ISSUES RELATED TO FINITE ELEMENT TECHNIQUES FOR TWO-DIMENSIONAL TRANSMISSION STRUCTURES

by

Charles F. Bunting

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

Electrical Engineering

APPROVED:

William A. Davis

Dr. William A. Davis, Chair

Gary S Brown

Dr. G. S. Brown

I F Besieris

Dr. I. M. Besieris

April, 1992

Blacksburg, Virginia

c.2

LD
5655
V855
1992
S867
c.2

ISSUES RELATED TO FINITE ELEMENT TECHNIQUES FOR TWO-DIMENSIONAL TRANSMISSION STRUCTURES

by
Charles F. Bunting

Committee Chair: Dr. William A. Davis
Electrical Engineering

(ABSTRACT)

This thesis addresses many issues associated with finite element techniques, concentrating on ideas that are not often emphasized in the literature. Pulling together the ideas of mesh generation, sparse storage solution techniques, and functional development, in a single volume, this work provides basic tools for implementation of finite element techniques for both static and dynamic problems in electromagnetics.

An automatic mesh generation scheme is developed by forming a Delaunay triangulation. A storage technique will be presented and used in conjunction with a conjugate gradient method to solve linear systems of equations. Application to electromagnetic problems will be demonstrated in the static, quasi-static, and full-field regimes. Laplace's equation is solved for various transmission line geometries to obtain capacitance and characteristic impedance. A finite element model that is a full field expression of Maxwell's equations is developed through a novel variational formulation involving the method of Lagrange multipliers, with attention given to the physical basis of the obtained functional. This model is then applied to the problem of determining the propagation constant of a waveguide.

ACKNOWLEDGEMENTS

I want to thank first of all my savior, the Lord Jesus Christ, who is my strength and life.

I want to express my deepest appreciation to my wife Lesly. Her love, vision, and sacrifice have made my education both possible and enjoyable.

Special thanks are due to my advisor, Dr. Davis, whose insight and patience were invaluable in leading me to dig deeper than I thought possible. I wish to thank Dr. Brown and Dr. Besieris for serving on my graduate committee, for their many helpful suggestions and asking the tough questions. Thanks are also due to Dr. Brown for the use of the Computational Electromagnetics Facility.

Thanks to the guys in the Theoretical Electromagnetics Laboratory: Mike Newkirk, Paul Werntz, and Bill Smith, for making EM "fun".

Thanks to Loretta Estes for her guidance and sense of humor.

Table of Contents

INTRODUCTION 1

THE FINITE ELEMENT METHOD 5

2.1 Introduction 5

2.2 Method of Weighted Residuals 6

2.3 Variational Methods 9

2.4 The Finite Element Method 15

2.4.1 Basic Concepts 16

2.4.2 Basis Functions 17

2.4.3 Shape Functions for Triangular Elements 19

2.4.4 Matrix Representation for Finite Element Problems 25

MESH GENERATION 28

3.1 Introduction to Mesh Generation 28

3.2 Delaunay Triangulation 31

FINITE ELEMENT TECHNIQUES APPLIED TO LAPLACE'S EQUATION 44

4.1 Introduction 44

4.2 Variational Formulation 44

4.3 Assembly of Element Matrices 49

4.4 Solution of Sparse Matrices 53

4.5 Results for Laplace's Equation 60

FULL-FIELD VARIATIONAL FORM 68
5.1 Introduction 68
5.2 Functional Development 70
5.3 Rectangular Waveguide Problem 77
5.4 Finite Element Implementation of Full Field Variational Form 87
5.5 Results and Discussion 95

SUMMARY AND CONCLUSIONS 101

REFERENCES 103

APPENDIX 1. FULL FIELD ELEMENT MATRIX COMPUTATIONS 111

APPENDIX 2. ALTERNATIVE EIGENVALUE FORMULATION 117

VITA 121

List of Figures

Figure 2.1. Pyramidal basis functions 18

Figure 2.2. Domain discretization 20

Figure 2.3. Element geometry in local coordinates 21

Figure 2.4. Linear shape functions 24

Figure 3.1. The Delaunay Triangulation/ Dirichlet Tessellation 33

Figure 3.2. The circle-check for the Delaunay Triangulation 35

Figure 3.3. Data structure for Watson's Algorithm 37

Figure 3.4. Cavendish point-in-plane generation 40

Figure 3.5. Laplace's equation for square trough with diagonal from upper-right to lower-left 42

Figure 3.6. Laplace's equation for square trough with diagonal from lower-left to upper-right 43

Figure 4.1. Element assembly 50

Figure 4.2. Coaxial geometry 62

Figure 4.3. Sample mesh 63

Figure 4.4. Equipotential contours for mesh of Figure 4.3 64

Figure 4.5. Stripline geometry 65

Figure 5.1. Waveguide geometry 81

Figure 5.2. Sample waveguide discretization 96

Figure 5.3. Error as a function of the initial guess 98

Chapter 1

Introduction

The analysis of two-dimensional transmission systems has been of interest for a number of years. Numerical analysis has taken on increased importance as the power of the computer has become readily available to design engineers while the increased frequency of operation of microwave devices has motivated the need for more thorough numerical techniques. Microstrip and tri-plate stripline transmission lines, used in many devices as the basic transmission line [Gupta,1979], require the designer to utilize appropriate circuit models to insure accurate results. For a low loss model, the results of static and quasi-static analysis are adequate to characterize the system [Silvester,1990; Itoh,1989; Hammond,1981; Pantic,1986].

A problem of interest in studying the propagation characteristics of transmission lines and waveguides is that of conductor loss and its effects on the propagation constant γ of the line or waveguide [Hayata,1988; Narayanan,1991]. This problem can be approached in a perturbational sense for small losses; however, when the losses are large the phase constant cannot be adequately calculated since the field is no longer only slightly altered by the loss and the tangential component of the electric field is no longer small [Harrington,1961]. Integral techniques have been formulated for this problem [Bucca,1989] with good success, but lack the flexibility to easily accommodate multiple dielectric regions.

A method that has become popular over the past twenty years is the finite element method. The finite element method was initially introduced in the area of structural mechanics [Clough,1960; Finlayson,1972], where the basis in variational principles [Fox,1950; Weinstock,1974] provided a solid mathematical foundation. The finite element method is essentially the Rayleigh-Ritz method [Reddy,1984; Lapidus,1982; Johnson,1982] with the domain of interest divided into subdomains or elements with interpolating functions defined over the subdomains or elements. Finite element techniques have been applied to both guided wave problems [Koshiba,1992; Lee,1991; Hayata,1988; Paulsen,1988] and electromagnetic scattering problems [McCartin,1990]. Results obtained for guided wave problems commonly require additional constraints, such as penalty methods, to achieve the desired solution.

This thesis will develop the mathematical basis of the finite element method and discuss the more commonly applied method of weighted residuals. The application to electromagnetic problems will be demonstrated in the static, quasi-static, and full-field regimes. A basic limitation to the use of finite element techniques is that of the discretization of the domain into suitable subdomains known as elements [Cendes,1985]. An automatic mesh generation scheme is developed where the domain is treated as a convex hull over which triangular elements are placed in a pseudo-random fashion and forming a Delaunay triangulation [Watson,1981; Cavendish,1974]. Laplace's equation will be solved using the finite element method to illustrate the application to static and quasi-static field problems. Another focus of this paper will be to develop a finite element model that is a full field expression of Maxwell's equations through a novel variational formulation involving the method of Lagrange multipliers. This

full-field approach will be used to attempt a prediction of the propagation constant of a rectangular waveguide. An important aspect of the use of finite element techniques is taking advantage of the sparsity of the matrices obtained in the numerical process. A storage technique will be presented and used in conjunction with a conjugate gradient method to solve linear systems of equations.

This thesis addresses many issues associated with finite element techniques that are often not emphasized in the literature. By pulling together the ideas of mesh generation, sparse storage solution techniques, and functional development in a single volume, the motivated reader, with this background, will be enabled to pursue these ideas further for specific applications.

The thesis is organized as follows: Chapter 2 will provide a background for the finite element method as related to the method of weighted residuals with its foundations in variational methods. Chapter 3 will introduce and develop automatic mesh generation concepts. A mesh generation scheme will be presented for use in later chapters. Chapter 4 will present and discuss the results for Laplace's equation to obtain the capacitance of a coaxial line and the characteristic impedance of a tri-plate stripline. The sparse storage and conjugate gradient techniques are developed for solving the resultant equations. These results provide useful static and quasi-static representations for many electromagnetic applications requiring a two-dimensional transmission analysis. In chapter 5, a full field functional will be developed using variational concepts and energy methods. This development will emphasize the physical basis for the minimization process associated with the variational problem. The full-field functional will be applied to the problem of obtaining the propagation constant

of a canonical problem in a Rayleigh-Ritz context with a discussion of the difficulty arising from the use of certain functional forms. The finite element implementation will be developed and also applied to a canonical problem with a discussion of spurious solutions for the eigenvalue problem.

Chapter 2

The Finite Element Method

2.1 INTRODUCTION

The finite element method has its origins in structural mechanics [Clough, 1960] for the numerical solution of partial differential equations governing the elasticity and bending of beams and plates. Since a variational form was readily obtainable in structural engineering, this led to variational methods in the formulation of the functional to be used as a minimizing principle in the solution of these equations. Finite element techniques were extended throughout the 60's to include many boundary value problems in physics and engineering. The purpose of this chapter is to address details concerning the finite element method and provide a comparison to commonly applied numerical methods. Consideration of related numerical techniques commonly applied in electromagnetics provides insight into the unique role that finite elements plays in the analysis of microwave structures.

The method of weighted residuals is a related numerical technique which seeks the solution in a weighted integral sense. Finite element techniques can be applied in a weighted residual format when a variational principle is unavailable as in the Navier-Stokes fluid flow equations. Some discussion of weighted residual methods will help in placing the finite element method in proper context.

2.2 METHOD OF WEIGHTED RESIDUALS

The need for numerical techniques in electromagnetics arises whenever direct analytical or related approximation techniques cannot, without difficulty, be employed. The determination of the characteristic impedance of a tri-plate stripline, or propagation characteristics of arbitrarily shaped waveguides are good examples of problems that are not amenable to analytical results. The method of weighted residuals (MWR) is one of the earliest developed techniques to allow for a numerical approach to solve analytically intractable problems [Galerkin,1915]. This approach consists of rewriting the differential equation (or integral equation) for which the solution is desired in a matrix form.

Consider the problem

$$Lu = f \tag{2.1}$$

where the linear operator L on a normed linear vector space S can be either an integral or differential operator with domain $\mathfrak{D}$ and range $\mathfrak{R}$, subject to prescribed boundary conditions. Evaluation of this vector space problem requires the notion of length of a vector and the distance between two vectors, thus defining a metric [Reddy,1986]. In a normed linear space this metric is defined in terms of a norm which is typically defined as an inner product. The right-hand side, f, is the forcing function or excitation, and u is the unknown to be determined. The inner product is denoted by the brackets $\langle\ ,\ \rangle$, typically referring to integration over the domain, and obeys the following:

$$\langle f, g \rangle = \langle g, f \rangle^*$$

$$\langle \alpha f + \beta g, h \rangle = \alpha \langle f, h \rangle + \beta \langle g, h \rangle \tag{2.2}$$

$$\begin{aligned} \langle f, f \rangle &> 0 \ , \ f \neq 0 \\ &= 0 \ , \ f = 0 \quad . \end{aligned}$$

An approximation to the true solution u can be written in a series form as $\hat{u}$ where

$$\hat{u} = \sum_i a_i \, u_i(\cdot) \ , \tag{2.3}$$

where the functions $u_i(\cdot)$ are known with the coefficients a_i to be determined. The functions $u_i(\cdot)$ are usually chosen to satisfy the boundary conditions of (2.1). Referring to Eq. (2.1), it is apparent that the residual can be defined as $R(\cdot) = L\hat{u} - f$, for a series terminated after N terms, thus forming a finite-dimensional subspace S_N of the original space S. The method of weighted residuals hinges on the concept of minimizing this residual in some sense. This is accomplished by forming the inner product with a weighting function, $w_j(\cdot)$, that typically satisfies the boundary conditions. This results in the expression,

$$\langle R(\cdot), w_j(\cdot) \rangle = \langle L\hat{u} - f, w_j(\cdot) \rangle = 0 \quad j = 1, 2 \ , \ \ldots \ , N \ , \tag{2.4}$$

(where, keep in mind that, $\hat{u}$ is summed over i) . This is expressed

$$\sum_{i=1}^{N} \langle L u_i(\cdot), w_j(\cdot) \rangle \, a_i = \langle f, w_j(\cdot) \rangle \quad j = 1,2, \ \ldots \ , N \ , \tag{2.5}$$

where Eq. (2.5) can be represented as the matrix problem

$$Ax = b, \tag{2.6}$$

with A given as a known $N \times N$ matrix $\left\langle Lu_i(\cdot), w_j(\cdot) \right\rangle$, x as the column vector of unknown coefficients a_i, and b given as the known right-hand side $\left\langle f, w_j(\cdot) \right\rangle$. The first term in Eq. (2.4) is referred as the weighted residual, hence the name method of weighted residuals. The method of moments (MOM) is equivalent to the description of the method of weighted residuals, though the operator is typically an integral operator. For the remainder of this thesis, the operator will be considered a differential operator in the context of solving boundary value problems.

The choice of weighting functions is important in determining the nature of the solution obtained. The simplest choice is that of $w_j(\cdot) = \delta(\cdot)$ which corresponds to the point-matching or collocation method. This choice is often used because the integral is trivial although the solution is sensitive to placement of these weighting functions. A commonly used weighting function used is that of the weighting function equal to the expansion function or $u_i(\cdot) = w_j(\cdot)$ resulting in a more difficult integration, but a symmetric matrix A. This technique is often called the Galerkin procedure and is closely related to the variational method [Finlayson,1972]. Galerkin [1915] developed the procedure as an alternative to Rayleigh-Ritz procedures when an appropriate functional is not readily available. Finlayson also observes that the Galerkin method forces the residual equal to zero by making it orthogonal to each member of the complete set in the limit as N approaches infinity.

2.3 VARIATIONAL METHODS

The discussion of the method of weighted residuals has already hinted at some aspects of variational concepts as being closely related and many authors blur the distinction between them and include the method of weighted residuals as a variational method. Wexler [1969] presents a very clear account of variational methods without considering weighted residuals. Often a point of semantics, variational principles, while seeking an extremum, are unique in the representation of the boundary value problem as is to be shown. Variational *principles* as distinguished from variational *methods* may be considered to represent a subset of weighted residuals due to the interpretation of the minimizing principle [Reddy,1986] for those problems which can be cast in weighted residual form.

The goal of obtaining a variational principle is that a more complete picture of the physical process can be achieved for *some* problems. Another benefit is that a variational principle may result in stationarity or minimization of some specific quantity of interest. Examples of specific quantities include resonant frequency [Harrington,1961], radiation resistance and characteristic impedance [Hammond,1981], radar cross section [Stakgold,1967], and cutoff frequency [Itoh,1989]. Itoh also indicates that variational principles are also useful in the derivation of perturbation formulas. With this as motivation, the variational principles developed in this section will serve as useful tools in analysis.

The calculus of variations is essential in discussing the concept of variational principles and a brief introduction will assist. A more thorough

discussion is to be found in Weinstock [1952], Fox [1950], or Stakgold [1968]. This summary is based on the work of Finlayson [1972] and Reddy [1986]. The solution to Eq. (2.1) can be recast in such a way as to obtain an extremum or stationary value of a related functional $F(u)$. The concept of extrema of a function is related to the derivative behavior, namely a vanishing first derivative and non-negative second derivative. Recall that a function transforms each element of one set of real numbers into one element of another, representing a one-to-one mapping in real numbers $\mathbb{R}$ (or $\mathbb{C}$). A linear functional maps the members of the vector space S into $\mathbb{R}$ (or $\mathbb{C}$). Thus a functional can be thought of as a correspondence of a function in some class and $\mathbb{R}$ (or $\mathbb{C}$), whose domain is formed by a space of admissible functions. So to claim stationarity, the following must be true:

$$\delta F(u) = 0, \tag{2.7}$$

where δ is an operator that is commonly called the first variation. The first variation is considered a variational operator which acts as a total differential operator with respect to the independent variables. Specifically

$$\delta F(u) = \epsilon \left\{ \left[\frac{d}{d\epsilon} F(u + \epsilon\eta) \right]_{\epsilon = 0} \right\} \tag{2.8}$$

where ϵ is an arbitrarily small real number and η is an arbitrary function [Reddy,1986; Stakgold,1968]. For the functional to be stationary about some solution u then the functional must be stationary about small perturbations in the solution represented by $\epsilon\eta$ [Hazel, 1972]. Minimization requires the second

variation be non-negative, or $\delta^2 F(u) \geq 0$ which in general is much harder to achieve.

Avoiding questions of existence and uniqueness for this discussion, consider a symmetric positive operator defined on a linear space $\mathfrak{D}_L$ in S, S being a complex Hilbert space [Stakgold,1968]. Given a slightly modified statement of (2.1) we have

$$Lu = f \ , \quad u \in \mathfrak{D}_L \ . \tag{2.9}$$

Consider an arbitrary admissible basis function $v(\cdot) \neq u(\cdot)$, and $v(\cdot) \in \mathfrak{D}_L$, satisfying the boundary conditions implied by (2.9). The following *minimum* theorem is stated without proof. Given a functional $F(v)$, with inner product $\langle \ , \ \rangle$ as defined in (2.2),

$$F(v) = \langle Lv, v \rangle - \langle f, v \rangle - \langle v, f \rangle \ , \tag{2.10}$$

then if (2.9) has a solution u , then u is the one and only element in $\mathfrak{D}_L$ which minimizes F for the operator, L, positive and symmetric. This can be restated that if among the elements in $\mathfrak{D}_L$ there exists one which minimizes F, then it is the solution of (2.9) [Stakgold,1968]. Which is to say for $u \neq v$, $F(u) < F(v)$ for all $u, v \in \mathfrak{D}_L$.

There are two useful consequences of a vanishing first variation. The first result is that the Euler equation is satisfied. The Euler equation is often discussed in variational calculus, and in this context corresponds to the statement of the operator equation $Lu = f$. A further consequence is that the

natural boundary conditions are satisfied. The natural boundary conditions correspond to those conditions on (2.9) that need not be imposed on the set of admissible functions in (2.10). Natural boundary conditions are also known as secondary boundary conditions. Boundary conditions required in (2.10) are known as essential or primary boundary conditions [Reddy,1986].

Typically the Neumann-type boundary conditions are formulated in terms of natural boundary conditions, but Hazel [1972] has extended these concepts to include specific Dirichlet boundary conditions. Natural boundary conditions often correspond to a physically meaningful quantity. In electrostatics the natural boundary condition corresponds to the flux density D_n; in magnetostatics, the magnetic flux B_n; in flow through porous media, the seepage [Reddy,1984]; and in electromagnetics it can represent the electric surface current $\vec{J}$ [Paulsen,1988].

To determine the Euler equation and the natural boundary conditions of a functional, find $\delta F(u)$ according to (2.8) and apply Green's identity to obtain a surface integral and line integral for problems in R^2 or a volume integral and surface integral in R^3. The Euler equation will appear as multiplying the function, η, in the integral of highest dimension, and the remaining integral will define the natural boundary condition. Wexler [1969] gives the following brief example for $L = -\nabla^2$, and the boundary condition, $u(s) = g(s)$. A real inner product (2.10) yields the following functional form:

$$F(u) = \int_{\Omega} \Big(\mid \nabla u \mid^2 \; - \; 2fu) \; d\Omega \; . \tag{2.11}$$

Finding $\delta F/\epsilon$ with no restrictions on η and using Green's identity, we obtain

$$\int_{\Omega} \eta \left(\nabla^2 u_0 + f \right) d\Omega \; - \int_{s} \eta \, \frac{\partial u_0}{\partial n} \, ds = 0. \tag{2.12}$$

Since η is arbitrary, Eq. (2.12) is only satisfied if $-\nabla^2 u_0 = f$, corresponding to the Euler equation, and $\dfrac{\partial u_0}{\partial n} = 0$, corresponding to the natural boundary condition[Hazel,1972]. This is to say that the expansion function used to approximate u_0 naturally satisfies the Neumann boundary condition.

The expression $\langle Lv, v \rangle$ may be reduced to an expression involving lower order derivatives via integration by parts suggesting that the functions sought for solution to (2.10) may have relaxed differentiability requirements among the functions of $\mathfrak{D}_L$ than originally required by the differential operator. The differentiability requirements in this discussion are those of a function theoretic point of view. The concept of generalized functions or distributions is useful, though not pursued in this thesis, as a means of extending the admissible functions for a particular linear vector space. An example of this will be examined in detail in Chapter 4 for the Laplace equation where the differential equation specifies $v \in C^2(\,\cdot\,)$ and the functional representation requires $v \in C^1(\,\cdot\,)$, where C^n represents continuous derivatives to order n.

The numerical implementation of a variational principle occurs in the form of the Rayleigh-Ritz or simply Ritz method. Consider an approximation to the true solution u as $\hat{u} = \sum_j a_j \, v_j(\,\cdot\,)$, where in this context the unknown coefficients are known as *Ritz-parameters* [Itoh,1989], and the expansion functions $v_j(\,\cdot\,)$ are typically required to satisfy the essential boundary conditions of the problem. This approximation corresponds to seeking $\hat{u}$ in the N-dimensional subspace S_N generated by the finite set $\{v_1, v_2, \ldots, v_N\}$. This approximation to u is then applied in Eq. (2.9) resulting in N unknowns in one

functional. The N needed equations are obtained by taking the variation of $F(\hat{u})$ with respect to the Ritz-parameters and setting this equal to zero, that is,

$$\frac{\partial F(\hat{u})}{\partial a_i} = 0 \ , \ i = 1,2, \ldots , N, \tag{2.13}$$

resulting in a matrix expression for the coefficients. Note that since the expansion functions are required to satisfy the essential boundary conditions the resulting matrix will be a full matrix for expansion functions that span the entire domain. Except for simple geometrical shapes, difficulties arise in specifying these expansion functions so that the essential boundary conditions are satisfied. These expansion functions are commonly sine and cosine functions, or other admissible functions exactly satisfying the boundary conditions.

Reddy [1986] writes that "A generalization of the Ritz method, which is applicable to positive definite (bounded-below) operators that admit a functional formulation or operator equations that permit a weak formulation which includes the natural boundary conditions to general operator equations is known as the Bubnov-Galerkin method." The Bubnov-Galerkin method is commonly referred to as simply the Galerkin method [Galerkin,1915]. As discussed previously this method determines an approximate solution of $Lu = f$, by seeking unknown coefficients that make a residual orthogonal to all expansion functions. If the boundary conditions can be included in the variational statement of the problem the expansion functions under consideration are not restricted to satisfy these boundary conditions. Thus the Galerkin method may be considered equivalent to the Ritz method, and identical results are obtained for either approach. To qualify this, however, Finlayson demonstrates that there is always a Galerkin

method that corresponds to a variational principle. There are several means by which a Galerkin method may be implemented only one of which will give this correspondence.

Reddy [1986] and Finlayson [1972] regard variational methods as a subset of the method of weighted residuals (and in particular Galerkin's method) as not all boundary value problems have been cast in a variational form. Finlayson set forth the ability to verify a variational form as obtaining the Euler equation by integrating by parts and applying Green's identity. The inability to find a variational form occurs for many problems in engineering such as in fluid mechanics where the Navier-Stokes equation, as well as some problems in heat and mass transfer. For problems when a variational statement is possible, it may be limited to stationarity, unable to demonstrate either maximum or minimum. It is this lack of generality that prompts the need for weighted residuals when no variational principle has been established.

2.4 THE FINITE ELEMENT METHOD

The finite element method (FEM) is a means by which the solution to a differential equation is approximated in accordance with a variational principle. Unlike classical variational methods where specific requirements are placed upon the admissible functions and only simple geometric configurations provide adequate results, the FEM uses piecewise polynomial approximation functions based on interpolation theory. Similar to classical variational methods, the FEM seeks the projection of the finite subspace S_N on the infinite-dimensional subspace S. Classical methods define expansion functions that are defined over the entire domain $\mathfrak{D}_L$, satisfy essential boundary conditions, and use a finite

series expression to approximate the solution. The dimension N in classical methods is determined by the number of terms retained in the series representation. The finite element method also generates a finite dimensional subspace, but does so by partitioning the solution domain into geometrically simple shapes, sometimes referred to as a mesh. This discretization allows for modeling of arbitrarily complex boundary shapes depending upon the choice of shapes used to form the mesh.

2.4.1 BASIC CONCEPTS

To construct S_N, the solution domain is discretized by choosing node points corresponding to the points in the domain at which the solution is to be computed. Shape functions connecting the nodes are chosen to be defined over a single finite element and are used to define basis functions. The basis functions are formed by enforcing the continuity of adjacent element shape functions. The dimension of S_N corresponds to the number of *nodes* chosen. Other nodal distributions over an element may be determined by the geometry of the element and the degree of the shape function desired to approximate the solution. Lapidus [1982], Reddy [1984], and Silvester [1990] discuss the various shape functions in significant detail. A shape function in an element is equal to unity over one node point and is zero over all others within the element.

The spatial discretization, as to be discussed in chapter 3, uses linear triangular elements. This choice represents one of the simplest forms of finite elements. A benefit of this choice is that the integration is exact and does not require a numerical algorithm as does the rectangular element. Another benefit is that triangular elements easily model complex geometrical shapes whereas

rectangular elements do not. The linear triangular element has nodes located at the vertices and is the minimum acceptable shape function for the second order operators considered in this paper. (Recall that the degree of expansion function required is reduced by one through integration by parts). This restriction is only necessary in a function theoretical sense, implied throughout this thesis. In a distributional sense, the continuity requirement can be relaxed. Higher order shape functions can be defined to enforce a desired derivative continuity between shape functions.

