THE FINITE ELEMENT SOLUTION OF INHOMOGENEOUS ANISOTROPIC AND LOSSY DIELECTRIC WAVEGUIDES

by

LU Yilong

A thesis submitted for the degree of
Doctor of Philosophy

Department of Electronic and Electrical Engineering
University College London
University of London

October 1991

ABSTRACT

This thesis presents a new variational finite element formulation and its implementation for the analysis of microwave and optical waveguide problem with arbitrarily- shaped cross section, inhomogeneous, transverse-anisotropic, and lossy dielectrics.

In this approach, the spurious, nonphysical solutions, which ordinarily appear interspersed with the correct results of earlier vectorial finite element methods and thus have been the most serious problem in finite element analysis of waveguides, are totally eliminated. In this formulation either the propagation constant or the frequency may be treated as eigenvalues of the resulting generalized eigenvalue problem. This formulation also has the capability to find complex modes of lossless waveguides. Furthermore, the numerical efficiency of the solution is maximized since this formulation uses the most economical representation of a problem, in terms of only two vector components. This is achieved without losing the sparsity of the matrices of the resultant eigenvalue equation, which only depends on the topology of mesh used. This property is very important for solving large-size problems by efficient sparse matrix algorithms.

In this work, a basic vector wave equation which involves only transverse components of magnetic field is straightforwardly derived from Maxwell equations. This differential equation incorporates the divergence condition $\nabla \cdot \mathbf{B} = 0$ and leads to a canonical form of the resultant eigenvalue equation. The Local Potential Method is used to obtain the variational formulation. When implementing the finite element method, the Rayleigh-Ritz procedure is used to find stationary values of the functional to get the resulting generalized matrix eigenvalue equation.

To show the validity and applicability of the method, a series of examples of microwave and optical waveguides including inhomogeneity, anisotropy and loss are studied. These examples show good accuracy and complete absence of spurious modes, demonstrating the effectiveness of the new formulation developed.

ACKNOWLEDGEMENTS

This thesis would not have been possible without the kind help of many people. I would like to express my special gratitude to Dr. Anibal Fernandez and Professor Brian Davies from University College London (UCL) for their guidance and inspiration that I found so valuable throughout this study. I would like to thank Dr. Robert Ettinger from UCL for providing both a willingness to share his expertise and patience during many discussions. Many thanks to my fellow-Chinese, Mr. Zhu Shouzheng, a visiting scholar at UCL from East China Normal University, for many helpful discussions and his efforts in developing an efficient sparse matrix solver software which greatly increases the efficiency of computations for many examples in this thesis.

I would also like to thank Professor P.C. Kendall from Sheffield University for providing important comments and suggestions. Dr. T.B. Koch from King's College and Dr. B.M.A. Rahman from City University are warmly thanked for their valuable help and humour. The permission from Alcatel Alsthom Recherche, France, for using their optical waveguide structure is also appreciated.

Finally, I would like to thank my loving wife, Xiaohui, for her encouraging and supporting, as well as many hours of preparation of the manuscript.

Financial supports from the British Council and the Chinese Government are gratefully acknowledged.

To my wife,

my mother, my father

and my son

CONTENTS

CHAPTER 1

INTRODUCTION

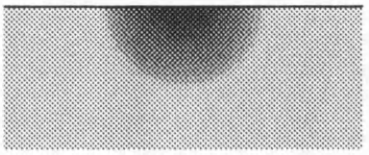
Computer simulation (or computer modelling) of physical problems is now widely accepted as the third investigative technique in science — alongside theory and experiment. Increasing complexity of modern wave functional devices has created a critical demand for accurate and efficient computer simulation of waveguides which are the fundamental components of these devices. The finite element method (FEM) is a powerful and versatile tool for this purpose. It provides unsurpassed accuracy in solving complicated problems while its flexibility allows the treatment of different structures without the need for device-dependent programming.

1.1 Dielectric Waveguides

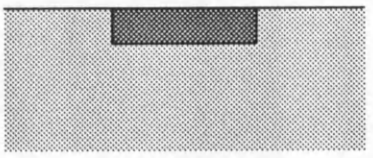
The development of new transmission methods for telecommunications during the last decade has been dominated by the evolution of optoelectronics. There is increasing interest in using optics to extend and replace electronics for some purposes. One major activity in research laboratories throughout the world is the demonstration of various *optical* integrated circuits to replace and enhance the performance of *electronic* integrated circuits, and also to perform novel functions particularly suited for optics. Dielectric waveguides play an essential role in the optoelectronics. They are widely used in optical fibre system, optical integrated circuits, and lasers.

Dielectric waveguides are the fundamental components of optoelectronic devices. As such, a full understanding of how electromagnetic waves propagate in dielectric waveguides is essential. On the other hand, the advance of material science and fabrication technology is continuously introducing more complicated dielectric guide structures.

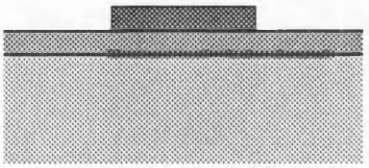
A glance at the materials and fabrication technologies used to make integrated optical waveguides may give us a better understanding of the practical demands for dielectric waveguide analysis.



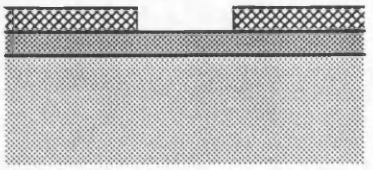
(a) Diffused waveguide [1]



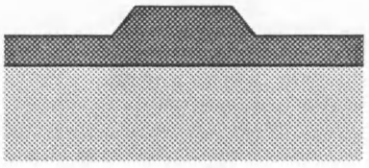
(b) Channel waveguide [1]



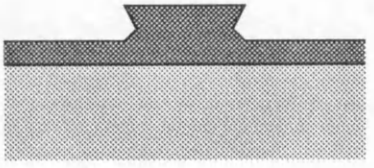
(c) Dielectric-film loaded strip waveguide [1]



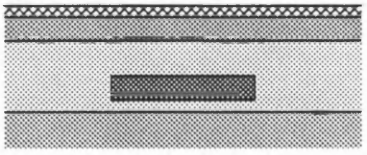
(d) Metal-film loaded strip waveguide [1]



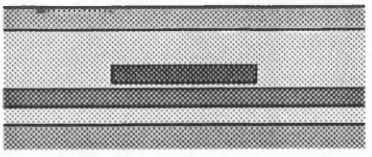
(e) Trapezoidal-rib waveguide [1]



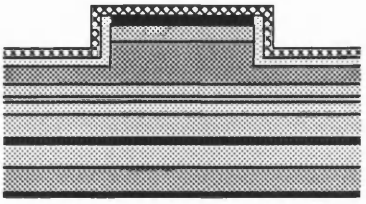
(f) Waisted-rib waveguide [1]



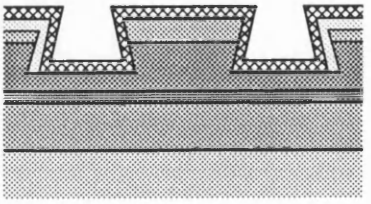
(g) Buried waveguide [132]



(h) Buried strip loaded waveguide [132]



(i) SLB GRIN-SCH ridge waveguide [4]



(j) DQW-SCH metal-clad ridge waveguide [5]

Fig. 1.1 Dielectric waveguides in optoelectronic technology

A variety of materials has been used for optical waveguides [1]-[5], such as gallium arsenide (GaAs), indium phosphide (InP), lithium niobate ($LiNbO_3$), lithium tantalate ($LiTaO_3$), silica (SiO_2), polymers, organic materials, and varieties of compounds, etc.. Many of these materials are anisotropic materials such as $LiNbO_3$, $LiTaO_3$, and most of organic materials. Quite often, significant losses need to be taken into account, for instance, optical waveguides with N^+InP or P^+InP layers, waveguides with metal claddings (as metals are highly optical absorbing), and laser waveguides working near their energy bands, etc. [1], [3]-[5].

Various microfabrication techniques are well-established to control very precisely the refractive index and geometry of dielectric waveguide [1]-[3]. *Diffusion and implantation techniques* alter waveguide refractive index by diffusion of a doping material in a substrate medium, such as thermal diffusion, ion implantation, ion exchange, and proton exchange. *Deposition and growth techniques* are in use to control layer thicknesses, such as coating, sputtering, thermal vapour deposition, chemical vapour deposition, polymerization, and epitaxial growth (liquid phase epitaxy (LPE), metal-organic vapour phase epitaxy (MOVPE), molecular beam epitaxy (MBE)). *Etching techniques* cause change in the geometry of structure and can also be used additionally in growth and diffusion processes (sputter-etching, reactive ion etching, and ion beam etching).

Fig. 1.1 shows several types of optical waveguides. The permittivity profile of optical waveguides can be arbitrarily inhomogeneous, anisotropic and lossy. This variety occurs either as a design preference or due to actual manufacturing processes. Most of such cases of waveguide arbitrariness do not lend themselves to analytical solutions. Besides, the fabrication of integrated optical waveguides requires various sophisticated technological steps and hence the production costs are still high. At the same time, the measuring techniques are difficult, expensive and very time consuming. There is, therefore, a great demand for more accurate and flexible computer simulation techniques which can be used for both the analysis and design of a wide range of waveguiding structures. This leads us to the study of computer simulation of inhomogeneous, anisotropic and lossy dielectric waveguides.

What do we need to know about the propagation characteristics of electromagnetic wave in optical waveguides? First of all, it is necessary

to establish how many modes the structure will support. Most applications will require the propagation of one or two modes, and small changes in dimensions or refractive indices can frequently result in the structure being either cut-off or supporting more than desired.

Secondly, it is often desirable to know the precise field distribution of the mode. This is important when designing devices for high coupling efficiency between planar waveguides and optical fibres. Also, the performance of some practical waveguides is limited by scattering losses caused by roughness induced by the fabrication process, and detailed information about the magnitude of the field at rough edges allows these losses to be assessed.

Thirdly, it is usually necessary to know the propagation constant of a mode in a waveguide and in some cases, quite accurately. For example, for most optical switching functions the operating principle is interference between two modes, and an precise knowledge of difference between propagation constants of two modes is necessary. This difference is usually a very small percentage of the value of the propagation constant, and so a precise calculation of the propagation constant for each guide is very important, and for this the most accurate techniques for calculating propagation constant are needed.

1.2 Computer Simulation and Finite Element Method

1.2.1 Numerical Methods and Finite Elements

There is a growing emphasis on numerical methods for engineering analysis because it is not possible to obtain analytical mathematical solutions for many engineering problems. An analytical solution is a mathematical expression that gives the values of the desired unknown quantity at any point in the problem domain. As a consequence, it is valid for an infinite number of points in that domain. However, analytical solutions can be obtained only for certain simple situations. For problems involving complex material properties and boundary conditions, engineers resort to numerical methods that provide approximate, but acceptable solutions. In most of the numerical methods, the solutions yield approximate values of unknown quantities only at a discrete number of points in the domain. The process of selecting only a certain number of

discrete points in the domain is termed as *discretization.* One of the ways to discretize a domain is to divide it into an equivalent system of smaller domains or units. The assemblage of such units then represents the original domain. Instead of solving the problem for the entire domain in one operation, the solutions are formulated for each constituent unit and combined to obtain the solution for the original domain. This approach is known as *going from part to the whole.* Although the analysis procedure is thereby considerably simplified, the amount of data to be handled depends on the number of subdivisions, and it is a formidable task to handle the volume of data manually. Consequently, recourse must be made to automatic electronic computation.

Before the advent of electronic computers, the applicability of many numerical methods were somewhat limited. With increasing development and widespread of computers, many of the numerical methods developed before the era of computers are now adapted for use with these machines. Perhaps the best known is the finite difference method [7]. Other types of classical methods that have been adapted to modern computation are such weighted residual methods as the least square method and such variational method as the Rayleigh-Ritz method [8].

In contrast to the techniques mentioned above, the finite element method [9], [10] is essentially a product of the computer age. It has developed simultaneously with the increasing use of high-performance computers and with the growing emphasis on numerical methods for engineering analysis. Although the approach shares many of the features common to the previous numerical approximations, it possesses certain characteristics that take advantage of the special facilities offered by high-speed computers. In particular, the method can be systematically programmed to accommodate such complicated and difficult problems as inhomogeneous anisotropic materials, and complicated geometries and boundary conditions. It is difficult to accommodate these complexities in the methods mentioned above.

The finite element method was originally developed for structural analysis, the general nature of the theory on which it is based has also made possible its successful application for solutions of problems on other fields of engineering, such as heat flow, fluid dynamics as well as electromagnetics, etc. As a result of this broad applicability and the

systematic generality of the associated computer codes, the method has gained wide acceptance by researchers and designers in computer simulation.

1.2.2 Perspective of Computer Power

Traditionally, finite element solutions were thought only to be achievable on mainframes. However, with increasing computer power available in relatively small machines, reasonably-large finite element solutions nowadays can be easily achieved on widely used workstations and PCs. For example, one can solve eigenvalue problems with matrix orders more than ten thousand on a medium-sized workstation (see Chapter 7). In fact, many current workstations are much more powerful than many mainframes years ago. For example, the widely used SUN SPARC 2 workstation has a speed of 28.5 MIPS (mega instruction per second) or 4.2 MFLOPS (mega floating-point operation per second) and with RAM up to 96 MB (mega bytes) [11], comparing a 16.5 MIPS, 24 MB IBM 3081-KX2 mainframe. Some top model workstations can reach 320 MFLOPS and with RAM up to 1.4 GB (giga bytes) [11].

In addition to ordinary computer power, there are and will be more and more supercomputer power available. The increase of supercomputer resources during the last a few years is surprisingly fast. For example, from 1985 to 1990, the academic supercomputing capacities in Japan, USA, and Germany had increased 3432%, 2032%, and 1182%, respectively [12]. There are reports of finite element solutions of very complicated mechanical problems with more than half million unknowns by use of supercomputers [13]. The fastest supercomputer can achieve 22 gigaflops (10^9 flops). And what will be more, the teraflop (10^{12} flops) computer is reported under way [12].

With the popularization of fibre-based high-speed local area networks (LANs) (100-megabit-per-second range LANs are now commercially available and 1 gigabit per second are being tested), metropolitan area networks (MANs), and wide area networks (WANs), access and data communication to supercomputer are getting more and more convenient and efficient [14].

No doubt, there will be more and more computer power available, therefore we should take advantage and make good use of it. Following this trend, if condition allows, it probably is more important and preferable to develop an accurate method even if it requires considerable computing

resources rather than to develop a rough approximation method which requires less computing resources.

1.2.3 Difficulties in Finite Element Analysis of Waveguides

Since the first papers on finite element solution of electrical engineering problems appeared in 1968 [25], the finite element method has grown to one that offers probably the most powerful and efficient numerical solution of the most general (i.e., arbitrarily-shaped, inhomogeneous, anisotropic, and lossy) electromagnetic waveguide problems [20]-[24]. However, the most serious difficulty in applying the finite element method to waveguide problems is the appearance of the spurious, nonphysical solutions which ordinarily appear interspersed with real solutions. Using a formulation which is not immune to spurious solutions, it is difficult and quite cumbersome to distinguish between spurious and physical modes.

Although the occurrence of spurious modes in finite element vector wave equation solutions has been known for some time, and the suppression of such undesirable erroneous solutions is still a subject of great interest, the development of a method to eliminate spurious solutions is a pressing need and research on this topic has been extensive in recent years.. The traditional approach to suppress spurious modes has been the penalty method [64], [65], [67]. The penalty method only partially cures the problem, and the effect is not entirely satisfactory. Other approaches resort to using either dense matrices [80], [81], [85] or more components [84] than are absolutely necessary for a full description of general hybrid mode situations. In addition, none of the existing finite element formulations has satisfactorily treated lossy waveguides. The demand for an efficient finite element formulation for general inhomogeneous, anisotropic and lossy dielectric waveguides plus the general lack of insight studies on spurious modes of finite element solutions have motivated us to this study.

1.3 Principal Aims of the Study

In view of the foregoing, the pull of demanding optical waveguide application and the push of advancing computer power stimulates this study.

The project was initially conceived to develop an efficient finite

element formulation able to eliminate spurious solutions which had been the most serious problems in finite element analysis of electromagnetic waveguides. If possible it was also hoped to include significant loss in dielectric waveguides, and by doing so establish a more realistic model and better computer simulation.

1.4 On the Layout of the Thesis

The first chapter includes an overview to the background on currently used dielectric waveguides, a brief introduction of fundamental features the finite element method, a perspective on the advancing of modern high-speed computers, and the major problems in finite element analysis of waveguides. This leads to an understanding of the importance to develop a better finite element method for dielectric waveguide problems.

The second chapter reviews past finite element formulations for electromagnetic dielectric waveguide, particularly focusing on the recent development of methods to suppress or eliminate spurious solutions encountered in vectorial finite element analysis of waveguide problems. Eight criteria are proposed for judging the appropriateness of a finite element formulation for dielectric waveguide problems. They are also used as the targets of deriving the new finite element formulation.

The next four chapters are the technical core of the thesis. In chapter 3, the mathematical definition of the dielectric waveguide problem is given and the mechanism behind spurious modes is discussed. Strategies to eliminate spurious solutions are proposed. A basic wave equation in terms of only transverse magnetic field components is derived from Maxwell's equations. This differential equation will be used as the starting point of deriving the new variational finite element formulation in chapter 5.

In chapter 4, the mathematical fundamentals of the finite element method are discussed rigorously in order to have an accurate understanding of the finite element method and thus to avoid any mistake in the latter derivation of the finite element formulation in chapter 5, and the following finite element implementation in chapter 6.

Having established a mathematical model of a dielectric waveguide and understood the fundamentals of finite elements, chapter 5 presents the detailed procedures for deriving the new variational finite element

formulation.

In chapter 6 the finite element implementation to the new formulation is detailed. Additionally, a highly efficient solver for large, sparse, nonsymmetrical and complex matrix eigenequation is also briefly described here [136].

Chapter 7 demonstrates the validity and effectiveness of the new finite element formulation by illustrating the computational results of a variety of waveguides with the present method. The examples are classified into four categories: isotropic lossless, anisotropic lossless, isotropic lossy, and anisotropic lossy dielectric waveguides. Statistics of the sparse matrix solver developed for the formulation are also shown.

Finally, conclusions are presented in chapter 8.

1.5 Notational Convention

1.5.1 Scalars, Vectors, Tensors, and Matrices

(i) The scalar, vector and tensor representation
Vector: bold-font letters, e.g., $\mathbf{H}_t$, $\mathbf{A}$;
Tensor: plain letters with symbol '=' on top, e.g., $\overline{\overline{\varepsilon}}$, $\overline{\overline{\kappa}}$
Scalar: plain letters, e.g., H_x, ψ

(ii) The matrix representation
Rectangular matrix: square bracketed letters, e.g., [A]
Column matrix (vector): braced letters, e.g., {A}
Row matrix (vector): braced letters with transposition, e.g., $\{A\}^T$

For tensor and matrix, the superscript 'T' indicates the transposition, and the superscript '–1' indicates the inversion.

1.5.2 Mode Designations

The following commonly used mode designations are simultaneously adopted in the thesis for convenience of comparison.

1) For homogeneous rectangular metallic waveguides [15]

TE_{mn} (or H_{mn}) if ($E_z = 0$)

TM_{mn} (or E_{mn}) if ($H_z = 0$)

2) For Dielectric-slab-loaded rectangular metallic waveguides [15]

LSE_{mn} if no electric field component normal to the interface

LSM_{mn} if no magnetic field component normal to the interface

3) For other inhomogeneous waveguides [16]-[18]

E^{x}_{mn} (or H^{y}_{mn}, or HE_{mn}) if $|E_x| > |E_y|$ (or $|H_y| > |H_x|$)

E^{y}_{mn} (or H^{x}_{mn}, or EH_{mn}) if $|E_y| > |E_x|$ (or $|H_x| > |H_y|$)

The indices m and n are used to designate the number of maxima of the dominant component in the guide region in the x and y directions, respectively.

1.6 Summary of Main Achievements

In this study, the whole original objectives have been achieved. The origins of spurious modes have been studied in a more general way. Insight comments about spurious modes in approximate Maxwell solutions are made. Strategies of eliminating spurious solutions are proposed.

