reprint.

[4] M. Becker, *The Principles and Applications of Variational Methods*, Cambridge, MA: MIT Press, 1964.

[5] B. A. Findlaysen, *The Method of Weighted Residuals and Variational Principles*, New York: Academic, 1972.

[6] L. V. Kantorovich and V. F. Krylov, *Approximate Methods of Higher Analysis*, New York: Wiley, 1964.

[7] L. V. Kantorovich and G. P. Akilov, *Functional Analysis*, Oxford: Pergamon, 1982.

[8] R. F. Harrington, "Origin and development of the method of moments for field computation," in *Applications of the Method of Moments to Electromagnetic Fields*, ed. B. Strait, St. Cloud, FL: SCEEE Press, 1981.

[9] S. G. Mikhlin, *Variational Methods in Mathematical Physics*, New York: Macmillan, 1964.

[10] O. C. Zienckiewicz and R. L. Taylor, *The Finite Element Method*, New York: McGraw-Hill, 1989.

[11] P. P. Silvester and R. L. Ferrari, *Finite Elements for Electrical Engineers*, Cambridge: Cambridge University Press, 1990.

[12] T. K. Sarkar, A. R. Djordjevic, and E. Arvas, "On the choice of expansion and weighting functions in the numerical solution of operator equations," *IEEE Trans. Antennas Propagat.*, vol. AP-33, pp. 988–996, Sept. 1985.

[13] D. Kahaner, C. Moler, and S. Nash, *Numerical Methods and Software*, Englewood Cliffs, NJ: Prentice-Hall, 1989.

[14] G. A. Thiele, "Wire antennas," in *Computer Techniques for Electromagnetics*, ed. R. Mittra, New York: Pergamon, 1973.

[15] E. K. Miller and F. J. Deadrick, "Some computational aspects of thin-wire modeling," in *Numerical and Asymptotic Techniques in Electromagnetics*, ed. R. Mittra, New York: Springer-Verlag, 1975.

[16] G. J. Burke and A. J. Poggio, *Numerical Electromagnetics Code (NEC)—Method of Moments*, Technical Document 116, Naval Ocean System Center, San Diego, Jan. 1981.

[17] B. Z. Steinberg and Y. Leviatan, "On the use of wavelet expansions in the method of moments," *IEEE Trans. Antennas Propagat.*, vol. 41, pp. 610–619, May 1993.

[18] G. Wang, "On the utilization of periodic wavelet expansions in the moment methods," *IEEE Trans. Microwave Theory Tech.*, vol. 43, pp. 2495–2498, Oct. 1995.

[19] J. C. Goswami, A. K. Chan, and C. K. Chui, "On solving first-kind integral equations using wavelets on a bounded interval," *IEEE Trans. Antennas Propagat.*, vol. 43, pp. 614–622, June 1995.

[20] R. Lee and A. C. Cangellaris, "A study of discretization error in the finite element approximation of wave solutions," *IEEE Trans. Antennas Propagat.*, vol. 40, pp. 542–549, May 1992.

[21] W. R. Scott, Jr., "Errors due to spatial discretization and numerical precision in the finite element method," *IEEE Trans. Antennas Propagat.*, vol. 42, pp. 1565–1570, Nov. 1994.

[22] A. F. Peterson and R. J. Baca, "Error in the finite element discretization of the scalar Helmholtz equation over electrically large regions," *IEEE Microwave Guided Wave Lett.*, vol. 1, pp. 219–222, Aug. 1991.

[23] R. Piessens, E. deDoncker-Kapenga, C. W. Uberhuber, and D. K. Kahaner, *QUADPACK: A Subroutine Package for Automatic Integration*, Berlin: Springer-Verlag, 1983.

[24] A. F. Peterson and R. E. Jorgenson, "Is Galerkin's method really better?" *Proceedings of the Sixth Annual Review of Progress in Applied Computational Electromagnetics*, Monterey, CA, The Applied Computational Electromagnetics Society, Monterey, CA, pp. 380–386, Mar. 1990.

[25] A. F. Peterson, D. R. Wilton, and R. E. Jorgenson, "Variational nature of Galerkin and non-Galerkin moment method solutions," *IEEE Trans. Antennas Propagat.*, vol. 44, pp. 500–503, Apr. 1996.

[26] D. Colton and R. Kress, *Integral Equation Methods in Scattering Theory*, New York: Wiley, 1983.

[27] K. E. Atkinson, *A Survey of Numerical Methods for the Solution of Fredholm Integral Equations of the Second Kind*, Philadelphia: SIAM, 1976.

[28] C. T. H. Baker, *The Numerical Treatment of Integral Equations*, Oxford: Clarendon, 1977.