2.4.2 BASIS FUNCTIONS

The basis functions used are pyramid-shaped basis functions [Steele,1987] with each basis function corresponding to one node. These basis functions have local (or compact) support, which is to say that they are nonzero only over those elements which contain the corresponding node point. Examining Fig. 2.1, the elements adjoining a particular node constitute the support of the basis function. Basis functions have a value of unity at their node point and have a nonzero gradient over their support. This nonvanishing gradient property allows for globally continuous elements and for computations corresponding to the gradient of the desired solution. The alternative pulse basis function fails in this regard, and though well suited to moment method source modeling, will not be considered adequate for finite element techniques used here. These pyramidal basis functions, $v_i(x,y)$, are linearly independent and as such can be represented over the space S_N as the linear combination

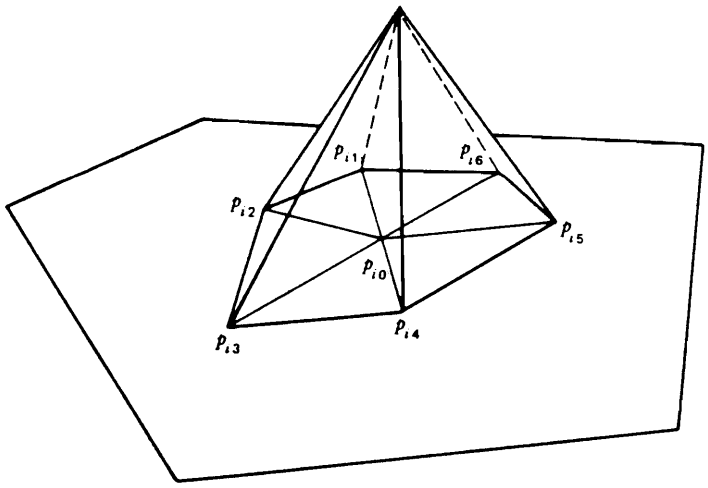


Fig. 2.1 Pyramidal basis functions [Prenter,1975] .

$$\hat{u}(x,y) = \sum_{i=1}^{N} u_i \, v_i(x,y) \tag{2.14}$$

where u_i are real numbers. It is important to keep in mind that since more than one element constitutes the support of $v_i(x,y)$, the basis functions are composed of more than one shape function.

2.4.3 SHAPE FUNCTIONS FOR TRIANGULAR ELEMENTS

The solution domain can be represented by Ω bounded by Γ and approximated by the union of nonoverlapping triangles Ω_e each with boundary Γ_e, as seen in Fig. 2.2. The solution can be thought of as the summation of individual elements by considering the contributions of each element adjoining, say, node i in Eq. (2.14). The domain of a shape function is a single finite element. The interpolation functions required take the form

$$\hat{u}^{(e)}(x,y) = \mathrm{a}_1 + \mathrm{a}_2 x + \mathrm{a}_3 y \ , \tag{2.15}$$

where the set $\{1,x,y\}$ is linearly independent. By writing expressions relating the values at the three nodes u_i and satisfying the expression $v(x_i,y_i) = u_i$, it is possible to obtain a matrix expression to evaluate the constants a_1, a_2, and a_3. Referring to Fig. 2.3, the values at the nodes u_i (i = 1,2,3) in terms of the coefficients can be written

$$\begin{Bmatrix} u_1 \\ u_2 \\ u_3 \end{Bmatrix} = \begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{bmatrix} \begin{Bmatrix} \mathrm{a}_1 \\ \mathrm{a}_2 \\ \mathrm{a}_3 \end{Bmatrix} , \tag{2.16}$$

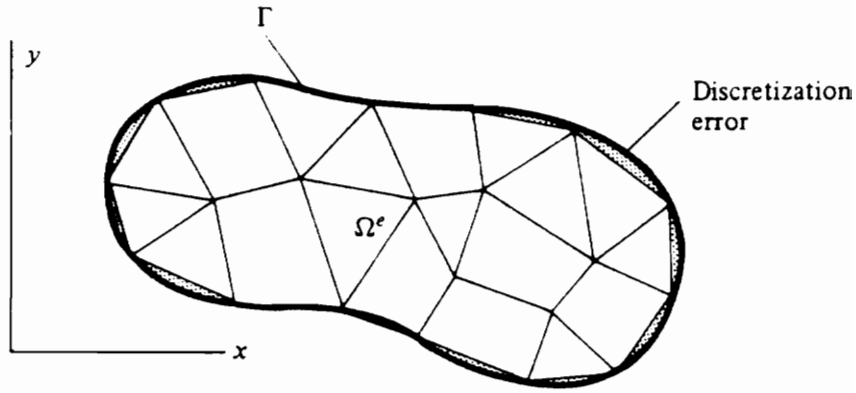


Fig. 2.2 Domain discretization [Reddy,1984].

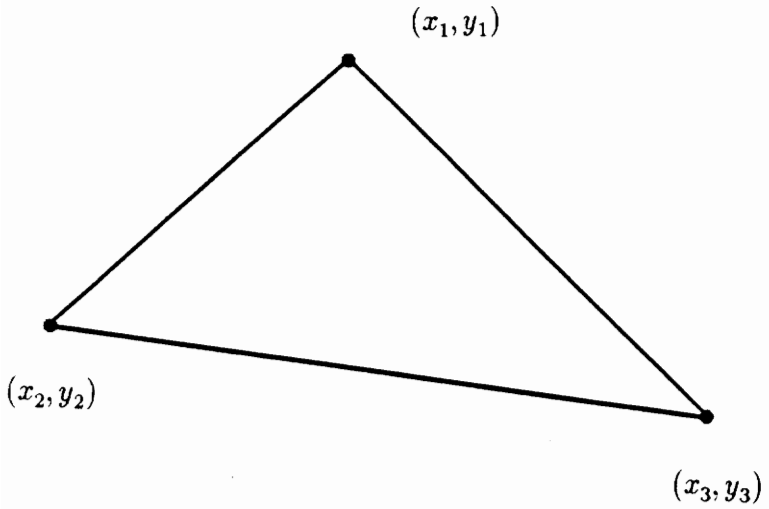


Fig. 2.3 Element geometry in local coordinates.

where the nodes are numbered clockwise. Solving (2.16) for a_i (i = 1,2,3) results in the following

$$a_1 = \frac{1}{2A_e}\left[u_1(x_2y_3 - x_3y_2) + u_2(x_3y_1 - x_1y_3) + u_3(x_1y_2 - x_2y_1)\right]$$

$$a_2 = \frac{1}{2A_e}\left[u_1(y_2 - y_3) + u_2(y_3 - y_1) + u_3(y_1 - y_2)\right] \qquad (2.17)$$

$$a_3 = \frac{1}{2A_e}\left[u_1(x_3 - x_2) + u_2(x_1 - x_3) + u_3(x_2 - x_1)\right] ,$$

where A_e is the area of triangle e given by the following determinant

$$2A_e = \begin{vmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{vmatrix} = (x_2y_3 - x_3y_2) + (x_3y_1 - x_1y_3) + (x_1y_2 - x_2y_1). \qquad (2.18)$$

The area relationship in Eq. (2.18) can be obtained by considering a cross product of two vectors, forming a parallelogram, and the factor of 1/2 recovers the triangle area. Note that the area given in Eq. (2.18) is positive for nodes numbered clockwise. The node ordering determines the handedness of the cross product and a "negative" area is possible.

Substituting the values of a_i from (2.17) into (2.15), the following is obtained:

$$\hat{u}^{(e)}(x,y) = u_1\psi_1(x,y) + u_2\psi_2(x,y) + u_3\psi_3(x,y) = \sum_{i=1}^{3} u_i\psi_i^{(e)} \qquad (2.19)$$

where $\psi_i^{(e)}$ represents a Lagrange interpolation function for triangular elements given by

$$\psi_i^{(e)} = \frac{1}{2A_e}(\alpha_i + \beta_i x + \gamma_i y) \qquad i = 1,2,3 \tag{2.20}$$

and α_i, β_i, and γ_i are constants obtained from local element geometry. These constants are given explicitly as

$$\begin{aligned} \alpha_i &= x_j y_k - x_k y_j \\ \beta_i &= y_j - y_k \\ \gamma_i &= x_k - x_j \end{aligned} \tag{2.21}$$

where $i \neq j \neq k$, and i, j, k permute in a natural order. The linear shape functions $\psi_i^{(e)}$ are shown in Fig. 2.4, and have the following properties:

$$\psi_i(x_j, y_j) = \delta_{ij} \qquad i, j = 1,2,3$$

and (2.22)

$$\sum_{i=1}^{3} \psi_i = 1 \ .$$

Note that the shaded region in Fig. 2.4 corresponds to the amplitude of the shape function. The dashed lines are hidden lines.

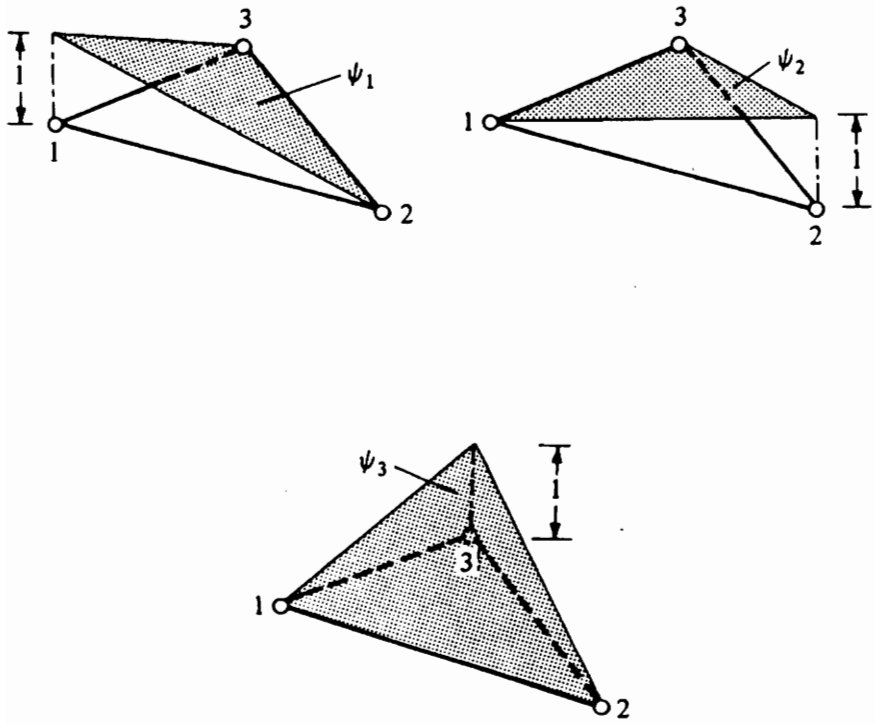


Fig. 2.4 Linear shape functions [Reddy,1984].

2.4.4 MATRIX REPRESENTATION FOR FINITE ELEMENT PROBLEMS

For an equation of the form of (2.1), it is possible to define possible expansion functions by

$$v(\cdot) = \sum_{i=1}^{N} a_i \, v_i(\cdot) \, , \tag{2.23}$$

and the approximation to the solution $\hat{u}$ as

$$\hat{u}(\cdot) = \sum_{i=1}^{N} u_i \, v_i(\cdot) \, . \tag{2.24}$$

For a real operator L and a real f, equation (2.10) may be written

$$F(v) = \langle Lv, v \rangle - 2 \langle f, v \rangle \, . \tag{2.25}$$

Taking the partial derivative with respect to the free variables as in (2.13) ,

$$\frac{\partial F(\hat{u})}{\partial a_i} = 0 \, , \quad \text{for } i = 1,2, \ldots , N, \tag{2.26}$$

and making use of the operator symmetry, leads to a system of equations of the form:

$$\sum_{i=1}^{N} \langle Lv_i(\cdot), v_j(\cdot) \rangle \, u_i = \langle f, v_j(\cdot) \rangle \quad \text{for } j = 1,2, \ldots , N \, , \tag{2.27}$$

similar to Eq. (2.5). Eq. (2.27) is a weighted equality of the desired equation with an appropriate expansion function. Integration by parts leads to the following form

$$\left\langle Lv_i(\cdot),v_j(\cdot)\right\rangle = B\left(v_i(\cdot),v_j(\cdot)\right) + \ell\left(v_j(\cdot)\right), \tag{2.28}$$

where $B(\cdot)$ is a bilinear functional representation of the solution and $\ell(\cdot)$ is a linear functional determined by the natural boundary conditions of the problem. These natural boundary conditions correspond to Neumann boundary conditions for the purposes of this thesis. The system of equations can then be written $Kx=b$, where

$$\sum_{i=1}^{N} B\left(v_i(\cdot),v_j(\cdot)\right) u_i = \left\langle f,v_j(\cdot)\right\rangle - \ell\left(v_j(\cdot)\right) \quad \text{for } j = 1,2, \ldots ,N. \tag{2.29}$$

The entries of the matrix K are determined by considering the support of the basis functions $v_i(\cdot)$ and $v_j(\cdot)$. Consider an index $n(i,j)$, the number of elements in the intersection of the basis functions $v_i(\cdot)$ and $v_j(\cdot)$. The entries of K can be *symbolically* written

$$k_{ij} = \sum_{e=1}^{n(ij)} B_e(\psi_i,\psi_j) \quad , \tag{2.30}$$

where the e subscript on B signifies integration over an element e. The entries for the matrix K are discussed more fully in Chapter 4, after a more explicit definition of the discretization is established and a representation of a proposed data structure is discussed. The first term on the right-hand side of (2.29) can be written *symbolically*

$$\left\langle f,v_j(\cdot)\right\rangle \Rightarrow \sum_{e=1}^{n(i)} \int_{\Omega_e} \psi_i \; f \; d\Omega_e, \tag{2.31}$$

with $n(i)$ interpreted as the intersection of the basis function v_i and Γ. Before defining the right-hand side vector b, it is important to note that the Dirichlet boundary conditions have not been imposed. These are known as essential boundary conditions, and before enforcement, the matrix K is singular. The Dirichlet boundary conditions are enforced by eliminating the equation that corresponds to the known boundary data. This elimination is performed in such a way as to preserve the symmetry of the system of equations. The enforcement of Dirichlet conditions is described in detail in section 4.3. The resulting form of K is no longer singular. The matrix K is sparse due to the compact support of the basis functions. The linear system can be solved using a number of methods, but with consideration given to sparsity, a large number of unknowns can be handled on a personal computer. The sparse solution techniques will be discussed in chapter 4.

Chapter 3

Mesh Generation

3.1 INTRODUCTION TO MESH GENERATION

In order to set up a problem for solution using techniques discussed in Chapter 2, the domain must be suitably discretized. In two dimensions this is commonly accomplished by breaking the domain into rectangular or triangular elements. In three dimensions this takes the form of brick elements and tetrahedrons, respectively. Rectangular elements can be readily used to discretize the solution domain, but are unable to easily model curved or complex boundaries. Triangular elements easily model complicated geometrical shapes, but care must be taken to assure a "good" triangulation. The notion of an adequate triangulation sometimes takes the form of philosophical questions since poor triangulations can still produce accurate results. Most references concerned with mesh generation [Cavendish,1974; Sibson,1978] indicate that an ideal triangle is an equilateral triangle, providing the best solution. This ideal is impossible to achieve in practice, so Cavendish[1974] and others propose a scheme that maximizes the sum of smallest angles [Lawson,1972]. The approach of maximizing the sum of smallest angles will be used for this thesis. The details concerning implementation will be postponed to section 3.2.

A brief discussion of the philosophical questions will serve to illustrate the concepts of the effects of the discretization on the obtained solution. The determination of an optimal mesh for a particular set of points was addressed by Lawson[1972] and considered the accuracy of interpolation functions on the mesh.

Lawson's work emphasized the use of the Delaunay triangulation to maximize the sum of minimum angles. Wilson [1990] has indicated that avoidance of thin triangles is the key concern when examining a mesh. Considering an interpolation over a triangular element, the resulting matrix representation has a condition number C that can be used in comparing the effect of a poor triangulation. Consider a triangle T with vertices $\{x_i, y_i\}_{i=1...3}$, and included angles $\{\theta_i\}_{i=1...3}$ corresponding to each vertex. Also let $\{s_i\}$ represent the side of T opposite θ_i and label the vertices such that $\theta_1 \leq \theta_2 \leq \theta_3$. Given a linear interpolant of the form

$$\tilde{u}(x, y) = a_1 + a_2 x + a_3 y \ , \tag{3.1}$$

corresponding to measured data a_i, i = 1,2,3 at the vertices $\{x_i, y_i\}_{i=1..3}$, the interpolant can be determined from the following system of equations:

$$A \begin{Bmatrix} a_1 \\ a_2 \\ a_3 \end{Bmatrix} = \begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{bmatrix} \begin{Bmatrix} a_1 \\ a_2 \\ a_3 \end{Bmatrix} = \begin{Bmatrix} u_1 \\ u_2 \\ u_3 \end{Bmatrix} \ . \tag{3.2}$$

The particular matrix A for a given system is determined by choosing a vertex from which to base the solution. It is apparent that the solution will not depend on the vertex chosen, however, the condition number will be affected. The condition number is a measure of the system's sensitivity to small changes in the data a_i and indicates the potential error in solving (3.2). Wilson gives the condition number as

$$C(A_i) = \frac{1 + \sqrt{1 - \lambda_i^2}}{\lambda_i}, \tag{3.3}$$

where

$$\lambda_i = \frac{4\ Area(T)}{s_j^2 + s_k^2}, \tag{3.4}$$

and $i, j,$, and k permute in natural order. It follows that due to the ordering of $\theta_1 \leq \theta_2 \leq \theta_3$ that $s_1 \leq s_2 \leq s_3$. From (3.3) and (3.4) it can be seen that $C(A_1) \geq C(A_2) \geq C(A_3)$, and that computing the condition number using the smallest angle θ will yield the maximum condition number for a particular triangle T. Controlling the minimum allowable angle in a triangle would appear to be an effective means of limiting the error in a given system. The triangle shape plays an important role in determining the condition number C, and through (3.3) and (3.4) it is seen that an equilateral triangle would yield a minimum condition number for a given area of coverage. Very thin triangles lead to larger condition numbers and increase the error bound for a given system of equations based on some interpolation scheme over triangular elements. Babuska [1976] qualifies this by indicating that the minimum angle condition may be amended to consider the effect of more than one small angle in a triangle. Babuska indicates that a maximum angle condition, not pursued in this thesis, may be more appropriate so as to avoid two narrow angles within the same triangle.

A more practical question arises in connection with modeling the solution in a particular region of space. For fields that are rapidly changing in a particular direction it may be that "thin" triangles are an appropriate representation. This local behavior would occur near sharp corners or just inside

conducting regions where the field attenuates rapidly in a direction normal to the boundary. Though not explicitly considered in this paper, there are several papers [Bank,1980; Cendes,1985] that address one means of handling rapidly changing fields by a form of adaptive mesh generation. Adaptive mesh generation is based on obtaining an error minimization over individual finite elements, and can alleviate the need for human intervention in "fine-tuning" a particular mesh.

The following section will describe a means by which an optimum triangulation, in the sense of minimizing the angles, may be obtained for a given set of data.

3.2 DELAUNAY TRIANGULATION

Lawson [1972] suggested that an equiangular triangulation criterion be obtained for use in evaluating the quality of a given mesh. The goal of a maximally equilateral triangulation formed the basis for the *max-min angle criterion.* This criterion was originally used to determine the location of a diagonal in a quadrilateral used to subdivide a planar region. Sibson [1977] showed that this requirement is met by use of the Delaunay triangulation [1934] which is the dual to the Dirichlet tessellation[1]. Delaunay has shown that there exists an optimal triangulation for a given set of points that maximizes the sum of minimum angles in each triangle. This section will examine the

[1]A tessellation may be defined as a covering of a geometric plane without gaps or overlaps by tiles or polygonal regions. These polygons are sometimes referred to as a Voronoi or Thiessen tessellation.

characteristics of the Delaunay triangulation and describe the particular implementation chosen for use in this work.

An understanding of the Delaunay triangulation is aided by considering the Dirichlet tessellation. Given a finite set of discrete points $\{P_i\}_{i=1...N}$ in a plane it is possible to assign to each point a control region [McCartin,1990] D_j such that each collection of points that make up the polygonal region D_j, is closer to P_j than to any other P_i for $i \neq j$. This is mathematically described by Sibson [1977] as

$$D_j = \{x : d(x, P_j) < d(x, P_i) \text{ for all } i \neq j\},$$

where d is the Euclidian distance. These control regions are known as Dirichlet regions, and are convex polygons whose union forms the tessellation. An example of a Dirichlet tessellation is seen in the top of Fig. 3.1. Note that the boundaries between nearest neighbors are the perpendicular bisectors of the line joining the two neighbors. By connecting the point P_j to adjacent Dirichlet regions whose boundaries form the perpendicular bisectors between adjacent generating points, the Delaunay triangulation is formed. The set of M Delaunay triangles $\{T_k\}_{k=1...M}$ is the optimum triangulation of the given point set $\{P_i\}_{i=1...N}$ in the sense that the sum of minimum angles in each triangle is maximized. The bottom figure in Fig. 3.1 demonstrates an example of the Delaunay triangulation of the tessellation shown at the top of Fig. 3.1. Miles [1970] has indicated that given the number of boundary nodes, η_b, and interior nodes, η_i, there are $2\eta_i + \eta_b - 2$ members in the set $\{T\}$. That is to say that the number of members M is

$$M = 2\eta_i + \eta_b - 2 .$$

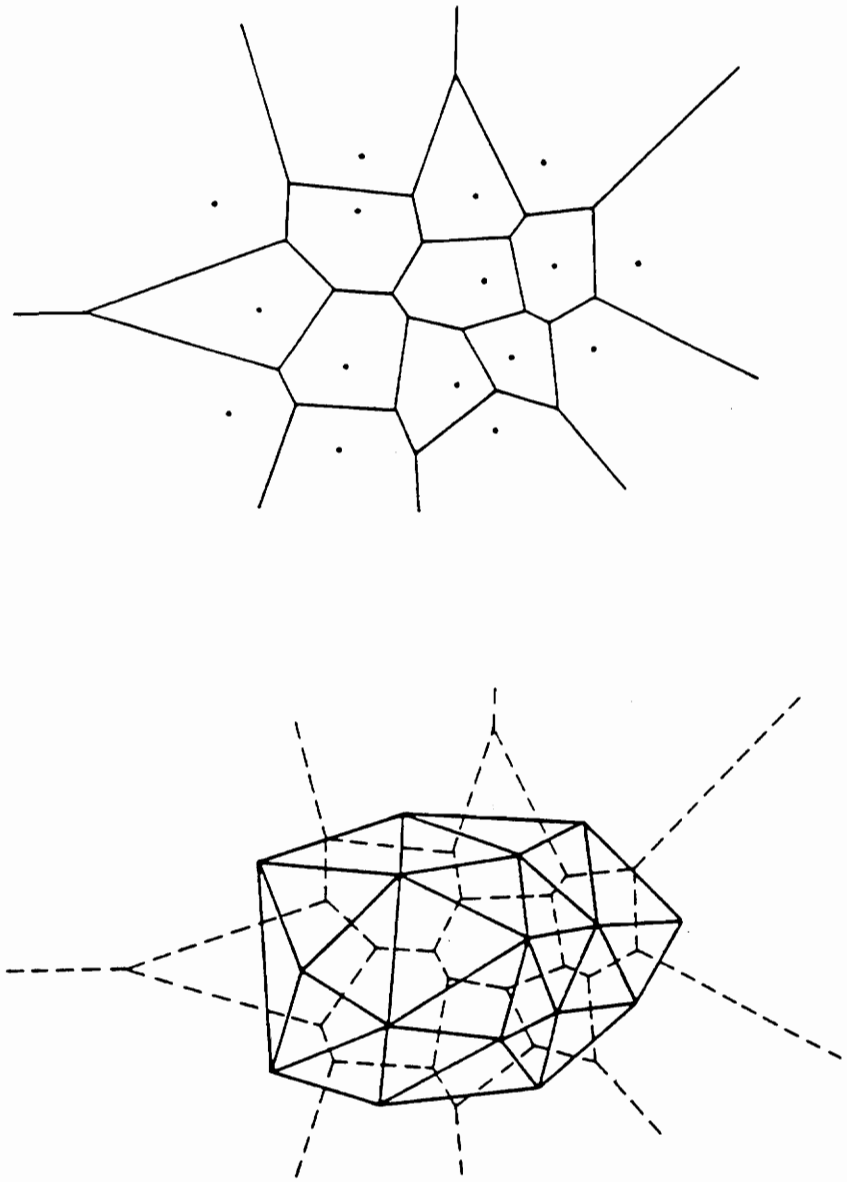


Figure 3.1 The Delaunay Triangulation/ Dirichlet Tessellation [McCartin,1990].

There are a number of different algorithms [Ho-Le,1988] that provide a means to obtain this Delaunay triangulation. Ho-Le's review paper describes the strategies that can be used in generating a mesh. Watson's [1981] algorithm appears to be sufficiently general and the most straightforward to implement. The algorithm is capable of extending into higher dimensions and provides a simple data structure. Many authors use complex tree data structures that typically require some form of optimal node ordering to reduce the computational bandwidth. This will not be of concern for the application in this thesis since an iterative method will be used in solving the system of linear equations that will not require any particular ordering. This iterative method will be described in detail in Chapter 4. Watson's triangulation algorithm is used in conjunction with the point generation scheme suggested by Cavendish [1972] to obtain a triangulation for a distribution of points in the plane. Cavendish's algorithm provides the capability of controlling the mesh density to allow for rapidly varying fields in a particular region.