An efficient variational finite element formulation for the full wave analysis of dielectric waveguides has been developed. This formulation provides five major contributions:

1) it can treat a wide range of dielectric waveguide problems with arbitrarily-shaped cross section, inhomogeneity, transverse-anisotropy, and significant loss (or gain);
2) it totally eliminates troublesome non-physical spurious solutions which ordinarily appear interspersed with the correct results of many other vectorial finite element formulations;
3) it allows direct solutions for (complex) propagation constants;
4) the numerical efficiency of solution is maximized since this formulation uses only two magnetic field components; this is achieved without losing the matrix sparsity which only depends on the topology of mesh used, and this property is of decisive importance for solving large-size problems;
5) it provides the capability to compute complex modes† in lossless waveguide, showing the completeness of the solutions.

---

† Complex modes or complex waves are modes existing in inhomogeneous lossless waveguide with complex propagation constant [19]

This formulation is believed to be the most efficient finite element formulation now available for inhomogeneous lossy waveguides, and the finite element solutions of complex modes presented here are ones which have not been achievable elsewhere.

This study also prompted and partly contributed to the development of an efficient matrix solver for general, large, sparse, nonsymmetrical and complex matrix eigenequations. This solver drastically reduces requirements for computing time and memory. No other comparable solver is available in standard computer libraries or even has been reported. Together with this highly efficient sparse matrix solver, the new finite element formulation has been coded in FORTRAN language. The whole FORTRAN program have been thoroughly tested with all categories of dielectric waveguides, and numerical results are satisfactory showing the effectiveness and robustness of the method presented in this thesis.

The computer software implementing all the algorithm allows us to make more realistic and accurate full wave analysis of complicated dielectric waveguide problems.

Six papers originating from this study have been published [133]-[138].

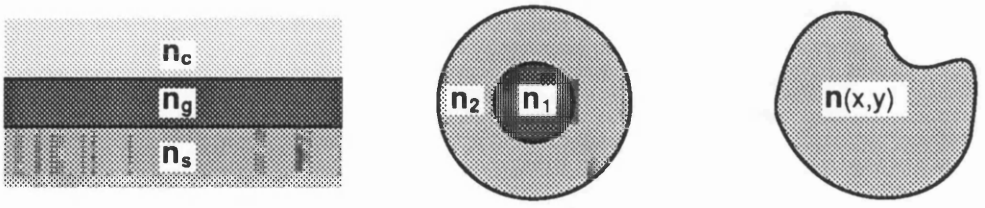
CHAPTER 2

REVIEW OF FINITE ELEMENT FORMULATIONS

2.1 Introduction

This chapter reviews the finite element formulations for the analysis of microwave and optical waveguides, particularly focusing on the recent development of methods to suppress or eliminate spurious solutions encountered in the vectorial finite element analysis of waveguide problems. Here, the dielectric waveguide, widely used from microwave to optical wavelength regions, is considered. No magnetic material is considered unless it is particularly mentioned.

The electromagnetic waveguide can be classified into two categories from its cross-section shape. One is the layered waveguide such as a layered film waveguide (planar waveguide, illustrated in Fig. 2.1a) or a layered circular waveguide (axially symmetric waveguide, illustrated in Fig. 2.1b), which can be treated as an equivalent one-dimensional problem; the other is the more general arbitrarily shaped waveguide (illustrated in Fig. 2.1c) that should be treated as a two-dimensional problem. We will only discuss the truly two-dimensional case and finite element techniques of general interest.



(a) (b) (c)

Fig. 2.1 Classification of waveguides

(a) Planar waveguide

(b) Axially symmetric waveguide

(c) Waveguide with arbitrarily shaped cross-section

Finite element formulations may be achieved by direct methods, variational methods or weighted residual methods [8]. It is advantageous to take a variational approach whenever it is possible, especially when one global parameter (like propagation constant) is needed. The weighted residual methods are useful for problems in which a variational functional may not be available, although they may be applied to any boundary value problem with established differential equations.

For a two dimensional problem, the waveguide is assumed to be uniform along its longitudinal z axis and the electric and magnetic fields can then be expressed as:

$$\begin{aligned} \mathcal{H}(x,y,z,t) &= (\mathbf{H}_t(x,y) + \mathbf{H}_z(x,y)) \, exp(j\omega t - \gamma z) \\ \mathcal{E}(x,y,z,t) &= (\mathbf{E}_t(x,y) + \mathbf{E}_z(x,y)) \, exp(j\omega t - \gamma z) \end{aligned} \tag{2.1}$$

where $\gamma = \alpha + j\beta$ is the propagation constant in the positive z-direction, α being the attenuation constant, β being the phase constant, and ω being the angular frequency, with the subscript t denoting "transverse to z".

When applying the standard finite element method to waveguide problems for propagation characteristics analysis, it is usually expected to arrive at a matrix eigenvalue equation of the canonical form:

$$[A] \{x\} = \lambda [B] \{x\} \tag{2.2}$$

where $\{x\}$ is the discretized nodal field vector, [A] and [B] are in general sparse matrices. The eigenvalue λ may correspond, for example, to ω^2 or γ^2.

According to the type of eigenvalue, finite element formulations may be classified into two types. One is frequency formulation (or simply noted as ω-type formulation), where the eigenvalue is an explicit known function of ω; the other is propagation constant formulation (simply noted as γ-type formulation), where the eigenvalue is an explicit known function of γ.

One important deficiency of an ω-type formulation is that for a given waveguide, it searches for the frequency of each mode corresponding to a selected value of the propagation constant while in practice the problem is usually the inverse, that is: one is interested in finding the propagation constant (possibly complex) at a given frequency. Consequently, iterations

are usually needed to solve a practical problem when using an ω-type formulation.

A γ-type formulation solves directly for the propagation constant at a given frequency. Unnecessary iterations can be avoided. In addition, only a γ-type formulation is applicable for lossy problems. Therefore, a γ-type formulation is in general preferable.

2.2 Scalar Finite Element Formulations

Finite element analysis for electromagnetic waveguides started with scalar formulations in the late 1960's [25], [26]. Since then, various scalar formulations have been developed. The scalar finite element analysis has been used for solving homogeneous waveguide problems [25]-[28], for approximate analysis of lossy guides [29], for open-boundary problems [30], [31], and for analysis of anisotropic waveguides [32], [33].

Spurious solutions are usually not involved in scalar finite element analysis, this is one special and redeeming feature of a scalar approach.

Although a single scalar formulation is inadequate for the inherently hybrid mode situation of anisotropic or genuinely two dimensional, inhomogeneous waveguide problems, depending on waveguiding structures or propagating modes, the quasi-TEM, quasi-TE, or quasi-TM mode approximations are practically available. Besides, scalar formulations take significantly lower computational cost, and hence they may be suitable for design procedures in CAD systems.

2.3 Vector Finite Element Formulations

To evaluate rigorously the propagation characteristics of an inhomogeneous anisotropic waveguide, vectorial wave analysis is necessary, with at least two field components. The vectorial formulation are fundamentally more accurate than scalar forms since they can represent true hybrid modes in dielectric waveguides.

Before reviewing the vectorial finite element formulations, it is worth mentioning a few articles [34]-[38] of purely theoretical study of general variational formulations for self-adjoint and non-self-adjoint operators. Berk [34], Morishita and Kumagai [35], [36], Jeng and Wexler [37], Chen and

Lien [38] proposed a number of variational ω-type and γ-type formulations for lossless anisotropic waveguide problems in terms of the magnetic field **H**, or the electric field **E**, or a combination of both. However, none of them have been found satisfactory success by direct finite element implementation, since they suffer either spurious solutions, or unacceptable complexity.

Finite element methods in terms of the longitudinal electric and magnetic field components (E_z,H_z) have been used for analyzing various microwave [39]-[46] and optical waveguides [47]-[54]. They have also been applied to anisotropic waveguides with diagonal permittivity tensor [55], and to lossy waveguides [56]. The (E_z,H_z) formulation cannot treat general anisotropic problems without destroying the canonical form of (2.2). Also, for a waveguide with arbitrary dielectric distribution, enforcing boundary conditions in this method can be quite difficult [52]. Another fundamental disadvantage for optical waveguide problems is that it is based on longitudinal components which are usually the least important of the six components of the vector fields. Additionally, this type of formulation is also affected by spurious solutions, and the techniques to reduce them [52] greatly increase the complexity of the program and the computing cost.

In early 1970's, English and Young [57], [58] applied a six component **E**-field and **H**-field formulation and a three component **E**-field formulation to cylindrical waveguides. However, spurious solutions were encountered. Besides, the boundary conditions on trial functions are restrictive, waveguides of shapes other than circular or rectangular can not be treated.

The full field **H** or **E** finite element analysis virtually started from the late 1970's [59], [60], but spurious solutions were encountered. From the early 1980's, more (H_x,H_y,H_z) or (E_x,E_y,E_z) formulations were applied to analyze various problems [61]-[63]. However, all of them were found to have spurious solutions and no remedies were used then. It was observed by Davies, Fernandez and Philippou [62] that for the **H** formulation [62], [63]:

$$\int_{\Omega} (\nabla\times\mathbf{H})^{*}\cdot\bar{\bar{\varepsilon}}^{-1}\cdot(\nabla\times\mathbf{H})\, d\Omega - \omega^2 \int_{\Omega} \mathbf{H}^{*}\cdot\bar{\bar{\mu}}^{-1}\cdot\mathbf{H}\, d\Omega = 0 \qquad (2.3)$$

the spurious solutions do not satisfy the divergence-free condition

$\nabla \cdot \mathbf{H} = 0$. In fact, for the formulation (2.3), the condition $\nabla \cdot \mathbf{H} = 0$ is neither implied nor forced. This causes the system to be 'under-determined' or excessively flexible, which in turn is believed [62], [63] to be one of the causes of spurious solutions.

Following this idea, the penalty method was applied to enforce the divergence-free condition to the **E**-field [64], [68] and to the **H**-field [65], [67] formulations.

The **H**-field formulation is more suitable to dielectric waveguide problems where the magnetic field is continuous everywhere. The **H**-field penalty formulation

$$\int_{\Omega} (\nabla \times \mathbf{H})^* \cdot \bar{\bar{\varepsilon}}^{-1} \cdot (\nabla \times \mathbf{H}) \, d\Omega - \omega^2 \int_{\Omega} \mathbf{H}^* \cdot \bar{\bar{\mu}}^{-1} \cdot \mathbf{H} \, d\Omega + (\rho/\varepsilon_0) \int_{\Omega} (\nabla \times \mathbf{H})^* \cdot \bar{\bar{\varepsilon}}^{-1} \cdot (\nabla \times \mathbf{H}) \, d\Omega = 0 \qquad (2.4)$$

has been extensively studied [65], [67] and applied to various types of waveguiding problems [69]-[76] in which the divergence-free condition is satisfied in the least square sense and the spurious solutions may be suppressed from the guided- or slow-wave region [23], [67].

In the penalty method, an arbitrary positive constant ρ, called the penalty coefficient is included, this penalty coefficient itself introduces a error. The accuracy of solution by the penalty method depends on the magnitude of the penalty coefficient. The penalty method only partially cures the spurious problem, and the effect is not entirely satisfactory. It requires the careful choice of the penalty coefficient in order to achieve adequate balance between the appearance of spurious modes and the amount of error introduced, rendering the programs less robust and friendly.

Hano [66] developed an ω-type vectorial finite element formulation in terms of all three components of either electric or magnetic fields. Special triangular elements ensure the continuity of the tangential components of field vectors only with no constraints on the normal components. No spurious modes appear but there are many needless zero non-physical solutions. The implementation of the special triangular elements in a finite element code is rather complicated.

Kobelansky and Webb [80] suggested a two stage procedure, solving first using the functional

$$G(\mathbf{H}) = \int_{\Omega} [(\nabla \cdot \mathbf{H})^{*} (\nabla \cdot \mathbf{H}) \, d\Omega - \lambda \, \mathbf{H}^{*} \cdot \mathbf{H}] \, d\Omega \qquad (2.5)$$

The solutions obtained will be divergence-free, and are the only allowed trial functions for the second stage in which the functional (2.3) is minimized. This approach has drawbacks for very large problems, especially as the advantage of matrix sparsity is lost.

Su [82] studied the origin of spurious modes and proposed a combined method using a finite element technique and a surface integral equation method for lossless isotropic dielectric waveguides [83]. Although the spurious solutions are eliminated, this method can only treat problems with isotropic and smoothly changing inhomogeneous materials.

Angkaew *et al.* [84] developed a γ-type formulation for lossless waveguides in terms of the transverse components of both electric and magnetic fields. With this approach, real eigenvalues are the distinctive physical solutions of a complex eigenvalue problem. In other words, spurious eigenvalues move off the real axis in the complex plane so that the real eigenvalues are genuine. One might expect difficulty in choosing the threshold figure to distinguish 'real' and 'complex' eigenvalues. This approach suffers considerable expense of increasing computing effort, as four complex unknowns per node are needed for lossless problems.

Using the Galerkin method, Hayata *et al.* [81] suggested an approach which can eliminate spurious solutions in terms of only the transverse magnetic field components for anisotropic lossless waveguide problems. By first applying standard finite element techniques to the **H**-field Helmholtz equation and the divergence equation, they arrive respectively at

$$[S] \{H\} - k_0^2 [T] \{H\} = \{0\} \qquad (2.6)$$

and

$$[D_z] \{H_z\} = [D_t] \{H_t\} \qquad (2.7)$$

where

$$\{H\} = [D] \{H_t\} \qquad (2.8)$$

$$[D] = \begin{bmatrix} [U] \\ [D_z]^{-1}[D_t] \end{bmatrix}, \text{ ([U] is a unit matrix)} \tag{2.9}$$

In (2.7), the phase constant β is implicitly included in both $[D_z]$ and $[D_t]$. Using (2.6) to (2.9) one straightforwardly arrives at

$$[\tilde{S}_u]\ \{H_t\} - k_0^2\ [\tilde{T}_u]\ \{H_t\} = \{0\} \tag{2.10}$$

where

$$[\tilde{S}_u] = [D]^T[S]\ [D] \tag{2.11}$$

$$[\tilde{T}_u] = [D]^T[T]\ [D] \tag{2.12}$$

Due to the introduction of matrix inversion and multiplication, the sparsity of the resultant matrices $[\tilde{S}_u]$ and $[\tilde{T}_u]$ have been sacrificed.

It is worth mentioning that in [81] the final formulation is expressed as

$$[\tilde{S}_u]\ \{H_t\} - (k_0/\beta)^2\ [\tilde{T}_u]\ \{H_t\} = \{0\} \tag{2.13}$$

This is a somewhat misleading. Although the eigenvalue is expressed as $(k_0/\beta)^2$, one is not at liberty to choose the frequency and calculate β since the matrices are also functions of β. The real eigenvalue of this problem is k_0^2.

Later, Hayata *et al.* extended their method to diagonal anisotropic and lossy waveguide problems [85]. Following the same procedure (2.6) to (2.10) at first, they next expressed (2.10) to be a complex quadratic eigenequation with eigenvalue $\lambda = -\gamma^2$ as:

$$\lambda^2\ [A]\ \{H_t\} + \lambda\ [B]\ \{H_t\} + [C]\ \{H_t\} = \{0\} \tag{2.14}$$

where matrices [B] and [C] are obtained by a series of matrix operations (including inversion, transposition, and addition). Finally, (2.14) is transformed into an eigenequation with double order

$$\begin{bmatrix} [0] & [U] \\ -[A]^{-1}[C] & -[A]^{-1}[B] \end{bmatrix} \begin{bmatrix} \{H_t\} \\ \lambda\{H_t\} \end{bmatrix} = \lambda \begin{bmatrix} \{H_t\} \\ \lambda\{H_t\} \end{bmatrix} \tag{2.15}$$

The formulation (2.15) which they claimed to be an 'efficient' standard formulation has two disadvantages: (i) it doubled the unknowns to $4N_p$, where N_p the number of nodal points; and more importantly, (ii) the complicated matrix operation will considerably increase the computing effort and lose sparsity which is fatal for large problems even with recourse to supercomputers. As shown in their examples, a simple mesh of

153 nodes requires 27 MB memory and about 40 seconds to obtain one point in the dispersion curve by using a Hitachi S-810/10 supercomputer (with peak speed about 800 MFLOPS) [97].

Following Chen and Lien's principles for non-self-adjoint problems [38], Chew and Nasir [86] proposed a four component variational formulation, which incorporates the divergence-free condition to eliminate spurious solutions, in terms of the transverse components of both electric and magnetic fields for anisotropic dielectric waveguides with permittivity tensor as

$$[\varepsilon] = \begin{bmatrix} \varepsilon_{xx} & \varepsilon_{xy} & 0 \\ \varepsilon_{yx} & \varepsilon_{yy} & 0 \\ 0 & 0 & \varepsilon_{zz} \end{bmatrix} \tag{2.16}$$

This formulation looks very attractive because (i) it can be reduced to be in terms of only (H_x,H_y), (ii) it is a γ-type formulation. However, the reduced formulation is derived by disregarding electric field discontinuities across dielectric interfaces. Furthermore, their special treatment of the wave equation results in the two components of the transverse field being inherently weakly coupled. The resultant matrix equation of their reduced (H_x,H_y) formulation has the form

$$\begin{bmatrix} [R] & [S] \\ -[S] & [R] \end{bmatrix} \begin{bmatrix} \{H_x\} \\ \{H_y\} \end{bmatrix} = \gamma^2 \begin{bmatrix} [0] & [T] \\ -[T] & [0] \end{bmatrix} \begin{bmatrix} \{H_x\} \\ \{H_y\} \end{bmatrix} \tag{2.17}$$

And [P] and [S] are found to be highly sensitive to the choice of type of finite element used. For instance, for square elements [R] = [0] so that $\{H_x\}$ and $\{H_y\}$ are totally decoupled, in other words, (2.17) is degraded into two independent scalar equations of H_x and H_y respectively, which can not represent a general hybrid problem. Because of their use of the same vector trial function for both **E** and **H** in their derivation, boundary conditions on electric and magnetic walls can not be properly forced; neither are they implied. Using special type of element and without taking any plane of symmetry, they only give examples of open waveguides with electric or magnetic wall far from the region of strong fields. Because Chew and Nasir's formulation is not robust and is not able to analyze waveguides with conductors (including symmetry planes), this method is not suitable to general waveguide problems.

Svedin [87] proposed a formulation in terms of all six components of the electric and magnetic fields. The divergence of electric and magnetic field vectors is fixed implicitly to zero, and all tangential and normal interface and boundary conditions are enforced, so no spurious modes appear. This method gives direct solution for propagation constant. And it can treat the most general anisotropic materials with full permittivity and permeability tensors. Using six variables is the main disadvantage of this method. It can be reduced to four, but then the sparsity is lost. Both cases affect the computational efficiency with considerable increase of computing time and storage requirement for large problems. In addition, to enforce normal component of electric field at the dielectric interfaces introduces complexity. Besides, Svedin's corresponding [B] matrix in the canonical form (2.2) is singular, introducing some zero non-physical solutions.

Bardi and Biro [89] proposed a finite element formulation for lossless anisotropic waveguides. This formulation is a four-component, ω-type formulation of the form

$$\begin{bmatrix} [M_{aa}] & [0] \\ [0] & [0] \end{bmatrix} \begin{bmatrix} \{A\} \\ \{V\} \end{bmatrix} = k_0^2 \begin{bmatrix} [N_{aa}] & [N_{av}] \\ [N_{va}] & [N_{vv}] \end{bmatrix} \begin{bmatrix} \{A\} \\ \{V\} \end{bmatrix} \quad (2.18)$$

where {A} represents the three components of a magnetic (or electric) vector potential and {V} the electric (or magnetic) scalar potential. The matrix on the left side of (2.18) is singular. Therefore, there are degenerate eigenvectors that correspond to $k_0^2 = 0$, these eigenvectors are non-physical spurious solutions though they may be easily distinguished. Besides, additional computing effort is needed to get electric or magnetic field distribution. This method is not efficient because it uses four variables.