[29] P. M. Anselone, *Collectively Compact Operator Approximation Theory*, Englewood Cliffs, NJ: Prentice-Hall, 1971.

[30] A. F. Peterson, "The 'interior resonance' problem associated with surface integral equations of electromagnetics: Numerical consequences and a survey of remedies," *Electromagnetics*, vol. 10, pp. 293–312, 1990.

[31] C. W. Steele, *Numerical Computation of Electric and Magnetic Fields*, New York: Van Nostrand Reinhold, 1987.

[32] G. Strang and G. J. Fix, *An Analysis of the Finite Element Method*, Englewood Cliffs, NJ: Prentice-Hall, 1973.

[33] M. H. Schultz, "L^2 error bounds for the Rayleigh-Ritz-Galerkin method," *SIAM J. Num. Anal.*, vol. 8, pp. 737–748, 1971.

[34] G. F. Roach, *Green's Functions*, Cambridge: Cambridge University Press, 1982.

[35] R. F. Harrington, *Time-Harmonic Electromagnetic Fields*, New York: McGraw-Hill, 1961, Chapter 7.

[36] J. R. Mautz, "Non-variational nature of the functional obtained by testing with a Dirac delta function," *Digest of the 1994 IEEE Antennas and Propagation International Symposium*, Seattle, WA, IEEE, NY, pp. 1169–1172, June 1994.

[37] J. H. Richmond, "On the variational aspects of the moment method," *IEEE Trans. Antennas Propagat.*, vol. 39, pp. 473–479, Apr. 1991.

[38] S. Wandzura, "Optimality of Galerkin method for scattering computations," *Microwave Opt. Technol. Lett.*, vol. 4, pp. 199–200, Apr. 1991.

PROBLEMS

P5.1 For real-valued functions defined on $0 \le x \le 1$, define inner products

$$\langle a, b\rangle_1 = \int_0^1 a(x)b(x)\,dx$$

$$\langle a, b\rangle_2 = \int_0^1 \sin(\pi x)a(x)b(x)\,dx$$

$$\langle a, b\rangle_3 = a\left(\tfrac{1}{2}\right)b\left(\tfrac{1}{2}\right)$$

and consider the three functions

$$f(x) = \sin(\pi x)$$

$$g(x) = \begin{cases} 2x & 0 \le x \le \frac{1}{2} \\ 2(1-x) & \frac{1}{2} \le x \le 1 \end{cases}$$

$$h(x) = 4x(1-x)$$

For each of the three inner products, evaluate the norm of each function, the metric $d(f, g)$ and the metric $d(f, h)$.

P5.2 (a) Given a set of functions $\{f_1, f_2, \ldots, f_N\}$, and an inner product $\langle f, g\rangle$, construct a set of orthogonal functions $\{g_1, g_2, \ldots, g_N\}$ from the set $\{f_n\}$. *Hint:* Choose $g_1 = f_1$, and let $g_2 = f_2 - \gamma_{21} g_1$, where γ_{21} is selected so that $\langle g_1, g_2\rangle = 0$. Continue this procedure and find a general expression for the coefficient γ_{ij}. This process is known as *Gram–Schmidt orthogonalization.*

(b) Modify the procedure in order to produce a set of orthonormal functions.

(c) Finally, use the procedure to produce a set of orthonormal functions from the set $\{1, x, x^2, x^3, x^4\}$, employing the inner product

$$\langle a, b\rangle = \int_0^1 a(x)b(x)\,dx$$

Sketch the resulting orthonormal functions.

P5.3 Consider the differential equation $Lf = g$, where

$$Lf = \frac{d^2 f}{dx^2} + \frac{df}{dx}$$

and

$$g = (1 - x)\cos x - (2 + x)\sin x$$

The equation is defined on the interval $(0 \le x \le \frac{1}{2}\pi)$ and is subject to the boundary conditions $f(0) = 0$, $f(\frac{1}{2}\pi) = 0$. The solution is $f(x) = x\cos x$.
Define an inner product

$$\langle a, b\rangle = \int_0^{\pi/2} a(x)b(x)\,dx$$

and consider the complete, orthonormal basis on $(0, \frac{1}{2}\pi)$ given by

$$B_n(x) = \sqrt{\frac{4}{\pi}}\sin(2nx) \qquad n = 1, 2, \ldots$$

(a) Calculate numerical values for the three coefficients α_1, α_2, and α_3 that give the "best" representation according to (5.9), that is,

$$\alpha_n = \int_0^{\pi/2} \sqrt{\frac{4}{\pi}}\sin(2nx)x\cos x\,dx$$

(b) Construct a method-of-moments solution using the identical functions B_1, B_2, and B_3 as basis and testing functions. In other words, let

$$f^N = \sum_{n=1}^{3} \beta_n \sqrt{\frac{4}{\pi}}\sin(2nx)$$

and construct the 3×3 system $\mathbf{L}\beta = \mathbf{b}$, where

$$L_{mn} = \langle B_m, LB_n\rangle$$

and

$$b_m = \langle B_m, g\rangle$$

Compare the numerical values obtained for β_1, β_2, and β_3 with those obtained in part (a) for α_1, α_2, and α_3. Are they the same? Discuss the implications.