The heart of Watson's [1981] algorithm is a means of evaluating a set of data points, for inclusion in an element, known as a "circle-check", originally described by Lawson in 1972. This circle-check requires a successful Delaunay triangle to contain no points within the circle that circumscribes the three vertices that form the triangle. A data point falling within the circle is considered to *fail* the circle check, and will result in the formation of new elements. These new elements will be formed in any triangle that has failed the circle-check. The circle check can be illustrated by considering the mesh in Fig. 3.2. Figure 3.2(a) shows a current element with the new data point x just

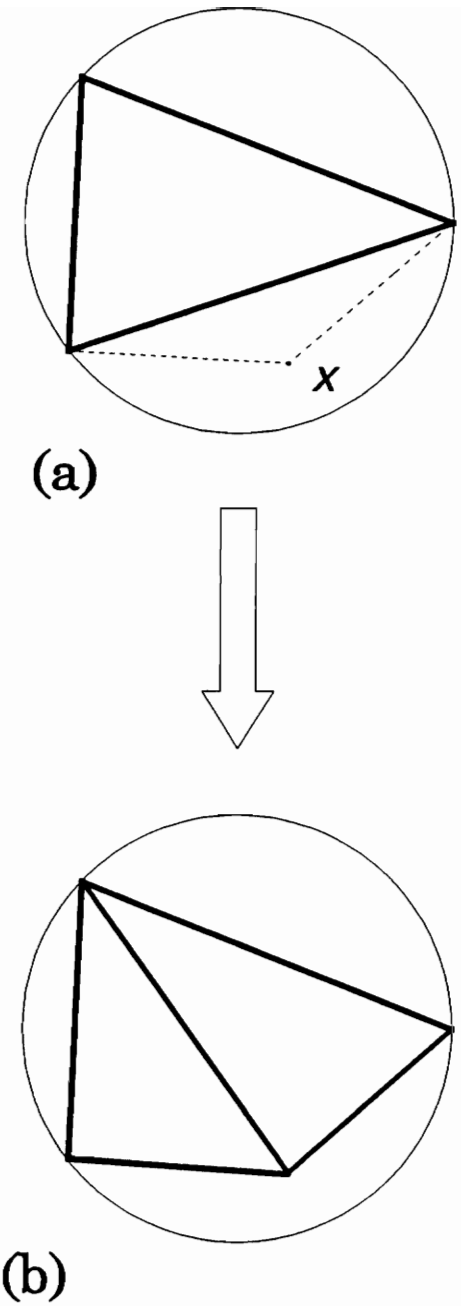
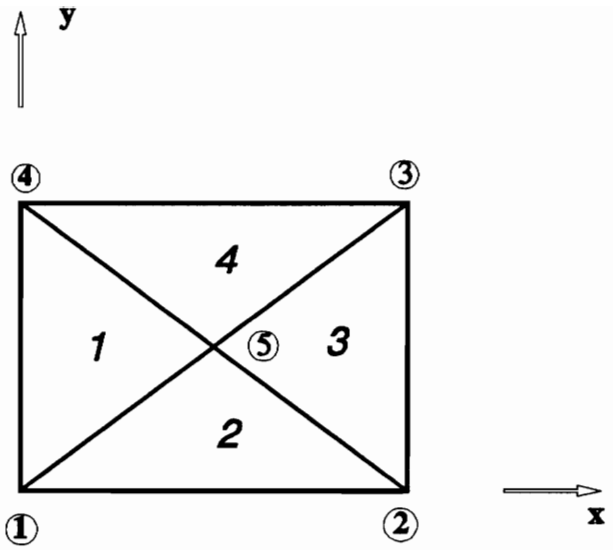


Figure 3.2 The circle-check for the Delaunay Triangulation.

inside the circle. The dotted line indicates a possible, though undesirable and non-Delaunay, triangulation. Figure 3.2(b) shows the new elements formed as a result of failing the circle check. The circle that circumscribes the new elements in Fig. 3.2b will contain no data points. The data structure of Figure 3.3 is convenient for the purpose of programming and is efficient in the use of computer memory. The element node matrix is an integer valued array containing the vertex labels that comprise each element. The vertex labels shown are the global node numbers that correspond to the vertex coordinate array. This real valued array contains the xy coordinates of each global vertex label.

Assume there exists some point distribution $\{P_i\}_{i=1...N}$ in the plane and there is some initial coarse triangulation $\{T\}^{(0)}$. This initial triangulation may be that of some arbitrary interior point and *all* the boundary nodes forming $M^{(0)}$ elements. The triangulation proceeds from $\{T\}^{(0)}$, $\{T\}^{(1)}$ $\cdots$ $\{T\}^{(F)}$ with the number of elements $M^{(0)}$, $M^{(1)}$ $\cdots$ $M^{(F)}$ as new points are considered for inclusion in the mesh. For each new point P_i to be introduced to the triangulation, search the list of $M^{(0)}$ elements to determine which elements fail the circle check and flag them. This search is accomplished by subtracting the distance from the new point, P_i, to the center of the circle and the radius of the circle corresponding to every element. If this operation yields a number less than zero for a particular element then that element has failed the circle check and is flagged. The algorithm continues by forming a vector stack containing the sides of each flagged element. This flagging process results in a stack of $3N_f$ members, where N_f is the number of flagged elements. After eliminating multiple occurrences of sides within the stack, corresponding to a shared side,



Element #	Global Node Number
1	4 1 5
2	1 2 5
3	2 3 5
4	3 4 5

Global Node Number	X Y Coordinates
1	(0.0, 0.0)
2	(1.0, 0.0)
3	(1.0, 1.0)
4	(0.0, 1.0)
5	(0.5, 0.5)

Figure 3.3 Data structure for Watson's Algorithm.

the stack is rewritten to the element list forming new elements composed of the new point P_i and each stacked side. This new element list and new vertex list now form an updated triangulation $\{T\}^{(1)}$ with $M^{(1)}$ elements. This process is repeated until each point $\{P_i\}_{i=1...N}$ has been included in the mesh resulting in the final triangulation $\{T\}^{(F)}$ with $M^{(F)}$ members. The extension to 3 dimensional elements involves a "sphere-check" and stacking the faces of the tetrahedral elements in the case of a failed sphere-check.

The Delaunay triangulation provides a unique triangulation for a given set of points. Degeneracies can occur in this triangulation scheme when the numerical round off of a given machine affects the determination of a point's intersection with the circle. Degeneracies lead to a nonunique triangulation. When a subtraction yields a number that is within numerical round off of machine precision, the point should be dropped from consideration as a new point. This point corresponds to a point that lies directly on the circle that circumscribes a current element and results in an ambiguity in the choice of new elements. Degeneracies can occur when an initial triangulation $\{T\}^{(0)}$ is obtained by subdividing squares.

Up to now it has been assumed that the point set $\{P_i\}_{i=1...N}$ has been given, and this typically will seldom occur in the case of a finite element discretization. This may not be the case for other modeling processes that utilize the Delaunay triangulation to model irregular data sets as in biological or sociological studies [Miles,1970]. A method has been developed following Cavendish's [1974] work to fill the domain with a statistically uniform distribution of points. After designating the fixed boundary nodes, a specified density of interior points, and a perturbing distance δ from the boundary, the

domain is automatically filled. It is important that the density of points and δ be of the same order as the distance between the given boundary nodes, or undesirable elements can result. Cavendish's scheme is outlined as follows:

Given a set of boundary nodes (that may be multiply connected) slightly perturb the region, $\Re$, contained by the boundary nodes, by an amount δ to obtain the perturbed region $\tilde{\Re}$. Overlay the region $\Re$ with an imaginary grid of squares whose dimensions correspond to the prescribed density of points, ρ_i as in Figure 3.4. Using a uniform random number generator begin in the first square (bottom-left) to pick a point that [1)] lies in the square [2)] lies in the perturbed region $\tilde{\Re}$, and [3)] is no closer than ρ_i to any previously randomly picked internal points. This process is continued until every imaginary square contains a point. The density of points in the plane can be varied by using zones to denote the area of differing point density.

Some comments concerning the idea of a random mesh are appropriate since it is not commonly used. For the choice of triangular Lagrange shape functions which are not required to have first derivative continuity between elements, the mesh orientation can have an undesirable effect on the obtained solution. For this reason some authors avoid using triangular elements and use rectangular elements instead [Reddy,1984]. The rectangular elements, though conceptually satisfactory, suffer from two disadvantages. The first and most commonly noted disadvantage is that it is difficult to model complex boundary shapes with rectangular elements. Secondly, the integration over rectangular elements is more difficult than for triangular elements, requiring a numerical integration scheme. This difficulty in integration arises from the more complex expansion functions required by using rectangular elements. The use of random

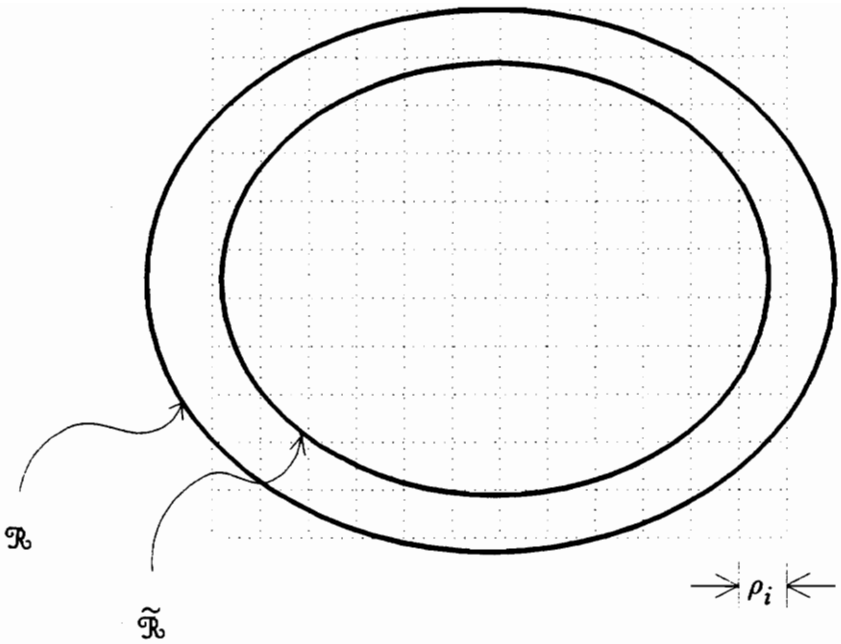


Figure 3.4 Cavendish point-in-plane generation.

mesh generation, though time consuming, provides an arbitrariness to this undesirable effect. Which is to say that the user will not force a "preferred" direction to the solution. Figure 4.3 demonstrates a random mesh for a coaxial transmission line. Figure 5.2 illustrates a waveguide cross section with a coarse triangulation.

Using the finite element method for Laplace's equation described in detail in Chapter 4, the preferred direction concept is portrayed in Figures 3.5 and 3.6. The physical problem considered is an infinite square trough with a top plate potential specified at some nonzero value and the other three sides are assumed to be grounded or at zero potential. These solutions are based on a simple mesh structure that modifies the direction of a diagonal of a rectangular mesh. It is noted that the two solutions form mirror images of one another. This observation is not critical when the physical quantity of interest is minimized such as the capacitance. The capacitance per unit length is identical for either mesh, but demonstrates the degeneracy that can occur when using squares or rectangles. The matrix equations for the two problems are identical, so the differences are due to the expansion function used for the particular mesh.

Possible future application to random rough media, where both the guided propagation and scattering problems are of great interest to many researchers, will also benefit from this approach, as a random surface can be generated using these ideas. The added complexity of a random mesh is felt a worthwhile compromise in light of the desire to model complex geometrical shapes with a minimum amount of human intervention.

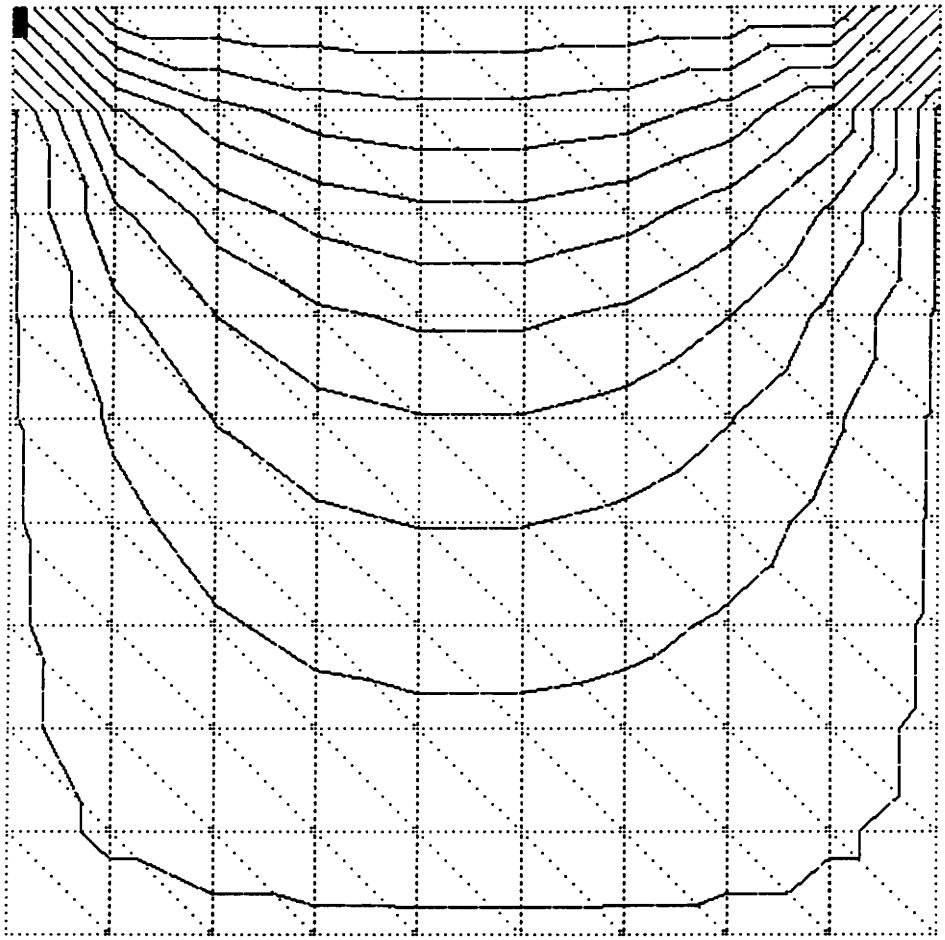


Figure 3.5 Laplace's equation for square trough with diagonal from upper-left to lower-right.

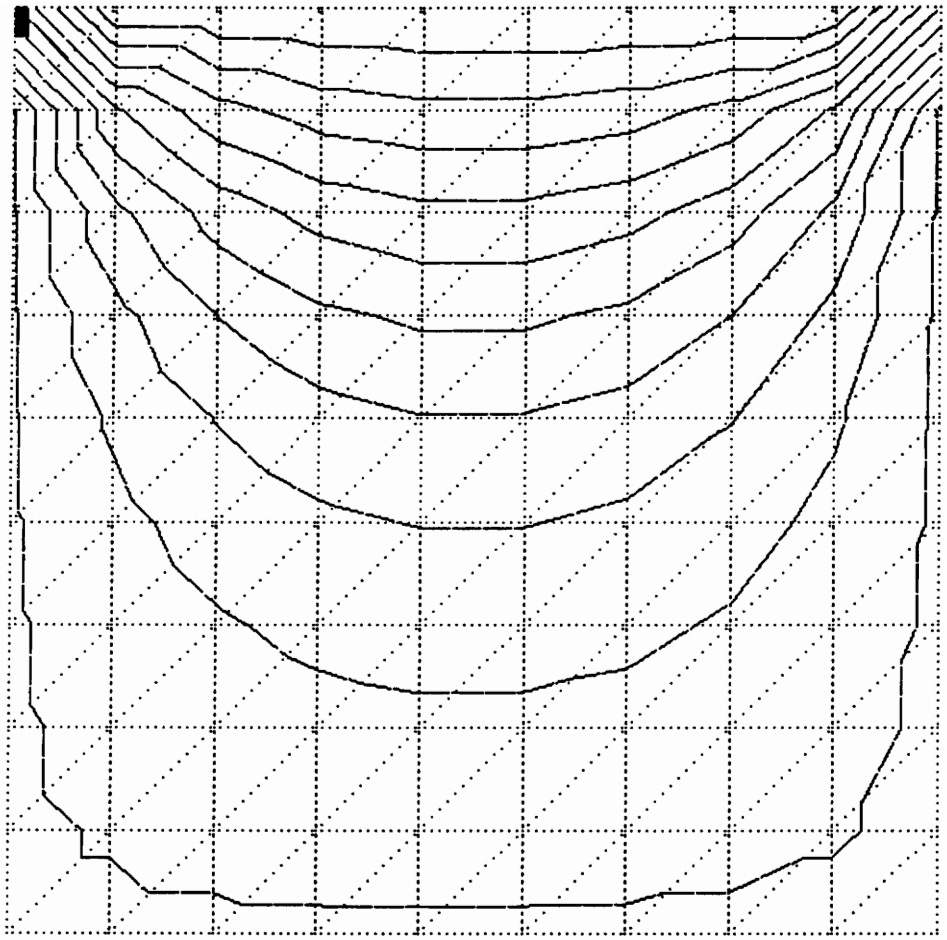


Figure 3.6 Laplace's equation for square trough with diagonal from lower-left to upper-right.

Chapter 4

Finite Element Techniques Applied to Laplace's Equation

4.1 INTRODUCTION

The purpose of this chapter is to detail the finite element formulation for Laplace's equation in two dimensions and the subsequent computer implementation for a canonical problem and then as applied to a stripline structure. Comparison is made to exact results for the canonical problem and approximate results from the literature for the stripline structure.

4.2 VARIATIONAL FORMULATION

Consider Laplace's equation in 2 dimensions

$$-\nabla^2 u = 0 \text{ in } \Omega \ , \tag{4.1}$$

where the operator $L = -\nabla^2$ in Cartesian coordinates is

$$\nabla^2 = -\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) \ ,$$

subject to Dirichlet boundary conditions $u(\vec{r}) = U(\vec{r})$, $\vec{r} \in \partial\Omega$ [Zachmanoglou, 1986]. The vector $\vec{r}$ is the position vector in the x-y plane. Though not considered for the sample calculations, Neumann boundary conditions could be applied where $u(\vec{r}) = \frac{\partial U(\vec{r})}{\partial n}$, $\vec{r} \in \partial\Omega$ is specified [Zachmanoglou, 1986]. The variable u represents the electrostatic potential within a given region Ω. To

obtain a variational expression, the results of section 2.4.4. are used, that is, from Eq. (2.27)

$$\sum_{m=1}^{N} \left\langle Lv_m(\cdot), v_n(\cdot) \right\rangle u_m = 0 \quad \text{for } n = 1,2, \ldots, N \,, \tag{4.2}$$

with the approximation for $u \Rightarrow \hat{u}$ given by Eq. (2.24) as

$$\hat{u}(\cdot) = \sum_{m=1}^{N} u_m \, v_m(\cdot) \,. \tag{2.24}$$

where the right hand side is zero since $f \equiv 0$ for Laplace's equation. To obtain the bilinear form corresponding to Eq. (2.28), it is possible to express the inner product $\left\langle L\hat{u}, \hat{u} \right\rangle$ as an integration over the domain Ω , given by

$$- \int_{\Omega} \hat{u} \nabla^2 \hat{u} \, d\Omega.$$

Applying Green's first identity (integration by parts) to this inner product one obtains

$$- \int_{\Omega} \hat{u} \nabla^2 \hat{u} \, d\Omega = \int_{\Omega} \nabla \hat{u} \cdot \nabla \hat{u} \, d\Omega - \oint_{\partial\Omega} \hat{u} \nabla \hat{u} \cdot \hat{n} \, dS \,,$$

with an outward normal $\hat{n}$, or equivalently

$$- \int_{\Omega} \hat{u} \nabla^2 \hat{u} \, d\Omega = \int_{\Omega} \nabla \hat{u} \cdot \nabla \hat{u} \, d\Omega - \oint_{\partial\Omega} \hat{u} \frac{\partial \hat{u}}{\partial n} \, dl \,.$$

Recalling Eq. (2.28)

$$\left\langle Lv_m(\,\cdot\,),v_n(\,\cdot\,)\right\rangle = B\left(v_m(\,\cdot\,),v_n(\,\cdot\,)\right) + \ell\left(v_n(\,\cdot\,)\right),$$

it is possible to identify the bilinear and linear terms

$$B\left(v_m(\,\cdot\,),v_n(\,\cdot\,)\right) = \int_{\Omega} \nabla v_m \cdot \nabla v_m \, d\Omega \ , \tag{4.3}$$

and

$$\ell\left(v_n(\,\cdot\,)\right) = - \oint_{\partial\Omega} v_n \frac{\partial v_m}{\partial n} \, dl \ . \tag{4.4}$$

A system of equations may be written as

$$\sum_{m=1}^{N} K_{mn} \, u_m \ = 0 \ , \ \text{for } n = 1, \ldots N \ , \tag{4.5}$$

where the matrix K is obtained by substituting Eq. (4.3) and (4.4) in Eq. (4.2).

The Neumann boundary condition in Eq. (4.4) corresponds to the natural boundary condition of the operator $L = -\nabla^2$. Steele [1987] indicates that the domain boundary $\partial\Omega$ can be considered to be the union of the boundary region, ∂E, over which the essential Dirichlet boundary conditions are to imposed, and the corresponding region ∂N for the Neumann boundary conditions. In general, the bounding surface $\partial\Omega$ comprises both types of boundaries, and is often a result of symmetry which can be exploited to reduce the problem size. The problem to be examined for this work is the Dirichlet problem so that specification of the normal derivative as required in Eq. (4.4) will lead to an ill posed problem. The contribution from this term can be removed by a judicious choice of expansion function $v_n\,(\,\cdot\,)$, namely $v_n = 0$ on $\partial\Omega$.

The bilinear expression over the entire domain Ω may be replaced by an appropriate expression in terms of the triangular interpolating shape functions as described in detail in Chapter 2. Recall that the only nonzero contribution to the matrix K arises from an intersection of basis functions. This intersection can be expressed in terms of the shape functions ψ_i as in Eq. (2.30) *symbolically* as follows:

$$k_{mn} = \sum_{e=1}^{n(mn)} B_e(\psi_i,\psi_j) \ . \tag{4.6}$$

The bilinear expression for an *element* is emphasized by the use of the subscript e in the expression for B_e and *local* indices i,j = 1,2,3. The discussion of element equations is satisfactory since the continuity of shape functions along interelement boundaries is to be enforced. Eq. (4.6) can be written as

$$k_{mn} = \sum_{p=1}^{n(mn)} \int_{\Omega_e} \nabla\psi_i^{(e)} \cdot \nabla\psi_i^{(e)} \, d\Omega_e \ . \tag{4.7}$$

The integral in Eq. (4.7) represents the element matrix $k_{ij}{}^e$ given as

$$k_{ij}^e = \int_{\Omega_e} \nabla\psi_i^{(e)} \cdot \nabla\psi_i^{(e)} \, d\Omega_e = \int_{\Omega_e} \left\{ \frac{\partial\psi_i^{(e)}}{\partial x} \frac{\partial\psi_j^{(e)}}{\partial x} + \frac{\partial\psi_i^{(e)}}{\partial y} \frac{\partial\psi_j^{(e)}}{\partial y} \right\} d\Omega_e \ , \tag{4.8}$$

where k^e is a 3 by 3 matrix. Recall that the elemental shape function is given by

$$\psi_i^{(e)} = \frac{1}{2A_e} (\alpha_i + \beta_i x + \gamma_i y) \qquad i = 1,2,3 \tag{4.9}$$

with α_i, β_i, and γ_i constants obtained the from local element geometry as in Eq. (2.21).

The partial derivatives are easily determined and lead to element matrices of the form

$$k_{ij}^e = \int_{\Omega_e} \left\{ \frac{\partial \psi_i^{(e)}}{\partial x} \frac{\partial \psi_j^{(e)}}{\partial x} + \frac{\partial \psi_i^{(e)}}{\partial y} \frac{\partial \psi_j^{(e)}}{\partial y} \right\} d\Omega_e \,, \quad \text{i,j} = 1,2,3 \tag{4.10a}$$

$$k_{ij}^e = \int_{\Omega_e} \left\{ \frac{\beta_i}{2A_e} \cdot \frac{\beta_j}{2A_e} + \frac{\gamma_i}{2A_e} \cdot \frac{\gamma_j}{2A_e} \right\} d\Omega_e \,, \tag{4.10b}$$

$$k_{ij}^e = \frac{1}{4(A_e)^2} (\beta_i \beta_j + \gamma_i \gamma_j) \int_{\Omega_e} d\Omega_e. \tag{4.10c}$$

Since the integration over a single element is simply A_e, we obtain the final form:

$$k_{ij}^e = \frac{1}{4A_e} (\beta_i \beta_j + \gamma_i \gamma_j) \,, \tag{4.11}$$

where the area, A_e, is given by (2.18).