A recent scheme for avoiding spurious solutions is the application of *edge elements* [90] and their generalization, *tangential elements* [91]. In this approach, the tangential field components between elements are forced to be continuous. The advantages of this approach are that (i) it imposes only the continuity of the tangential components of the electric and magnetic fields, as required physically; (ii) the interface boundary

conditions are automatically obtained through the natural boundary conditions built into the variational principle [92]. Most recently, Lee and Cendes presented a first order tangential elements method [92] for lossless dielectric waveguide problems. Their resultant matrix eigenvalue equation has the form:

$$\begin{bmatrix} [A] & [0] \\ [0] & [0] \end{bmatrix} \begin{bmatrix} \{E_t\} \\ \{E_z\} \end{bmatrix} = -\beta^2 \begin{bmatrix} [B] & [C]^T \\ [C] & [D] \end{bmatrix} \begin{bmatrix} \{E_t\} \\ \{E_z\} \end{bmatrix} \qquad (2.19)$$

Similar to (2.18), the matrix on the left side of (2.19) is singular. There are degenerate zero non-physical solutions.

Another defect of this method is that it cannot treat lossy waveguides neither can it find complex modes in lossless waveguides. Because the matrices on both sides of (2.19) are real and symmetric, no complex eigenvalues can appear. Therefore, this formulation is not complete.

2.4 Open-Boundary Problems

Open dielectric waveguide structures are becoming increasingly important for integrated optical devices and optical communication systems. In optical waveguides, the region of interest extends to infinity outside some guide 'core', where the field decay is roughly exponential.

The crudest approach is the simple truncation at a certain distance [47], [52], which sets artificial electric or magnetic walls enclosing the waveguide. But this approach either introduces a significant error when the boundary is too close, or needs to consider an excessively large domain. One adaptive technique involves shifting the virtual boundary wall recursively to satisfy a criterion for maximum field strength at that wall [51].

Open-boundary problems may be solved more accurately by several techniques, all having in common the idea that the open infinitely-extending region is divided into interior and exterior regions [93]-[96]. It is possible to find the internal and external solutions and to match them on an imaginary boundary, some choice of integral solution being possible for the outside region [83]. But it is quite complicated to consider an integral equation method for this unbounded region without losing the canonical form of the matrix problem. Another approach is to use a recursion technique [30], [93] to represent the region outside the

main domain. Wu and Chen [31] used conformal mapping to condense the exterior region, but this technique can only be applied to open guides having a plane of symmetry.

Yeh *et al.* [50] used sectorial infinite elements, with radial exponential decay outside the guide core, but because of nonconformity between the two coordinate systems, inter-element conditions cannot be satisfied exactly along the interface between the standard elements and sectorial infinite elements.

Infinite elements have also been used with a Cartesian coordinate system using suitable exponentially decaying functions [51], [63], [69], [95]. Iterative procedures are proposed in [51] that allow a self-consistent determination of the optimum decay length by using the previous eigenvalue or eigenvector.

A different approach is proposed in [95], one that removes the need to iterate for an optimum decay parameter. Instead, a set of decay lengths has to be chosen by the user. These infinite elements extend the domain of explicit field representation to infinity without increasing the matrix order, so that computational time is virtually unchanged. The shape functions for such an element should be realistic to represent the fields and should be square integrable over the infinite area to satisfy the radiation condition. Adding these infinite elements along the outer boundary of orthodox finite elements, any open-type optical waveguide cross-sectional domain can be represented very conveniently.

2.5 Remarks

2.5.1 Lossy Waveguides

To date most of the applications of the finite element method have been restricted to lossless dielectric waveguide problems. For waveguide with significant loss, the full wave vectorial analysis is necessary.

In an ω-type formulation, a complex eigenvalue problem is solved iteratively until a real eigenvalue (frequency) is obtained [56]. However, in a lossy case, complex propagation constants have to be guessed, while the guess of a complex number is difficult. Using this approach, considerable computing time is needed in iteration. In addition, eigenvalues may be very slow or may even not be able to converge.

Obviously, a direct formulation for γ is the only efficient and reliable way to obtain solutions for the propagation constants in the cases of lossy waveguides.

Summarizing all the recent γ-type formulations which can eliminate spurious solutions, there are only three possible formulations [85], [86], [87] for lossy waveguide problems. Among them, only Hayata *et al.* [85] have shown application to lossy waveguide problems. However Hayata uses a four-component, dense-matrix equation, which is not efficient and cannot be applied to large problems. Chew and Nasir's formulation [86] is ill-conditioned and cannot be used as a general method. Svedin's six-component formulation [87] is not efficient and suffers considerable cost of computing resource.

2.5.4 Criteria of Judgments

In judging the appropriateness of a finite element formulation for dielectric waveguide, the following 8 criteria may be adopted.

1) The formulation should be robust and capable of including as many waveguide features, such as arbitrarily shaped cross-section, inhomogeneity, anisotropy, and significant loss (or gain), as possible.
2) The formulation should be immune from spurious solutions which often plague the finite element solution of vector variational formulations.
3) The resultant matrix equation of the formulation should also be well-conditioned.
4) The electromagnetic field should be represented in terms of only vector magnetic field **H**, which is more suitable to dielectric waveguide problems, because the magnetic field is continuous everywhere, and no special treatment is needed to enforce normal component of electric field at dielectric interfaces.
5) The solution should be direct for complex propagation constant in terms of real frequency, so as to be more efficient and reliable.
6) If possible, the formulation variables should be represented by only two field components, the least number necessary, thus minimizing the unknowns.
7) The formulation should lead to a canonical eigenvalue equation which can be solved efficiently.

8) The resultant matrices of the formulation should be highly sparse and able to utilize a fast and efficient matrix equation solver. This is of decisive importance for large problems, even on a supercomputer.

Criterion 1 refers to the problem coverage; criteria 2 and 3 refer to the applicability; criteria 4 and 5 refer to simplicity and user friendliness; criteria 5 to 8 refer to ability to treat large problems.

We will show in the following chapters that the formulation presented in this thesis satisfies all the above criteria except it is for transverse-anisotropic rather than for the most general anisotropic cases. Nevertheless, this is enough for most of applications.

CHAPTER 3

PROBLEM DEFINITION

3.1 Introduction

Computer simulation of the electromagnetic fields within a dielectric waveguide begins with a mathematical description of the problem. In this chapter the primary concern is the interpretation of the physical problem in mathematical terms. We will describe in section 3.2 the mathematical model of dielectric waveguide problems. In section 3.3, we will discuss in general the origin of spurious modes in approximate solutions of vector Maxwell boundary-value problems; based on this discussion we will propose the strategies to eliminate spurious modes in vector finite element solutions of Maxwell boundary-value problems. Then in section 3.4 we will derive an appropriate expression in terms of transverse magnetic field components only, which is immune from spurious solutions, and also is suitable and economic to represent dielectric waveguide problems. From this boundary-value problem definition, we will derive a new variational finite element formulation in Chapter 5.

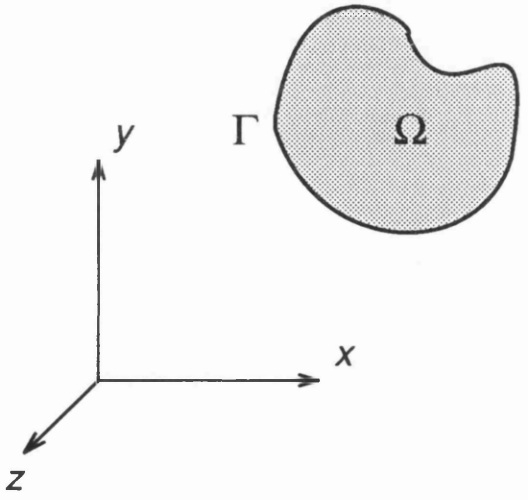


Fig. 3.1 The arbitrary dielectric waveguide structure

3.2 Boundary-Value Problem

3.2.1 Description of the Dielectric Waveguide Problem

In this thesis we consider a dielectric waveguide as depicted in Fig. 3.1, with arbitrary cross section in the *x-y* plane and uniformity in the *z*-direction. The structure region Ω, which may include guide, substrate and cladding, etc., consists of linear dielectric material(s) and electric conductors. The dielectric material in Ω may be arbitrarily inhomogeneous, anisotropic, and dissipative. The structure of the waveguide may be open or closed. We assume that Γ, the boundary of Ω, is divided into three parts: the perfect electric conductor (PEC), the perfect magnetic conductor (PMC), and infinity (INF).

The assumption that the permeability of all dielectric materials is the constant scalar μ_0 everywhere is also made. The relative permittivity profile is assumed by the complex tensor

$$\bar{\bar{\varepsilon}}(x,y,\omega) = (\bar{\bar{\varepsilon}}' - j\bar{\bar{\varepsilon}}'') = \begin{bmatrix} \varepsilon_{xx} & \varepsilon_{xy} & 0 \\ \varepsilon_{xy} & \varepsilon_{yy} & 0 \\ 0 & 0 & \varepsilon_{zz} \end{bmatrix} \tag{3.1}$$

The assumption of (3.1) implies that the geometry in Fig. 3.1 has reflection symmetry about *z* axis; i.e., a mode propagating in the $+z$ direction is degenerate with a mode propagating in the $-z$ direction [15].

3.2.2 General Boundary-Value Problem Definition

In order to discuss the origin and elimination of spurious modes in general, let us describe the general boundary-value problem first. Assuming a harmonic time dependence of the form $exp(j\omega t)$, where ω is the real angular frequency, the governing source-free Maxwell's equations for a general boundary-value problem are

$$\nabla \times \mathbf{E} = -j\,\omega\,\mathbf{B} = -j\,\omega\,\mu_0\bar{\bar{\mu}}\cdot\mathbf{H} \tag{3.2a}$$

$$\nabla \times \mathbf{H} = j\,\omega\,\mathbf{D} = j\,\omega\,\varepsilon_0\bar{\bar{\varepsilon}}\cdot\mathbf{E} \tag{3.2b}$$

$$\nabla \cdot \mathbf{D} = \nabla \cdot (\varepsilon_0\bar{\bar{\varepsilon}}\cdot\mathbf{E}) = 0 \tag{3.2c}$$

$$\nabla \cdot \mathbf{B} = \nabla \cdot (\mu_0 \bar{\bar{\mu}} \cdot \mathbf{H}) = 0 \qquad (3.2d)$$

where $\mathbf{E}$ the electric field intensity vector

$\mathbf{H}$ the magnetic field intensity vector

$\mathbf{D}$ the electric displacement intensity vector

$\mathbf{B}$ the magnetic induction intensity vector

$\bar{\bar{\varepsilon}}$ the relative permittivity tensor

$\bar{\bar{\mu}}$ the relative permeability tensor

ε_0 the vacuum permittivity scalar

μ_0 the vacuum permeability scalar

The field **E, D, H, B** should satisfy the associated boundary conditions. At a discontinuous interface of two contiguous media *a* and *b*, the boundary conditions are

$$\mathbf{n} \times (\mathbf{E}_a - \mathbf{E}_b) = \mathbf{0} \qquad (3.3a)$$

$$\mathbf{n} \times (\mathbf{H}_a - \mathbf{H}_b) = \mathbf{0} \qquad (3.3b)$$

$$\mathbf{n} \cdot (\mathbf{D}_a - \mathbf{D}_b) = \varepsilon_0 \, \mathbf{n} \cdot (\bar{\bar{\varepsilon}}_a \cdot \mathbf{E}_a - \bar{\bar{\varepsilon}}_b \cdot \mathbf{E}_b) = 0 \qquad (3.3c)$$

$$\mathbf{n} \cdot (\mathbf{B}_a - \mathbf{B}_b) = \mu_0 \, \mathbf{n} \cdot (\bar{\bar{\mu}}_a \cdot \mathbf{H}_a - \bar{\bar{\mu}}_b \cdot \mathbf{H}_b) = 0 \qquad (3.3d)$$

where **n** is a normal unit vector at the interface between two media *a* and *b*, the direction of **n** is from medium *a* towards medium *b*.

On PEC, the boundary conditions are

$$\mathbf{n} \times \mathbf{E} = \mathbf{0} \qquad (3.3e)$$

$$\mathbf{n} \cdot \mathbf{B} = 0 \quad (\text{or} \quad \mathbf{n} \cdot \bar{\bar{\mu}} \cdot \mathbf{H} = 0) \qquad (3.3f)$$

On PMC, the boundary conditions are

$$\mathbf{n} \times \mathbf{H} = \mathbf{0} \qquad (3.3g)$$

$$\mathbf{n} \cdot \mathbf{D} = 0 \quad (\text{or} \quad \mathbf{n} \cdot \bar{\bar{\varepsilon}} \cdot \mathbf{E} = 0) \qquad (3.3h)$$

And at INF, all the fields vanish:

$$\mathbf{E} = \mathbf{D} = \mathbf{H} = \mathbf{B} = \mathbf{0} \tag{3.3i}$$

For simplicity, we will ignore (3.3i) in the following discussion.

Classically, the boundary-value problem is unambiguously defined by the two curl equations ((3.2a), (3.2b)), and either the tangential boundary conditions ((3.3a), (3.3b), (3.3e), (3.3g)) or the normal boundary conditions ((3.3c), (3.3d), (3.3f), (3.3h)). Solutions to these implicitly satisfy the two divergence equations ((3.2c) and (3.2d)) and their corresponding complementary boundary conditions.

Alternatively, the boundary-value problem can be defined by only one field (i.e., the electric field or the magnetic field) and the associated boundary conditions. This may simplify the problem and increase the efficiency of solutions. The other field can be obtained later by one of the two curl relations if necessary. In this way, the two curl equations (3.2a) and (3.2b) can be transformed into a double-curl equation in terms of the magnetic field or electric field only. For example, the double-curl magnetic field equation:

$$\nabla \times (\bar{\bar{\varepsilon}}^{-1} \cdot \nabla \times \mathbf{H}) - \omega^2 \varepsilon_0 \mu_0 \bar{\bar{\mu}} \cdot \mathbf{H} = \mathbf{0} \tag{3.4}$$

and its tangential boundary conditions:

$$\mathbf{n} \times (\mathbf{H}_a - \mathbf{H}_b) = \mathbf{0} \tag{3.5a}$$

$$\mathbf{n} \times (\bar{\bar{\varepsilon}}_a^{-1} \cdot \nabla \times \mathbf{H}_a - \bar{\bar{\varepsilon}}_b^{-1} \cdot \nabla \times \mathbf{H}_b) = \mathbf{0} \tag{3.5b}$$

$$\mathbf{n} \times \mathbf{H} = \mathbf{0} \qquad (\text{ on PMC }) \tag{3.5c}$$

$$\mathbf{n} \times (\bar{\bar{\varepsilon}}^{-1} \cdot \nabla \times \mathbf{H}) = \mathbf{0} \qquad (\text{ on PEC }) \tag{3.5d}$$

or its normal boundary conditions:

$$\mathbf{n} \cdot (\bar{\bar{\mu}}_a \cdot \mathbf{H}_a - \bar{\bar{\mu}}_b \cdot \mathbf{H}_b) = 0 \tag{3.6a}$$

$$\mathbf{n} \cdot (\nabla \times \mathbf{H}_a - \nabla \times \mathbf{H}_b) = 0 \tag{3.6b}$$

$$\mathbf{n} \cdot \bar{\bar{\mu}} \cdot \mathbf{H} = 0 \qquad (\text{ on PEC }) \tag{3.6c}$$

$$\mathbf{n} \cdot \nabla \times \mathbf{H} = 0 \qquad (\text{ on PMC }) \tag{3.6d}$$

Similarly one can have the double-curl electric field equation:

$$\nabla \times (\bar{\bar{\mu}}^{-1} \cdot \nabla \times \mathbf{E}) - \omega^2 \varepsilon_0 \mu_0 \bar{\bar{\varepsilon}} \cdot \mathbf{E} = \mathbf{0} \tag{3.7}$$

and its tangential boundary conditions:

$$\mathbf{n} \times (\mathbf{E}_a - \mathbf{E}_b) = \mathbf{0} \tag{3.8a}$$

$$\mathbf{n} \times (\bar{\bar{\mu}}_a^{-1} \cdot \nabla \times \mathbf{E}_a - \bar{\bar{\mu}}_b^{-1} \cdot \nabla \times \mathbf{E}_b) = \mathbf{0} \tag{3.8b}$$

$$\mathbf{n} \times \mathbf{E} = \mathbf{0} \qquad (\text{ on PEC }) \tag{3.8c}$$

$$\mathbf{n} \times (\bar{\bar{\mu}}^{-1} \cdot \nabla \times \mathbf{E}) = \mathbf{0} \qquad (\text{ on PMC }) \tag{3.8d}$$

or its normal boundary conditions:

$$\mathbf{n} \cdot (\bar{\bar{\varepsilon}}_a \cdot \mathbf{E}_a - \bar{\bar{\varepsilon}}_b \cdot \mathbf{E}_b) = 0 \tag{3.9a}$$

$$\mathbf{n} \cdot (\nabla \times \mathbf{E}_a - \nabla \times \mathbf{E}_b) = 0 \tag{3.9b}$$

$$\mathbf{n} \cdot \bar{\bar{\varepsilon}} \cdot \mathbf{E} = 0 \qquad (\text{ on PMC }) \tag{3.9c}$$

$$\mathbf{n} \cdot \nabla \times \mathbf{E} = 0 \qquad (\text{ on PEC }) \tag{3.9d}$$

3.3 Origin and Elimination of Spurious Modes

Non-physical spurious solutions have been observed in finite element [62], [63], [65] and finite difference [100] formulations based on the above mentioned two-curl equation or double-curl equation definitions. Spurious solutions have been found mainly in modal analysis (see Chapter 2) and finite element formulations for driven problems have largely been perceived to be free of these computational difficulties. In fact, this presumed immunity of driven problems has been suggested as a possible remedy to the eigenvalue dilemma [102], [103]. Recent studies, however, show that some selections of the forcing term can lead to completely erroneous finite element solutions of some simple double-curl boundary value problems [104]-[107]. Above findings imply that for some double-curl finite element formulations, even if spurious solutions are not normally observed, the formulations themselves are not inherently immune from spurious modes, they just incidentally avoid the appearance of spurious

solutions.

As reviewed in Chapter 2, spurious modes are observed in formulations where the divergence condition $\nabla \cdot \mathbf{B} = 0$ is not imposed [62], [63], [67]. Consequently, the introduction of the divergence condition is suggested and applied [65], [67], [81], [84], [86]. However, the questions of 'Why should we have to introduce the divergence condition' and 'Is it sufficient to eliminate all spurious modes' have never been explained convincingly.

Based on the general lack of insight studies of spurious modes in finite element solutions, it is worth trying to explain the inherent origin of spurious modes. To begin with, let us investigate the analytical and approximate solutions of Maxwell boundary-value problems.

3.3.1 Analytical Solutions to Maxwell's Equations

In analytical approaches, the two-curl or the double-curl definition mentioned in section 3.2 has widely been adopted. For most analytical methods, the curliness of the field solution can be guaranteed at almost every point in the problem domain. Because of the curliness of the field solution, the tangential and normal boundary conditions are derivable from each other. This also implies that the field solutions satisfy the divergence equations $\nabla \cdot \mathbf{B} = 0$ and $\nabla \cdot \mathbf{D} = 0$, which are the remaining equations governing the electromagnetic field and should be, but are not included in the problem definition. When the complete solutions can be fulfilled by the two-curl or the double-curl definitions, it is, of course, unnecessary to include the divergence conditions. Therefore, using the two-curl or the double-curl definition is usually sufficient to achieve true analytical solutions of the fields.

3.3.2 Approximate Solutions to Maxwell's Equations

In numerical methods, however, the story is different. For such weak approximation as the finite element solution, the curlinesses of field solutions can not be guaranteed. For the double-curl equation definition, derivative boundary-conditions such as (3.5b), (3.5d), (3.6b), (3.6d) cannot be strictly imposed. Hence, the tangential and normal boundary conditions are no longer automatically derivable from each other and the divergence conditions cannot be implied in the two curl equations or the

double-curl equation definition. As a consequence, the problem is underdetermined and non-physical, spurious solutions may appear with the two-curl or these problem definitions. Therefore the two-curl equation or the double-curl equation definitions of boundary-value problems are no longer sufficient to determine completely a boundary-value problem.

3.3.3 Elimination of Spurious Solutions

Based on the discussion in section 3.3.2, three Maxwell source-free boundary-value problem definitions are proposed in order to eliminate spurious solutions.