P5.4 Apply the method-of-moments procedure to obtain approximate solutions for the integral equation (D. R. Wilton and S. Govind, "Incorporation of edge condition in moment method solutions," *IEEE Trans. Antennas Propagat.*, vol. AP-25, pp. 845–850, Nov. 1977).

$$\int_{-1}^{1} f(x') \ln|x - x'|\, dx' = -\pi \ln 2$$

Use pulse basis functions and Dirac delta testing functions as defined in Section 5.3. Compare your numerical results with the solution

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

P5.5 Consider the equation

$$\frac{d^2 f}{dx^2} + k^2 f = g$$

on some domain $0 < x < M\Delta$, subject to boundary conditions $f(0) = f_L$ and $f(M\Delta) = f_R$.

(a) Construct a second-order finite-difference discretization using cells of dimension Δ and the central-difference formula

$$\frac{d^2 f}{dx^2} \cong \frac{f_{n+1} - 2f_n + f_{n-1}}{\Delta^2}$$

to replace the second derivative. Identify the entries of the tridiagonal finite-difference matrix.

(b) Construct a method-of-moments discretization using the subsectional triangles in (5.27) to represent $f(x)$ and pulse testing functions to enforce the equation. The basis functions are to straddle two cells of dimension Δ, while the testing functions are defined between the center of adjacent cells. How does the (mn)th entry of the resulting matrix compare with the entries obtained in part (a)? Discuss the implications.

P5.6 Use the three quadratic Lagrangian functions

$$\phi_1(x) = \frac{1}{2}\frac{x}{\Delta}\left(\frac{x}{\Delta} - 1\right)$$

$$\phi_2(x) = \left(1 + \frac{x}{\Delta}\right)\left(1 - \frac{x}{\Delta}\right)$$

$$\phi_3(x) = \frac{1}{2}\frac{x}{\Delta}\left(\frac{x}{\Delta} + 1\right)$$

to approximate the cubic polynomial

$$f(x) = a + bx + cx^2 + dx^3$$

on the interval $-\Delta \leq x \leq \Delta$. In other words, construct

$$f_{\text{ap}}(x) = f(-\Delta)\phi_1(x) + f(0)\phi_2(x) + f(\Delta)\phi_3(x)$$

(a) Show that the error in this approximation is

$$f - f_{\text{ap}} = d(x^3 - \Delta^2 x)$$

(b) Plot the error function over the interval $-\Delta \leq x \leq \Delta$, and determine the peak error. At what value of x does the peak error occur?

(c) Identify the integer exponent p that best characterizes the interpolation error as $O(\Delta^p)$ as $\Delta \to 0$.

P5.7 Repeat Prob. P5.6 using the cubic Hermitian functions defined in (5.35)–(5.38) to represent

$$f(x) = a + bx + cx^2 + dx^3 + ex^4$$

on the interval $-\Delta \leq x \leq \Delta$.

P5.8 (a) Show that the central finite-difference formula (Prob. P5.5), if applied to the one-dimensional scalar Helmholtz equation in (5.48), produces

$$2E_m - E_{m-1} - E_{m+1} - k^2\Delta^2 E_m = 0$$

(b) Carry out a dispersion analysis similar to that of Section 5.5 for this discrete equation. Show that a solution of the form of (5.50) satisfies the equation as long as

$$\beta = \frac{1}{\Delta}\cos^{-1}\left(1 - \frac{(k\Delta)^2}{2}\right)$$

(c) Tabulate the resulting error as a function of Δ, comparing your results with those in Table 5.1. Is the finite-difference approach more accurate than the first-order finite-element discretization?

P5.9 Quadratic interpolation functions can be defined in one dimension so that three functions

$$\Phi_1(x) = \frac{x(x-\Delta)}{2\Delta^2}$$

$$\Phi_2(x) = 1 - \left(\frac{x}{\Delta}\right)^2$$

$$\Phi_3(x) = \frac{x(x+\Delta)}{2\Delta^2}$$

overlap a cell spanning the interval $-\Delta < x < \Delta$, with Φ_2 interpolative at the center of the cell. Show that element matrices associated with a discretization of the one-dimensional scalar Helmholtz equation in (5.48) have the form