The resulting system of equations is

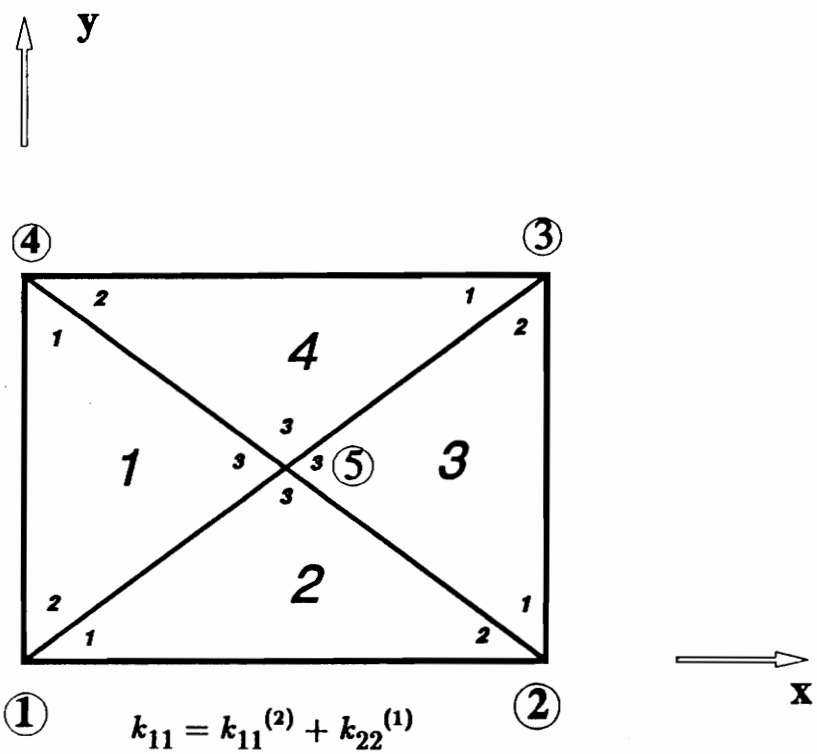
$$\sum_{m=1}^{N} \left\{ \sum_{m \cup n \in e} \left(k_{i=l(m), j=l(n)}^e \right) \right\} u_m = 0, \quad n = 1, ..., N \tag{4.12}$$

where $\ell(p)$ is the column that contains node p in the data structure of Fig. 3.3. , resulting in an $N \times N$ system of equations. This unique and formal representation performs the practical implementation of the summation written symbolically in Eq. (4.7). Equation (4.12) enforces the continuity between

elements by considering only the union of global nodes m and n in an element as providing a nonzero contribution. This practical matter is mentioned to clarify an occasional point of confusion when employing finite element techniques. The element integrations are used to form *element matrices* whose linear dimension is equal to the number of degrees of freedom per node corresponding to the number of vector components, multiplied by the number of nodes per element. Some authors refer to a connection matrix idea to heuristically represent the assembly process, but this notion obscures the fundamental basis of the finite element method which is rooted in interpolation theory. The shape functions developed in Chapter 2 were developed on an element by element basis simply because the domain of each shape function is *only* over a single element, not over the entire domain. These shape functions combine to form the support of the pyramidal basis functions.

4.3 ASSEMBLY OF ELEMENT MATRICES

This section provides an alternative to the formal representation of Eq. (4.12) and performs the symbolic summation of Eq. (4.7). By considering the "sum" or assembly of element matrices the global matrix K can also be obtained. This matrix is created element by element by considering the contributions from each element in the support of the pyramid basis functions. Recall that only those triangles in the support of a particular basis function contribute to the solution at its corresponding node point. Note that only those points directly connected to the node of interest will have nonzero contribution. The contribution to a node from other adjacent nodes is weighted by the elements containing a shared side. Fig 4.1 demonstrates the practical assembly



$$k_{11} = k_{11}^{(2)} + k_{22}^{(1)}$$

$$k_{12} = k_{12}^{(2)}$$

$$k_{13} = 0$$

$$k_{14} = k_{21}^{(1)}$$

$$k_{15} = k_{13}^{(2)} + k_{23}^{(1)}$$

$$\vdots$$

$$k_{22} = k_{22}^{(2)} + k_{11}^{(3)}$$

$$k_{24} = 0$$

$$k_{21} = k_{21}^{(2)}$$

$$\vdots$$

$$k_{55} = k_{33}^{(1)} + k_{33}^{(2)} + k_{33}^{(3)} + k_{33}^{(4)}$$

Figure 4.1. Element assembly.

process. The nodes numbered internally to each triangle represent the local element matrix subscripts $k_{ij}^{(e)}$, the circled node numbers represent the global nodes, and the internal bold numbers are element numbers. Consider the k_{11} entry in the global matrix where direct connection to local node 2 of element 1 and local node 1 of element 2 yields an entry of $k_{11}=\left(k_{11}^{(2)}+k_{22}^{(1)}\right)$. Entry k_{12} is given by simply $k_{12}^{(2)}$. The entry k_{13} is equal to zero since there is no direct connection between nodes 1 and 3. Further entries are examined in Fig. 4.1.

The resulting matrix is a sparse matrix where the nonzero entries correspond to adjacent global nodes as a result of the local support of the expansion functions (recall Fig. 2.1). The symmetry of the global stiffness matrix, commonly called the stiffness matrix (so denoted due to a structural mechanics origin of the finite element method to describe a system of trusses), results from the variational formulation. The resulting system of linear equations in matrix form appears as

$$\begin{bmatrix} k_{11} & k_{12} & k_{13} & \cdots \\ k_{21} & k_{22} & k_{23} & \cdots \\ k_{31} & k_{32} & k_{33} & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \cdots \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ \cdots \end{bmatrix} \tag{4.13}$$

where the right hand side is given for the general case of a nonzero forcing function f in (4.1). The global stiffness matrix is singular prior to the imposition of boundary conditions in (4.13) [Reddy,1984]. The boundary conditions to be enforced are essential boundary conditions and are applied to the solution $\vec{u}$.

For $u_m = U_m$, U_m representing the known boundary data, the following practical procedure is utilized in numerical application [Reddy,1984]:

1. For boundary nodes, the variational expression is invalid for a constrained function, since it is identically zero. The equation corresponding to the constrained boundary node is eliminated, and replaced with the boundary condition by setting the diagonal entry to unity and right hand side to the boundary value, or

$$k_{mm} = 1 \qquad f_m = U_m \ . \tag{4.14a}$$

2. If desired and to preserve the symmetry, the contribution of the boundary value to all other nodes may be moved to the right hand side by

$$f_i \rightarrow f_i \ - k_{im} U_m . \tag{4.14b}$$

3. To completely eliminate the equation from the left hand side the corresponding row and column are set to zero except for the diagonal, or

$$k_{mi} = k_{im} = 0 \ , \tag{4.14c}$$

where

$$i = 1,\ 2,...,\ m-1,\ m+1,...,\ n, \qquad i \neq k \ .$$

To summarize, this procedure essentially replaces the row and column corresponding to the boundary node m with zeros, except for the diagonal entry, which becomes unity. The right hand side is also modified by (4.14b) to preserve symmetry. After application of boundary conditions the desired system of equations $Ax = b$ is obtained.

4.4 SOLUTION OF SPARSE MATRICES

General solution techniques for solving the linear system $Ax = b$ can be divided into two categories. The first method is the direct method which generally involves some form of Gauss elimination or LU decomposition [Johnson,1982; Press,1989]. The direct methods require matrices of special forms such as banded, diagonal, or block diagonal forms, which often lead to excessive memory requirements for large sparse systems. Although faster for a particular system of equations, the direct methods limit the number of unknowns for which a solution can be obtained on a personal computer due to the memory requirements. A motivation for using a personal computer is the need for interactive computation from mesh design and boundary data implementation to post processing the particular solution for a given geometry. For double precision computations using banded storage the number of unknowns is typically limited to around 300. If care is given to the numbering of nodes, the resulting matrix in (4.13) is block diagonal and simple algorithms can transform this form into a banded form where nonzero entries are minimized but not altogether eliminated from the solution. The special handling required in numbering the nodes is treated by a number of authors [Cuthill,1972; Durocher,1979; Rivara,1984]. It is prohibitively time consuming to process the ordering of nodes for a solution when the number of unknowns are in the thousands. Alternatives to direct methods are the indirect or iterative methods.

Iterative methods are based on schemes by which a sequence of vectors, $\{x^{(k)}\}$, is generated starting from some initial guess $x^{(0)}$. The particular methods vary in the details, but all generally provide some means of updating the vector sequence until the approximate solution is found, provided convergence has been

established. An advantage of iterative methods is that the solution can be terminated when the desired accuracy is achieved, whereas the direct methods will find the "exact" solution. This is to say that the direct methods are finding the true solution to within round-off errors due to machine precision. Iterative methods used to solve a system of linear equations include Jacobi, Gauss-Seidel, successive overrelaxation (SOR), and alternating-direction implicit (ADI) methods [Johnson,1982]. These particular methods generally require some sort of *splitting*, involving a decomposition of the matrix A into more than one matrix. The sparse nature of the system leads to the consideration of a more specialized approach. Another form of iterative solution is that of optimization. Optimization methods include the method of steepest descent and the conjugate gradient method.

The method of steepest descent and the conjugate gradient method fall into the category of line minimization algorithms [Press,1989]. These algorithms provide an improvement to quasi-Newton methods for the solution of nonlinear equations as discussed by Johnson [1982], and are easily extended to the solution of a system of linear equations.

The description of the method of steepest descent is taken from Press [1989]. In two dimensions the steepest descent concept is analogous to finding the bottom of a valley from some point on a mountain. The direction of travel is governed by the direction of steepest slope. This direction is maintained until the path begins to ascend, and then a new direction is chosen in the direction of steepest descent. This process continues until the bottom is reached. The method of steepest descent may be summarized as follows: Consider the starting point to be $\vec{P}_0$. Move from the point $\vec{P}_i$ to the point $\vec{P}_{i+1}$ by minimizing along

a line from $\vec{P}_i$ in the direction of the fastest downhill slope. The slope in any direction of a function at a particular point in N dimensions is given by the gradient of the function. The gradient provides the direction outward normal to the surface at $\vec{P}_i$, and since a descent path is desired, the negative gradient or $-\nabla f(\vec{P}_i)$ is used. A difficulty arises when this approach is applied to a complicated function. The steepest descent method requires a perpendicular direction of travel at each step chosen arising from computation of the gradient. Many iterations may be required to reach the minimum of a for some functions which change rapidly in a particular direction, but not in other directions.

Press [1989] indicates that a better approach is that of choosing a *conjugate* direction of travel. A conjugate, or non-interfering, direction does not spoil the minimization as does the method of steepest descent. Given that at some step, a direction $\hat{d}_i$ was traveled to a minimum. The new direction, $\hat{d}_{i+1}$, will not spoil the minimization along $\hat{d}_i$ if the gradient remains perpendicular along the path indicated by $\hat{d}_{i+1}$. This is to say that the change in the gradient remains perpendicular to $\hat{d}_i$. The gradient of a general function, $f(\vec{x})$, of N variables can be calculated be considering a Taylor series about the origin, $\vec{P}$, to second order. By taking a first derivative of the Taylor expansion, the gradient is approximated by

$$\nabla f = H\vec{x} - \vec{b}, \tag{4.15}$$

where H represents a matrix whose components are the second partial derivatives, at the origin, known as the Hessian matrix. The vector $\vec{b}$ represents

the gradient evaluated at the origin. The gradient given by Eq. (4.15) changes as a particular path is traveled by the amount

$$\delta(\nabla f) = H(\delta\vec{x}). \tag{4.16}$$

Choosing a conjugate direction of travel is the same as requiring

$$\hat{d}_i \delta(\nabla f) = \hat{d}_i^T \; H \; \hat{d}_{i+1} = 0. \tag{4.17}$$

Two vectors $\hat{d}_i$ and $\hat{d}_{i+1}$ are said to be conjugate when Eq. (4.17) is satisfied. A set of vectors is a conjugate set when this relation holds for each vector in the set. It is desirable to choose a direction of travel $\hat{d}_{i+1}$ that is conjugate to *all* previous directions to provide a non-interfering direction of travel. Conjugate gradient methods accomplish this construction and provide an approach to function minimization.

The conjugate gradient method is based on the following theorem given by Press [1989]: Let A be a symmetric positive definite $n \times n$ matrix. Let g_0 be an arbitrary vector and $h_0 = g_0$. Define the following sequence of vectors for $i = 1,2, \ldots$

$$g_{i+1} = g_i - \lambda_i A \cdot h_i \qquad h_{i+1} = g_{i+1} + \gamma_i \cdot h_i \tag{4.18}$$

where λ_i and γ_i are chosen to make $g_{i+1} \cdot g_i = 0$ and $h_{i+1} \cdot A \cdot h_i = 0$. Premultiplying by g_i and $h_i \cdot A$ respectively in Eq. (4.18) allows an explicit representation for γ_i and λ_i , namely

$$\lambda_i = \frac{g_i \cdot g_i}{g_i \cdot A \cdot h_i} \qquad \text{and} \qquad \gamma_i = -\frac{g_{i+1} \cdot A \cdot h_i}{h_i \cdot A \cdot h_i}. \tag{4.19}$$

For all $i \neq j$ we have that

$$g_i \cdot g_j = 0 \text{ and } h_i \cdot A \cdot h_j = 0. \tag{4.20}$$

Press indicates that the procedure of Eq. (4.18) is a sort of Gram-Schmidt bi-orthogonalization that makes each g_i orthogonal, and each h_i conjugate, to its immediate predecessor. This procedure generates a sequence of orthogonal vectors, g_i, and conjugate set of vectors, h_i. Press details a theorem that shows that the gradient of $f(P_i)$ is obtained without explicit knowledge of the Hessian. Instead the slope information can be obtained from Eq. (4.18). The calculation of g_i represents the steepest descent direction, orthogonal to previous g_i 's, but *not* conjugate. The vector h_i represents the actual directions along which the minimization proceeds. The new x_{i+1} if found by moving a distance λ_i along h_i from the previous point x_i, that is, $x_{i+1} = x_i + \lambda_i h_i$.

To consider the use of conjugate gradient methods for the solution of a system of linear equations $Ax = b$, the following function is minimized:

$$f(x) \equiv \frac{1}{2} |Ax - b|^2. \tag{4.21}$$

The gradient of f for this problem is determined by

$$\nabla f(x) = A^T \cdot (A \cdot x - b). \tag{4.22}$$

The subroutine SPARSE in Press [1989] was modified for implementation on sparse systems, and was successful in finding the solution of the linear system of equations corresponding to Eq. (4.13). The modification focused on forming the products Ax and $A^T y$, required for computing the gradient in Eq. (4.22), where $y = (A \cdot x - b)$ Using row-wise compressed pointer storage, the products can be formed without storing the large number of zeros arising in Eq. (4.13).

The direct and indirect techniques take advantage of the sparsity by using a special storage scheme as described by Young [1981] with compressed row pointer storage. The advantage to this storage method over banded storage methods is that the number of zeros that are stored can be greatly reduced if not completely eliminated. The use of compressed row pointer storage has been applied in both direct [Duff,1986] and iterative methods [Young, 1971]. The representation of the matrix A is reduced to a row-wise description where the nonzero entries are stored as real values in a *vector* $A(*)$. The entries corresponding to nonzero locations are mapped by two vectors $jA(*)$ and $iA(*)$. Column numbers are stored in the integer vector $jA(*)$ where $jA(k)$ is the column number for the entry $A(k)$. The vector $iA(*)$ is used to denote the starting locations of contiguous blocks where the starting block for row i is given by $iA(i)$, and the end by $iA(i+1)-1$. The number of nonzero entries in row i is obtained by forming the difference $iA(i+1)-iA(i)$. As an example of this storage method, consider the real nonsymmetric matrix M

$$\mathrm{M} = \begin{bmatrix} 21 & 0 & 5 & -1 \\ 0 & 23 & 5 & 0 \\ 2 & 11 & 15 & 0 \\ 1 & 0 & 0 & 31 \end{bmatrix},$$

whose nonzero entries are given by

$$A(*) = [21, 5, -1, 23, 5, 2, 11, 15, 1, 31].$$

The column locations for the nonzero entries is given by

$$jA(*) = [1, 3, 4, 2, 3, 1, 2, 3, 1, 4]$$

and the locations of contiguous row blocks given by

$$iA(*) = [1, 4, 6, 9, 11] \quad .$$

The length of $iA(*)$ for N entries in $A(*)$ is $(N+1)$ in order to denote the full length of $A(*)$.

To close this section some brief observations can be made. The use of the conjugate gradient method using compressed pointer storage removes the need for a particular node ordering as required by Reddy [1984] to obtain a banded form. The $iA(*)$ and $jA(*)$ vectors contain valuable information about the mesh itself. To determine the number of nodes connected to node k, examine the difference $iA(k+1) - iA(k)$. The global nodes connected to node k are the nodes $jA(iA(k))$ through $jA(iA(k+1)-1)$. This information is most useful in the actual formation of these vectors and are obtained from the mesh generated as described in Chapter 3. After the mesh is generated and boundary conditions are applied, the location of the entries in the vector $A(*)$ are completely determined from the mesh structure.

4.5 RESULTS FOR LAPLACE'S EQUATION

A problem of interest to demonstrate the techniques developed thus far is that of determining the electrostatic capacitance of various geometries. The variational statement of the problem can be written so that the solution becomes stationary about the capacitance [Itoh,1989; Daly,1984]. For some geometrical configurations application of a quasi-static approximation is made to determine the characteristic impedance, where comparison to calculations based on published results can be made. The electrostatic functional can be written

$$\mathrm{F}(\phi) = \int_{\Omega} \epsilon \, |\nabla\phi|^2 \, d\Omega \; = 0 \;\; , \tag{4.23}$$

where the unknown is the potential ϕ. The permittivity is given as $\epsilon = \epsilon_r \epsilon_0$, where ϵ_r is the relative permittivity and ϵ_0 is the permittivity of free space. As was indicated previously Eq. (4.23) represents the energy stored in the electric field and can be equated to the capacitance by the following [Daly,1984] :

$$C = \frac{\epsilon_0}{V^2} \int_{\Omega} \epsilon_r \, |\nabla\phi|^2 \, d\Omega \; . \tag{4.24}$$

V is the potential difference on the boundary. The exact value of capacitance will be obtained for only the exact potential distribution. The value of capacitance found from (4.24) will be an upper bound to the true value since any approximation $\hat{\phi}$ to ϕ, meeting the boundary conditions, will satisfy $\mathrm{F}(\hat{\phi}) \geq \mathrm{F}(\phi)$. Using the finite element solution, the value of capacitance can be found directly from using the assembled global matrix K and the solution ϕ in the following manner [Silvester,1990] :

$$\frac{C}{\epsilon_0} = \vec{\phi}^T K \, \vec{\phi} \, , \tag{4.25}$$

where $\vec{\phi}$ is the solution vector for a potential difference of 1.0 volt between conducting surfaces $\partial\Omega_1$ and $\partial\Omega_2$. The ϵ_r is that of the medium separating the conductors, and is built into the matrix K. The characteristic impedance of a two dimensional transmission line [Collin,1966] is $Z_o = \sqrt{\mu\epsilon}/C$, where $\mu = \mu_0 = 4\pi \cdot 10^{-7}$ is the permeability of free space for a nonmagnetic media.

Consider the coaxial air line with the cross-section of Fig. 4.2 where the ratio of b, the outer radius, to a, the inner radius, is given by $b/a = 2$. The mesh is shown in Fig. 4.3 where the number of elements is 384 and the total number of nodes is 222, with 60 of these nodes being boundary nodes. Solving by finite element techniques, the potential distribution of Fig. 4.4 is obtained, and the capacitance from Eq. (4.25) is found to be C/ϵ_0= 9.1180. This can be compared to the value obtained from basic electromagnetic concepts as

$$\frac{C}{\epsilon_0} = \frac{2\pi\epsilon_r}{\ell n(b/a)} \, ,$$

resulting in $C/\epsilon_0 = 9.0647$ for a 0.584 % difference between the approximate and exact value. For a ratio of $b/a = 10$, the error increases to 3.49 % since more of the space is discretized for the same number of elements and nodes.

More interesting is the result for stripline transmission lines where a solution for the characteristic impedance is not available in closed exact form. Fig. 4.5 shows the geometry of the striplines under consideration. There are a number of different formulae for computing the characteristic impedance of striplines [Gupta,1979]. These are, however, limited in scope and subject to

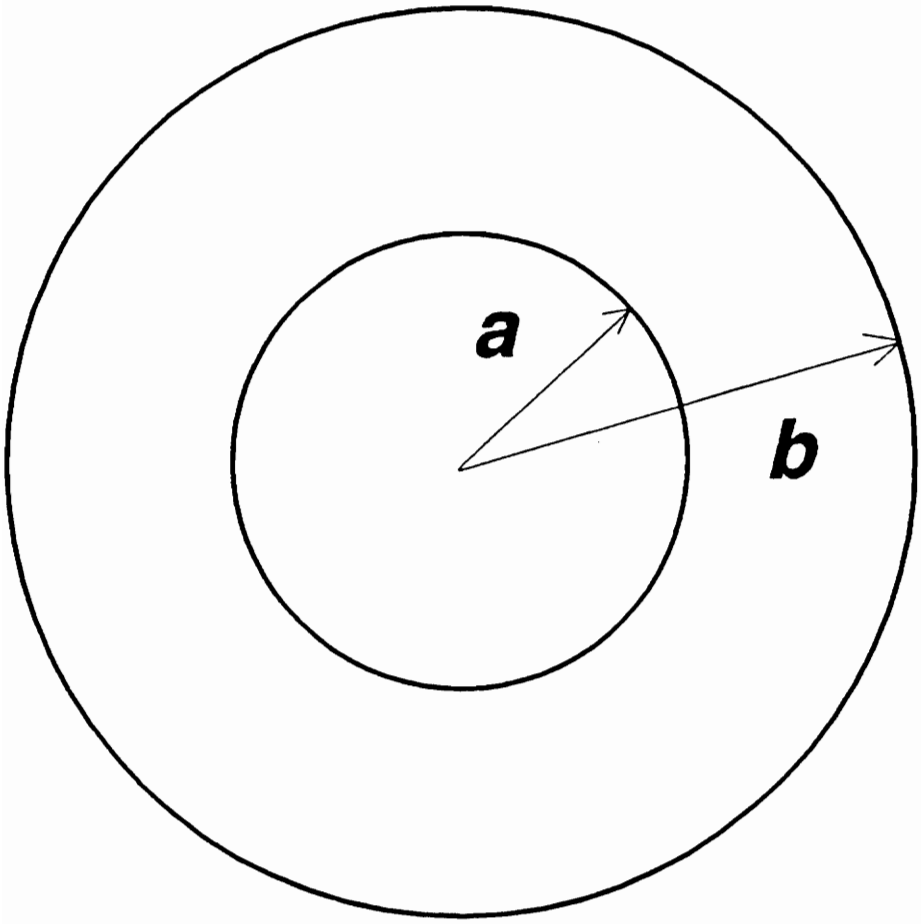


Figure 4.2. Coaxial geometry.

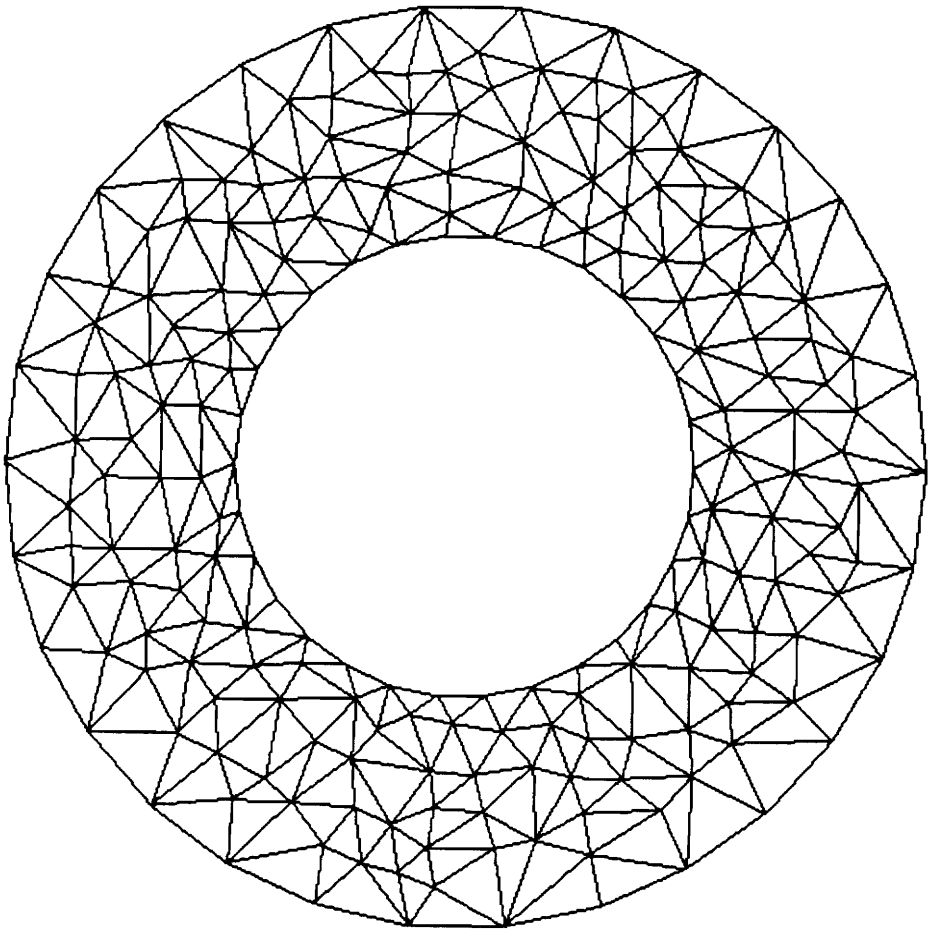


Figure 4.3. Sample mesh.