Strategy 1

For full field (**E**,**H**) *approximation, all four equations* ((3.2a-d)), *and both tangential and normal boundary conditions* ((3.3a-h)) *should be used in the problem definition in order to eliminate the spurious modes.*

Strategy 2

For magnetic field **H** *approximation, magnetic double-curl equation* (3.4), *magnetic field divergence equation* (3.2d), *and both associated magnetic field tangential and normal boundary conditions* ((3.5a-d) *and* (3.6a-d)) *should be used in the problem definition in order to eliminate the spurious modes.*

Strategy 3

For electric field **E** *approximation, electric double-curl equation* (3.7), *electric field divergence equation* (3.2c), *and both associated electric field tangential and normal boundary conditions* ((3.8a-d) *and* (3.9a-d)) *should be used in problem definition in order to eliminate the spurious modes.*

Note that from (3.2a-d) and vector identity $\nabla\cdot(\nabla\times\mathbf{A}) \equiv 0$, the following is always true:

$$\nabla\cdot\mathbf{D} = \nabla\cdot(\bar{\bar{\varepsilon}}\cdot\mathbf{E}) = \frac{1}{j\,\omega\,\varepsilon_0}\,\nabla\cdot(\nabla\times\mathbf{H}) \equiv 0 \qquad (3.10)$$

$$\nabla\cdot\mathbf{B} = \nabla\cdot(\bar{\bar{\mu}}\cdot\mathbf{H}) = \frac{-1}{j\,\omega\,\varepsilon_0}\,\nabla\cdot(\nabla\times\mathbf{E}) \equiv 0 \qquad (3.11)$$

(3.10) and (3.11) show that the only remaining condition for problem definition in *Strategy 2 and Strategy 3* is always automatically satisfied.

Remark 3.1

Any one of the definitions described in Strategies 1, 2, and 3 is a complete and sufficient definition for a Maxwell source-free boundary-value problem. It is also sufficient to eliminate the spurious modes.

For the **E** or **H** approximation, the derivatives of the fields are used in the boundary conditions. It is in general difficult to impose them strictly. However, for such weak approximation as the finite element method, they may be fulfilled in a wide mean-value sense, or they may not have all to be imposed for the mostly desired one field first-order approximation where we only have to impose the essential tangential and normal boundary conditions of the field itself. This is reasonable and acceptable, and should not deteriorate the original definition in the weak approximation sense and therefore should not introduce spurious modes in the weak solutions.

3.4 Basic Differential Equation

For the dielectric waveguide problem described in section 3.2.1, the permeability is the constant scalar μ_0. As $\overline{\overline{\mu}}$ in this case is an unit tensor, from the interface conditions (3.3a)-(3.3d), we can see that at the interface of two contiguous media both tangential and normal components of the magnetic field are continuous while the normal component of electric field is not. This means that in the whole region Ω all components of the magnetic field vector **H** are continuous everywhere while all components of the electric field vector are not. Obviously, in this case it is much more convenient to define the dielectric waveguide problem in terms of the magnetic field only. Hence, the definition described in *Strategy 2* will be adopted.

However, as the permittivity tensor is of the special form (3.1), we may simplify the boundary-value problem definition (3.4)-(3.6) further to include only two transverse components of the magnetic field, which are the minimum number of components required to represent a general problem.

Denoting $\overline{\overline{\varepsilon}}_{tt}$ as the 2×2 tensor

$$\bar{\bar{\varepsilon}}_{tt} = \begin{bmatrix} \varepsilon_{xx} & \varepsilon_{xy} \\ \varepsilon_{xy} & \varepsilon_{yy} \end{bmatrix} \tag{3.12}$$

tensor $\bar{\bar{\varepsilon}}$ can be represented as:

$$\bar{\bar{\varepsilon}} = \bar{\bar{\varepsilon}}_{tt} + \varepsilon_{zz}\mathbf{zz} \tag{3.13}$$

The magnetic field double-curl equation (3.4) has all three components of the magnetic field which are more than the minimum of two desired. Incorporating the divergence-free condition (3.2d) into (3.4), we can reduce the number of components in the field equation to the two transverse components H_x and H_y only. To achieve this purpose, we next proceed to separate the transverse and longitudinal components of equation (3.6a). Because the waveguide is assumed uniform in the *z*-direction, we may assume that the field has a *exp(-γz)* dependence so that the operator

$$\frac{\partial}{\partial z} = -\gamma \tag{3.14}$$

We define

$$\mathbf{H}(x,y,z) = [\ \mathbf{H}_t(x,y) + \mathbf{H}_z(x,y)\ ]\ e^{-\gamma z} \tag{3.15}$$

and
$$\nabla = \nabla_t + \mathbf{z}\frac{\partial}{\partial z} = \nabla_t - \mathbf{z}\,\gamma \tag{3.16}$$

where the subscript *t* denotes a vector transverse to *z*.

Introducing (3.14) and (3.15) into equation (3.4), the wave equation (3.4) can be separated into two equations. One is the transverse component of (3.4), the other is the longitudinal component of (3.4). The transverse component of equation (3.4) becomes:

$$\nabla_t \times (\kappa_{zz}\nabla_t \times \mathbf{H}_t) - \gamma\,\mathbf{z} \times [\ \bar{\bar{\kappa}}_{tt} \cdot \nabla_t \times \mathbf{H}_z\ ) - \omega^2\mu_0\varepsilon_0\,\mathbf{H}_t + \gamma^2\mathbf{z} \times [\ \bar{\bar{\kappa}}_{tt} \cdot (\mathbf{z} \times \mathbf{H}_t)\ ] = \mathbf{0} \tag{3.17}$$

where we have defined

$$\bar{\bar{\kappa}} = \bar{\bar{\varepsilon}}^{-1} \tag{3.18a}$$

and consequently

$$\bar{\bar{\kappa}}_{tt} = \bar{\bar{\varepsilon}}_{tt}^{-1} \tag{3.18b}$$

$$\kappa_{zz} = \varepsilon_{zz}^{-1} \tag{3.18c}$$

We can remove H_z in (3.17) by incorporating the divergence-free condition (3.2d), from which we have

$$H_z = \frac{\nabla_t \cdot \mathbf{H}_t}{\gamma} \tag{3.19}$$

Substituting (3.19) into (3.17), we reduce (3.17) to an equation involving only transverse magnetic field components $\mathbf{H}_t$, viz.,

$$\nabla_t \times (\kappa_{zz} \nabla_t \times \mathbf{H}_t) - \mathbf{z} \times [\bar{\bar{\kappa}}_{tt} \cdot \nabla_t \times (\mathbf{z}\, \nabla \cdot \mathbf{H}_t)] - \omega^2 \mu_0 \varepsilon_0 \mathbf{H}_t$$
$$+ \gamma^2 \mathbf{z} \times [\bar{\bar{\kappa}}_{tt} \cdot (\mathbf{z} \times \mathbf{H}_t)] = \mathbf{0} \tag{3.20}$$

The above is an eigenvalue problem with eigenvalue γ^2. The dependence on γ^2 implies that *exp(±γz)* modes are degenerate. This would not have been possible without assuming the form of $\bar{\bar{\varepsilon}}$ in (3.1)

Introducing (3.15), (3.16) and (3.19) into the complete boundary conditions (3.5) and (3.6) of **H**-definition, after performing separation, we can get the complete boundary conditions for wave equation (3.20) in terms of the transverse magnetic field component $\mathbf{H}_t$ only as follows:

$$\mathbf{n} \cdot (\mathbf{H}_{ta} - \mathbf{H}_{tb}) = 0 \tag{3.21}$$

$$\mathbf{n} \times (\mathbf{H}_{ta} - \mathbf{H}_{tb}) = \mathbf{0} \tag{3.22a}$$

$$\nabla_t \cdot \mathbf{H}_{ta} - \nabla_t \cdot \mathbf{H}_{tb} = 0 \tag{3.22b}$$

$$\mathbf{n} \cdot \Big\{ [\nabla_t \times (\mathbf{z} \nabla_t \cdot \mathbf{H}_{ta}) - \gamma^2 \mathbf{z} \times \mathbf{H}_{ta}]$$
$$- [\nabla_t \times (\mathbf{z} \nabla_t \cdot \mathbf{H}_{tb}) - \gamma^2 \mathbf{z} \times \mathbf{H}_{tb}] \Big\} = 0 \tag{3.23}$$

$$\mathbf{n} \times \Big\{ \bar{\bar{\varepsilon}}_{tta}^{-1} \cdot [\nabla_t \times (\mathbf{z} \nabla_t \cdot \mathbf{H}_{ta}) - \gamma^2 \mathbf{z} \times \mathbf{H}_{ta}]$$
$$- \bar{\bar{\varepsilon}}_{ttb}^{-1} \cdot [\nabla_t \times (\mathbf{z} \nabla_t \cdot \mathbf{H}_{tb}) - \gamma^2 \mathbf{z} \times \mathbf{H}_{tb}] \Big\} = \mathbf{0} \tag{3.24a}$$

$$\mathbf{n} \times (\varepsilon_{zza}^{-1} \nabla_t \times \mathbf{H}_{ta} - \varepsilon_{zzb}^{-1} \nabla_t \times \mathbf{H}_{tb}) = 0 \tag{3.24b}$$

$$\mathbf{n} \cdot \mathbf{H}_t = 0 \qquad \text{(on PEC)} \qquad (3.25)$$

$$\mathbf{n} \times \mathbf{H}_t = 0 \qquad \text{(on PMC)} \qquad (3.26a)$$

$$\nabla \cdot \mathbf{H}_t = 0 \qquad \text{(on PMC)} \qquad (3.26b)$$

$$\mathbf{n} \cdot \left\{ [\nabla_t \times (\mathbf{z} \nabla_t \cdot \mathbf{H}_t) - \gamma^2 \mathbf{z} \times \mathbf{H}_t] \right. \qquad \text{(on PMC)} \qquad (3.27)$$

$$\mathbf{n} \times \bar{\bar{\varepsilon}}_{tt}^{-1} \cdot [\nabla_t \times (\mathbf{z} \nabla_t \cdot \mathbf{H}) - \gamma^2 \mathbf{z} \times \mathbf{H}_t] = \mathbf{0} \qquad \text{(on PEC)} \qquad (3.28a)$$

$$\mathbf{n} \times \nabla_t \times \mathbf{H}_t = 0 \qquad \text{(on PEC)} \qquad (3.28b)$$

Eqs. (3.21) and (3.25) are the normal magnetic induction boundary conditions. Eq. (3.21) is equivalent to eq. (3.3d) or eq. (3.6a). And eq. (3.25) is equivalent to eq. (3.3f) or eq. (3.6c).

Eqs. (3.22) and (3.26) reflect the tangential magnetic field boundary conditions. Eq. (3.22a) plus eq. (3.22b) are equivalent to eq. (3.3b) or eq. (3.5a). And eq. (3.26a) plus eq. (3.26b) are equivalent to eq. (3.3g) or eq. (3.5c).

Eqs. (3.23) and (3.27) are the normal electric displacement boundary conditions. Eq. (3.23) is equivalent to eq. (3.3c) or eq. (3.6b). And eq. (3.27) is equivalent to eq. (3.3h) or eq. (3.6d).

Eqs. (3.24) and (3.28) are the tangential electric field boundary conditions. Eq. (3.24a) plus eq. (3.24b) are equivalent to eq. (3.3a) or eq. (3.5b). And eq. (3.28a) plus eq. (3.28b) are equivalent to eq. (3.3e) or eq. (3.5d).

The differential equation (3.20) is the basic wave equation in terms of only two transverse magnetic field components for the problem defined in section 3.2. We will use eq. (3.20) and its associated boundary conditions to derive a new variational finite element formulation in chapter 5.

CHAPTER 4

MATHEMATICAL FUNDAMENTALS OF THE FINITE ELEMENT METHOD

4.1 Introduction

The purpose of this chapter is to help us to establish an accurate understanding of the finite element method by summarizing its mathematical fundamentals before applying finite element method to our waveguide problem described in the previous chapter. The emphasis of this chapter is on the philosophy of finite element formulation, the relationship of finite element method to a variety of classic approximation methods, and the techniques for formulating various types of finite element models of boundary-value problems.

In order to establish the generality and flexibility of the finite element method, we begin with a typical example whose notation will used in the entire chapter. Considering an open bounded domain Ω in n-dimension Euclidean space E^n, $\mathbf{x} = (x_1, x_2, ..., x_n)$ being a point in Ω. Let $C^m(\Omega)$ denote the space of functions $u(\mathbf{x})$ with continuous derivatives of order $k \leq m$ on Ω. If $\mathbb{L}$ is a linear partial differential operator, we consider the problem of finding those functions $u(\mathbf{x})$ for which

$$\mathbb{L}(u) = s \tag{4.1}$$

at every $\mathbf{x} \in \Omega$, where s is a fixed function. To make possible the existence of unique solutions to (4.1), we may also impose conditions on $u(\mathbf{x})$ and various of its derivatives at points on the boundary $\partial\Omega = \Gamma$ of the form

$$\mathbb{B}(u) = t \tag{4.2}$$

where $\mathbb{B}$ is also a linear operator and t is a fixed function on Γ. The problem of finding functions u which simultaneously satisfy (4.1) and (4.2) is a linear boundary-value problem.

4.2 Strong and Weak Solutions of Boundary-Value Problems

The finite element method is one of a large class of approximate methods designed to give approximations to *weak* solutions of boundary and initial value problems of mathematical physics. It is necessary to make clear the distinction between strong and weak solutions at first.

4.2.1 *Strong Form of Problem* ([108], [115])

The strong form of the boundary-value problem, (S), is stated as follows:

$$(S)\ \begin{cases} \text{Given } s \text{ and } t, \text{ find } u^* \in C^m(\Omega),\ \forall\ \mathbf{x} \in \Omega \text{ such that} \\ \qquad \mathbb{L}(u) = s \quad \text{on } \Omega \\ \qquad \mathbb{B}(u) = t \quad \text{on } \Gamma \end{cases} \tag{4.3}$$

The functions $u^*(\mathbf{x})$ which satisfy (4.1) and (4.2) at every $\mathbf{x}$ in Ω and Γ are called *strong* solutions of the boundary-value problem.

Some methods of approximation begin directly with the strong statement of the problem. The most notable example is the finite difference method [7]. The finite element method usually starts from a weak form of the problem.

4.2.2 *Weak Form of Problem* ([108], [115], [116])

In general, we can expand the class of functions in which we seek solutions to boundary-value problems by regarding $u(\mathbf{x})$ as an element of Hilbert space $\mathcal{H}$. In other words, we can define for any pair of functions $u,\ v \in \mathcal{H}$, a real number, denoted $\langle u,v \rangle$, that satisfies all of the rules required for inner products. For example

$$\langle u,v \rangle = \int_{\Omega} u \cdot v \, d\Omega \tag{4.4}$$

wherein Lebesgue integration is implied. The associated norm of a function $u \in \mathcal{H}$ is then $\|u\|^2 = \langle u,u \rangle < \infty$ (often mentioned as *square integrable*), and the resulting space (generally denoted $L_2(\Omega)$) is complete in the norm.

Let $\mathcal{H}(\Omega)$ ($\mathcal{H}(\Omega) \subset \mathcal{H}$) denote the *support space* of Ω, i.e, for any $h(\mathbf{x}) \in \mathcal{H}(\Omega)$, the closure of the set of points on which $h(\mathbf{x}) \neq 0$ is contained in Ω. Then the weak form of the boundary-value problem, (W), can be stated as:

$$(W)\ \begin{cases} \text{Given } s,\ \text{find } u^* \in \mathcal{H}(\Omega),\ \forall\ h \in \mathcal{H}(\Omega) \text{ such that} \\ \quad \langle h, \mathbb{L}(u) \rangle = \langle h, s \rangle \end{cases} \tag{4.5}$$

The solutions u^* of weak form of problem are called *weak* (or *generalized*) solutions of the boundary-value problem (4.1). The class of all weak solutions of the boundary-value problem is often much larger than that of the strong solutions, since (4.5) requires only that the integral of $h \cdot \mathbb{L}(u)$ be the same as that of $h \cdot s$. The definition given to a weak formulation is not unique.

In (4.5) it is implicitly assumed that the integrals are capable of being evaluated. This places certain restrictions on the possible families to which the functions h and u must belong. In general we shall seek to avoid functions which result in any term in the integrals becoming infinite. Thus, in (4.5) we limit the choice of h to single, finite value functions without restricting the validity of previous statements. The restriction placed on the functions depends obviously on the order of differentiation implied in the equation $\mathbb{L}(u)$. If n*th*-order derivatives occur in any terms of $\mathbb{L}$ then the function has to be such that its n – 1 derivatives are of C^{n-1} continuity.

On many occasions it is possible to perform an integration by parts on (4.5) and replace it by an alternative statement of the form

$$(W)\ \begin{cases} \text{Given } s,\ \text{find } u^* \in \mathcal{H}(\Omega),\ \forall\ h \in \mathcal{H}(\Omega) \text{ such that} \\ \quad \langle \mathbb{C}(h), \mathbb{D}(u) \rangle + \text{b.t.} = \langle h, s \rangle \end{cases} \tag{4.6}$$

In the above, 'b.t.' stands for boundary terms which we disregard at moment, the operators $\mathbb{C}$ and $\mathbb{D}$ contain lower order derivatives than those occurring in operators $\mathbb{L}$. Now a lower order of continuity is required in the choice of the u function at a price of higher continuity for h. The statement (4.6) is now more 'permissive' than the original problem posed by (4.3), or (4.5), and can also be called the weak form of (4.1).

It is a somewhat surprising fact that often the weak forms are more realistic physically than the original differential differential equation which implied an excessive 'smoothness' of the true solution.

There are two distinct procedures available for obtaining the

approximation in weak forms. The first is the *method of weighted residuals* [8], the second the determination of *variational functional* for which stationary is sought [8], [109], [110].

4.3 The Weighted Residual - Galerkin Method

4.3.1 Dual and Conjugate Space ([108], [111])

The notion of weak solutions of boundary-value problems can be put in a different setting by introducing the notion of dual spaces. Let $\mathcal{U}$ denote a linear vector space, the elements of which may be regarded as functions $u(\mathbf{x})$ of a certain type (say, square integrable), defined on the region $\Omega \subset E^n$. Let $\mathcal{V}$ denote another linear vector space defined over the same field, and suppose that there exists a mapping $\mathcal{S}$: $\mathcal{U} \otimes \mathcal{V} \longrightarrow E$ (i.e., a mapping of ordered pairs $[u,v]$ of vectors into the real numbers) such that

$$\left.\begin{aligned}
&\text{(a) } (u, \alpha v_1 + \beta v_2) = \alpha(u, v_1) + \beta(u, v_2), \\
&\text{(b) } (\alpha u_1 + \beta u_2, v) = \alpha(u_1, v) + \beta(u_2, v), \\
&\text{(c) } (u, v^*)=0, \text{ for fixed } v^* \text{ and all } u \Rightarrow v^*=0 \text{ (almost everywhere)}, \\
&\text{(d) } (u^*, v)=0, \text{ for fixed } u^* \text{ and all } v \Rightarrow u^*=0 \text{ (almost everywhere)},
\end{aligned}\right\} \quad (4.7)$$

where α and β are scalars and (u,v) is the real number associated with the pair of vectors u and v. Then $\mathcal{V}$ is called the *dual space* of $\mathcal{U}$, and the mapping $\mathcal{S}[u,v] \equiv (u,v)$ is called the *scalar inner product* of u and v.

A linear mapping ℓ of $\mathcal{U}$ into E is called a *linear functional*, and the set $\mathcal{U}^*$ of all linear functional on $\mathcal{U}$ is itself a linear vector space called the *conjugate space* of $\mathcal{U}$. For every continuous linear functional ℓ on a real space $\mathcal{U}$ there exists an element $v \in \mathcal{V}$ such that $\ell(u) = \langle u,v \rangle$ and the vector v is uniquely determined by ℓ. As a consequence, we can generally treat the dual space $\mathcal{V}$ as algebraically the same (isomorphic to) the conjugate space $\mathcal{U}^*$.