$$\left[\int \frac{d\Phi_i}{dx}\frac{d\Phi_j}{dx}\,dx\right] = \begin{bmatrix} \frac{7}{6\Delta} & \frac{-8}{6\Delta} & \frac{1}{6\Delta} \\ \frac{-8}{6\Delta} & \frac{16}{6\Delta} & \frac{-8}{6\Delta} \\ \frac{1}{6\Delta} & \frac{-8}{6\Delta} & \frac{7}{6\Delta} \end{bmatrix}$$

and

$$\left[\int \Phi_i \Phi_j\,dx\right] = \begin{bmatrix} \frac{4\Delta}{15} & \frac{2\Delta}{15} & \frac{-\Delta}{15} \\ \frac{2\Delta}{15} & \frac{16\Delta}{15} & \frac{2\Delta}{15} \\ \frac{-\Delta}{15} & \frac{2\Delta}{15} & \frac{4\Delta}{15} \end{bmatrix}$$

P5.10 (a) Extend the dispersion analysis of Section 5.5 to the quadratic Lagrangian interpolation functions in Prob. P5.9 in order to obtain a numerical phase constant

$$\beta = \frac{1}{2\Delta}\cos^{-1}\left(\frac{15 - 26(k\Delta)^2 + 3(k\Delta)^4}{15 + 4(k\Delta)^2 + (k\Delta)^4}\right)$$

(b) Identify the region where β is real valued.

(c) Tabulate the error per wavelength as a function of Δ, comparing your result with Table 5.1. Identify the integer exponent q that best fits the error, assuming that the error decreases as $O(\Delta^q)$ as $\Delta \to 0$. Is this a superconvergent result?

P5.11 Review Prob. P3.21 in order to derive the analytical eigenvalues presented in Equations (5.82), (5.87), and (5.91).

P5.12 Obtain an analytical expression for the eigenvalues of the TM MFIE (Prob. P2.9) applied to a circular conducting cylinder of radius a.

P5.13 For a circular conducting cylinder of fixed radius, show that as the number of basis functions N is increased, the ratio of the largest to smallest eigenvalue in the method-of-moments matrix behaves as $O(1)$ for the TE MFIE, $O(1/\Delta)$ for the TE EFIE, and $O(1/\Delta)$ for the TM EFIE. Assume that the cell size is given by $\Delta = ka/N$ and that no eigenvalue is identically zero for this value of ka. Use the analytical expressions in (5.82), (5.87), and (5.91) to approximate the first N matrix eigenvalues. Finally, assume that N is large enough to justify use of the asymptotic approximations

$$J_n(z) \approx \sqrt{\frac{1}{2\pi n}}\left(\frac{ez}{2n}\right)^n \qquad \text{as } n \to \infty$$

$$H_n^{(2)}(z) \approx j\sqrt{\frac{2}{\pi n}}\left(\frac{2n}{ez}\right)^n \qquad \text{as } n \to \infty$$

for the Bessel functions.

P5.14 Consider the two-dimensional scalar Helmholtz equation

$$L\Psi = -\nabla^2\Psi - k^2\Psi = g$$

on the rectangular domain $0 < x < a, 0 < y < b$, subject to a homogeneous Dirichlet boundary condition.

(a) Show that the eigenvalues of L are given by

$$\lambda_{mn} = \left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2 - k^2$$

(b) Assuming that this Helmholtz equation is discretized using a uniform finite-difference or finite-element grid, with cells of dimension $\Delta \times \Delta$, and that the matrix eigenvalues are well approximated by the λ_{mn} in part (a), show that the ratio of the largest to smallest matrix eigenvalue behaves as $O(\Delta^{-2})$ as $\Delta \to 0$.

P5.15 Consider the equation $Lf = g$, where L has eigenvalues $\{\lambda_n\}$ and eigenfunctions $\{e_n\}$ that form a complete, orthonormal set. The excitation g can be expressed as

$$g = \sum_n \langle g, e_n \rangle e_n$$

The solution can be sought in the form

$$f = \sum_n \alpha_n e_n$$

and determined by substitution, since

$$Lf = \sum_n \alpha_n L e_n = \sum_n \alpha_n \lambda_n e_n$$

Therefore, the solution can be written formally as

$$f = \sum_n \frac{\langle g, e_n \rangle}{\lambda_n} e_n$$

However, if L is a compact operator such as the TM EFIE in Equation (5.79), $\lambda_n \to 0$ as $n \to \infty$. Consequently, the summation for f may be divergent. For the TM EFIE applied to a circular cylinder, determine the asymptotic order of the nth term in this summation when $g(\phi)$ is (a) a delta function, (b) a subsectional triangle function with support limited to a 90° portion of the circle, and (c) a constant function over the entire circle. Use the eigenfunctions and eigenvalues presented in (5.81)–(5.84). What constraint on g is necessary to ensure convergence?