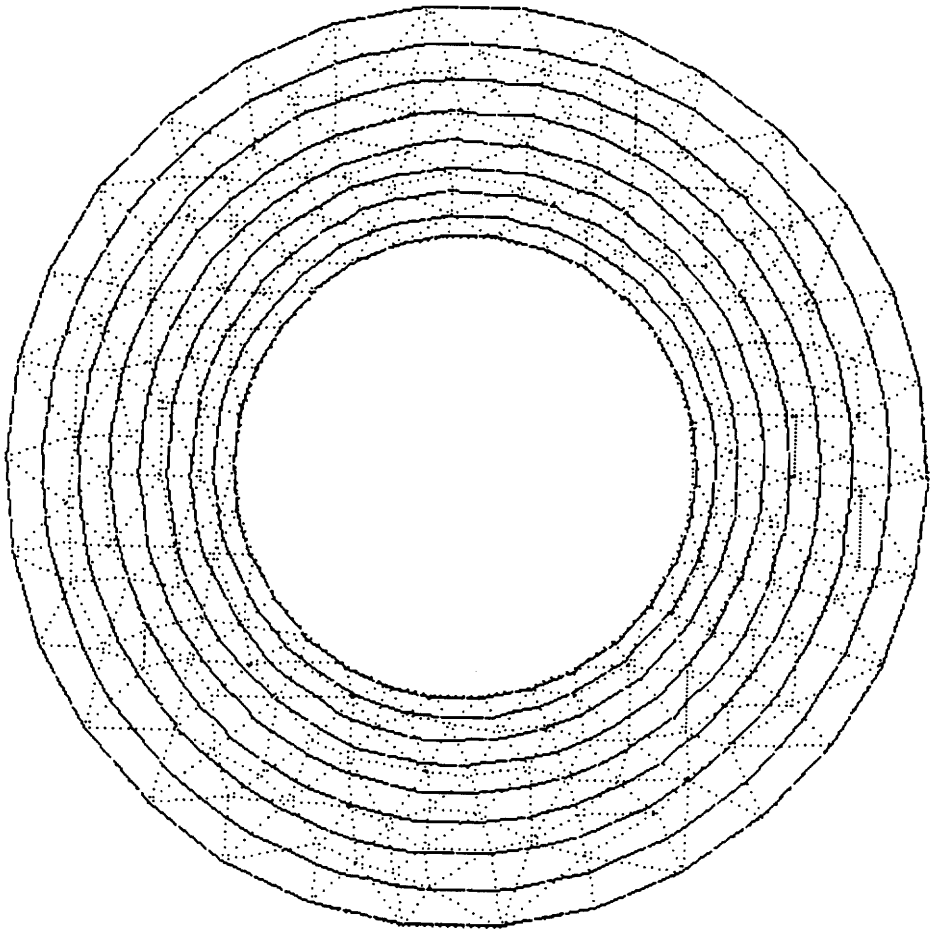


Figure 4.4. Equipotential contours for mesh of Fig. 4.3.

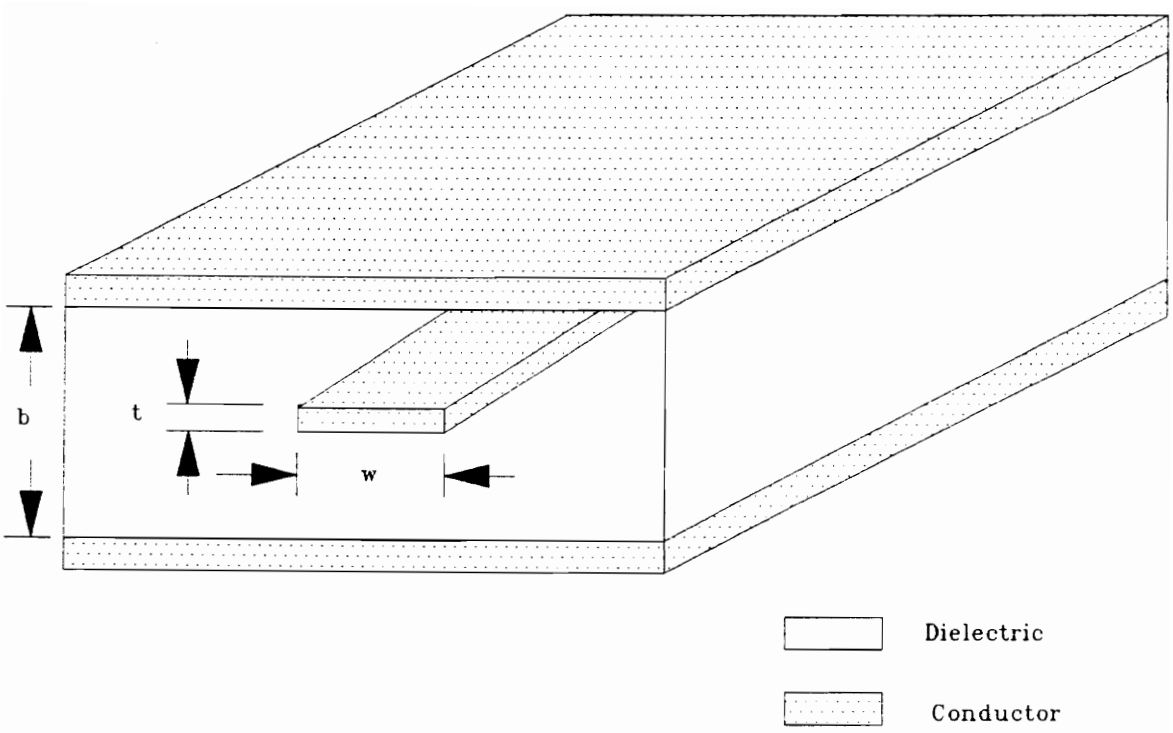


Figure 4.5. Stripline geometry.

varying degrees of accuracy due to the approximations involved. The finite element method will be seen as useful in determining the Z_o. Referring to the geometry of Fig. 4.5, w is the width of the line, t is the thickness of the center conductor, and b is the separation between the top and bottom ground plates with all dimensions in μm. The materials used in the striplines are low loss, low permittivity materials typical of those used in industry for the construction of microwave components and devices. The results obtained for various geometries and materials are summarized in Table 4.1 with the number of elements and nodes also listed. Z_o is found from $Z_o = \frac{\sqrt{\epsilon\mu}}{C}$ [Silvester,1990].

Table 4.1

ϵ_r	w	t	b	# elem.	# nodes	$Z_{o(FEM)}$	$Z_{o(ref.)}$
3.0	500	70	270	258	175	15.9	16.02
3.0	250	70	270	348	220	25.3	25.4
9.1	750	9.6	74.2	2162	1075	2.66	2.57

The characteristic impedance reference value, $Z_{o(ref)}$, was computed from the formulas of Cohn [1955] for ϵ_r of 3.0, while the higher relative permittivity value of 9.1 dielectric required the formulations of Gupta [1979] due to the extreme dimensions. It can be seen that the results are very good as compared to the approximate calculations found in Cohn and Gupta. The finite element method can be employed as a tool to enable the determination of useful modeling parameters for complex geometrical configurations where analytical results are not available.

The static and quasi-static approach to determine circuit parameters such as capacitance or characteristic impedance provide good results when the loss is small. The quasi-static result is also useful for analyzing structures with multiple dielectric regions, though not pursued for this work. The static and quasi-static approach are only valid for a small component of the propagating field in the direction of propagation. Typically the quasi-static approach is limited to structures which can support a nearly TEM field. The effects of a small loss can be incorporated in a perturbational sense in TEM structures with good results. The effects of small losses and interfacial/surface roughness in stripline structures have been modeled by a quasi-static approximation to solve the inverse problem of determining the dielectric constant of low loss dielectrics by Davis [1992] with good results. To consider the modeling of non-TEM structures such as cylindrical waveguides or high loss striplines, the finite element functional must be extended beyond the representation given in Eq. (4.2), written explicitly in Eqs. (4.3) and (4.4). The next chapter will address the "full-field" functional and consider the subsequent numerical implementation to find the propagation characteristics of guiding structures as well as address some difficulties that arise in solving the vector field problem.

Chapter 5

Full-Field Variational Form

5.1 INTRODUCTION

This chapter addresses the need for a more thorough physical understanding of basic functionals used in solving electromagnetic problems. Specifically a full field functional is developed from basic electromagnetic principles without applying a methodology such as a simple quadratic functional approach [Stakgold,1968] used by several authors [Jeng,1984; Hazel,1972; Konrad,1976]. These functionals are used to obtain a field solution using a variational process making the functional either minimum or stationary. A commonly used approach is the Galerkin or weighted residual method where a weighting function is used to provide a weak equality between an operation on a set of the expansion or shape functions and a corresponding forcing function in an integral sense [Hayata,1988; Costache,1987; Biro,1989]. Although Jeng and Wexler [1984] has shown that the Galerkin method is equivalent to the quadratic functional form under certain conditions, both forms fail to provide adequate insight as to the physical process of stationarity or minimization. The soundness of either methodology is not questioned, as the mathematical basis for each is well established and for many years have yielded excellent results in application to electromagnetic problems for a wide variety of interesting problems. This chapter will provide a connection between the various methods and the physical process in terms of energy relations for the electromagnetic field derived from a fundamental point of view [Davis,1992].

Chen [1980], Morgan [1977], and Paris [1965] indicate a useful connection between the energy relationships and the particular functional form obtained. Chen [1980] pointed to the necessity of making the reactive powers stationary as the key to functional minimization where the reactive powers are related to the difference of the average stored energy plus the source reactive power. Morgan [1977] provides additional insight in specifying that the functional obtained is associated with the difference of the time varying and time average radiated power densities, but indicated that further consideration of the physical mechanisms is required. The properties of the electromagnetic field energy in the electrostatic, magnetostatic, and time-varying cases were considered by Paris [1965]. Paris indicated that the stationarity of a functional for the static cases results in charges and direct currents, for the corresponding electrostatic and magnetostatic cases respectively, that distribute themselves in such a way as to minimize the static energy of the resultant field solutions. For the time varying case, Paris suggests the charges and currents orient themselves in such a way as to minimize the transfer of power.

The approach used in this chapter will start with Maxwell's equations and utilize a development of the related energy to yield an expression that can be used in a finite element or related method to minimize an energy functional. The energy in the system is minimized, or at least made stationary, in a sense we shall examine in detail, while satisfying Maxwell's equations. Specifically it is desired to show a more reasonable physical basis for the particular functional typically used in numerical techniques in electromagnetics. Finite element techniques will then be used to find propagation constant of a waveguide, using the obtained functional, with a discussion of spurious solutions.

5.2 FUNCTIONAL DEVELOPMENT

An appropriate starting point for determining an electromagnetic functional is Maxwell's equations in the time domain

$$\nabla \times \vec{\mathcal{H}}(\vec{r},t) = \epsilon \frac{\partial \vec{\mathcal{E}}(\vec{r},t)}{\partial t} + \mathcal{J}(\vec{r},t) \tag{5.1a}$$

$$\nabla \times \vec{\mathcal{E}}(\vec{r},t) = -\mu \frac{\partial \vec{\mathcal{H}}(\vec{r},t)}{\partial t}\,, \tag{5.1b}$$

which have been restricted to simple, linear, isotropic, and homogeneous media for our treatment. It is possible to obtain a form of conservation of energy by scalar multiplication (1a) with $\vec{\mathcal{E}}$ and (1b) with $\vec{\mathcal{H}}$ and taking their difference to obtain

$$\nabla \cdot (\vec{\mathcal{E}} \times \vec{\mathcal{H}}) = -\frac{1}{2}\left(\mu \frac{\partial}{\partial t}(\vec{\mathcal{H}} \cdot \vec{\mathcal{H}}) + \epsilon \frac{\partial}{\partial t}(\vec{\mathcal{E}} \cdot \vec{\mathcal{E}})\right) - \vec{\mathcal{J}} \cdot \vec{\mathcal{E}} \tag{5.2}$$

Integrating this expression over an arbitrary volume V containing electric sources $\vec{\mathcal{J}}$, the following is obtained after applying the divergence theorem:

$$\oint_{\partial V} (\vec{\mathcal{E}} \times \vec{\mathcal{H}}) \cdot d\vec{s} + \int_V \left[\frac{1}{2}\left(\mu \frac{\partial}{\partial t}|\vec{\mathcal{H}}| + \epsilon \frac{\partial}{\partial t}|\vec{\mathcal{E}}|\right) + \vec{\mathcal{J}} \cdot \vec{\mathcal{E}}\right] dV \equiv 0\,, \tag{5.3}$$

where ∂V represents the surface bounding V. Eq. (5.3) represents the integral form of the conservation of energy by recalling that

$$\text{Energy} = \int_0^t (\text{Power})\, dt \tag{5.4}$$

resulting in

$$\int_0^t \oint_{\partial V} (\vec{\mathcal{E}} \times \vec{\mathcal{H}}) \cdot d\vec{s}\, dt + \int_0^t \int_V \vec{\mathcal{J}} \cdot \vec{\mathcal{E}}\; dv dt + \frac{1}{2} \int_0^t \int_V (\mu \frac{\partial}{\partial t} |\vec{\mathcal{H}}|^2 + \epsilon \frac{\partial}{\partial t} |\vec{\mathcal{E}}|^2)\; dv dt = 0 \tag{5.5}$$

It is possible to describe the integrals in (5.5) respectively as

$$I_1 = \int_0^t \oint_{\partial V} (\vec{\mathcal{E}} \times \vec{\mathcal{H}}) \cdot d\vec{s}\; dt \text{ , the energy leaving } V \text{ through } \partial V \text{ ,} \tag{5.6}$$

$$I_2 = \int_0^t \int_V \vec{\mathcal{J}} \cdot \vec{\mathcal{E}}\; dv dt \text{ , the energy supplied by sources } \vec{\mathcal{J}} \text{ ,} \tag{5.7}$$

and

$$I_3 = \frac{1}{2} \int_0^t \int_V (\mu \frac{\partial}{\partial t} |\vec{\mathcal{H}}|^2 + \epsilon \frac{\partial}{\partial t} |\vec{\mathcal{E}}|^2)\; dv dt \text{ , the energy stored in } V \text{ .} \tag{5.8}$$

For the time-harmonic case, the field terms may be replaced with

$$\vec{\mathcal{E}}(\vec{r}, t) = \Re\left\{ \vec{E}(\vec{r})\, e^{j\omega t} \right\} = \frac{1}{2} (\vec{E} e^{j\omega t} + \vec{E}^* e^{-j\omega t}) \text{ ,} \tag{5.9a}$$

$$\vec{\mathcal{H}}(\vec{r}, t) = \frac{1}{2} (\vec{H} e^{j\omega t} + \vec{H}^* e^{-j\omega t}) \text{ ,} \tag{5.9b}$$

and

$$\vec{\mathcal{J}}(\vec{r}, t) = \frac{1}{2} (\vec{J} e^{j\omega t} + \vec{J}^* e^{-j\omega t}) \text{ ,} \tag{5.9c}$$

where $*$ represents the complex conjugate. Considering the time-harmonic form of each integral separately and performing the time integration we obtain

$$I_1 = \frac{1}{4} \left\{ \frac{e^{j2\omega t} - 1}{j2\omega} \oint_{\partial V} \vec{E} \times \vec{H} \cdot d\vec{s} - \frac{e^{-j2\omega t} - 1}{j2\omega} \oint_{\partial V} \vec{E}^* \times \vec{H}^* \cdot d\vec{s} \right.$$

$$\left. + t \oint_{\partial V} (\vec{E} \times \vec{H}^* + \vec{E}^* \times \vec{H}) \cdot d\vec{s} \right\} , \tag{5.10}$$

for the energy flowing across the bounding surface denoted by ∂V. For the energy due to sources in V we obtain

$$I_2 = \frac{1}{4}\left\{ \frac{e^{j2\omega t} - 1}{j2\omega} \int_V \vec{J} \cdot \vec{E}\, dv - \frac{e^{-j2\omega t} - 1}{j2\omega} \int_V \vec{J}^* \cdot \vec{E}^*\, dv \right.$$

$$\left. + t \int_V (\vec{J} \cdot \vec{E}^* + \vec{J}^* \cdot \vec{E})\, dv \right\}. \tag{5.11}$$

Similarly, the energy stored in V can be rewritten as

$$I_3 = \frac{1}{8} \int_V \left\{ \mu[\vec{H} \cdot \vec{H}(e^{j2\omega t} - 1) + \vec{H}^* \cdot \vec{H}^*(e^{-j2\omega t} - 1)] \right.$$

$$\left. + \epsilon[\vec{E} \cdot \vec{E}(e^{j2\omega t} - 1) + \vec{E}^* \cdot \vec{E}^*(e^{-j2\omega t} - 1)] \right\} dv. \tag{5.12}$$

We may sum the three integrals and group terms by time dependence to obtain

$$\frac{1}{4}\left(e^{j2\omega t} \left\{ \frac{1}{j2\omega} \oint_{\partial V} \vec{E} \times \vec{H} \cdot d\vec{s} + \frac{1}{j2\omega} \int_V \vec{J} \cdot \vec{E}\, dv + \frac{1}{2} \int_V (\mu \vec{H} \cdot \vec{H} + \epsilon \vec{E} \cdot \vec{E})\, dv \right\} \right.$$

$$- e^{-j2\omega t} \left\{ \frac{1}{j2\omega} \oint_{\partial V} \vec{E}^* \times \vec{H}^* \cdot d\vec{s} + \frac{1}{j2\omega} \int_V \vec{J}^* \cdot \vec{E}^* dv - \frac{1}{2} \int_V (\mu \vec{H}^* \cdot \vec{H}^* + \epsilon \vec{E}^* \cdot \vec{E}^*) dv \right\}$$

$$+ t \left\{ \oint_{\partial V} (\vec{E} \times \vec{H}^* + \vec{E}^* \times \vec{H}) \cdot d\vec{s} + \int_V (\vec{J} \cdot \vec{E}^* + \vec{J}^* \cdot \vec{E})\, dv \right\}$$

$$- \left\{ \frac{1}{j2\omega} \oint_{\partial V} (\vec{E} \times \vec{H} - \vec{E}^* \times \vec{H}^*) \cdot d\vec{s} + \frac{1}{j2\omega} \int_V (\vec{J} \cdot \vec{E} - \vec{J}^* \cdot \vec{E}^*) dv \right.$$

$$\left. \left. + \frac{1}{2} \int_V \left(\mu(\vec{H} \cdot \vec{H} + \vec{H}^* \cdot \vec{H}^*) + \epsilon(\vec{E} \cdot \vec{E} + \vec{E}^* \cdot \vec{E}^*) \right) dv \right\} \right) = 0 . \tag{5.13}$$

The first two terms with the $e^{\pm j2\omega t}$ time dependence may be associated with the oscillatory transfer of energy between the electric and magnetic fields. This

oscillatory energy form is similar to that obtained by Jeng and Chen [1984] in forming a variational principle for solving exterior electromagnetic problems by the finite element method. The third term with a simple t dependence represents the energy available for dissipation/supply with both source and radiated energy shown explicitly. The fourth and fifth terms, with no time dependence, represent the total energy stored in the system in a time average sense and are closely related to the oscillatory energy terms. It is important to note that each of the grouped terms must independently equate to zero due to the independent time-variation. By examining the field/supply energy terms with no explicit time dependence we are able to form a variational principle similar to Jeng and Chen.

Examining the fourth and fifth terms of (5.13), we identify each of the energy terms separately in order to more easily develop an appropriate functional. The oscillatory energy $\mathcal{E}_F$ may be denoted by

$$\mathcal{E}_F = \frac{1}{4\omega^2\mu} \int_V \left\{ k^2 \, \vec{E} \cdot \vec{E} \; - (\nabla \times \vec{E}) \cdot (\nabla \times \vec{E}) \right\} dv \; , \tag{5.14}$$

where $k^2 = \omega^2 \mu\epsilon$ and the time harmonic form of (5.1b) was used to obtain (5.14). The time average energy stored in the fields within the volume V, given by the real part of $\mathcal{E}_F$ denoted $\Re(\mathcal{E}_F)$, is given by

$$\Re(\mathcal{E}_F) = \frac{1}{8} \int_V \left\{ \mu(\, \vec{H} \cdot \vec{H} + \, \vec{H}^* \cdot \vec{H}^*) \; + \; \epsilon(\vec{E} \cdot \vec{E} \; + \; \vec{E}^* \cdot \vec{E}^*) \right\} dv \; , \tag{5.15}$$

The stored energy from the source, denoted by $\Re(\mathcal{E}_J)$, is represented by

$$\Re(\mathcal{E}_J) = \frac{1}{8\omega} \int_V \left(\frac{\vec{J} \cdot \vec{E} - \vec{J}^* \cdot \vec{E}^*}{j} \right) dv = \frac{1}{4} \, \Re \int_V \frac{\vec{J} \cdot \vec{E}}{j\omega} \, dv, \tag{5.16}$$

and energy stored outside the volume, $\Re(\mathcal{E}_R)$, is

$$\Re(\mathcal{E}_R) = \frac{1}{4}\,\Re \oint_{\partial V} \frac{\vec{E}\times\vec{H}}{j\omega}\cdot d\vec{s} = \frac{1}{4}\,\Re \oint_{\partial V} \left[\, \frac{1}{\omega^2\mu}\,\vec{E}\times(\nabla\times\vec{E})\,\right]\cdot d\vec{s}\ . \tag{5.17}$$

In order to solve for the field, it is common to approximate $\vec{E}$ by a form such as

$$\vec{E} \approx \sum_n (a_n + jb_n)\ \vec{E}_n\ , \tag{5.18}$$

where a_n and b_n are real unknown coefficients to be determined and the $\vec{E}_n$ are known approximation functions.

The proposed approach to developing an appropriate functional for finding a variational solution leads to the minimization of the energy stored in the fields under the constraint that the total stored energy is identically zero. The method of Lagrange multipliers is well suited to such problems and can be written as

$$F(\vec{E}) = \Re(\mathcal{E}_F + \mathcal{E}_R) + \lambda\ \Re(\mathcal{E}_F + \mathcal{E}_R + \mathcal{E}_J)\ , \tag{5.19}$$

where $\Re(\mathcal{E}_F + \mathcal{E}_R + \mathcal{E}_J) = 0$ is the desired constraint. The Lagrange multiplier λ may be determined by taking partial derivatives of F with respect to a_n, b_n, and λ to obtain a set of independent equations. For $\Re(\mathcal{E}_F)$, we obtain the derivatives as

$$\frac{\partial}{\partial a_n}\,\Re(\mathcal{E}_F) = 2\ \Re(\mathcal{E}_{Fn}) = \frac{2}{4\omega^2\mu}\,\Re \int_V \left\{\, k^2\ \vec{E}_n\cdot\vec{E}\ -\ (\nabla\times\vec{E}_n)\cdot(\nabla\times\vec{E})\right\} dv \tag{5.20a}$$

$$\frac{\partial}{\partial \mathrm{b}_n} \Re(\mathcal{E}_F) = 2\ \Im(\mathcal{E}_{Fn}) = -\frac{2}{4\omega^2\mu}\ \Im \int_V \left\{ k^2\ \vec{E}_n \cdot \vec{E}\ -\ (\nabla \times \vec{E}_n) \cdot (\nabla \times \vec{E}) \right\} dv. \qquad (5.20b)$$

For $\Re(\mathcal{E}_J)$, we likewise obtain

$$\frac{\partial}{\partial \mathrm{a}_n} \Re(\mathcal{E}_J) = \Re(\mathcal{E}_{J_n}) = \frac{1}{4}\ \Re \int_V \frac{\vec{J} \cdot \vec{E}_n}{j\omega}\ dv \qquad (5.21a)$$

$$\frac{\partial}{\partial \mathrm{b}_n} \Re(\mathcal{E}_J) = \Im(\mathcal{E}_{J_n}) = -\frac{1}{4}\ \Im \int_V \frac{\vec{J} \cdot \vec{E}_n}{j\omega}\ dv\ . \qquad (5.21b)$$

Finally, for $\Re(\mathcal{E}_R)$ we obtain

$$\frac{\partial}{\partial \mathrm{a}_n} \Re(\mathcal{E}_R) = 2\ \Re(\mathcal{E}_{Rn}) = \frac{2}{4\omega^2\mu}\ \Re \left\{ \frac{1}{2} \oint_{\partial V} \left[\vec{E}_n \times (\nabla \times \vec{E}) + \vec{E} \times (\nabla \times \vec{E}_n) \right] \cdot d\vec{s} \right\} \qquad (5.22a)$$

$$\frac{\partial}{\partial \mathrm{b}_n} \Re(\mathcal{E}_R) = 2\ \Im(\mathcal{E}_{Rn}) = -\frac{2}{4\omega^2\mu}\ \Im \left\{ \frac{1}{2} \oint_{\partial V} \left[\vec{E}_n \times (\nabla \times \vec{E}) + \vec{E} \times (\nabla \times \vec{E}_n) \right] \cdot d\vec{s} \right\}. \qquad (5.22b)$$

Enforcing both $\frac{\partial F}{\partial \mathrm{a}_n}$ and $\frac{\partial F}{\partial \mathrm{b}_n}$ to be zero results in the combined equation

$$(2\ +\lambda)\ (\mathcal{E}_{Fn} + \mathcal{E}_{Rn}) = -\lambda\ (\mathcal{E}_{Fn} + \mathcal{E}_{Rn} + \mathcal{E}_{Jn})\ \text{, for all } n. \qquad (5.23)$$

In addition, the derivative with respect to the Lagrangian multiplier gives

$$\frac{\partial F}{\partial \lambda} = \Re(\mathcal{E}_F + \mathcal{E}_R + \mathcal{E}_J) = 0 \qquad (5.24)$$

which is the original constraint. Summing over n in (5.23), the right hand side is zero due to (5.24) if we impose the same constraint on the imaginary part to require $\lambda = -2$. The resultant functional obtained with $\lambda = -2$ is

$$F = \mathcal{E}_F + \mathcal{E}_R - 2\,(\mathcal{E}_F + \mathcal{E}_R + \mathcal{E}_J), \tag{5.25}$$

or

$$F = -\frac{1}{4\omega^2\mu}\int_V \left\{ k^2\,\vec{E}\cdot\vec{E} - (\nabla\times\vec{E})\cdot(\nabla\times\vec{E})\right\} dv$$

$$-\frac{1}{4\omega^2\mu}\oint_{\partial V}\left\{\vec{E}\times(\nabla\times\vec{E})\right\}\cdot d\vec{s} - \frac{2}{4}\int_V \frac{\vec{J}\cdot\vec{E}}{j\omega}\,dv\,, \tag{5.26}$$

where the use of the real part is no longer required. This form is identical to the form suggested by Collin [1991], with the symmetric product $\vec{E}\cdot\vec{E}$ instead of $\vec{E}\cdot\vec{E}^*$. This form is used by Silvester [1990], Paulsen [1988], Chew [1989], and Webb [1983] in the context of solving electromagnetic problems by the finite element method.