Let $\mathcal{U}_1$ denote a subspace of $\mathcal{U}$. The set of elements in $\mathcal{V}$ which have the property that

$$(u,v) = 0, \quad \forall\ u \in \mathcal{U}_1 \quad (4.8)$$

is called the *orthogonal complement* of $\mathcal{U}_1$ and is denoted $\mathcal{U}_1^{\perp}$. In (4.8), v

is said to be orthogonal to u. Generally, if $\mathcal{U}$ is the direct sum of two subspaces $\mathcal{U} = \mathcal{U}_1 \oplus \mathcal{U}_2$, $\mathcal{V}$ is the direct sum $\mathcal{U}_1^{\perp} \oplus \mathcal{U}_2^{\perp}$, where $\mathcal{U}_1^{\perp}$ and $\mathcal{U}_2^{\perp}$ are the orthogonal complements of $\mathcal{U}_1$ and $\mathcal{U}_2$ respectively. Then it can be shown to follow that $\mathcal{U}_1$, $\mathcal{U}_2^{\perp}$ and $\mathcal{U}_2$, $\mathcal{U}_1^{\perp}$ are dual pairs.

We can interpret the idea of weak solutions in the concept of dual spaces, and let $\mathbb{L}$: $\mathcal{H} \longrightarrow \mathcal{U}$ denote a linear mapping of a functional space $\mathcal{H}$ into $\mathcal{U}$. Let $\mathbb{L}(u)$ denote the image of u under the mapping and let s denote a fixed element in $\mathcal{U}$. The function

$$r = \mathbb{L}(u) - s \tag{4.9}$$

is called the *residual of* $\mathbb{L}(u)$ *with respect to* s. Clearly, the residual $r = r(\mathbf{x})$ belongs to $\mathcal{U}$ since $\mathcal{U}$ is, by hypothesis, a linear space.

An element $u \in \mathcal{H}$ *is a weak solution of* (4.1) *if the residual of* $\mathbb{L}(u)$ *with respect to* s *is orthogonal to the entire dual space* $\mathcal{V}$; i.e., u is a weak solution if and only if

$$(r,h) = 0 \tag{4.10}$$

for all $h \in \mathcal{V}$. According to the condition (c) of (4.7), this means $r = 0$ (weakly).

4.3.2 Methods of Weighted Residuals ([8], [10], [108])

In weighted residual methods, we seek approximations to solutions of (4.5) in a finite dimensional subspace $\mathcal{S}_G$ of $\mathcal{U}$ spanned by a linearly independent set of basis functions $T(\mathbf{x})_1$, $T(\mathbf{x})_2$, ..., $T(\mathbf{x})_G$. Every element $g(\mathbf{x})$ in $\mathcal{S}_G$ is, therefore, of form

$$u(\mathbf{x}) = \sum_{1}^{G} a_i T_i(\mathbf{x}) \tag{4.11}$$

where a_1, a_2, ..., a_G are scalars. Setting a dual or conjugate space $\mathcal{S}_G^*$, in which an element $v(\mathbf{x})$ is assumed to be of the form

$$v(\mathbf{x}) = \sum_{1}^{G} b_i W_i(\mathbf{x}) \tag{4.12}$$

where $W(\mathbf{x})_1$, $W(\mathbf{x})_2$, ..., $W(\mathbf{x})_G$ are G linearly independent functions which provide a basis for $\mathcal{S}_G^*$ and which may be unrelated to the functions $T_i(\mathbf{x})$.

Weighted residual methods consist of selecting the coefficients a_i so

that $u(\mathbf{x})$ approximates a weak solution of (4.1); i.e., $\tilde{u}(\mathbf{x})$ of (4.11) is a weighed residual approximation of the solution $u(\mathbf{x})$ of (4.5) if

$$<v,\mathbb{L}(\tilde{u}) - s> = 0 \tag{4.13}$$

for all $v \in \mathcal{S}_G^*$. If $\mathbb{L}$ is linear, (4.13) leads to a system of linear equations for the coefficients a_i, that is

$$<\sum_j^G b_j W_j(\mathbf{x}),\mathbb{L}(\sum_i^G a_i T_i(\mathbf{x})) - s> = \sum_j^G b_j <W_j(\mathbf{x}),\mathbb{L}(\sum_i^G a_i T_i(\mathbf{x})) - s> = 0 \tag{4.14}$$

for each j, so that (4.14) implies

$$<W_j(\mathbf{x}),\mathbb{L}(\sum_i^G a_i T_i(\mathbf{x})) - s> = 0 \tag{4.15}$$

Consequently, the coefficient a_i of weighted residual approximation $\tilde{u}(\mathbf{x})$ of $u(\mathbf{x})$ must satisfy

$$\sum_i^G L_{ij}\, a_i - s_j = 0 \tag{4.16}$$

where

$$L_{ij} = <W_j,\mathbb{L}(T_i)> \quad \text{and} \quad s_j = <W_j,s> \tag{4.17}$$

In weighted residuals methods, the basis functions $T_i(\mathbf{x})$ of $\mathcal{S}_G$ are usually called *trial functions*, and the basis functions of $\mathcal{S}_G^*$ are usually called *weighting functions*. Clearly, the weighting functions can be chosen in many ways and each choice corresponds to a different criterion of method of weighted residuals.

4.3.3 The Galerkin (Bubnov-Galerkin) Method ([108], [116])

Among the most important methods for obtaining approximate solutions to (4.5) is *Galerkin method* which is a special case of methods of weighted residuals. In Galerkin method, the dual space of $\mathcal{S}_G$ is itself, i.e., $\mathcal{S}_G = \mathcal{S}_G^*$, in another words, the weighting functions are chosen to be the trial functions

$$W_j = T_j \tag{4.18}$$

For L_2-approximations, it is easy to show that the Galerkin

approximation $\tilde{u}(\mathbf{x})$ is the best L_2-approximation of the solution $u^*(\mathbf{x})$ of (4.5); i.e.

$$\| u^* - \tilde{u} \| = \inf_{u \in \mathcal{S}_G} \| u^* - u \| \tag{4.19}$$

Moreover, it is clear that Galerkin method chooses the coefficients a_i so that the residual (error), $r(\mathbf{x}) = \mathbb{L}(\tilde{u}) - s$, is orthogonal to (lies in the orthogonal complement of) the linear mainfold $\mathcal{S}_G$.

4.3.4. Subdomain Method ([116])

We could divide the domain Ω into G smaller subdomains, Ω_j, and choose

$$W_j = \begin{cases} 1 & \text{in } \Omega_j \\ 0 & \text{elsewhere.} \end{cases} \tag{4.20}$$

As G increase, the differential equation is satisfied on the average in smaller and smaller subdomains, and the integral of error presumably approaches zero everywhere.

4.3.5 Least Square Method ([8], [116])

In the least square method, the weighting functions are chosen as

$$W_j = \mathbb{L}(T_j) \tag{4.21}$$

The weighting functions so defined are the derivatives of the error $r(\mathbf{x}) = \mathbb{L}(\tilde{u}) - s$ with respect to the parameters a_i, i.e., $\partial r/\partial a_i$, so that the functional (which represents the square of the error norm)

$$I(a_i) = \langle r,r \rangle = \|r\|^2 \tag{4.22}$$

is minimized with respect to the parameters a_i. This method often leads to cumbersome equations, but it has been applied to complicated problems. The mean square residual (4.22) has theoretical significance since error bounds can be derived in terms of it. Thus, minimization of (4.22) gives the best possible bounds for the error.

4.3.6 General Galerkin (Petrov-Galerkin) Method ([116])

If the weighting functions are not chosen as the trial functions, such that

$$W_j \neq T_j \tag{4.24}$$

Then the method is often called *General Galerkin Method* or *Petrov-Galerkin Method.*

4.3.7 Generalized Finite Element Method ([116])

The name of 'weighted residuals' is much older than that of the 'finite element method'. The latter uses mainly locally based (element) functions in expansion (4.11), but the general procedures are identical. As the process leads always to equations which, being of integral form, can be obtained by summation of contributions from various subdomains, all weighted residual approximations are usually called *generalized finite element method.* Frequently, simultaneous use of both local and global trial functions will be useful.

4.4 Variational Principles

4.4.1 Variational Method ([8], [10], [108], [110])

An alternative and rewarding way to view the idea of weak solutions of boundary-value problems is from the variational principles. To review quickly some of the important features of variational methods, let $\mathbb{P}$ denote an operator mapping a Hilbert space $\mathcal{U}$ into $\mathcal{U}^*$, $\mathbb{P}$ being not necessarily linear. If α is a scalar and h is an arbitrary element of $\mathcal{U}$, the *Gateaux differential* of $\mathbb{P}$ at u is the function $D\mathbb{P}(u,h)$ such that

$$\lim_{\alpha \to 0} \left\| \frac{1}{\alpha} [\mathbb{P}(u + \alpha h) - \mathbb{P}(u)] - D\mathbb{P}(u,h) \right\| = 0 \tag{4.24}$$

It is meaningful to refer to $D\mathbb{P}(u,h)$ as the *Gateaux derivative* of $\mathbb{P}$ in the "direction" h, or to the operator $D\mathbb{P}(u)$ on h as the Gateaux derivative of $\mathbb{P}$ at u. If $\mathcal{U} = \mathcal{U}^* = E$, the real numbers, $D\mathbb{P}(u)$ is the ordinary derivative of a real-valued function $\mathbb{P}(u) = p(\mathbf{x})$ (i.e., $D\mathbb{P}(u,h) = (dp/dx)h$). If $\mathbb{P}$: $E^n \longrightarrow E$ and $h = (1, 0, 0, ..., 0)$, then $D\mathbb{P}(u,h) = \partial\mathbb{P}/\partial x_1$, etc. However, in

more general settings the Gateaux differential of a continuous operator need not be continuous and it need not be linear in h. If, on the other hand, $D\mathbb{P}(u,h)$ exists in some neighbourhood $\|u - u_0\| < r$ of u_0, is continuous in u in this neighbourhood, and if it is continuous in h at zero element $h = 0$, then $D\mathbb{P}(u,h)$ is linear in h. We shall henceforth assume that $D\mathbb{P}(u,h)$ exists and is continuous in u and h, and is linear in h.

A functional $\mathbb{K}$: $\mathcal{U} \longrightarrow E$ takes elements of the space $\mathcal{U}$ into real numbers. If a given functional is Gateaux differentiable at u, we can compute

$$D\mathbb{K}(u,h) = \lim_{\alpha \to 0} \frac{1}{\alpha} [\mathbb{K}(u + \alpha h) - \mathbb{K}(u)] \tag{4.25}$$

Now $D\mathbb{K}(u,h)$ is, for each u, a linear functional on $\mathcal{U}$ which can be written in the form $\langle \mathbb{P}(u),h \rangle$, $\mathbb{P}(u)$ being a possibly nonlinear mapping from $\mathcal{U}$ into $\mathcal{U}^*$ which is equivalent to the Gateaux derivative of $\mathbb{K}(u)$. Consequently, the operator $\mathbb{P}(u)$ given by

$$\langle h,\mathbb{P}(u) \rangle = \frac{\partial}{\partial \alpha} \mathbb{K}(u + \alpha h) \Big|_{\alpha \to 0} \tag{4.26}$$

is called the *gradient* of the functional $\mathbb{K}(u)$, and we write $\mathbb{P}(u) = \text{grad } \mathbb{K}(u)$. If at a particular point u^*, $\text{grad}\mathbb{K}(u^*) = 0$, then u^* is called a *stationary point* or *critical point* of $\mathbb{K}(u)$ and it is said that $\mathbb{K}(u)$ assumes a *stationary value* at u^*. If, for a given $\mathbb{P}(u)$, there exists a functional $\mathbb{K}(u)$ such that $\mathbb{P}(u) = \text{grad } \mathbb{K}(u)$, then $\mathbb{P}(u)$ is referred as a *potential operator.*

Let us now appreciate the concept of variational formulations of boundary-value problems. Take, for example, the case of boundary-value problem $\mathbb{P}(u) = \mathbb{L}(u) - s = 0$, where s is fixed and $\mathbb{P}(u)$ is a potential operator. By definition, there exists a functional $\mathbb{K}(u)$ for which $\mathbb{P}(u) = \text{grad } \mathbb{K}(u)$. Indeed, $D\mathbb{K}(u,h) = \langle h,\mathbb{P}(u) \rangle = \langle h,\mathbb{L}(u) - s \rangle$. If u^* is a stationary point of the functional $\mathbb{K}(u)$, then $\langle h,\mathbb{L}(u^*) - s \rangle = \langle h,\mathbb{P}(u^*) \rangle = 0$; that is, *stationary points of the functional* $\mathbb{K}(u)$ *which has the property* $\text{grad } \mathbb{K}(u) = \mathbb{P}(u)$ *are weak solutions of the problem* $\mathbb{P}(u) = 0$. This, in fact, is the essence of the variational method; *to obtain weak solutions of boundary value problems by determining stationary points of an associated functional.*

It is clear that the *inverse problem* of the calculus of variations

(i.e., given a potential operator $\mathbb{P}(u)$, find a functional $\mathbb{K}(u)$ such that grad $\mathbb{K}(u) = \mathbb{P}(u)$) is of crucial importance in applying the variational formulation.

In the next a few subsections, we will briefly refer, in a more engineering language, to the techniques of applying variational method.

4.4.2 Rayleigh-Ritz Procedure ([10], [116])

If a variational formulation can be found, then a standard procedure, which is known as *Rayleigh-Ritz procedure*, can be used immediately for obtaining approximate solutions in weak form suitable for finite element analysis. Assume a variational formulation is specified as

$$F = \langle u,\ \mathbb{L}(u) - s \rangle = \langle u,\ \mathbb{P}(u) \rangle \tag{4.27}$$

The solution to the problem is a function u which makes F stationary with respect to small changes δu. Thus, for a solution to the problem, the variation is

$$\delta F = 0 \tag{4.28}$$

Assuming a trial function expansion in the usual form

$$u(\mathbf{x}) = \sum_{i=1}^{G} a_i T_i(\mathbf{x}) \in \mathcal{S}_G$$

we can insert this into (4.27) and write

$$\delta F[u(\mathbf{x})] = \sum_{i=1}^{G} \frac{\partial F[u(\mathbf{x})]}{\partial a_i} \delta a_i = 0 \tag{4.29}$$

This being true for any variation $\delta \mathbf{a}$ yields a set of equations

$$\frac{\partial F}{\partial a_i} = 0 \tag{4.30}$$

from which parameters a_i are found. The equations are of a form involving integration necessary for the finite element approximation as the original specification of F was given in terms of domain and boundary integrals.

The process of finding stationarity with respect to trial function parameters a_i is an old one and is associated with the names of Rayleigh and Ritz [8], [10]. It has become extremely important in finite element

analysis which is typified as a 'variational process'.

4.4.3 Euler Equations ([116])

If we consider the definitions of (4.27) and (4.28) we observe that for stationarity we can write, after performing some differentiations,

$$\delta F = \langle \delta u, \mathbb{P}(u) \rangle = 0 \tag{4.31}$$

As the above has to be true for any variations δu, we must have

$$\mathbb{P}(u) = 0 \tag{4.32}$$

If $\mathbb{P}$ corresponds precisely to the differential equations governing the problem, then the variational principle is a *natural variational principle.* Equations (4.31) and (4.32) are known as the *Euler Equations* corresponding to the variational principle requiring the stationarity of F. For any variational principle a corresponding set of Euler equations can be established. The reverse is unfortunately not true, i.e., only certain forms of differential equations are Euler equations of a variational functional.

4.4.4 Relation of the Galerkin Method to Variational Principles ([108])

We can observe that the approximation obtained by the use of a natural variational principle and by the use of the Galerkin weighting process is identical. That this is the case follows directly from equation (4.27), in which the variation was derived in terms of the original differential equations and the associated boundary conditions.

If we consider the usual trial function expansion (4.11)

$$u(\mathbf{x}) \approx \tilde{u}(\mathbf{x}) = \sum_{1}^{G} a_i T_i(\mathbf{x})$$

we can write the variation of this approximation as

$$\delta\tilde{u} = \sum_{1}^{G} \delta a_i T_i(\mathbf{x}) \tag{4.33}$$

and inserting the above into (4.31) yields

$$\delta F = \sum_{j}^{G} \delta a_j \langle T_j, \mathbb{P}(\sum_{i}^{G} a_i T_i) \rangle = 0 \qquad (4.34)$$

The above form, being true of all δa_j, requires that the expression under the integrals should be zero. We will immediately recognize this as simply the Galerkin form of the weighted residual statement discussed in section 4.3.2 and 4.3.3, and the identity is hereby proved.

We need to underline, however, that this is only true if the Euler equations of the variational principle coincide with the governing equations of the original problems. The Galerkin process thus retains its greater range of applicability.

4.4.5 Variational Formulation for Self-Adjoint Problems ([109])

General rules for deriving natural variational principles from non-linear differential equations are complicated. For linear differential equations the situation is much simpler and a thorough study is available in [109], and this section we summarize such rules.

We shall consider here only the establishment of variational principles for a linear equation with *forced* boundary conditions, implying only variation of functions which yield $\delta u = 0$ on the boundaries The extension to include the natural boundary conditions is simple and will be omitted.

Consider a boundary-value problem of form (4.1)

$$\mathbb{L}(u) = s$$

in which $\mathbb{L}$ is a linear differential operator; it can be shown that natural variational principles require that the operator $\mathbb{L}$ be such that

$$\langle v, \mathbb{L}(u) \rangle = \langle \mathbb{L}(v), u \rangle + \text{b.t.} \qquad (4.35)$$

for any two function sets u and v. The property required in the above operator is called one of *self-adjointness* or *symmetry*.

If the operator $\mathbb{L}$ is self-adjoint, the variational principle can be expressed immediately as

$$F = \langle u, \mathbb{L}(u) \rangle - 2 \langle u, s \rangle + \text{b.t.} \qquad (4.36)$$

In fact, a variation of (4.36) can be written as

$$\delta F = \langle \delta u, \mathbb{L}(u) \rangle + \langle u, \delta \mathbb{L}(u) \rangle - 2 \langle \delta u, s \rangle + \text{b.t.} \tag{4.37}$$

Because of the linearity of the operator $\mathbb{L}$, there always is

$$\delta \mathbb{L}(u) \equiv \mathbb{L}(\delta u) \tag{4.38}$$

and because u and δu can be treated as any two independent functions, by identity (4.35) we can write (4.37) as

$$\delta F = 2 \langle \delta u , [\mathbb{L}(u) - s] \rangle + \text{b.t.} \tag{4.39}$$

We observe immediately that the term in the square brackets, i.e., the Euler equation of the functional, is identical with the original equation postulated, and therefore the variational principle is verified.

4.4.6 Variational Formulation for Non-Self-Adjoint Problems ([38])

In the previous section, we discussed how to establish a variational formulation of a self-adjoint problem. In this section we shall discuss how to obtain the variational formulation of a non-self-adjoint problem

$$\mathbb{L}(u) = s \tag{4.40}$$

where $\mathbb{L}$ is a non-self-adjoint linear operator, u is the unknown vector field function to be determined, and s is a known vector source function.

A method of solving the original non-self-adjoint problem (4.40) is to introduce an auxiliary problem, the adjoint problem as follows:

$$\mathbb{L}^a(u^a) = s^a \tag{4.41}$$

where $\mathbb{L}^a$ is the adjoint operator of $\mathbb{L}$, u^a is another unknown vector function to be determined, and s^a is another known vector function.

It can be proved that the problem of solving u and u^a simultaneously from (4.40) and (4.41) is completely equivalent to that of determining the stationary functions (both u and u^a) from the following variational equation:

$$\delta F(u, u^a) = 0$$

$$F(u, u^a) = \langle u^a, \mathbb{L}(u)\rangle - \langle s^a, u\rangle - \langle u^a, s\rangle \tag{4.42}$$

In fact, taking a variation of (4.42) it arrives at

$$\begin{aligned}\delta F(u, u^a) &= \langle \delta u^a, \mathbb{L}(u)\rangle + \langle u^a, \delta\mathbb{L}(u)\rangle - \langle s^a, \delta u\rangle - \langle \delta u^a, s\rangle \\ &= \langle \delta u^a, \mathbb{L}(u)\rangle + \langle \mathbb{L}^a(u^a), \delta u\rangle - \langle s^a, \delta u\rangle - \langle \delta u^a, s\rangle \\ &= \langle \delta u^a, [\mathbb{L}(u) - s]\rangle + \langle [\mathbb{L}^a(u^a) - s^a], \delta u\rangle \\ &= 0 \end{aligned} \tag{4.43}$$

As the above has to be true for any variations δu and δu^a, we must have the two terms in the two pairs of brackets equal to zero, they are the Euler equations

$$\mathbb{L}(u) - s = 0$$

$$\mathbb{L}^a(u^a) - s^a = 0$$

which are identical with the original problem and its adjoint problem shown in (4.40) and (4.41) respectively, and therefore the variational principle (4.42) has been proved.