Applying a Ritz procedure by taking partial derivatives with respect to the free complex variable $c_n(=\mathrm{a}_n + j\mathrm{b}_n)$, Eq. (5.26) yields an expression that includes both the real and imaginary parts by satisfying the complex expression

$$\frac{\partial F}{\partial c_n} = 2\,(\mathcal{E}_{Fn} + \mathcal{E}_{Rn}) - 2\,[2\,(\mathcal{E}_{Fn} + \mathcal{E}_{Rn}) + \mathcal{E}_{Jn}] = 0\,, \tag{5.27a}$$

or

$$2\,(\mathcal{E}_{Fn} + \mathcal{E}_{Rn} + \mathcal{E}_{Jn}) = 0\,. \tag{5.27b}$$

Writing (5.27b) in terms of the expansion functions, we obtain

$$\frac{2}{4\omega^2\mu}\int_V \left\{k^2\ \vec{E}_n\cdot\vec{E}\ -(\nabla\times\vec{E})\cdot(\nabla\times\vec{E}_n)\right\}dv\ +\frac{2}{4}\int_V \frac{\vec{J}\cdot\vec{E}_n}{j\omega}\,dv$$

$$+\ \frac{1}{4\omega^2\mu}\oint_{\partial V}\left\{\ \vec{E}_n\times(\nabla\times\vec{E})\ +\ \vec{E}\times(\nabla\times\vec{E}_n)\cdot d\vec{s}\ \ =\ 0\ , \qquad (5.28)$$

which may be rearranged into common weighted residual forms as

$$\frac{1}{\omega^2\mu}\int_V \vec{E}_n\cdot\left\{(k^2+\nabla^2)\vec{E}-j\omega\mu\vec{J}-\nabla(\nabla\cdot\vec{E})\right\}dv$$

$$+\ \frac{1}{2\omega^2\mu}\oint_{\partial V}[\vec{E}\times(\nabla\times\vec{E}_n)-\vec{E}_n\times(\nabla\times\vec{E})]\cdot d\vec{s}]=0, \qquad (5.29a)$$

or

$$\frac{1}{\omega^2\mu}\int_V \vec{E}_n\cdot\left\{k^2\ \vec{E}-\nabla\times\nabla\times\vec{E}\ -j\omega\mu\vec{J}\right\}dv$$

$$+\ \frac{1}{2\omega^2\mu}\oint_{\partial V}[\vec{E}\times(\nabla\times\vec{E}_n)-\vec{E}_n\times(\nabla\times\vec{E})]\cdot d\vec{s}\ \ =\ 0. \qquad (5.29b)$$

The boundary surface integral may not be set to zero a priori, but may become the fundamental part of the problem to be solved. The treatment of this boundary integral depends on the specific method to be used in solution.

5.3 RECTANGULAR WAVEGUIDE PROBLEM

This section will examine the functional derived in section 5.2 as applied to a rectangular waveguide where the expansion functions used are not finite

element polynomials of compact support, but entire domain functions that satisfy the boundary conditions of the waveguide exactly. The source free functional given by Eq. (5.26) may be rewritten as

$$F = \int_V \left\{ k^2 \, \vec{E} \cdot \vec{E} - (\nabla \times \vec{E}) \cdot (\nabla \times \vec{E}) \right\} dv + \oint_{\partial V} \left\{ \vec{E} \times (\nabla \times \vec{E}) \right\} \cdot d\vec{s} \, . \qquad (5.30)$$

Consider a slice of volume of width Δz along the direction of propagation z of a waveguide where the bounding surface δV is given by $\delta V = \mathrm{S}_1 + \mathrm{S}_2 + \mathrm{S}_3$, S_1 is the bounding curve comprised of $\mathrm{C} \cdot \Delta z$, S_2 is the front face at $\frac{\Delta z}{2}$, and S_3 is the back face at $-\frac{\Delta z}{2}$. Eq. (5.30) then becomes

$$F = \int_S \int_{\Delta z} \left\{ k^2 \, \vec{E} \cdot \vec{E} - (\nabla \times \vec{E}) \cdot (\nabla \times \vec{E}) \right\} ds \, dz$$

$$+ \int_c \int_{\Delta z} \vec{E} \times (\nabla \times \vec{E}) \cdot \hat{n} \, d\ell \, ds$$

$$+ \int_{S_2} \vec{E} \times (\nabla \times \vec{E}) \cdot \hat{z} \, ds + \int_{S_3} \vec{E} \times (\nabla \times \vec{E}) \cdot (-\hat{z}) \, ds. \qquad (5.31)$$

Due to the z-independence it is convenient to express the electric field as

$$\vec{E}(x,y,z) = \vec{e}(x,y) \, \mathrm{e}^{-\gamma z}$$

with $\gamma = \alpha + j\beta$, the propagation constant of the guide. Using the decomposed operator $\nabla = \nabla_t + \hat{z}\frac{\partial}{\partial z}$, the representation for $\vec{E}$ becomes an integral over the cross sectional surface S as

$$F = \int_S \int_{\Delta z} e^{-2\gamma z} \left\{ k^2 \, \vec{e} \cdot \vec{e} \; - (\nabla_t \times \vec{e} \; - \gamma \hat{z} \times \vec{e}) \cdot (\nabla_t \times \vec{e} \; - \gamma \hat{z} \times \vec{e}) \right\} ds \, dz$$

$$+ \int_c \int_{\Delta z} e^{-2\gamma z} \left\{ \vec{e} \times (\nabla_t \times \vec{e} \; - \gamma \hat{z} \times \vec{e}) \right\} \cdot \hat{n} \, d\ell \, dz$$

$$+ \int_S (e^{-\gamma \Delta z} - e^{+\gamma \Delta z}) \, \vec{e} \times (\nabla_t \times \vec{e} \; - \gamma \hat{z} \times \vec{e}) \cdot \hat{z} \, ds \; . \qquad (5.32)$$

Integrating the dz integrals Eq. (5.32) is rewritten

$$F = \frac{\sinh(\gamma z)}{\gamma} \int_S \left\{ k^2 \, \vec{e} \cdot \vec{e} \; - (\nabla_t \times \vec{e} \; - \gamma \hat{z} \times \vec{e}) \cdot (\nabla_t \times \vec{e} \; - \gamma \hat{z} \times \vec{e}) \right\} ds$$

$$+ \frac{\sinh(\gamma z)}{\gamma} \int_c \vec{e} \times (\nabla_t \times \vec{e} \; - \gamma \hat{z} \times \vec{e}) \cdot \hat{n} \, d\ell$$

$$- \; 2 \sinh(\gamma z) \int_S \vec{e} \times (\nabla_t \times \vec{e} \; - \gamma \hat{z} \times \vec{e}) \cdot \hat{z} \, ds \; . \qquad (5.33)$$

Multiplying through by the *reciprocal* of $\frac{\sinh(\gamma z)}{\gamma}$ and employing vector identities Eq. (5.33) leads to a redefinition of F as

$$F = \int_S \left\{ k^2 \, \vec{e} \cdot \vec{e} \; - (\nabla_t \times \vec{e} \;) \cdot (\nabla_t \times \vec{e} \;) + \gamma^2 (\hat{z} \times \vec{e}) \cdot (\hat{z} \times \vec{e}) \right\} ds$$

$$+ \int_c \left\{ (\hat{n} \times \vec{e}) \cdot (\nabla_t \times \vec{e} \; - \gamma \hat{z} \times \vec{e}) \right\} d\ell \; . \qquad (5.34)$$

A significant feature of this formulation is the absence of the propagation constant γ in a linear form as several cancellations were made in obtaining Eq. (5.34). At first appearance this form is beneficial since the solution for the

propagation constant appears as a generalized eigenvalue problem $Ax = \lambda Bx$ with $\lambda = \gamma^2$. The second term in the functional given in Eq. (5.34) can be set to zero for the special case of perfectly conducting waveguide walls. This particular term in the functional representation will allow the inclusion of imperfect conductors through an impedance boundary condition type formulation. Such a formulation replaces $(\nabla_t \times \vec{e} \; - \gamma \hat{z} \times \vec{e})$ in Eq. (5.34) with $- jk_c(\hat{n} \times \vec{e})$, where k_c is the wave number in the metal related to the skin effect and the $\hat{n} \times \vec{e}$ term appears as part of the solution, contributing to the eigenvector corresponding to the obtained eigenvalue γ.

For the discussion to follow, we consider a perfectly conducting rectangular waveguide filled with a homogeneous, isotropic, dielectric medium with properties ϵ and μ, with dimension a × b as shown in Fig. 5.1. The desired quantity to be determined is the γ of the guide. This problem was chosen since the exact solution is well known and provides a verification of the variational form suggested for numerical implementation. The functional is explicitly given by

$$F = k^2 \int_S (e_x^2 + e_y^2 + e_z^2) \, ds \; + \gamma^2 \int_S (e_x^2 + e_y^2) \, ds$$

$$- \int_S \left\{ \left(\frac{\partial e_z}{\partial y}\right)^2 + \left(\frac{\partial e_z}{\partial x}\right)^2 + \left(\frac{\partial e_y}{\partial x}\right)^2 + \left(\frac{\partial e_x}{\partial y}\right)^2 - 2\frac{\partial e_x}{\partial y}\frac{\partial e_y}{\partial x} \right\} ds \; . \qquad (5.35)$$

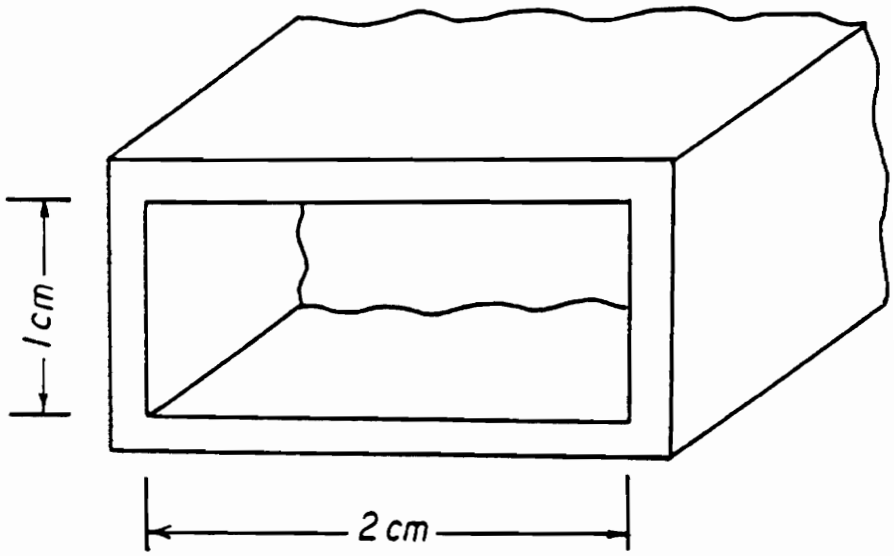


Figure 5.1 Waveguide geometry [Bucca,1989].

The surface integration in Eq. (5.35) can be written

$$\int_S ds \rightarrow \int_0^a dx \int_0^b dy \ .$$

Trial functions that satisfy the boundary conditions exactly are

$$e_z = \sum_m \sum_n A_{mn} \sin\left(\frac{m\pi x}{a}\right) \sin\left(\frac{n\pi y}{b}\right) \ , \tag{5.36a}$$

$$e_y = \sum_m \sum_n B_{mn} \sin\left(\frac{m\pi x}{a}\right) \cos\left(\frac{n\pi y}{b}\right) \ , \tag{5.36b}$$

and

$$e_x = \sum_m \sum_n C_{mn} \cos\left(\frac{m\pi x}{a}\right) \sin\left(\frac{n\pi y}{b}\right) \ . \tag{5.36c}$$

For convenience F can be written

$$\begin{aligned} F = & k^2 \int_S {e_z}^2 \, ds \ + (k^2+\gamma^2) \int_S (\, {e_x}^2 + {e_y}^2 \,) \, ds \\ & - \int_S \left\{ \left(\frac{\partial e_y}{\partial x}\right)^2 - 2\frac{\partial e_x}{\partial y}\frac{\partial e_y}{\partial x} + \left(\frac{\partial e_x}{\partial y}\right)^2 \right\} ds \\ & - \int_S \left\{ \left(\frac{\partial e_z}{\partial y}\right)^2 + \left(\frac{\partial e_z}{\partial x}\right)^2 \right\} ds \ . \end{aligned}$$

Using the expansion of (5.36), we obtain

$$F=\sum_p \sum_q \Bigg\{ k^2 \int_S \left[A_{pq} \sin\left(\frac{p\pi x}{a}\right) \sin\left(\frac{q\pi y}{b}\right) \right]^2 ds$$

$$+ (k^2+\gamma^2) \left\{ \int_S \left[B_{pq} \sin\left(\frac{p\pi x}{a}\right) \cos\left(\frac{q\pi y}{b}\right) \right]^2 ds + \int_S \left[C_{pq} \cos\left(\frac{p\pi x}{a}\right) \sin\left(\frac{q\pi y}{b}\right) \right]^2 ds \right\}$$

$$- \int_S \Bigg\{ \left[B_{pq} \left(\frac{p\pi}{a}\right) \cos\left(\frac{p\pi x}{a}\right) \cos\left(\frac{q\pi y}{b}\right) \right]^2 + \left[C_{pq} \left(\frac{q\pi}{b}\right) \cos\left(\frac{p\pi x}{a}\right) \cos\left(\frac{q\pi y}{b}\right) \right]^2$$

$$- 2 \left[C_{pq} \left(\frac{q\pi}{b}\right) \cos\left(\frac{p\pi x}{a}\right) \cos\left(\frac{q\pi y}{b}\right) \right] \left[B_{pq} \left(\frac{p\pi}{a}\right) \cos\left(\frac{p\pi x}{a}\right) \cos\left(\frac{q\pi y}{b}\right) \right] \Bigg\} ds \qquad (5.37)$$

$$- \int_S \left\{ \left[A_{pq} \left(\frac{p\pi}{a}\right) \cos\left(\frac{p\pi x}{a}\right) \sin\left(\frac{q\pi y}{b}\right) \right]^2 + \left[A_{pq} \left(\frac{q\pi}{b}\right) \sin\left(\frac{p\pi x}{a}\right) \cos\left(\frac{q\pi y}{b}\right) \right]^2 \right\} ds \Bigg\} .$$

The first variation with respect to the free variables can be obtained by setting

$$\frac{\partial F}{\partial A_{mn}} = \frac{\partial F}{\partial B_{mn}} = \frac{\partial F}{\partial C_{mn}} = 0 ,$$

and making use of orthogonality. Omitting details, the results are

$$\frac{\partial F}{\partial A_{mn}} = \left(\frac{ab}{2}\right) A_{mn} \left[k^2 - \left(\frac{m\pi}{a}\right)^2 - \left(\frac{n\pi}{b}\right)^2 \right] = 0 \qquad (5.38)$$

$$\frac{\partial F}{\partial B_{mn}} = \left(\frac{ab}{2}\right) \left[(k^2+\gamma^2) B_{mn} - \left(\frac{m\pi}{a}\right)^2 B_{mn} - \left(\frac{mn\pi^2}{ab}\right) C_{mn} \right] = 0 \qquad (5.39)$$

$$\frac{\partial F}{\partial C_{mn}} = \left(\frac{ab}{2}\right) \left[(k^2+\gamma^2) C_{mn} - \left(\frac{n\pi}{b}\right)^2 C_{mn} - \left(\frac{mn\pi^2}{ab}\right) B_{mn} \right] = 0 . \qquad (5.40)$$

The resulting system of equations is of the form $Mx = 0$, or

$$\begin{bmatrix} k^2 - \left(\frac{m\pi}{a}\right)^2 - \left(\frac{n\pi}{b}\right)^2 & 0 & 0 \\ 0 & (k^2 + \gamma^2) - \left(\frac{m\pi}{a}\right)^2 & \left(\frac{mn\pi^2}{ab}\right) \\ 0 & \left(\frac{mn\pi^2}{ab}\right) & (k^2 + \gamma^2) - \left(\frac{n\pi}{b}\right)^2 \end{bmatrix} \begin{bmatrix} A_{mn} \\ B_{mn} \\ C_{mn} \end{bmatrix} = 0, \quad (5.41)$$

with the solution corresponding to that which makes the determinant of M vanish or

$$k^2 - \left(\frac{m\pi}{a}\right)^2 - \left(\frac{n\pi}{b}\right)^2 = 0 \,, \quad (5.42)$$

and

$$(k^2 + \gamma^2)\left[(k^2 + \gamma^2) - \left(\frac{m\pi}{a}\right)^2 - \left(\frac{n\pi}{b}\right)^2\right] = 0 \,. \quad (5.43)$$

This result indicates that one solution is the zero eigenvalue which corresponds to cutoff for the field component associated with the A_{mn} term, namely, the e_z component. This false requirement for A_{mn} would suggest that the TM mode could not be modeled by the functional denoted by Eq. (5.30). A TEM can also be supported from the first bracketed term in Eq. (5.43). The other components of the electric field propagate with propagation for the m, n mode according to

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

which is the classic form for rectangular guides. These results indicate that use of the non-conjugated form will not necessarily yield the correct eigenvalue.

An interesting result is obtained if the functional is slightly modified to allow for a product of the form $\vec{E}\cdot\vec{E}^*$ as described by

$$\widetilde{F} = \int_V \left\{ k^2\, \vec{E}\cdot\vec{E}^* \;-\; (\nabla\times\vec{E})\cdot(\nabla\times\vec{E}^*)\right\} dv + \oint_{\partial V} \left\{ \vec{E}^*\times(\nabla\times\vec{E})\right\}\cdot d\vec{s}\;, \qquad (5.44)$$

to suggest the use of power rather than energy from Eq. (5.13). This expression is a weighted residual form that results from

$$\widetilde{F} = \int_V \left\{ k^2\, \vec{E} \;-\; \nabla\times\nabla\times\vec{E}\right\}\cdot\vec{E}^* dv.$$

The use of Eq. (5.44) is also discussed by Collin[1991] and others [Davies,1982; Hayata,1989; Konrad,1976] as one typically used in finite element codes. Following a similar procedure of reducing the dimensionality of the integral as accomplished earlier, with the stipulation

$$\vec{E}(x,y,z) = \vec{e}(x,y)\, \mathrm{e}^{-\gamma z}, \qquad (5.45a)$$

and

$$\vec{E}^*(x,y,z) = \vec{e}^*(x,y)\, \mathrm{e}^{-\gamma^* z}, \qquad (5.45b)$$

results in

$$F = \frac{\alpha}{\sinh\alpha\Delta}\,\widetilde{F} \text{ or}$$

$$F = \int_S \Big\{ k^2\, \vec{e}\cdot\vec{e}^* \;-(\nabla_t\times\vec{e}\,)\cdot(\nabla_t\times\vec{e}^*\,)$$

$$+\,\gamma\left[(\hat{z}\times\vec{e})\cdot(\nabla_t\times\vec{e}^*) - (\hat{z}\times\vec{e}^*)\cdot(\nabla_t\times\vec{e})\right] + \gamma^2(\hat{z}\times\vec{e})\cdot(\hat{z}\times\vec{e}^*)\Big\}\, ds$$

$$+ \int_c \left\{(\hat{n}\times\vec{e}^*)\cdot(\nabla_t\times\vec{e}\; - \gamma\hat{z}\times\vec{e})\right\} d\ell\;. \qquad (5.46)$$

Neglecting the line integral for perfectly conducting walls, this functional corresponds to English's [1971] functional for $\gamma = jk_c$ or a lossless medium, $\alpha = 0$. Taking the first variation with respect to the real and imaginary parts of the free variables and setting them equal to zero results in the following system of equations:

$$\begin{bmatrix} k^2 - \left(\frac{m\pi}{a}\right)^2 - \left(\frac{n\pi}{b}\right)^2 & -\gamma\left(\frac{n\pi}{b}\right) & -\gamma\left(\frac{m\pi}{a}\right) \\ +\gamma\left(\frac{n\pi}{b}\right) & (k^2+\gamma^2) - \left(\frac{m\pi}{a}\right)^2 & \left(\frac{mn\pi^2}{ab}\right) \\ +\gamma\left(\frac{m\pi}{a}\right) & \left(\frac{mn\pi^2}{ab}\right) & (k^2+\gamma^2) - \left(\frac{n\pi}{b}\right)^2 \end{bmatrix} \begin{bmatrix} A_{mn} \\ B_{mn} \\ C_{mn} \end{bmatrix} = 0. \quad (5.47)$$

Finding the determinant of the matrix of Eq. (5.47) and setting this equal to zero results in the single eigenvalue

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

The appearance of the cutoff mode does not appear in this case and suggests that use of the functional of Eq. (5.44) will provide for the modeling of both TE and TM modes.

This conclusion is suggestive of some error introduced in the energy formulation approach of section 5.2 as well as adding a note of caution to the use of a functional form which neglects the use of a complex inner product definition in applications for guided wave structures. The use of a real inner product causes no difficulty in modeling cavity type structures [Davies,1982] where the boundary term of Eq. (5.30) can be set to zero provided that $\hat{n} \times \vec{e} = 0$ or $\nabla \times \hat{n} \times \vec{e} = 0$ on the cavity walls. Although it was possible to specify $\hat{n} \times \vec{e} = 0$

on the bounding curve of the waveguide there is no such specification on the front and back faces of the bounding surface δV. It is not apparent at this time as to the fundamental difficulty in expressing the minimization in terms of energy concepts as in section 5.2, but work is continuing in the hope of providing insight into the actual minimization process in terms of the energy.

5.4 FINITE ELEMENT IMPLEMENTATION OF FULL FIELD VARIATIONAL FORM

This section will attempt to model Eq. (5.46) in a finite element context. The problem of interest is that of two dimensional waveguides with perfectly conducting walls. The idea of a two dimensional waveguide has been introduced in the previous section by considering the longitudinal field to have z dependence as in Eq. (5.45a,b). The domain of interest will be modeled by a mesh of three node triangular elements as discussed previously. Neglecting the line integral in Eq. (5.46), we obtain

$$F = \int_S \Big\{ k^2\, \vec{e}\cdot\vec{e}_n^{\,*} - (\nabla_t \times \vec{e})\cdot(\nabla_t \times \vec{e}_n^{\,*}) + \gamma\left[(\hat{z}\times\vec{e})\cdot(\nabla_t\times\vec{e}_n^{\,*}) - (\nabla_t\times\vec{e})\cdot(\hat{z}\times\vec{e}_n^{\,*})\right] + \gamma^2(\hat{z}\times\vec{e})\cdot(\hat{z}\times\vec{e}_n^{\,*})\Big\}\, ds. \tag{5.48}$$

Hayata [1988] has suggested a useful implementation, based on the element by element formulation discussed in chapter 4. The present work is a modification of Hayata's work. The electric field can be represented by

$$\vec{E} \rightarrow [\mathrm{N}]^T \{e\}^{(el)}\, exp(-\gamma z)\,. \tag{5.49}$$

The superscript (el) is used to emphasize that the matrices are defined per element. The matrix $[\mathrm{N}]^T$ is a 3×9 matrix given by

$$[\mathrm{N}]^T = \begin{bmatrix} \{\Psi\}^T & \{0\}^T & \{0\}^T \\ \{0\}^T & \{\Psi\}^T & \{0\}^T \\ \{0\}^T & \{0\}^T & \{\Psi\}^T \end{bmatrix}. \tag{5.50}$$

The vector $\{\Psi\}$ is a 3×1 vector of element shape functions,

$$\{\Psi\} = \begin{bmatrix} \psi_1(x,y) \\ \psi_2(x,y) \\ \psi_3(x,y) \end{bmatrix}, \tag{5.51}$$

where $\psi_i(x,y)$ is given by Eq. (4.9). The vector $\{e\}^{(el)}$ is a 9×1 vector containing the x, y, and z components of the three nodes comprising a particular element, that is,

$$\{e\}^{(el)} = \begin{bmatrix} \{e_x\} \\ \{e_y\} \\ \{e_z\} \end{bmatrix}, \tag{5.52}$$

and

$$\{e_i\} = \begin{bmatrix} e_{i1} \\ e_{i2} \\ e_{i3} \end{bmatrix}, \tag{5.53}$$

a 3×1 vector of i components of nodes 1,2, and 3 of a particular element.

The curl of the electric field can be represented by

$$\nabla \times \vec{E} \rightarrow [\mathrm{G}]^T \{e\}^{(el)} exp(-\gamma z). \tag{5.54}$$

Where the matrix $[\mathrm{G}]^T$ is a 3×9 matrix given by,

$$[\mathrm{G}]^T = \begin{bmatrix} \{0\}^T & \gamma\{\Psi\}^T & \frac{\partial\{\Psi\}^T}{\partial y} \\ -\gamma\{\Psi\}^T & \{0\}^T & -\frac{\partial\{\Psi\}^T}{\partial x} \\ -\frac{\partial\{\Psi\}^T}{\partial y} & \frac{\partial\{\Psi\}^T}{\partial x} & \{0\}^T \end{bmatrix}. \tag{5.55}$$

The curl operation in Eq. (5.55) above can be expressed as a transverse curl and a cross product, each required in the representation of (5.48). The transverse curl can be written

$$(\nabla_t \times \vec{e}) = \begin{bmatrix} \{0\}^T & \{0\}^T & \frac{\partial\{\Psi\}^T}{\partial y} \\ \{0\}^T & \{0\}^T & -\frac{\partial\{\Psi\}^T}{\partial x} \\ -\frac{\partial\{\Psi\}^T}{\partial y} & \frac{\partial\{\Psi\}^T}{\partial x} & \{0\}^T \end{bmatrix} \{e\}^{(el)} = [\mathrm{G}_t]^T \{e\}^{(el)}. \tag{5.56}$$

The cross product can be written

$$(\hat{z} \times \vec{e}) = [\mathrm{Cr}]^T \{e\}^{(el)} = \begin{bmatrix} \{0\}^T & -\{\Psi\}^T & \{0\}^T \\ \{\Psi\}^T & \{0\}^T & \{0\}^T \\ \{0\}^T & \{0\}^T & \{0\}^T \end{bmatrix} \{e\}^{(el)}. \tag{5.57}$$

Eq. (5.48) can be arranged in powers of γ to yield an expression

$$\{e_n^*\}^{T(el)} \left\{ \gamma^2 A_0^{(el)} + \gamma A_1^{(el)} + A_2^{(el)} \right\} \{e\}^{(el)} = 0 \, . \tag{5.58}$$

The expressions for $A_0^{(el)}$, $A_1^{(el)}$, and $A_2^{(el)}$ are

$$A_0^{(el)} = \int_\Omega [\mathrm{Cr}]^* [\mathrm{Cr}]^T \, d\Omega \, , \tag{5.59a}$$

$$A_{1a}^{(el)} = \int_\Omega \left\{ [\mathrm{G}_t]^* [\mathrm{Cr}]^T - [\mathrm{Cr}]^* [\mathrm{G}_t]^T \right\} d\Omega \, , \tag{5.59b}$$

and

$$A_2^{(el)} = \int_\Omega \left\{ k^2 \, [\mathrm{N}^*][\mathrm{N}]^T - [\mathrm{G}_t]^* [\mathrm{G}_t]^T \right\} d\Omega \, . \tag{5.59c}$$

The 9×9 element matrices are

$$A_0^{(el)} = \begin{bmatrix} A & 0 & 0 \\ 0 & A & 0 \\ 0 & 0 & 0 \end{bmatrix}, \tag{5.60}$$

$$A_{1a}^{(el)} = \begin{bmatrix} 0 & 0 & B \\ 0 & 0 & C \\ -B^T & -C^T & 0 \end{bmatrix}, \tag{5.61}$$

and

$$A_2^{(el)} = \begin{bmatrix} k^2 A - D & E & 0 \\ E^T & k^2 A - F & 0 \\ 0 & 0 & k^2 A - D - F \end{bmatrix}. \tag{5.62}$$

The entries $A, B, C, D, E,$ and F in Eqs. (5.60)-(5.62) are each 3×3 matrices that result from the integration over an element given by,

$$A = \int_{\Omega} \{\Psi\}^* \{\Psi\}^T dxdy \ , \tag{5.63}$$

$$B = \int_{\Omega} \{\Psi\}^* \frac{\partial \{\Psi\}^T}{\partial x} dxdy \ , \tag{5.64}$$

$$C = \int_{\Omega} \{\Psi\}^* \frac{\partial \{\Psi\}^T}{\partial y} dxdy \ , \tag{5.65}$$

$$D = \int_{\Omega} \frac{\partial \{\Psi\}^*}{\partial y} \frac{\partial \{\Psi\}^T}{\partial y} dxdy \ , \tag{5.66}$$

$$E = \int_{\Omega} \frac{\partial \{\Psi\}^*}{\partial y} \frac{\partial \{\Psi\}^T}{\partial x} dxdy \ , \tag{5.67}$$

and

$$F = \int_{\Omega} \frac{\partial \{\Psi\}^*}{\partial x} \frac{\partial \{\Psi\}^T}{\partial x} dxdy \ . \tag{5.68}$$

The evaluation of Eqs. (5.63)-(5.68) is described in detail in Appendix I.

The expression for $A_{1a}^{(el)}$ in Eq. (5.61) can be symmetrized to ease the eigenvalue computation by redefining $\{e\}$ as:

$$\{e\} = \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & jI \end{bmatrix} \{f\}. \tag{5.69}$$

The matrix $A_{1a}^{(el)}$ in Eq. (5.61) then becomes $A_1^{(el)}$ where

$$A_1^{(el)} = \begin{bmatrix} 0 & 0 & B \\ 0 & 0 & C \\ B^T & C^T & 0 \end{bmatrix}. \tag{5.70}$$

After factoring the j from $\{e\}$,Eq. (5.58) can be written

$$\{f_n^*\}^{T(el)} \left\{ \gamma^2 A_0^{(el)} + j\gamma A_1^{(el)} + A_2^{(el)} \right\} \{f\}^{(el)} = 0 \,, \tag{5.71}$$

where f relates to e by a phase shift in the z component.

To obtain the desired eigenvalue-eigenvector problem the elements in Eq. (5.71) must be appropriately summed over all the elements in a particular mesh. This summation was discussed at length in Chapter 4, and will be symbolically represented as the summation

$$\sum_e \{f_n^*\}^T \left\{ \gamma^2 A_0^{(el)} + j\gamma A_1^{(el)} + A_2^{(el)} \right\} \{f\}^{(el)} = 0 \,. \tag{5.72}$$

This sum leads to the complex quadratic eigenvalue problem with $\gamma = jk_z$ as

$$\left\{ -k_z{}^2 A_0 \; - \; jk_z A_1 + A_2 \right\} \{f\} = 0 \tag{5.73}$$

after the appropriate boundary conditions are imposed on the tangential components of the electric field. Eq. (5.73) is sometimes called a lambda-matrix eigenvalue problem [Lancaster,1966] of order $\ell = 2$ where $D_\ell(\lambda)$ is

$$D_\ell(\lambda) \equiv A_0 \, \lambda^\ell + A_1 \lambda^{\ell-1} + \cdots + A_{\ell-1}\lambda + A_\ell \,. \tag{5.74}$$

The solution of eigenvalue problems is an interesting topic in its own right, and is the subject of several excellent reviews [Haley,1988; Peters,1970; Ruhe,1973]. As for the solution of linear equations discussed in Chapter 4, the solution of eigenvalue problems takes the form of either direct or indirect methods.

The direct methods usually involve similarity transforms to obtain a form that has the same eigenvalues but is easier to solve [Johnson,1982]. Direct methods include reduction to Hessenburg form, Householder transformations, and QR factorization. Direct methods are not well suited to the large sparse systems that are obtained from general finite element techniques. Appendix 2 describes the formulation to use direct methods, which are useful for the inverse problem of determining the dielectric constant for a known propagation constant.

The indirect methods are iterative methods that include the power method, inverse power method, Rayleigh quotient techniques [Johnson,1982], Lanczos method [Peters,1970; Lee,1991], and the conjugate gradient method [Chen,1986; Narayanan,1991; Yang,1991]. Iterative methods will be exclusively considered for the solution of Eq. (5.73). The solution of the quadratic complex eigenvalue problem in Eq. (5.73) is often handled by an appropriate change of variables [Lee,1991; Hayata,1988]. This variable transformation can introduce nonphysical solutions which require special consideration. The goal of the variable transformation is to reduce the quadratic eigenvalue problem to the generalized form $Ax = \lambda Bx$. Borri [1977] recommends a *factorization* that converts (5.73) into a generalized eigenvalue problem of the form $Ax = \lambda Bx$. Borri asserts that under certain conditions the standard eigenvalue problem can be obtained of the form $Ax = \lambda x$. Factorization is not practical for the same reason that the direct methods were discarded since the large matrices in (5.73) are irregularly sparse.

Lancaster [1961,1964,1966] proposed an eigenvalue-eigenvector algorithm that can be applied directly to Eq. (5.73). This algorithm is based on a Rayleigh quotient iteration scheme that utilizes derivative information to optimize the

eigenvalue update λ_i at each step of iteration. Lancaster's generalized Rayleigh quotient can be summarized as follows: For arbitrary vectors $\vec{w}$ and $\vec{z}$, form

$$\vec{x}_i = [D_2(\lambda_i)]^{-1}\ \vec{w} \qquad \vec{y}_i = [D_2^T(\lambda_i)]^{-1}\ \vec{z}\ , \tag{5.75}$$

then

$$\lambda_{i+1} = \lambda_i\ -\ R(\vec{x}_i,\ \vec{y}_i,\ \lambda_i). \tag{5.76}$$

The Rayleigh quotient $R(\vec{x}_i,\ \vec{y}_i,\ \lambda_i)$ at step i is given explicitly by

$$R(\vec{x}_i,\ \vec{y}_i,\ \lambda_i) = \frac{\vec{y}_i^T\ [D_2(\lambda_i)]\ \vec{x}_i}{\vec{y}_i^T\ [D_2^{(1)}(\lambda_i)]\ \vec{x}_i}\ . \tag{5.77}$$

The superscript (1) in the denominator represents the first derivative with respect to λ. The procedure requires an initial guess eigenvalue λ_0. The vectors $\vec{x}_i$ and $\vec{y}_i$ are the left and right eigenvectors at step i for a general matrix $D_\ell(\lambda)$. This is a Newton-Raphson iteration step if $D_\ell(\lambda)$ is regular and simple [Lancaster,1966] and quadratic convergence can be guaranteed. The algorithm, upon convergence, produces the final eigenvalue λ_f, and right and left eigenvectors $\vec{x}_f$ and $\vec{y}_f$. Eq. (5.73) can be written in a form so as to meet regularity [Bridges,1984], however, the system is not simple, so rapid convergence cannot be guaranteed. The system in Eq. (5.73) is symmetric so that $\vec{x}_i = \vec{y}_i$ since $D_2(\lambda_i) = D_2^T(\lambda_i)$. This reduces the amount of work required to implement Lancaster's algorithm. This algorithm was programmed and tested on several systems of equations from the literature with fast convergence and accurate results.

5.5 RESULTS AND DISCUSSION

The method of the previous section was applied to a rectangular waveguide filled with a vacuum ($\epsilon_r = 1.0$, $\mu_r = 1.0$) with dimension $a = 2$ cm and $b = 1$ cm as in Fig. 5.1. For a frequency of 10Ghz the only propagating mode is the dominant mode. The propagation constant for this dominant mode is purely imaginary and is given by

$$\beta_{10} = \sqrt{k_0^2 - \left(\frac{\pi}{2}\right)^2} \ ,$$

with $k_0^2 = \omega^2 \epsilon_0 \mu_0 = 4.39256$ so that $\beta_{10} = 1.387503$ rad/cm. The waveguide cross section can be discretized using the ideas of Chapter 3. A sample mesh is shown in Fig. 5.2. The problem statement of Eq. (5.73) can be recast for this example so that the eigenvalue will appear as purely real to simplify the computations. The problem becomes

$$\left\{ -\beta^2 A_0 \ - \ \beta A_1 + A_2 \right\} \{f\} = 0 \ . \tag{5.78}$$

The solution of the waveguide problem exhibits the so-called spurious modes that have plagued the application of finite element techniques to the full-field problem [Paulsen,1991; Hayata,1988]. Tabulated results are in Table 5.1 for three mesh configurations. As the problem size increases the number of solutions to the eigenvalue problem increases as well with the desired β embedded in these solutions. For 88 nodes and 154 elements, the closest β is 1.373, so convergence can be seen, though it is difficult to isolate the "desired" β. Spurious solutions can be seen in Fig. 5.3 where the x axis represents the guess values for the start of Lancaster's eigenvalue solution. The y axis represents the error based on the computation of the Rayleigh quotient as in Eq. (5.77) for the initial guess

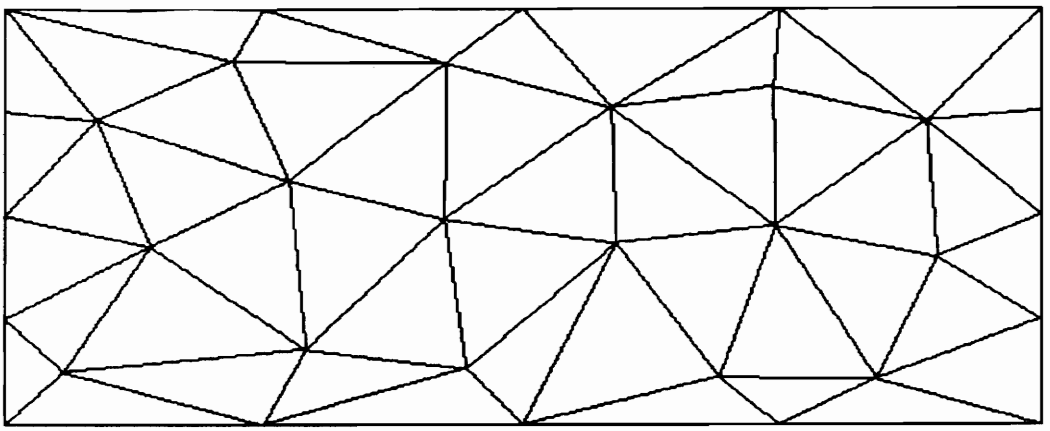


Figure 5.2 Sample waveguide discretization. (48 elements, 33 total nodes)

Table 5.1 Phase constant (β) for rectangular waveguide.

NEM: number of elements, NNM: number of nodes.

NEM	*NNM*	β
4	5	1.180
10	10	1.299 3.810
48	33	0.924 1.364 1.847 2.148 2.703 3.352 3.476 4.129

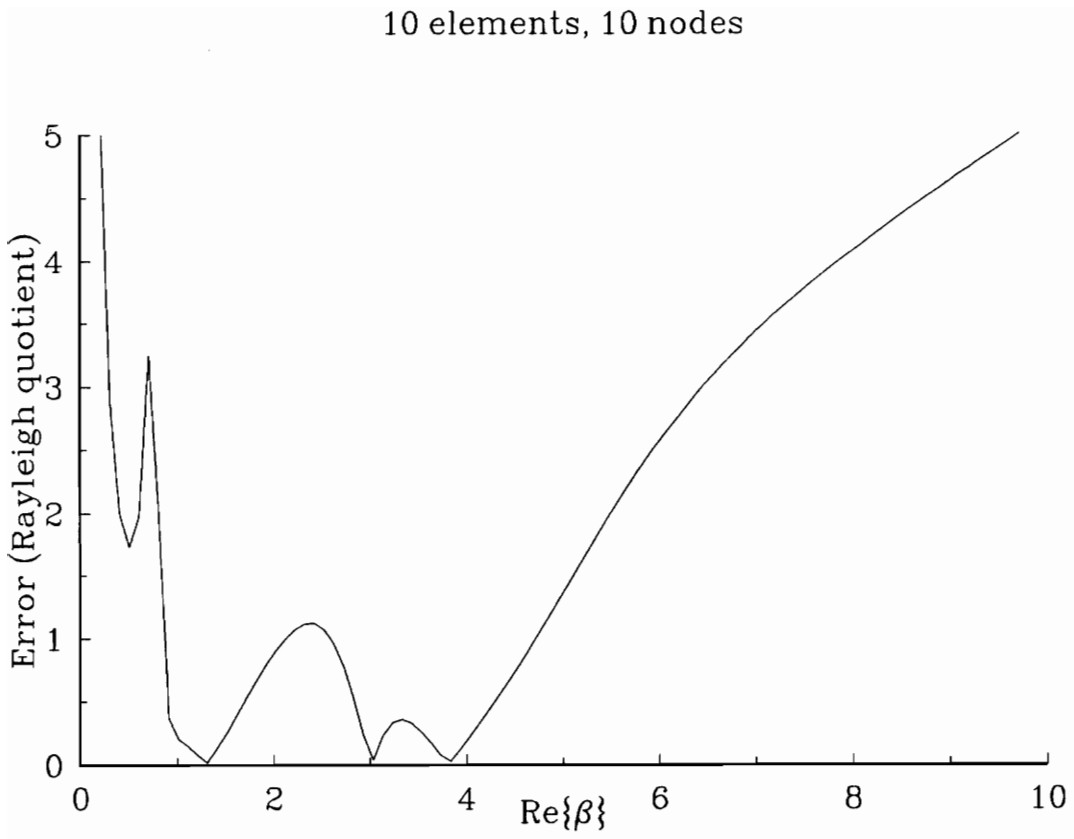


Figure 5.3 Error as a function of the initial guess.

λ_0, given by

$$Error = R(\vec{x}_1, \vec{x}_1, \lambda_0) = \frac{\vec{x}_1^{\dagger} [D_2(\lambda_0)] \vec{x}_1}{\vec{x}_1^{\dagger} [D_2^{(1)}(\lambda_0)] \vec{x}_1},$$

where † represents the conjugate transpose, and $D_2(\lambda_0)$ given by Eq. (5.74). The minima of the curve in Fig. 5.3 represent the eigenvalue closest to the initial guess along the real axis. The eigenvalue solver will accurately pick out many of these minima as the solution to the desired accuracy (typically 5 or 6 decimal places). Some minima are a result of the initial guess eigenvector and will change when randomizing this eigenvector. The failure to select a single accurate eigenvalue suggests that further work is required.

The issue of spurious modes has been of interest for a number of years [Paulsen,1991]. The two primary methods of eliminating these modes center around the use of penalty methods or the use of edge elements as basis functions.

Penalty methods [Paulsen,1991; Hayata,1988; Koshiba,1985] are methods used to force the divergence of the field quantity to be zero over the solution domain. The divergence is cast as part of the weighted residual statement of the problem with a unknown experimental parameter used as a weight. The parameter depends upon the geometry and domain discretization and is usually determined through numerical experimentation. The penalty method was motivated by the observation of a nonzero divergence for the field obtained from the eigenvectors whose eigenvalues were considered as spurious solutions [Koshiba,1985]. The divergence for each eigenvalue in the solution to Eq. (5.78) was computed. The divergence was of the same order for each eigenvalue found, though not particularly small for any eigenvalue, suggesting that the cause of these undesirable solutions is not based on a divergence criteria.

The use of edge elements was introduced recently [Bossavit,1988] as a means of properly enforcing the continuity requirements along interelement boundaries. The node based elements used in this thesis enforce the continuity of both the normal and tangential components at interelement boundaries. This assumption is tacitly made when performing the assembly process of section 4.3. The spurious modes are attributed to the overspecification of the field at interelement boundaries. Edge elements are elements based on enforcement of the continuity of the tangential component of the field along element edges. Edge elements have been successfully applied to the problem of spurious modes [Koshiba,1992; Lee,1991] with the added benefit of reducing the overall system size.

The use of penalty methods and edge elements provide a means of obtaining good results. The penalty method seems to lack a theoretical basis to provide a fundamental insight into the problem of spurious modes. Edge elements provide some insight into the fundamental problem. Lee [1991] indicates that the node based elements provide an inappropriate mapping of the null space of the curl operator into the solution domain. The basis functions of section 5.3 did not require any special handling to obtain a correct solution.

Future work will involve a more thorough functional development to attempt to answer some of the physical questions raised in section 5.3 regarding the breakdown of the non-conjugate form of the variational formulation. The concerns raised by the proper function space requirements implied by edge elements also need to be addressed.

Chapter 6

Summary and Conclusions

This thesis has addressed the many issues associated with finite element techniques that are often mentioned in the electromagnetics literature, but not emphasized. By pulling together the ideas of mesh generation, sparse storage solution techniques, and functional development in a single volume, this work provides basic tools for implementation of finite element techniques for both static and dynamic problems in electromagnetics.

This thesis has developed the mathematical basis of the finite element method and discussed the application of the method of weighted residuals in a finite element context. The finite element techniques were applied to several problems of interest in electromagnetics.

The use of mesh generation techniques was explored and applied to achieve a discretization that was random in spatial variation. The concepts underlying the ideas of characterizing mesh quality were introduced as a means of insuring good numerical results based on a particular mesh. This thesis applied the Delaunay triangulation to achieve a quality mesh focusing on the angles in the triangular regions as a criteria for determining a quality mesh.

The use of sparse storage techniques was examined in an effort to minimize the required computational storage. A compressed row pointer storage technique was used in conjunction with a conjugate gradient method to solve linear systems of equations which arise in the static and quasi-static problem.

The storage technique was also used to obtain the eigenvalues/eigenvectors associated with the full-field problem.

The techniques were applied to electromagnetic problems in the static and quasi-static regimes with little difficulty. The values of capacitance and impedance were calculated for a coaxial transmission line and a stripline structure with very good results.

A finite element model was developed through a novel variational formulation involving the method of Lagrange multipliers representing a full field expression of Maxwell's equations. This development emphasized the physical basis for the minimization process associated with the variational problem. The full-field functional was applied to the problem of obtaining the propagation constant of a canonical problem in a Rayleigh-Ritz context and indicated the difficulty arising from the use of certain functional forms. Lossy conducting regions were addressed by suggesting the use of impedance boundary conditions; though not used in this thesis, as more fundamental questions arose. The finite element implementation was developed and applied to a canonical problem with a discussion of spurious solutions for the eigenvalue problem.

Future work will involve a closer examination of the spurious solutions with an emphasis on obtaining a fundamental view of the minimizing process and function space requirements. The use of edge elements will also be investigated as a possible means of avoiding these false solutions. Also the effect of loss in the guiding structure needs to be addressed as related to the functional minimization and the underlying fundamentals.

REFERENCES

Babuska I. and A. K. Aziz (1976), "On the angle condition in the finite element method," *SIAM J. Numer. Anal.*, vol. 12, no. 2, April.

Bank R. E. (1980), "The use of adaptive grid refinement for badly behaved elliptical partial differential equations," *Math. Comput. in Simulation XXII*, pp. 18-24.

Biro O. and K. Preis (1989), "On the use of the magnetic vector potential in the finite element analysis of three-dimensional eddy currents," *IEEE Trans. Magnetics*, vol. 25, pp. 3145-3159.

Borri M. and P. Mantegazza (1977), "Efficient solution of quadratic eigenproblems arising in dynamic analysis of structures," *Comp. Appl. Mech. and Eng.*, vol. 12, pp. 19-31.

Bossavit A. (1988), "A rationale for edge-elements in 3-D fields computations," *IEEE Trans. Magnetics*, vol. 24, pp. 74-79, Jan.

Bridges T. J. and P. J. Morris (1984), "Differential eigenvalue problems in which the parameter appears nonlinearly," *J. Computat. Phys.*, vol. 55, pp. 437-460.

Bucca S. E. (1989) , *Wave Propagation in Lossy Waveguide Structures*, Ms. Thesis, Virginia Polytechnic Institute and State University.

Cavendish J. C. (1974), "Automatic triangulation of arbitrary planar domains for the finite element method," *Int. J. Numer. Meth Eng.*, vol. 8, pp. 679-696.

Cendes Z. J. and D. N. Shenton (1985), "Complementary error bounds for foolproof finite element mesh generation," *Mathematics and Computers in Simulation*, vol. 27, pp. 295-305, Aug.

Cendes Z. J. and D. N. Shenton (1985), "Adaptive mesh refinement in the finite element computation of magnetic fields," *IEEE Trans. Magnetics*, vol. 21, no. 5, pp. 1811-1816, Sept.

Chen C. H. and C. Lien (1980), "The variational principle for non-self adjoint electromagnetic problems," *IEEE Trans. Microwave Theory Tech.*, MTT-28, pp. 878-886.

Chen H. , T. K. Sarkar, J. Brule, and S. A. Dianat (1986), "Adaptive spectral estimation by the conjugate gradient method," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, pp. 272-284, April.

Chew W. C. and M. A. Nasir (1989), "A variational analysis of anisotropic inhomogeneous dielectric waveguides," *IEEE Trans. Microwave Theory Tech.*, vol. 37, pp. 661-668, April.

Clough, R. W. (1960), "The finite element method in plane stress analysis," *J. Struct. Div., ASCE, Proc. 2d Conf. Electronic Computation*, pp. 345-378.

Collin R. E. (1966), *Foundations for Microwave Engineering.* New York: McGraw-Hill.

Collin R. E. (1991), *Field Theory of Guided Waves, 2nd Ed.* New York: IEEE Press.

Cohn S. B. (1955), "Characteristic impedance of the shielded-strip transmission line," *Trans. IRE* , vol. 2, p. 52.

Costache G. I. (1987), "Finite element method applied to skin effect problems in strip transmission lines," *IEEE Trans Microwave Theory Tech.*, MTT-35, pp. 1009-1013.

Cuthill E. H. (1972) , "Several strategies for reducing the bandwidth of matrices," *Sparse Matrices and their Applications.* Eds. D. J. Rose and R.

A. Willoughby, New York: Plenum Press, pp. 157-160.

Durocher L. L. (1979), "A versatile two-dimensional mesh generator with automatic bandwidth reduction," *Comp. Struct.*, vol. 10, pp. 561-575.

Daly P. (1984), "Upper and lower bounds to the characteristic impedance of transmission lines using the finite element method," *Int. J. Comput. Math. Electr. Electron. Eng. (COMPEL)*, vol. 3, pp. 65-78.

Davies W. A. and C. F. Bunting (1992), "Variational formulation in Electromagnetics: Basic Concepts," *AEÜ*, vol. 46, no. 4.

Davies J. B. , F. A. Fernandez, and G. Y. Philippou (1982), "Finite-element analysis of all modes in cavities with circular symmetry," *IEEE Trans Microwave Theory Tech.*, vol. 30, pp. 1975-1980.