Note that for the problems defined by differential operators $\mathbb{L}$ and $\mathbb{L}^a$ with their boundary conditions $\mathbb{B}(u) = 0$ and $\mathbb{B}^a(u^a) = 0$ regarded as essential ones, the stationary functions u and u^a of (4.42) should also be subject to the constraints $\mathbb{B}(u) = 0$ and $\mathbb{B}^a(u^a) = 0$, respectively. However, if these boundary conditions should be made on (4.42) with the stationary functions u and u^a subject to no constraints on the boundary. The expression (4.42), of course, includes that discussed in (4.36) of the self-adjoint problem.

It looks as though both u (the desired function) and u^a (the adjoint function or auxiliary function) have to be solved simultaneously in the variational problem (4.42). However, the process of determining both u and u^a can be decoupled when Rayleigh-Ritz procedure is employed in the solution.

Let us express the solution in the form

$$u = \sum_{i=1}^{N} c_i \phi_i \tag{4.44a}$$

$$u^a = \sum_{i=1}^{N} c_i^a \phi_i^a \tag{4.44b}$$

where ϕ_i and ϕ_i^a are known functions, and c_i and c_i^a are parameters to be determined. Both ϕ_i and ϕ_i^a may or may not form the bases of the domains of $\mathbb{L}$ and $\mathbb{L}^a$, respectively, however they should be linearly independent and should form the complete sets as N approaches infinity. By inserting (4.44) into (4.42) and applying Rayleigh-Ritz procedure, one obtains two decoupled systems as follows

$$\sum_{i=1}^{N} <\phi_j^a, \mathbb{L}(\phi_i)> c_i = <\phi_j^a, s> \qquad (j = 1 \text{ to } N) \tag{4.45a}$$

$$\sum_{i=1}^{N} <\phi_j, \mathbb{L}^a(\phi_i^a)> c_i^a = <\phi_j, s^a> \qquad (j = 1 \text{ to } N) \tag{4.45b}$$

The positive integer N in (4.44) and (4.45) may be finite or infinite if an approximate or exact solution is to be determined. The fact that c_i and c_i^a are decoupled in (4.45) can greatly simplify the process of determining the stationary functions u and u^a from (4.42).

The discrete systems (4.45a) and (4.45b) from the Rayleigh-Ritz procedure are identical in form to those from of weighted residuals (or Petrov-Galerkin) method of solving simultaneously the original and adjoint problems (4.40) and (4.41). If u^a in (4.44b) is expanded into a series of ϕ_i instead of ϕ_i^a, then the resultant systems obtained will be identical to those from the Galerkin (or Bubnov-Galerkin) method.

The introduction of the auxiliary problem (4.41) for supplementing the original problem (4.40) has an interesting physical interpretation as follows

$$<u^a, s> = <u^a, \mathbb{L}u> = <\mathbb{L}^a u^a, u> = <s^a, u> \tag{4.46}$$

This is the generalized reciprocity theorem which states that the generalized reaction of the adjoint fields u^a on the source s of the original problem is identical to that of the original field u on the source s^a of the adjoint problem. The term $<u^a, s>$, for example, may be

interpreted as a generalized reaction if the real-type inner product is employed.

The general reciprocity theorem (4.46), in the case of a self-adjoint problem using the real-type inner product, has an important result as follows. By setting $u = u_1$, $s = s_1$, $u^2 = u_2$, $s^a = s_2$, one then has the conventional reciprocity theorem of relating the reactions between two different problem:

$$<u_2, s_1> = <u_1, s_2> \tag{4.47}$$

One has to note that for many non-self-adjoint practical boundary-value problems, their adjoint problems may not correspond to physical problems, it may also be extremely difficult to decide the corresponding adjoint boundary conditions. Therefore there may be two problems to apply (4.42) in practice. One is that it may be impossible to decouple the original and the adjoint equation, and one has to solve them simultaneously; the more serious problem is that it may simply make the formulation (4.42) impractical.

The *local potential method* to be discussed below may be used to extend the variational principles to some non-self-adjoint problems which are impossible to accomplish by (4.42)

4.4.7 Local Potential Method ([113])

Let us consider a non-self-adjoint problem

$$\mathbb{L}(u) = s$$

$$\mathbb{L}(u) = \mathbb{L}_1(u) + \mathbb{L}_2(u) = s \tag{4.48}$$

where the non-self-adjoint operator $\mathbb{L}$ is the sum of two parts of operators, the first part, $\mathbb{L}_1$, corresponds to a sum of self-adjoint operators, the second parts, $\mathbb{L}_2$, corresponds to a sum of non-self-adjoint ones, we may still be able to get a variational expression by using the idea of *the local potential method* [112], [113].

For the problem (4.48), the utilization of the local potential method can be interpreted below.

Consider the non-self-adjoint part of the problem, let us assume that

$\mathbb{L}_2(u)$ is displaced but infinitesimally from the stationary state and define u_0 as the function at the stationary state. We now suppose that for such small displacement from the stationary state

$$\mathbb{L}_2(u) \approx \mathbb{L}_2(u_0) \tag{4.49}$$

so that

$$\mathbb{L}_1(u) = s - \mathbb{L}_2(u_0) \tag{4.50}$$

The (4.50) is disguised as a self-adjoint problem if we take $s - \mathbb{L}_2(u_0)$ as the known function, then (4.36) can be applied to (4.50) obtaining the variational expression

$$F(u) = \langle u, \mathbb{L}_1(u) \rangle - 2 \langle u, s - \mathbb{L}_2(u_0) \rangle + \text{b.t.} \tag{4.51}$$

The quantity of $F(u)$ at the displacement infinitesimally off the stationary state is called *local potential.* During the next process, one must remember that we have two classes of unknown functions in the variational formulation. One of these is u, and we are at liberty to manipulate as in our previous discussion. The second class of unknown function is disguised, in the sense that this particular quantity plays the same role as a stationary solution. In other words, we must assume that u_0 is a known function of position; this dual personality must be maintained until the function is identified as that occurring at the stationary state. Thus, the necessary conditions which must be satisfied if F is to be a minimum or maximum are found by determining

$$\left.\frac{\partial F(u)}{\partial u}\right|_{u_0} = 0 \tag{4.52}$$

with the subsidiary condition that

$$u = u_0 \tag{4.53}$$

The constraints (4.53) is to be released after minimization, making

$$u_0 = u \tag{4.54}$$

In fact, the variation of (4.51) can be written as

$$\delta F(u)\Big|_{u_0} = \langle \delta u,\ \mathbb{L}_1(u)\rangle + \langle u,\ \delta\mathbb{L}_1(u)\rangle - 2\ \langle \delta u,\ s - \mathbb{L}_2(u_0)\rangle$$

$$= \langle \delta u,\ \mathbb{L}_1(u)\rangle + \langle u,\ \mathbb{L}_1(\delta u)\rangle - 2\ \langle \delta u,\ s - \mathbb{L}_2(u_0)\rangle$$

$$= \langle \delta u,\ \mathbb{L}_1(u)\rangle + \langle \mathbb{L}_1(u),\ \delta u\rangle - 2\ \langle \delta u,\ s - \mathbb{L}_2(u_0)\rangle$$

$$= 2\ \langle \delta u,\ [\mathbb{L}_1(u) + \mathbb{L}_2(u_0) - s]\rangle\Big|_{u_0 = u} = 0$$

$$\longrightarrow \quad 2\ \langle \delta u,\ [\mathbb{L}_1(u) + \mathbb{L}_2(u) - s]\rangle = 0 \qquad (4.55)$$

The terms in the pair of bracket in (4.55) is the Euler equation of original problem (4.48).

It is essential to distinguish between the stationary function u_0 and the local function u until the process of variation is complete. Otherwise incorrect results will arise.

4.5 *The Finite Element Method* ([9], [10], [108], [115], [116])

4.5.1 Introductory Remarks

Before the mid-1950's approximate methods of Rayleigh-Ritz and Galerkin types found limited applications in more difficult problem areas of mathematical physics because of the difficulty in generating appropriate trial functions $T_i(\mathbf{x})$ in (4.11). This was particularly true in problems involving irregular domains and mixed boundary conditions. Moreover, the conditioning of Rayleigh-Ritz and Galerkin equations

$$\langle T_j(\mathbf{x}),\ \mathbb{L}(\sum_i^G a_i T_i(\mathbf{x})) - s\rangle = 0$$

is highly sensitive to the choice of the functions $T_i(\mathbf{x})$ and the considerable effort required to generate such equations for significant problems was, in the past, a serious disadvantage.

The finite element method is a systematic technique for constructing the basis functions $T_i(\mathbf{x})$ for Rayleigh-Ritz and Galerkin methods for irregular domains. In addition to a number of other advantages, the finite element method overcomes all of the traditional disadvantages of Rayleigh-Ritz and Galerkin methods mentioned above. The basis functions $T_i(\mathbf{x})$ are generated in a straightforward and systematic manner, irregular

domains and mixed boundary conditions are easily accommodated, the resulting equations describing the discrete model are generally well-conditioned, and the finite element method is exceptionally well suited for implementation via electronic computers.

In order to represent finite element method concisely and compactly, we first introduce the standard multi-integer notation [108]: Let Z_+^n denote the set of all n-tuples of nonnegative integers (i.e., if $a \in Z_+^n$, then $a = (a_1, a_2, ..., a_n)$, a_i being integers ≥ 0); then the multi-integer conventions are defined as follows:

$$|a| = a_1 + a_2 + \ldots + a_n ; \tag{4.56a}$$

$$a! = \prod_{i=1}^{n} a_i! ; \tag{4.56b}$$

$$\mathbf{x}^a = x^{a_1} x^{a_2} x^{a_n} x^{a_i} = \prod_{i=1}^{n} x_i^{a_i} ; \tag{4.56c}$$

$$D_a = \partial^{|a|} / \partial x^{a_1} \partial x^{a_2} \ldots \partial x^{a_n} . \tag{4.56d}$$

Recall the problem defined in section 4.1., the Taylor-type expansions of $u(\mathbf{x}) \in C^{p+1}(\Omega)$ about $\mathbf{x} \in \Omega$ can be written concisely as

$$u(\mathbf{x} + \mathbf{y}) = \sum_{|a| \leq p} \frac{\mathbf{y}}{a!} D_a u(\mathbf{x}) + R_{p+1}(u) \tag{4.57a}$$

where $R_{p+1}(u)$ is the remainder

$$\frac{1}{(p+1)!} \sum_{i_1}^{n} \sum_{i_2}^{n} \cdots \sum_{i_{p+1}}^{n} \frac{\partial^{p+1} u(\mathbf{x} + \theta \mathbf{y})}{\partial x_{i_1} \partial x_{i_2} \ldots \partial x_{i_{p+1}}} (y_{i_1} - x_{i_1}) \ldots (y_{i_{p+1}} - x_{i_{p+1}})$$

$$0 \leq \theta \leq 1 \tag{4.57b}$$

4.5.2 Finite Element Method

We consider a finite-element model $\tilde{\Omega}$ of region Ω ($\Omega = \Omega + \Gamma$) which is the union of E closed and bounded subregions Ω_e of E^n. The subregions Ω_e, where Ω_e is the closure of an open region Ω_e ($\Omega_e = \Omega_e + \Gamma_e$), are called finite elements, and the region Ω_e are disjoint:

$$\tilde{\Omega} = \bigcup_{e=1}^{E} \Omega_e \tag{4.58a}$$

$$\Omega_e \cap \Omega_f = \varnothing, \quad e \neq f \tag{4.58b}$$

Conceptually, the finite elements are considered to be connected together at a number G of nodal points labeled $\mathbf{x}^g$, $g = 1, 2, ..., G$. Locally, it is meaningful to label the nodal points belonging to element Ω_e by $\mathbf{x}_e^N$, $N = 1, 2, ..., N_e$, N_e being the number of nodal points belonging to element Ω_e. For simplicity, we shall henceforth assume that the global and local coordinate systems coincide, thereby avoiding the necessity of introducing a coordinate transformation for each element. Then, assuming the nodal compatibility conditions are satisfied (i.e., there exists a one-to-one correspondence between all nodal points $\mathbf{x}_e^N$ in Ω_e and all points $\mathbf{x}^g$ in the *connected* model $\tilde{\Omega}$), the connectivity and decomposition of the model are established by the respective incidence mappings:

$$\mathbf{x}^g = \sum_{N=1}^{N_e} {}^{(e)}\Lambda_N^g \, \mathbf{x}_e^N \quad (e \text{ fixed}), \qquad \mathbf{x}_e^N = \sum_{g=1}^{G} {}^{(e)}\Xi_g^N \, \mathbf{x}^g \tag{4.59}$$

where

$${}^{(e)}\Lambda_N^g = \begin{cases} = 1 & \text{if node } g \text{ of the connected model } \tilde{\Omega} \text{ is} \\ & \text{coincident with node } N \text{ of element } \Omega_e \\ = 0 & \text{otherwise} \end{cases} \tag{4.60a}$$

and ${}^{(e)}\Xi_g^N$ simply the transpose of ${}^{(e)}\Lambda_g^N$

$${}^{(e)}\Xi_g^N = \begin{cases} = 1 & \text{if node } N \text{ of the element } \Omega_e \text{ is coincident} \\ & \text{with node } g \text{ of the connected model } \tilde{\Omega} \\ = 0 & \text{otherwise} \end{cases} \tag{4.60b}$$

We can use the mappings (4.60) to form identity mappings through the compositions:

$$\sum_{g=1}^{G} {}^{(e)}\Xi_g^N \, {}^{(e)}\Lambda_M^g = \delta_M^N \quad (e \text{ fixed}) \tag{4.61a}$$

$$\sum_{N=1}^{N_e} {}^{(e)}\Lambda_N^g \, {}^{(e)}\Xi_h^N = \begin{cases} \delta_h^g & \text{if } \mathbf{x}^g, \mathbf{x}^h \in \Omega_e \\ 0 & \text{if } \mathbf{x}^g, \mathbf{x}^h \notin \Omega_e \end{cases} \tag{4.61b}$$

where δ_M^N, δ_h^g are Kronecker deltas. The incidence mapping Λ of (4.59) is said to establish the *connectivity* of the discrete model $\tilde{\Omega}$, while Ξ of

(4.59) establishes a *decomposition* of $\tilde{\Omega}$ into finite elements, (see Fig.4.1).

A function $U(\mathbf{x})$ with domain $\tilde{\Omega}$ is called *a finite element representation of order q if and only if*

$$U(\mathbf{x}) = \bigcup_{e=1}^{E} U_e(\mathbf{x}), \quad U_e(\mathbf{x}) = \begin{cases} 0 & \text{if } \mathbf{x} \notin \Omega_e \\ \sum_{N=1}^{Ne} \sum_{|a| \leq q-1} c_a^{N(e)} \psi_N^{a(e)}(\mathbf{x}) \end{cases} \tag{4.62}$$

where $\psi_N^{a(e)}(\mathbf{x})$ are local interpolation functions corresponding to element Ω_e which are defined so as to have the properties

$$D_b \psi_N^{a(e)}(\mathbf{x}) \equiv 0, \quad \mathbf{x} \notin \Omega_e, \quad b \in Z_+^n \tag{4.63a}$$

$$D_b \psi_N^{a(e)}(\mathbf{x}^M) \equiv \delta_N^M \delta_{b_1}^{a_1} \delta_{b_2}^{a_2} \ldots \delta_{b_n}^{a_n} \tag{4.63b}$$

where δ_N^M, ..., $\delta_{b_n}^{a_n}$ are Kronecker deltas, $\mathbf{x}^M \in \Omega_e$, and a, $b \in Z_+^n$.

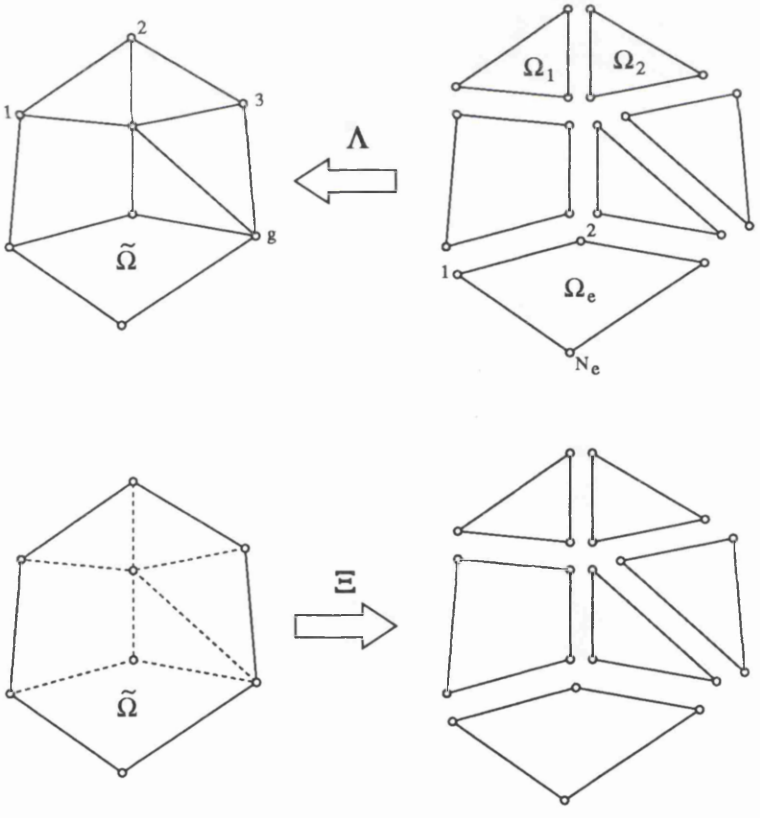


Fig. 4.1 Connection and decomposition of a finite element model

Globally, we write

$$U(\mathbf{x}) = \sum_{|a| \leq q-1} \sum_{g=1}^{G} A_a^g \, \phi_g^a(\mathbf{x}) \tag{4.64}$$

where

$$c_a^{N(e)} = \sum_{g=1}^{G} \Xi_g^{(e)N} A_a^g \tag{4.65a}$$

$$\phi_g^a(\mathbf{x}) = \bigcup_{e=1}^{E} \sum_{N=1}^{Ne} \Xi_g^{(e)N} \psi_N^{a(e)}(\mathbf{x}) \tag{4.65b}$$

If $U(\mathbf{x})$ is a *first-order representation*, we need only

$$U(\mathbf{x}) = \sum_{g=1}^{G} A^g \, \phi_g(\mathbf{x}) = \bigcup_{e=1}^{E} \sum_{N=1}^{Ne} c_e^N \, \psi_N^{(e)}(\mathbf{x}) \tag{4.66}$$

where

$$\phi_g(\mathbf{x}) = \bigcup_{e=1}^{E} \sum_{N=1}^{Ne} \Xi_g^{(e)N} \psi_N^{(e)}(\mathbf{x}) \tag{4.67a}$$

$$c_e^N = \sum_{g=1}^{G} \Xi_g^{(e)N} A^g \tag{4.67b}$$

$$U_e(\mathbf{x}) = \sum_{N=1}^{Ne} c_e^N \, \psi_N^{(e)}(\mathbf{x}) \tag{4.67c}$$

Remark 4.1. Finite element representations (4.64) and (4.66) are linear combinations of functions which have *compact* and *almost disjoint support.* Recall that the *support* of a function $f(\mathbf{x})$ is the closure of the set of points $\mathbf{x}$ in the domain of f such that $f(\mathbf{x}) \neq 0$. If $f(\mathbf{x}) = 0$ for $|\mathbf{x}| > \alpha$, the support of $f(\mathbf{x})$ is *compact* and if the support (written supp.) of function $\phi_1(\mathbf{x})$ and $\phi_2(\mathbf{x})$ is such that supp $\phi_1 \cap$ supp $\phi_2 = \varnothing$, except possibly at a finite number of points, lines, or surfaces, then ϕ_1 and ϕ_2 have "almost" disjoint support.