Duff I. S. , A. M. Erisman, and J. K. Reid (1986), *Direct Methods for Sparse Matrices.* Oxford: Clarendon Press.

Fox C. (1987), *An Introduction to the Calculus of Variations.* Dover Pub., Inc. (1950).

Finlayson B. (1972), *The Method of Weighted Residuals and Variational Principles.* New York: Academic Press.

Galerkin B. G. (1915), "Rods and plates. Series in some problems of elastic equilibrium of rods and plates." *Vestn. Inzh. Tech. (USSR)* vol. 19, pp. 897-908. Translation 63-18924, Clearinghouse, Fed. Sci. Tech. Info., Springfield, Va.

Gupta K. C. , R. Carg, and I. J. Bahl (1979), *Microstrip Lines and Slotlines.* Artech House.

Haley S. B. (1988), "The generalized eigenproblem: pole-zero computation," *Proc. IEEE*, vol. 76, no. 2, pp. 103-120, Feb.

Hammond P. (1981), *Energy Methods in Electromagnetism.* Oxford: Oxford University Press.

Harrington R. F. (1961), *Time-Harmonic Electromagnetic Fields.* New York: McGraw-Hill.

Harrington R. F. (1990), "Origin and development of the method of moments for field computation," *IEEE Antennas and Propagat. Mag.*, vol. 32, no. 3, pp. 31-36, June.

Hayata K. and K. Miura and M. Koshiba (1988), "Finite element formulation for lossy waveguides," *IEEE Trans. Microwave Theory Tech.*, vol. 36, pp. 268-275.

Hazel T. G. and A. Wexler (1972), "Variational formulation of the Dirichlet boundary condition," *IEEE Trans. Microwave Theory Tech.*, MTT-20 , pp. 385-390.

Ho-Le K. (1988), "Finite element mesh generation methods: a review and classification," *Comp. Aided Des.*, vol. 20, pp. 27-38.

Itoh T. (1989) *Numerical Techniques for Microwave and Millimeter-Wave Passive Structures.* New York: John Wiley and Sons.

Jeng S. and C. Chen (1984), "On variational electromagnetics: Theory and application," *IEEE Trans. Antennas Propagat.*, AP-32 , pp. 878-886.

Jeng G. and A. Wexler (1978) "Self-Adjoint Variational formulation of problems having non-self-adjoint operators," *IEEE Trans. Microwave Theory Tech.*, MTT-26 , pp. 91-94.

Johnson L. W. and R. D. Riess (1982), *Numerical Analysis, 2nd Ed*, Massachusetts: Addison-Wesley Pub. Co.

Konrad A. (1976) "Vector variational formulation of electromagnetic fields in anisotropic media," *IEEE Trans. Microwave Theory Tech.*, MTT-24 , pp. 553-559.

Koshiba M. and K. Inoue (1992), "Simple and efficient finite element analysis of microwave and optical waveguides," *IEEE Trans. Microwave Theory Tech.*, vol. 40, Feb.

Lancaster P. (1961), "A generalized Rayleigh quotient iteration for lambda-matrices," *Arch. Rat. Mech. Anal.*, vol. 8, pp. 309-322.

Lancaster P. (1964), "Algorithms for lambda-matrices," *Num. Math.*, vol. 6, pp. 388-394.

Lancaster P. (1966), *Lambda-Matrices and Vibrating Systems.* New York: Pergamon Press.

Lapidus L. and G. Pinder (1982), *Numerical Solution of Partial Differential Equations in Science and Engineering.* New York: John Wiley and Sons.

Lawson C. L. (1972), "Generation of a triangular grid with application to contour plotting," Cal. Tech. JPL Tech. Memorandum No. 299.

Lee J. , D. Sun, and Z. J. Cendes (1991), "Full-wave analysis of dielectric waveguides using tangential vector finite elements," *IEEE Trans. Microwave Theory Tech*, vol. 39, pp. 1262-1271, Aug.

McCartin B. J. , L. J. Bahrmasel, and G. Meltz (1990), "Application of the control region approximation to two dimensional scattering," *PIERS 2, Finite Element and Finite Difference Methods in Electromagnetic Scattering*, Ed. M. A. Morgan, New York: Elsevier Sci. Pub., pp. 175-210.

Miles R. E. (1970), "On the homogeneous planar Poisson point process," *Math. Biosci.*, vol. 6, pp. 85-127.

Morgan M. A., S. Chang, K. K. Mei (1977), "Coupled azimuthal potentials for electromagnetic problems in inhomogeneous axially symmetric media," *IEEE Trans. Antennas Propagat.*, AP-25, pp. 413-417.

Narayanan V. and R. K. Lade (1991), "Determination of the phase constant of closed transmission line systems using the finite difference and the conjugate gradient method," *Application of the Conjugate Gradient Method to Electromagnetic Signal Analysis, PIER 5*, Ed. T. Sarkar, New York: Elsevier Science Pub., pp. 539-565.

Pantic Z. and R. Mittra (1986), "Quasi-TEM analysis of microwave transmission lines by the finite element method," *IEEE Trans. Microwave Theory Tech.*, MTT-34, pp. 1096-1103.

Paulsen K. D. , D. R. Lynch , and T. W. Strohbehn (1988), "Three-dimensional finite, boundary, and hybrid element solution for lossy dielectric media," *IEEE Trans. Microwave Theory Tech.*, vol. 36, pp. 682-693, April.

Paulsen K. D. and D. R. Lynch (1991)," Elimination of vector parasites in finite element Maxwell solutions," *IEEE Trans. Microwave Theory Tech.*, vol. 39, March.

Paris D. and F. Hurd (1965), "Relaxation Properties of Fields and Circuits," *Proc IEEE*, vol. 53, pp. 150-156.

Peters G. and J. H. Wilkinson (1970), "$Ax = \lambda Bx$ and the generalized eigenproblem," *SIAM J. Numer. Anal.* vol. 7, Dec.

Prenter P. M. (1975), *Splines and Variational Methods*, New York: John Wiley and Sons.

Press W. H. , B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling (1989), *Numerical Recipes: The Art of Scientific Computing.* New York: Cambridge University Press.

Reddy J. N. (1984), *An Introduction to the Finite Element Method.* New York: McGraw-Hill.

Reddy J. N. (1986), *Applied Functional Analysis and Variational Methods in Engineering.* New York: McGraw-Hill.

Rivara M. C. (1984), "Design and data structure of fully adaptive, multigrid, finite element software," *ACM Trans. Math. Software*, vol. 10, no. 3, pp. 242-264.

Ruhe A. (1973), "Algorithms for the nonlinear eigenvalue problem,"*SIAM J. Numer. Anal.*, vol. 10, pp. 674-689.

Sibson R. (1978), "Locally equiangular triangulations," *The Comp. J.*, vol 21, no. 3, pp. 243-245.

Silvester P. P. and R. L. Ferrari (1990), *Finite Elements for Electrical Engineers, 2nd Ed.* New York: Cambridge University Press.

Stakgold I. (1968), *Boundary Value Problems of Mathematical Physics*, Vol. 2. New York: The Macmillan Company.

Steele C. W. (1987), *Numerical Computation of Electric and Magnetic Fields.* New York: Van Nostrand Reinhold Co.

Svedin J. A. M. (1989), "A numerically efficient finite-element formulation for the general waveguide problem without spurious modes," *IEEE Trans. Microwave Theory Tech.*, vol. 37, pp. 1708-1715.

Watson D. F. (1981), "Computing the n-dimensional Delaunay tessellation with application to Voronoi polytopes", *The Comp. J.*, Vol. 24, no. 2, pp. 167-172.

Webb J. B. , G. L. Maile, and R. L. Ferrari (1983), Finite element solution of three dimensional electromagnetic problems," *IEE Proc.*, vol. 130, pt. H, no. 2, pp. 153-159, March.

Weinstock R. (1974), *Calculus of Variations.* New York: Dover Pub., Inc. (1952).

Wexler A. (1969) "Computation of electromagnetic fields", *IEEE Trans. Microwave Theory Tech.* vol. 17, no. 8, pp. 416-439, Aug.

Wilson F. W. , R.K. Goodrich, and W. Spratte (1990), "Lawson's triangulation is nearly optimal for controlling error bounds," *SIAM J. Numer. Anal.* , vol. 27, no. 1, pp. 190-197, Feb.

Yang X. (1991), "A survey of various conjugate gradient algorithms for iterative solution of the largest/smallest eigenvalue of a symmetric matrix," *Application of the Conjugate Gradient Method to Electromagnetic Signal Analysis, PIER 5*, Ed. T. Sarkar, Elsevier Science Pub., pp. 567-588.

Young D. K. (1971) , *Iterative Solution of Large Linear Systems.* New York: Academic Press.

Young D. K. and D. R. Kincaid (1981), "The ITPACK package for large linear sparse linear systems," *Elliptic Problem Solvers.* Ed. by M. Schultz, New York: Academic Press.

Appendix 1

Full Field Element Matrix Computations

This appendix will illustrate the calculation of the element matrices $A, B, C, D, E,$ and F, in section 5.4. Recall that each element matrix is a 3×3 matrix that results from integration over an element given by,

$$A = \int_{\Omega} \{\Psi\}^* \{\Psi\}^T dxdy \ , \tag{A1.1}$$

$$B = \int_{\Omega} \{\Psi\}^* \frac{\partial \{\Psi\}^T}{\partial x} dxdy \ , \tag{A1.2}$$

$$C = \int_{\Omega} \{\Psi\}^* \frac{\partial \{\Psi\}^T}{\partial y} dxdy \ , \tag{A1.3}$$

$$D = \int_{\Omega} \frac{\partial \{\Psi\}^*}{\partial y} \frac{\partial \{\Psi\}^T}{\partial y} dxdy \ , \tag{A1.4}$$

$$E = \int_{\Omega} \frac{\partial \{\Psi\}^*}{\partial y} \frac{\partial \{\Psi\}^T}{\partial x} dxdy \ , \tag{A1.5}$$

and

$$F = \int_{\Omega} \frac{\partial \{\Psi\}^*}{\partial x} \frac{\partial \{\Psi\}^T}{\partial x} dxdy \ . \tag{A1.6}$$

The vector Ψ represents the elemental shape functions given by

$$\Psi = \begin{bmatrix} \psi_1{}^{(e)}(x,y) \\ \psi_2{}^{(e)}(x,y) \\ \psi_3{}^{(e)}(x,y) \end{bmatrix}. \tag{A1.7}$$

The shape function $\psi_i^{(e)}(x,y)$ is

$$\psi_i^{(e)} = \frac{1}{2A_e}\left(\alpha_i + \beta_i x + \gamma_i y\right) \qquad i = 1,2,3 \tag{A1.8}$$

and α_i, β_i, and γ_i are constants obtained from local element geometry. These constants are given explicitly as

$$\begin{aligned} \alpha_i &= x_j y_k - x_k y_j \\ \beta_i &= y_j - y_k \\ \gamma_i &= x_k - x_j \end{aligned} \tag{A1.9}$$

where $i \neq j \neq k$, and i, j, k permute in a natural order. The area A_e is given by

$$A_e = \frac{1}{2}\begin{vmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{vmatrix} = \frac{1}{2}\,det \quad . \tag{A1.10}$$

The integration over triangular shape functions is exact, and is given by the following useful formulas [Reddy,1984; Silvester,1990]:

$$\int_a^b {\psi_1}^m\, {\psi_2}^n\, ds = \frac{m!n!}{(m+n+1)!}\,(b-a) \tag{A1.11}$$

and

$$\int_{area}\int {\psi_1}^m\, {\psi_2}^n\, {\psi_3}^p\, dA = \frac{m!n!p!}{(m+n+p+2)!}\,2\,|\,A_e\,| \quad . \tag{A1.12}$$

The superscripts $m, n,$ and p are integers indicating the power on ψ. The area A is the area over which the integration is performed which is the area of the triangular region itself given by Eq. (A1.10).

The computation of A proceeds as

$$A = \int_{\Omega} \{\Psi\}^* \{\Psi\}^T dxdy = \int_{\Omega} \begin{bmatrix} \psi_1 \\ \psi_2 \\ \psi_3 \end{bmatrix} \begin{bmatrix} \psi_1 & \psi_2 & \psi_3 \end{bmatrix} dxdy,$$

or

$$A = \int_{\Omega} \begin{bmatrix} {\psi_1}^2 & \psi_1\psi_2 & \psi_1\psi_3 \\ \psi_2\psi_1 & {\psi_2}^2 & \psi_2\psi_3 \\ \psi_3\psi_1 & \psi_3\psi_2 & {\psi_3}^2 \end{bmatrix} dxdy.$$

Integrating this over an element using Eq. (A1.12) results in

$$A = \frac{|A_e|}{12} \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}. \qquad \text{(A1.13)}$$

The computation of B is

$$B = \int_{\Omega} \{\Psi\}^* \frac{\partial \{\Psi\}^T}{\partial x} dxdy = \frac{1}{2A_e} \int_{\Omega} \begin{bmatrix} \psi_1 \\ \psi_2 \\ \psi_3 \end{bmatrix} \begin{bmatrix} \beta_1 & \beta_2 & \beta_3 \end{bmatrix} dxdy,$$

where if we examine the 1,1 entry

$\int_{\Omega} \psi_1\beta_1 dxdy = \frac{\beta_1 |A_e|}{3}$. This leads to the element matrix for B of

$$B = \frac{1}{6}\begin{bmatrix} \beta_1 & \beta_2 & \beta_3 \\ \beta_1 & \beta_2 & \beta_3 \\ \beta_1 & \beta_2 & \beta_3 \end{bmatrix} \frac{|A_e|}{A_e} . \tag{A1.14}$$

The computation of C is similar to B and is found from

$$C = \int_{\Omega} \{\Psi\}^* \frac{\partial \{\Psi\}^T}{\partial y} dxdy = \frac{1}{2A_e} \int_{\Omega} \begin{bmatrix} \psi_1 \\ \psi_2 \\ \psi_3 \end{bmatrix} \begin{bmatrix} \gamma_1 & \gamma_2 & \gamma_3 \end{bmatrix} dxdy,$$

where if we examine the 1,1 entry

$\int_{\Omega} \psi_1 \gamma_1 dxdy = \frac{\gamma_1 |A_e|}{3}$. This leads to the element matrix for C of

$$C = \frac{1}{6}\begin{bmatrix} \gamma_1 & \gamma_2 & \gamma_3 \\ \gamma_1 & \gamma_2 & \gamma_3 \\ \gamma_1 & \gamma_2 & \gamma_3 \end{bmatrix} \frac{|A_e|}{A_e} . \tag{A1.15}$$

D is computed as follows:

$$D = \int_{\Omega} \frac{\partial \{\Psi\}^*}{\partial y} \frac{\partial \{\Psi\}^T}{\partial y} dxdy = \frac{1}{4{A_e}^2} \int_{\Omega} \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{bmatrix} \begin{bmatrix} \gamma_1 & \gamma_2 & \gamma_3 \end{bmatrix} dxdy,$$

or

$$D = \frac{1}{4|A_e|}\begin{bmatrix} {\gamma_1}^2 & \gamma_1\gamma_2 & \gamma_1\gamma_3 \\ \gamma_2\gamma_1 & {\gamma_2}^2 & \gamma_2\gamma_3 \\ \gamma_3\gamma_1 & \gamma_3\gamma_2 & {\gamma_3}^2 \end{bmatrix} . \tag{A1.16}$$

E is computed as follows:

$$E = \int_{\Omega} \frac{\partial\{\Psi\}^*}{\partial y} \frac{\partial\{\Psi\}^T}{\partial x}\, dxdy = \frac{1}{4{A_e}^2} \int_{\Omega} \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{bmatrix} \begin{bmatrix} \beta_1 & \beta_2 & \beta_3 \end{bmatrix} dxdy,$$

or

$$E = \frac{1}{4 \mid A_e \mid} \begin{bmatrix} \gamma_1\beta_1 & \gamma_1\beta_2 & \gamma_1\beta_3 \\ \gamma_2\beta_1 & \gamma_2\beta_2 & \gamma_2\beta_3 \\ \gamma_3\beta_1 & \gamma_3\beta_2 & \gamma_3\beta_3 \end{bmatrix}. \tag{A1.17}$$

F is computed as follows:

$$F = \int_{\Omega} \frac{\partial\{\Psi\}^*}{\partial x} \frac{\partial\{\Psi\}^T}{\partial x}\, dxdy = \frac{1}{4{A_e}^2} \int_{\Omega} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \begin{bmatrix} \beta_1 & \beta_2 & \beta_3 \end{bmatrix} dxdy,$$

or

$$F = \frac{1}{4 \mid A_e \mid} \begin{bmatrix} {\beta_1}^2 & \beta_1\beta_2 & \beta_1\beta_3 \\ \beta_2\beta_1 & {\beta_2}^2 & \beta_2\beta_3 \\ \beta_3\beta_1 & \beta_3\beta_2 & {\beta_3}^2 \end{bmatrix}. \tag{A1.18}$$

These results can be summarized as

$$A = \frac{\mid A_e \mid}{12} \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}. \tag{A1.13}$$

$$B = \frac{1}{6} \begin{bmatrix} \beta_1 & \beta_2 & \beta_3 \\ \beta_1 & \beta_2 & \beta_3 \\ \beta_1 & \beta_2 & \beta_3 \end{bmatrix} \frac{\mid A_e \mid}{A_e}. \tag{A1.14}$$

$$C = \frac{1}{6} \begin{bmatrix} \gamma_1 & \gamma_2 & \gamma_3 \\ \gamma_1 & \gamma_2 & \gamma_3 \\ \gamma_1 & \gamma_2 & \gamma_3 \end{bmatrix} \frac{\mid A_e \mid}{A_e}. \tag{A1.15}$$

$$D = \frac{1}{4\,|\,A_e\,|}\begin{bmatrix} \gamma_1{}^2 & \gamma_1\gamma_2 & \gamma_1\gamma_3 \\ \gamma_2\gamma_1 & \gamma_2{}^2 & \gamma_2\gamma_3 \\ \gamma_3\gamma_1 & \gamma_3\gamma_2 & \gamma_3{}^2 \end{bmatrix}. \tag{A1.16}$$

$$E = \frac{1}{4\,|\,A_e\,|}\begin{bmatrix} \gamma_1\beta_1 & \gamma_1\beta_2 & \gamma_1\beta_3 \\ \gamma_2\beta_1 & \gamma_2\beta_2 & \gamma_2\beta_3 \\ \gamma_3\beta_1 & \gamma_3\beta_2 & \gamma_3\beta_3 \end{bmatrix}. \tag{A1.17}$$

$$F = \frac{1}{4\,|\,A_e\,|}\begin{bmatrix} \beta_1{}^2 & \beta_1\beta_2 & \beta_1\beta_3 \\ \beta_2\beta_1 & \beta_2{}^2 & \beta_2\beta_3 \\ \beta_3\beta_1 & \beta_3\beta_2 & \beta_3{}^2 \end{bmatrix}. \tag{A1.18}$$

Appendix 2

Alternative Eigenvalue Formulation

This appendix presents the approach of Hayata [1988] to set up Eq. (5.48) for finite element techniques. For this approach the eigenvalue problem will be for a specified γ to determine the k of the guide. This problem becomes of interest for the inverse problem of determining the dielectric constant for a given propagation constant (perhaps from some measurements). The z component of the discretized electric field is coupled to the imaginary x and y components by defining

$$[\mathrm{N}] = \begin{bmatrix} \{\Psi\} & \{0\} & \{0\} \\ \{0\} & \{\Psi\} & \{0\} \\ \{0\} & \{0\} & j\{\Psi\} \end{bmatrix}. \tag{A2.1}$$

The electric field can be represented

$$\vec{E} = [\mathrm{N}]^T \{e\}^{(el)} \, exp(\text{-}j\gamma) \, , \tag{A2.2}$$

The vector $\{\Psi\}$ is a 3×1 vector of element shape functions,

$$\{\Psi\} = \begin{bmatrix} \psi_1(x,y) \\ \psi_2(x,y) \\ \psi_3(x,y) \end{bmatrix}, \tag{A2.3}$$

where $\psi_i(x,y)$ is given by Eq. (4.9). The vector $\{e\}^{(el)}$ is a 9×1 vector containing the x,y, and z components of the three nodes comprising a particular element given by (5.52) and (5.53). The curl can be represented by

$$\nabla\times\vec{E} = [\mathrm{G}]^T\{e\}^{(el)}exp(\text{-}j\gamma). \tag{A2.4}$$

Where the matrix [G] is a 9×3 matrix given by,

$$[\mathrm{G}]^T = \begin{bmatrix} \{0\}^T & \gamma\{\Psi\}^T & j\dfrac{\partial\{\Psi\}^T}{\partial y} \\ -\gamma\{\Psi\}^T & \{0\}^T & -j\dfrac{\partial\{\Psi\}^T}{\partial x} \\ -\dfrac{\partial\{\Psi\}^T}{\partial y} & \dfrac{\partial\{\Psi\}^T}{\partial x} & \{0\}^T \end{bmatrix}. \tag{A2.5}$$

Equation (5.48) can be written over an element Ω, as

$$\int_\Omega\left(k^2\left\{[\mathrm{N}^*]^T\{e_n^*\}^{(el)}\right\}^T\left\{[\mathrm{N}]^T\{e\}^{(el)}\right\}-\left\{[\mathrm{G}^*]^T\{e_n^*\}^{(el)}\right\}^T\left\{[\mathrm{G}]^T\{e\}^{(el)}\right\}\right)ds = 0, \tag{A2.6}$$

or, simplifying the transpose,

$$\int_\Omega\left(k^2\ \{e_n^*\}^{T(el)}\ [\mathrm{N}^*][\mathrm{N}]^T\ \{e\}^{(el)}\ -\ \{e_n^*\}^{T(el)}\ [\mathrm{G}^*][\mathrm{G}]^T\{e\}^{(el)}\right)ds = 0.$$

Factoring this expression leads to

$$\{e_n^*\}^{T(el)} \left[\int_\Omega k^2 \, [\mathrm{N}^*][\mathrm{N}]^T ds \; - \int_\Omega [\mathrm{G}^*][\mathrm{G}]^T ds \right] \{e\}^{(el)} = 0. \qquad \text{(A2.7)}$$

The eigenvalue problem is of the form $Ax = \lambda Bx$ after a summation over all elements to form the pyramidal basis set.

Let

$$[\mathrm{T}] = \int_\Omega k^2 \, [\mathrm{N}^*][\mathrm{N}]^T ds \, , \qquad \text{(A2.8)}$$

and

$$[\mathrm{S}] = \int_\Omega [\mathrm{G}^*][\mathrm{G}]^T ds \, , \qquad \text{(A2.9)}$$

where [T] and [S] are 9×9 matrices that form the basic element equation analogous to the 3×3 element matrices for the scalar problem discussed in Chapter 4.

[T] and [S] can be written out as follows:

$$[\mathrm{T}] = k^2 \begin{bmatrix} A & 0 & 0 \\ 0 & A & 0 \\ 0 & 0 & A \end{bmatrix} \qquad \text{(A2.10)}$$

and

$$[\mathrm{S}] = \begin{bmatrix} D + \gamma^2 A & -E & \gamma B \\ -E^T & F + \gamma^2 A & \gamma C \\ \gamma B^T & \gamma C^T & F + D \end{bmatrix}. \qquad \text{(A2.11)}$$

The entries $A, B, C, D, E,$ and F in Eqs. (A2.10) and (A2.11) are each 3×3 matrices that result from the integration over an element given by,

$$A = \int_{\Omega} \{\Psi\}^* \{\Psi\}^T dxdy \ , \tag{A2.12}$$

$$B = \int_{\Omega} \{\Psi\}^* \frac{\partial \{\Psi\}^T}{\partial x} dxdy \ , \tag{A2.13}$$

$$C = \int_{\Omega} \{\Psi\}^* \frac{\partial \{\Psi\}^T}{\partial y} dxdy \ , \tag{A2.14}$$

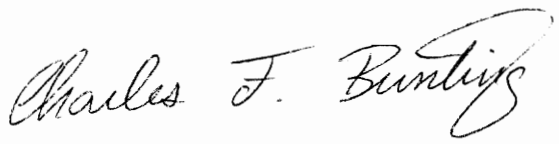
$$D = \int_{\Omega} \frac{\partial \{\Psi\}^*}{\partial y} \frac{\partial \{\Psi\}^T}{\partial y} dxdy \ , \tag{A2.15}$$

$$E = \int_{\Omega} \frac{\partial \{\Psi\}^*}{\partial y} \frac{\partial \{\Psi\}^T}{\partial x} dxdy \ , \tag{A2.16}$$

and

$$F = \int_{\Omega} \frac{\partial \{\Psi\}^*}{\partial x} \frac{\partial \{\Psi\}^T}{\partial x} dxdy \ . \tag{A2.17}$$

The evaluation of Eqs. (A2.12)-(A2.17) is described in detail in Appendix I. Note that the problem of spurious modes arises in the solution of $Ax = \lambda Bx$ for the system as described above and further work will be required to investigate the possible remedy.

VITA

Charles F. Bunting was born in Virginia Beach, Virginia on February 27, 1962. He graduated from Kempsville High School in 1980. He was employed for 8 years at the Naval Aviation Depot, Norfolk, Virginia, where he completed an apprenticeship as an electronics mechanic in 1985, and then worked in calibration and repair of electronic test equipment. He received his A.A.S. in Engineering Technology from Tidewater Community College in 1985, and the B.S. degree in Engineering Technology from Old Dominion University in 1989. In 1990 he was awarded a Bradley Fellowship, and a Dupont Fellowship, and is working toward a Doctoral degree in Electrical Engineering. His chief interests are microwave material characterization and numerical methods in electromagnetics.