Remark 4.2. The properties (4.63) of the local interpolation functions $\psi_N^{a(e)}(\mathbf{x})$ are preserved under the incidence mappings Ξ; i.e., the global functions $\phi_g^a(\mathbf{x})$ have the properties

$$D_b\,\phi_g^a(\mathbf{x}^h) = \delta_g^h\,\delta_{b_1}^{a_1}\,\delta_{b_2}^{a_2}\ldots\delta_{b_n}^{a_n} \tag{4.68}$$

(or, for first order representations, $D\phi_g(\mathbf{x}^h) = \delta_g^h$) for all g, h = *1, 2, ..., G.*

Remark 4.3. As a result of (4.63), (4.65), and (4.68), the coefficients A_a^g and $c_a^{N(e)}$ have a special interpretation:

$$D_v U(\mathbf{x}^h) = \sum_{|a| \le q-1} \sum_{g=1}^{G} A_a^g\, D_v \phi_g^a\,(\mathbf{x}^h) = A_v^h \tag{4.69}$$

$$D_v U_e(\mathbf{x}^M) = \sum_{|a| \le q-1} \sum_{N=1}^{Ne} c_a^{N(e)}\, D_v \psi_N^{a(e)}(\mathbf{x}^M) = c_v^{M(e)} \tag{4.70}$$

That is, if (4.63) and (4.68) hold, the coefficients A_v^h are the values of the derivatives D_v of $U(\mathbf{x})$ at node $\mathbf{x}^h$ and $c_v^{M(e)}$ are the values of the derivatives D_v of $U_e(\mathbf{x})$ at node $\mathbf{x}^M$ of element *e.* For first-order representations, $U(\mathbf{x}^g) = A^g$ and $U(\mathbf{x}_e^N) = c_e^N$.

4.5.3 Finite Element Approximation

Let us now combine the concepts of Galerkin and Rayleigh-Ritz methods discussed in sections 4.3 and 4.4 and the notion of finite element representations of functions to obtain approximations of weak solutions of a general boundary-value problem. Considering (4.64) as a Galerkin approximation of the solution of the boundary-value problem (4.5), (i.e., associating (4.64) with (4.11)), we seek coefficients A_a^g which satisfy (4.15), i.e.

$$< \mathbb{L}\Big(\sum_{|a|} \sum_{g} A_a^g\, \phi_g^a \Big) - s\,,\ \phi_h^b > = 0 \tag{4.71}$$

where $a, b \in Z_+^n$; $g, h = 1, 2, \ldots, G$. Because $\mathbb{L}$ is assumed linear, we have

$$\sum_{|a|} \sum_{g} L_{gh}^{ab}\, A_a^g \; - S_h^b = 0 \tag{4.72}$$

where

$$L^{ab}_{gh} = < \mathbb{L}(\phi^a_g) , \phi^b_h > , \quad S^b_h = < s , \phi^b_h > \tag{4.73}$$

Remark 4.4. The Fundamental Property of Finite Element Approximations. The success and utility of finite element concept as a method of approximations primarily is due to the following *fundamental property: the finite element approximation can be completely formulated locally, one element at a time and each element independent of the others, and global approximation can then be obtained by simple transformations of local equations.*

Conditions (4.63) and (4.65b) are responsible for this property of finite element approximations. As a result of this local character of approximation, it is possible to design large-scale computer programs in which local approximations of a given class of boundary-value problems are automatically generated for a typical element of a certain type; then, by appropriately connecting elements together, global models are easily generated for whatever domain and boundary conditions required.

In fact, by introducing (4.65b) into (4.71) we have

$$<\mathbb{L}\Big(\sum_{|a|} \sum_{g} A^g_a \bigcup_{e=1}^{E} [\sum_{N=1}^{Ne} \Xi^{(e)N}_g \psi^{a(e)}_N(\mathbf{x})] \Big) - s, \bigcup_{e=1}^{E} [\sum_{M=1}^{Nf} \Xi^{(e)M}_h \psi^{b(f)}_M(\mathbf{x})] > = 0 \tag{4.74}$$

Introducing (4.63), noting that $<\psi^{a(e)}_N(\mathbf{x}), \psi^{b(f)}_M(\mathbf{x})> = 0$ if $e \neq f$, as $\phi^a_g(\mathbf{x})$ have almost disjoint support, we have

$$\sum_{e=1}^{E} \sum_{M}^{Ne} \Big(\sum_{N=1}^{Ne} \sum_{g}^{G} \sum_{|a|} \Xi^{(e)N}_g \Xi^{(e)M}_h <\mathbb{L}(\psi^{a(e)}_N), \psi^{b(e)}_M> A^g_a - \Xi^{(e)M}_h <s, \psi^{b(e)}_M> \Big) = 0 \tag{4.75}$$

In this way, the local finite element model of weak form of the problem (4.5) corresponding to element Ω_e can written in the form

$$\sum_{N=1}^{Ne} \sum_{|a|} Q^{ab(e)}_{NM} c^{N(e)}_a - R^{b(e)}_M = 0 \tag{4.76}$$

where

$$Q^{ab(e)}_{NM} = <\mathbb{L}(\psi^{a(e)}_N), \psi^{b(e)}_M>, \quad R^{b(e)}_M = <s, \psi^{b(e)}_M> \tag{4.77}$$

The global equation (4.72) can be obtained by computing

$$L_{gh}^{ab} = \sum_{e=1}^{E} \sum_{N=1}^{Ne} \sum_{M=1}^{Ne} \Xi_g^{(e)N} \Xi_h^{(e)M} Q_{NM}^{ab(e)} \tag{4.78a}$$

$$S_g^b = \sum_{e=1}^{E} \sum_{M=1}^{Ne} \Xi_g^{(e)M} R_M^{b(e)} \tag{4.78b}$$

4.6 Concluding Remarks

The main constituents of the finite element method for the solution of a boundary-value problem are

(i) The finite element formulation of the problem; and

(ii) The approximate solution of the formulation through the use of "finite element functions."

The finite element formulations may be achieved by variational methods or weighted residual methods.

Many problems in engineering can be characterized by variational principles. The variational principles may succinctly summarize the equations, and allow insight into the effect of parameters. A variational integral is made stationary, and possibly minimized or maximized with respect to the undetermined constants. The results are identical to those obtained by the Galerkin method.

In addition to the variational approaches, finite element equations can be formulated by employing the weighted residual methods. The weighted residual methods are particularly useful for problems in which a variational formulation may not be available, although they may be applied to any boundary-value problem with established differential equations.

CHAPTER 5

A NEW VARIATIONAL FORMULATION

5.1 Introduction

Based on the preliminary theoretical discussion in the previous chapter 3, this chapter details the procedure of deriving a new variational finite element formulation, which complies with the criteria given in chapter 2, for inhomogeneous, anisotropic and lossy dielectric waveguides.

5.2 Derivation of the Variational Formulation

5.2.1 Prototype of the Variational Formulation

Starting with the differential equation (3.20), we may write it in the following three equivalent operator forms:

$$\mathbb{L}\ \mathbf{H}_t = \mathbf{0} \tag{5.1a}$$

$$\mathbb{A}\ \mathbf{H}_t + \gamma^2\ \mathbb{B}\ \mathbf{H}_t = \mathbf{0} \tag{5.1b}$$

$$\mathbb{A}_1\ \mathbf{H}_t + \mathbb{A}_2\ \mathbf{H}_t + \mathbb{A}_3\ \mathbf{H}_t + \gamma^2\ \mathbb{B}\ \mathbf{H}_t = \mathbf{0} \tag{5.1c}$$

with the operator relations:

$$\mathbb{L}\ \_ = \mathbb{A}\ \_ + \gamma^2\ \mathbb{B}\ \_ \tag{5.2a}$$

$$\mathbb{A}\ \_ = \mathbb{A}_1\ \_ + \mathbb{A}_2\ \_ + \mathbb{A}_3\ \_ \tag{5.2b}$$

The individual operators are expressed as:

$$\mathbb{A}_1\ \_ = \nabla_t \times (\kappa_{zz} \nabla_t \times \_) \tag{5.3a}$$

$$\mathbb{A}_2\ \_ = - \mathbf{z} \times [\bar{\bar{\kappa}}_{tt} \cdot \nabla_t \times (\mathbf{z}\ \nabla \cdot \_)] \tag{5.3b}$$

$$\mathbb{A}_3\ \_ = - \omega^2 \mu_0 \varepsilon_0\ \_ \tag{5.3c}$$

$$\mathbb{B}\,\_ = \mathbf{z} \times [\, \overline{\overline{\kappa}}_{tt} \cdot (\mathbf{z} \times \_) \,] \qquad (5.3d)$$

It can be proved, according to the definition presented in the previous chapter, that the operator $\mathbb{L}$ is not self-adjoint (see Appendix A).

A variational expression may be derived for this problem using a general method discussed in section 4.4.5, chapter 4, but it requires consideration of the adjoint field $\mathbf{H}_t^a$ which does not correspond to a physical field here. Therefore, the method discussed in section 4.4.5 is not applicable for eq. (5.1)

However, we can observe that in expression (5.1c), $\mathbb{A}_1$, $\mathbb{A}_3$, and $\mathbb{B}$ correspond to individual self-adjoint operators, only $\mathbb{A}_2$ is not self-adjoint (see Appendix A). Based on this fact, we can apply the *local potential method*, discussed in section 4.4.6 in chapter 4 to eq. (5.1c) to get a variational formulation involving only $\mathbf{H}_t$, the transverse components of magnetic field.

Using the local potential method, we now consider the magnetic field $\mathbf{H}_t$ in this non-self-adjoint term as a given *suffix-zero* function

$$\mathbf{H}_t = \mathbf{H}_t^0 \qquad (5.4)$$

assuming the value corresponding to the stationary state, i.e., to the solution of eq. (5.1c). With this assumption, eq. (5.1c) becomes

$$\mathbb{L}_{self}\mathbf{H}_t = - \mathbb{A}_2\mathbf{H}_t^0 \qquad (5.5)$$

where the right hand side is an assumed known function and the left hand side corresponds to a self-adjoint operator

$$\mathbb{L}_{self}\,\_ = \mathbb{A}_1\,\_ + \mathbb{A}_3\,\_ + \gamma^2\mathbb{B}\,\_ \qquad (5.6)$$

Formulating the problem in this way, we can now apply a standard method, which is discussed in section 4.4.4, chapter 4., to eq. (5.5) in such a way as to obtain a variational expression from (5.1).

In brief, for a self-adjoint problem

$$\mathbb{L}\,\mathbf{f} = \mathbf{s} \qquad (5.7)$$

where $\mathbf{f}$ is the field (unknown) vector function, and $\mathbf{s}$ is the given (known)

source vector function, the variational formulation is

$$\Pi = < \mathbf{f} , \mathbb{L}\, \mathbf{f} > - 2 < \mathbf{f} , \mathbf{s} > \tag{5.8}$$

where $< \cdot , \cdot >$ is the real-type inner product defined by

$$< \mathbf{u} , \mathbf{v} > = \int_S \mathbf{u} \cdot \mathbf{v}\, ds \tag{5.9}$$

where S is the cross-section of waveguide Ω defined in chapter 3.

Applying (5.8) to (5.5) it leads to the following functional:

$$\Pi = <\mathbf{H}_t, \mathbb{L}_{self}\mathbf{H}_t> + 2<\mathbf{H}_t, \{\mathbb{A}_2\mathbf{H}_t^0\}> = 0 \tag{5.10a}$$

introducing (5.6) into (5.10a) we have a more explicit expression of Π

$$\Pi = <\mathbf{H}_t,\mathbb{A}_1\mathbf{H}_t> + 2<\mathbf{H}_t,\{\mathbb{A}_2\mathbf{H}_t^0\}> + <\mathbf{H}_t,\mathbb{A}_3\mathbf{H}_t> + \gamma^2<\mathbf{H}_t,\mathbb{B}\,\mathbf{H}_t> = 0 \tag{5.10b}$$

The term $\{\mathbb{A}_2\mathbf{H}_t^0\}$ is considered as a known function and consequently will not be subjected to variations when extremizing the functional. This constraint is to be released after extremization, making $\mathbf{H}_t^0 = \mathbf{H}_t$.

Although (5.10b) is the weak form of the boundary-value problem, which is defined in chapter 3, ready for finite element implementation, it is not suitable for the most popular first-order finite elements which are only of C^0 continuity while the operators $\mathbb{A}_1$ and $\mathbb{A}_2$ contain second order derivatives which require finite elements of C^1 continuity. However, we can remove the second order derivatives by integration by parts, discussed in the previous chapter, with the help of vector identities [98].

5.2.2 Reduction of Continuity Requirement

Recall the fundamental property of finite element approximations discussed in section 4.5.3, the finite element approximation can be completely formulated locally, one element at a time and each element independent of the others, and global approximation can then be obtained by simple transformations of local equations.

In this way, we only need to pay attention on a typical element S_e, which is the closure of open region S_e ($S_e = S_e + C_e$). We also assume that the permittivity tensor inside S_e is constant. The surface integral over region S in (5.10) is simply the sum of the surface integral over each

element S_e:

$$\int_S (\cdot)\, ds = \sum_{e=1}^{E} \int_{S_e} (\cdot)\, ds \tag{5.11}$$

where we have assumed a finite element model with E elements.

We now perform integration by part to the first and second term in (5.10b) to reduce the continuity requirement for finite element shape function.

The first term can be transformed as:

$$\begin{aligned}
\langle \mathbf{H}_t, \mathbb{A}_1 \mathbf{H}_t \rangle &= \int_{S_e} [\mathbf{H}_t \cdot \nabla_t \times (\kappa_{zz} \nabla_t \times \mathbf{H}_t)]\, ds \\
&= \int_{S_e} \kappa_{zz} (\nabla_t \times \mathbf{H}_t) \cdot (\nabla_t \times \mathbf{H}_t)\, ds + \oint_{C_e} [(\kappa_{zz} \nabla_t \times \mathbf{H}_t) \times \mathbf{H}_t] \cdot \mathbf{n}\, dl \\
&= \int_{S_e} \kappa_{zz} (\nabla_t \times \mathbf{H}_t) \cdot (\nabla_t \times \mathbf{H}_t)\, ds + \oint_{C_e} (\kappa_{zz} \nabla_t \times \mathbf{H}_t) \cdot (\mathbf{H}_t \times \mathbf{n})\, dl
\end{aligned} \tag{5.12}$$

In a similar way, for the second term we have

$$\begin{aligned}
2 \langle \mathbf{H}_t, \{\mathbb{A}_2 \mathbf{H}_t\}_0 \rangle &= -2 \int_{S_e} \mathbf{H}_t \cdot \{\mathbf{z} \times [\bar{\bar{\kappa}}_{tt} \cdot \nabla_t \times (\mathbf{z} \nabla_t \cdot \mathbf{H}_t^0)]\}\, ds \\
&= 2 \int_{S_e} (\mathbf{z} \times \mathbf{H}_t) \cdot \{\bar{\bar{\kappa}}_{tt} \cdot [\nabla_t \times (\mathbf{z} \nabla_t \cdot \mathbf{H}_t^0)]\}\, ds \\
&= 2 \int_{S_e} [\bar{\bar{\kappa}}_{tt} \cdot (\mathbf{z} \times \mathbf{H}_t)] \cdot [\nabla_t \times (\mathbf{z} \nabla_t \cdot \mathbf{H}_t^0)]\, ds \\
&= 2 \int_{S_e} \{\nabla_t \times [\bar{\bar{\kappa}}_{tt} \cdot (\mathbf{z} \times \mathbf{H}_t)]\} \cdot \mathbf{z}\, \nabla_t \cdot \mathbf{H}_t^0\, ds \\
&\quad + 2 \oint_{C_e} \nabla_t \cdot \mathbf{H}_t^0\, \{\mathbf{z} \times [\bar{\bar{\kappa}}_{tt} \cdot (\mathbf{z} \times \mathbf{H}_t)]\} \cdot \mathbf{n}\, dl
\end{aligned} \tag{5.13}$$

Because there are no singularities in the integrands of the third and the fourth terms in (5.10b) over the whole closed region S_e, they simply yield

$$\begin{aligned}
\langle \mathbf{H}_t, \mathbb{A}_3 \mathbf{H}_t \rangle &= -\int_{S_e} \mathbf{H}_t \cdot \omega^2 \mu_0 \varepsilon_0 \mathbf{H}_t\, ds \\
&= -\int_{S_e} k_0^2\, \mathbf{H}_t \cdot \mathbf{H}_t\, ds
\end{aligned} \tag{5.14}$$

and

$$\langle \mathbf{H}_t, \mathbb{B}\, \mathbf{H}_t \rangle = \int_{S_e} \mathbf{H}_t \cdot \{ \mathbf{z} \times [\bar{\bar{\kappa}}_{tt} \cdot (\mathbf{z} \times \mathbf{H}_t)] \}\, ds$$

$$= - \int_{S_e} [\bar{\bar{\kappa}}_{tt} \cdot (\mathbf{z} \times \mathbf{H}_t)] \cdot (\mathbf{z} \times \mathbf{H}_t)\, ds \qquad (5.15)$$

Summarizing (5.12) to (5.15) and extending the integrals from one element to the whole waveguide region of E elements, we have the variational formulation of form:

$$\Pi = A + \gamma^2 B = 0 \qquad (5.16a)$$

where

$$A = \sum_{e=1}^{E} \int_{S_e} \kappa_{zz} (\nabla_t \times \mathbf{H}_t) \cdot (\nabla_t \times \mathbf{H}_t)\, ds$$

$$+ \sum_{e=1}^{E} \oint_{C_e} (\kappa_{zz} \nabla_t \times \mathbf{H}_t) \cdot (\mathbf{H}_t \times \mathbf{n})\, dl$$

$$+ \sum_{e=1}^{E} \int_{S_e} 2\, \nabla_t \cdot \mathbf{H}_t^0\, \mathbf{z} \cdot \nabla_t \times [\bar{\bar{\kappa}}_{tt} \cdot (\mathbf{z} \times \mathbf{H}_t)]\, ds$$

$$+ \sum_{e=1}^{E} \oint_{C_e} 2\, \nabla_t \cdot \mathbf{H}_t^0\, \{ \mathbf{z} \times [\bar{\bar{\kappa}}_{tt} \cdot (\mathbf{z} \times \mathbf{H}_t)] \} \cdot \mathbf{n}\, dl$$

$$- \sum_{e=1}^{E} \int_{S_e} k_0^2\, \mathbf{H}_t \cdot \mathbf{H}_t\, ds \qquad (5.16b)$$

$$B = - \sum_{e=1}^{E} \int_{S_e} (\mathbf{z} \times \mathbf{H}_t) \cdot [\bar{\bar{\kappa}}_{tt} \cdot (\mathbf{z} \times \mathbf{H}_t)]\, ds \qquad (5.16c)$$

The closed element boundaries C_e consist of a number of line sections which may be classified as following two types:

(i) exterior waveguide wall sections L_e^w : $L_e^w \subset C_e$ and $L_e^w \cap C = L_e^w$;

(ii) interior element interface sections L_e^i : $L_e^i \subset C_e$ and $L_e^i \cap C = \varnothing$;

where C is the cross-section of waveguide boundary Γ.

As a result, the overall contributions of contour integral on C_e in (5.16) can be rearranged as

$$\sum_{e=1}^{E} \oint_{C_e} \left\{ (\kappa_{zz} \nabla_t \times \mathbf{H}_t) \cdot (\mathbf{H}_t \times \mathbf{n}) + 2\, \nabla_t \cdot \mathbf{H}_t^0 \, \{\mathbf{z} \times [\bar{\bar{\kappa}}_{tt} \cdot (\mathbf{z} \times \mathbf{H}_t)]\} \cdot \mathbf{n} \right\} dl$$

$$= \sum_{p=1}^{P} A_p^w + \sum_{q=1}^{Q} A_q^i \qquad (5.17a)$$

where

$$A_p^w = \int_{L_p^w} \left\{ (\kappa_{zz} (\nabla_t \times \mathbf{H}_t) \cdot (\mathbf{H}_t \times \mathbf{n}) + 2\, \nabla_t \cdot \mathbf{H}_t^0 \, \{\mathbf{z} \times [\bar{\bar{\kappa}}_{tt} \cdot (\mathbf{z} \times \mathbf{H}_t)]\} \cdot \mathbf{n} \right\} dl \qquad (5.17b)$$

$$A_q^i = \int_{L_q^i} \left\{ \kappa_{zz}^{(+)} (\nabla_t \times \mathbf{H}_t^{(+)}) \cdot (\mathbf{H}_t^{(+)} \times \mathbf{n}^{(+)}) \right.$$

$$+ \kappa_{zz}^{(-)} (\nabla_t \times \mathbf{H}_t^{(-)}) \cdot (\mathbf{H}_t^{(-)} \times \mathbf{n}^{(-)})$$

$$+ 2\, \nabla_t \cdot \mathbf{H}_t^{0(+)} \{\mathbf{z} \times [\bar{\bar{\kappa}}_{tt}^{(+)} \cdot (\mathbf{z} \times \mathbf{H}_t^{(+)})]\} \cdot \mathbf{n}^{(+)}$$

$$\left. + 2\, \nabla_t \cdot \mathbf{H}_t^{0(-)} \{\mathbf{z} \times [\bar{\bar{\kappa}}_{tt}^{(-)} \cdot (\mathbf{z} \times \mathbf{H}_t^{(-)})]\} \cdot \mathbf{n}^{(-)} \right\} dl \qquad (5.17c)$$

In (5.17), P is the total number of element boundary sections on the waveguide wall, Q is the total number of inter-element sections, A_p^w is the line integral contribution on the pth wall section, A_q^i is the line integral contribution on the qth inter-element section, the symbol (+) denotes the values on L_q^i from the element on one side of the inter-element interface section L_q^i, (–) denotes the values on L_q^i from element on the other side of L_q^i.

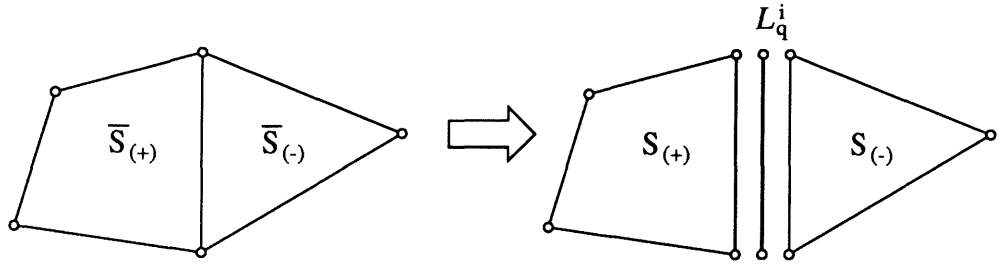


Fig. 5.1 Inter-element interface

5.2.3 *Investigation of the Line Integrals*

The line integral on exterior walls

For the first term $\kappa_{zz}(\nabla_t \times \mathbf{H}_t)\cdot(\mathbf{H}_t \times \mathbf{n})$ in (5.17b), we note from the boundary condition (see Appendix A) that $\kappa_{zz}\nabla_t \times \mathbf{H}_t$ vanishes on PEC (perfect electric conductor), and $\mathbf{H}_t \times \mathbf{n}$ vanishes on PMC (perfect magnetic conductor), therefore, the contribution of $(\kappa_{zz}\nabla_t \times \mathbf{H}_t)\cdot(\mathbf{H}_t \times \mathbf{n})$ is null on PEC, PMC, and obviously at INF (infinity) as well.

For the second term $2\,\nabla_t \cdot \mathbf{H}_t^0\,\{\mathbf{z}\times[\bar{\bar{\kappa}}_{tt}\cdot(\mathbf{z}\times\mathbf{H}_t)]\}\cdot\mathbf{n}$ in (5.17b), $\nabla_t \cdot \mathbf{H}_t^0$ vanishes on PMC, and $\{\mathbf{z}\times[\bar{\bar{\kappa}}_{tt}\cdot(\mathbf{z}\times\mathbf{H}_t)]\}\cdot\mathbf{n}$ vanishes on PEC if the dielectric in the element has

(i) isotropy ($\varepsilon_{xy} = 0$, $\varepsilon_{xx} = \varepsilon_{yy} = \varepsilon_{zz}$) or
special diagonal anisotropy ($\varepsilon_{xy} = 0$, $\varepsilon_{xx} = \varepsilon_{yy} \neq \varepsilon_{zz}$); (5.18)

or (ii) arbitrary diagonal anisotropy (only $\varepsilon_{xy} = 0$)
with the element edge in x or y direction. (5.19)

In fact, for isotropy or special diagonal anisotropy case (5.18) in which $\varepsilon_{xx} = \varepsilon_{yy} = \varepsilon_t$, ($\bar{\bar{\kappa}}_{tt} = \varepsilon_t^{-1}\bar{\bar{I}}$, $\bar{\bar{I}}$ is the unit tensor), we have

$$\{\mathbf{z}\times[\bar{\bar{\kappa}}_{tt}\cdot(\mathbf{z}\times\mathbf{H}_t)]\}\cdot\mathbf{n} = \varepsilon_t^{-1}[\mathbf{z}\times(\mathbf{z}\times\mathbf{H}_t)]\cdot\mathbf{n} = \varepsilon_t^{-1}\mathbf{H}_t\cdot\mathbf{n} \tag{5.20}$$

where $\mathbf{H}_t\cdot\mathbf{n}$ vanishes on PEC.

For arbitrary diagonal anisotropy case in which $\bar{\bar{\kappa}}_{tt} = \kappa_{xx}\mathbf{xx} + \kappa_{yy}\mathbf{yy}$, we have

$$\{\mathbf{z}\times[\bar{\bar{\kappa}}_{tt}\cdot(\mathbf{z}\times\mathbf{H}_t)]\}\cdot\mathbf{n} = \{\mathbf{z}\times[\kappa_{xx}\mathbf{xx} + \kappa_{yy}\mathbf{yy}\cdot(H_x\mathbf{y} - H_y\mathbf{x})]\}\cdot\mathbf{n}$$

$$= \{\mathbf{z}\times[\kappa_{yy}H_x\mathbf{y} - \kappa_{xx}H_y\mathbf{x})]\}\cdot\mathbf{n} = -\kappa_{yy}H_x\mathbf{x}\cdot\mathbf{n} - \kappa_{xx}H_y\mathbf{y}\cdot\mathbf{n} \tag{5.21}$$

If the PEC is in the y-direction, i.e., $\mathbf{n} = \pm\,\mathbf{x}$, then $\{\mathbf{z}\times[\bar{\bar{\kappa}}_{tt}\cdot(\mathbf{z}\times\mathbf{H}_t)]\}\cdot\mathbf{n} = \mp\,\kappa_{yy}H_x = 0$, since H_x vanishes; if the PEC is in the x-direction, $\mathbf{n} = \pm\,\mathbf{y}$, in this case $\{\mathbf{z}\times[\bar{\bar{\kappa}}_{tt}\cdot(\mathbf{z}\times\mathbf{H}_t)]\}\cdot\mathbf{n} = \mp\,\kappa_{xx}H_y = 0$, where H_y is zero.

Summarizing above investigation, it arrives at

$$A_p^w = \int_{L_p^w}\left\{(\kappa_{zz}(\nabla_t \times \mathbf{H}_t)\cdot(\mathbf{H}_t \times \mathbf{n}) + 2\,\nabla_t \cdot \mathbf{H}_t^0\,\{\mathbf{z}\times[\bar{\bar{\kappa}}_{tt}\cdot(\mathbf{z}\times\mathbf{H}_t)]\}\cdot\mathbf{n}\right\} dl$$

$$= \begin{cases} 0 & \text{(if on PMC, PEC*, INF)} \\ \displaystyle\int_{L_p^w} 2\,\nabla_t \cdot \mathbf{H}_t^0\,\{\mathbf{z}\times[\bar{\bar{\kappa}}_{tt}\cdot(\mathbf{z}\times\mathbf{H}_t)]\}\cdot\mathbf{n}\, dl & \text{(if otherwise)} \end{cases} \tag{5.22}$$

where * means in the cases (5.18) and (5.19).

Because the line integrals A_p^w may only have contributions on PEC, the line integrals A_p^w are naturally reduced only on PEC and we can write $A_p^w = A_p^{pec}$, $L_p^w = L_p^{pec}$, and reduce P to P^* ($P^* \leq P$).

The line integral on element interfaces

Note that in (5.17c) the inter-element interface unit normal vectors $\mathbf{n}^{(+)}$ and $\mathbf{n}^{(-)}$ from both sides of the interface L_q^i are just opposite in direction for any point on L_q^i, namely

$$\mathbf{n}^{(-)} = -\mathbf{n}^{(+)} \tag{5.23}$$

From the field interface conditions (see Appendix A), we have

$$\mathbf{H}_t^{(+)} = \mathbf{H}_t^{(-)} = \mathbf{H}_t \tag{5.24}$$

$$\kappa_{zz}^{(+)}\nabla_t \times \mathbf{H}_t^{(+)} = \kappa_{zz}^{(-)}\nabla_t \times \mathbf{H}_t^{(-)} \tag{5.25}$$

$$\nabla_t \cdot \mathbf{H}_t^{0(+)} = \nabla_t \cdot \mathbf{H}_t^{0(-)} = \nabla_t \cdot \mathbf{H}_t^0 \tag{5.26}$$

Eqs. (5.23) to (5.25) show that the first two terms in (5.17c) cancel each other, that is

$$\int_{L_q^i} [\kappa_{zz}^{(+)}(\nabla_t \times \mathbf{H}_t^{(+)}) \cdot (\mathbf{H}_t^{(+)} \times \mathbf{n}^{(+)}) + \kappa_{zz}^{(-)}(\nabla_t \times \mathbf{H}_t^{(-)}) \cdot (\mathbf{H}_t^{(-)} \times \mathbf{n}^{(-)})]\, dl = 0 \tag{5.27}$$

Introducing (5.23), (5.24), (5.26), and (5.27) into (5.17c), A_q^i can be expressed as

$$A_q^i = \int_{L_q^i} 2\, \nabla_t \cdot \mathbf{H}_t^0 \, \{\mathbf{z} \times [(\bar{\bar{\kappa}}_{tt}^{(+)} - \bar{\bar{\kappa}}_{tt}^{(-)}) \cdot (\mathbf{z} \times \mathbf{H}_t)]\} \cdot \mathbf{n}^{(+)}\, dl \tag{5.28}$$

The above line integral will not vanish only if $\bar{\bar{\kappa}}_{tt}^{(+)} \neq \bar{\bar{\kappa}}_{tt}^{(-)}$. For A_q^i, we therefore only need to take into account the line integral on interfaces between different dielectrics.

5.2.4 Final Expression of the Variational Expression

Summarizing the eqs. (5.16), (5.17), (5.22), (5.28), and the investigations in subsections 5.2.2 to 5.2.3, we finally obtain the finite element variational formulation

$$\Pi = A + \gamma^2 B = 0 \tag{5.29a}$$

where

$$A = \sum_{e=1}^{E} \int_{S_e} \kappa_{zz} (\nabla_t \times \mathbf{H}_t) \cdot (\nabla_t \times \mathbf{H}_t)\, ds$$

$$+ \sum_{e=1}^{E} \int_{S_e} 2\, \nabla_t \cdot \mathbf{H}_t^0\, \mathbf{z} \cdot \nabla_t \times [\bar{\bar{\kappa}}_{tt} \cdot (\mathbf{z} \times \mathbf{H}_t)]\, ds$$

$$- \sum_{e=1}^{E} \int_{S_e} k_0^2\, \mathbf{H}_t \cdot \mathbf{H}_t\, ds$$

$$+ \sum_{p=1}^{P^*} \delta_p^w \int_{L_p^{pec}} 2\, \nabla_t \cdot \mathbf{H}_t^0\, \{\mathbf{z} \times [\bar{\bar{\kappa}}_{tt} \cdot (\mathbf{z} \times \mathbf{H}_t)]\} \cdot \mathbf{n}\, dl$$

$$+ \sum_{q=1}^{Q^*} \int_{L_q^{int}} 2\, \nabla_t \cdot \mathbf{H}_t^0\, \{\mathbf{z} \times [(\bar{\bar{\kappa}}_{tt}^{(+)} - \bar{\bar{\kappa}}_{tt}^{(-)}) \cdot (\mathbf{z} \times \mathbf{H}_t)]\} \cdot \mathbf{n}^{(+)}\, dl \tag{5.29b}$$

$$B = - \sum_{e=1}^{E} \int_{S_e} (\mathbf{z} \times \mathbf{H}_t) \cdot [\bar{\bar{\kappa}}_{tt} \cdot (\mathbf{z} \times \mathbf{H}_t)]\, ds \tag{5.29c}$$

In order to distinguish the three types of elements, S_e, L_p^{pec}, L_q^{int}, we may call S_e the eth *area element*, L_p^{pec} the pth *PEC line element*, L_q^{int} the qth *interface line element*. In (5.29b), E, P^* and Q^* are the total numbers of area elements, PEC line elements, dielectric interface line elements, respectively. The δ_p^w is defined as

$$\delta_p^w = \begin{cases} 0 & \text{(if in the cases (5.18) or (5.19))} \\ 1 & \text{(if otherwise)} \end{cases} \tag{5.29d}$$

One should keep in mind that in the formulation (5.29) $\mathbf{H}_t^0$ is considered as a known function and consequently will not be subject to variations when extremizing the functional. This constraint is to be released after extremization, making $\mathbf{H}_t^0 = \mathbf{H}_t$.

5.3. Comments on the New Formulation

Equation (5.29) is the new variational finite element formulation which is a weak form of the boundary-value problem. The second order derivatives

have been removed, allowing the use of C^0 continuous first order or higher order finite elements. Otherwise, one may have to use the C^1 continuous Hermitian interpolating finite elements to solve the formulation (5.10). The use of Hermite elements would greatly increase the number of unknowns and result in increased complexity for numerical computation. The removal of the second order derivatives is at the expense of including the line integrals. Fortunately, we only need to take into account the line integrals on the dielectric interfaces between different media in most cases of interest. Rarely we also need to add the line integrals on PEC.

This formulation satisfies all the two equations, and the essential interface and boundary conditions for **H** approximation discussed in chapter 3. It also satisfied implicitly the longitudinal tangential conditions of H_z and E_z components. This ensures the elimination of spurious solutions.

We have obtained a formulation in terms of only two components rather than three [63]-[67], four [84]-[86], or even six component [88] formulations. Unlike the two component formulation of Hayata *et al.* [81] which resort to dense matrices, our formulation is achieved without losing the sparsity of matrices of the resultant eigenvalue equation. This will substantially reduce the amount of computing storage and time. This is of decisive importance for large problems, even on a supercomputer. Therefore, the gain is worth the penalty of including the line integral.

CHAPTER 6

FINITE ELEMENT IMPLEMENTATION

6.1 Introduction

In this chapter, we will implement the variational formulation derived in Chapter 5 with the use of finite elements. The basic procedure follows the theory discussed in Chapter 4. The representations of field component, minimization, local and global matrix elements will be expressed in spirit and form of practical computation procedures. We will also describe briefly the quadrilateral and infinite elements adopted in the computer program. Finally we will give an account of the principles of a unique efficient eigenvalue solver for large, sparse, non-symmetric complex general eigenvalue equations.

6.2 Finite Element Representation

Dividing the cross-section $\bar{S}$ ($\bar{S} = S + C$) of Fig. 3.1 into a mesh of G area elements, we here denote $\bar{S}$, S, and C as the cross-section of $\bar{\Omega}$, Ω, and Γ, respectively. We assume that there are P^* PEC line sections (elements), and Q^* dielectric interface line sections (elements) in the mesh. The finite element model is established by the respective mappings

$$\overset{(e)}{\Lambda}_{g}^{N} = \begin{cases} 1 & \text{if node } g \text{ of global model } \bar{S} \text{ is coincident with} \\ & \text{node } N \text{ of element } S_e, \\ 0 & \text{otherwise,} \end{cases} \tag{6.1a}$$

$$\overset{(e)}{\Delta}_{N}^{g} = \begin{cases} 1 & \text{if node } N \text{ of the element } \bar{S}_e \text{ is coincident with} \\ & \text{node } g \text{ of the connected model } \bar{S}, \\ 0 & \text{otherwise,} \end{cases} \tag{6.1b}$$

when concerning line sections on PEC and dielectric interface, mapping $\overset{(e)}{\Delta}_{N}^{g}$ can be reduced to $\overset{(p)}{\Xi}_{N}^{g}$, $\overset{(q)}{\Theta}_{N}^{g}$ respectively

$$ {}^{(p)}\Xi_N^g = \begin{cases} 1 & \text{if node } N \text{ of line element } L_p^{pec} \subset \mathcal{C}_e \subset \mathcal{S}_e \text{ is} \\ & \text{coincident with node } g \text{ of the connected model } \mathcal{S}, \\ 0 & \text{otherwise,} \end{cases} \tag{6.1c} $$

$$ {}^{(q)}\Theta_N^g = \begin{cases} 1 & \text{if node } N \text{ of line element } L_q^{int} \subset \mathcal{C}_e \subset \mathcal{S}_e \text{ is} \\ & \text{coincident with node } g \text{ of the connected model } \mathcal{S}, \\ 0 & \text{if otherwise,} \end{cases} \tag{6.1d} $$

Referring to section 4.5, the finite element representation of $\mathbf{H}_t$ can be expressed as

$$ \mathbf{H}_t(x,y) = \sum_{g=1}^{G} \mathbf{h}_g \, \phi_g(x,y) $$

$$ = \bigcup_{e=1}^{E} \mathbf{H}_t^{(e)}(x,y) = \bigcup_{e=1}^{E} \sum_{N=1}^{Ne} \mathbf{h}_N^{(e)} \psi_N^{(e)}(x,y) \tag{6.2a} $$

where

$$ \phi_g(x,y) = \bigcup_{e=1}^{E} \sum_{N=1}^{Ne} {}^{(e)}\Delta_g^N \, \psi_N^{(e)}(x,y) \tag{6.2b} $$

$$ \mathbf{h}_N^{(e)} = \sum_{g=1}^{G} {}^{(e)}\Delta_g^N \, \mathbf{h}_g \tag{6.2c} $$

$$ \mathbf{H}_t^{(e)}(x,y) = \sum_{N=1}^{Ne} \mathbf{h}_N^{(e)} \psi_N^{(e)}(x,y) \tag{6.2d} $$

in which $\psi_N^{(e)}(x,y)$ are the trial functions of local element S_e,

$$ \mathbf{h}_g = \mathbf{x} \, h_{x_g} + \mathbf{y} \, h_{y_g} \tag{6.2e} $$

are unknown vector coefficients to be determined. Introducing (6.2c) into (6.2d) we have

$$ \mathbf{H}_t^{(e)} = \sum_{N=1}^{Ne} \sum_{g=1}^{G} {}^{(e)}\Delta_g^N \, \mathbf{h}_g \, \psi_N^{(e)} \tag{6.3} $$

6.3 Extremizing the Functional

There are four steps in applying the Rayleigh-Ritz procedure discussed in chapter 4 to formulation (5.29) to extremize the functional Π:

Step 1: introducing the finite representation (6.3) into the finite formulation (5.29);

Step 2: taking derivatives with respect to hx_h and hy_m respectively;

Step 3: releasing the local potential constraints after step 2;

Step 4: simplifying mapping $\overset{(e)}{\Delta}_N^g$ to $\overset{(p)}{\Xi}_N^g$, $\overset{(q)}{\Theta}_N^g$ respectively on PEC and dielectric interfaces.

Following the above 4 steps, we can obtain the following systems of equations:

$$\sum_{e=1}^{E}\sum_{M=1}^{Ne}\sum_{N=1}^{Ne}\sum_{g=1}^{G}\overset{(e)}{\Delta}_h^M\,\overset{(e)}{\Delta}_g^N\int_{S_e}2\Big\{$$

$$\kappa_{zz}\nabla_t\times(\mathbf{x}\psi_M^{(e)})\cdot\nabla_t\times(\mathbf{h}_g\psi_N^{(e)})$$

$$+\nabla_t\cdot(\mathbf{h}_g\psi_N^{(e)})\;\mathbf{z}\cdot\nabla_t\times[\overline{\overline{\kappa}}_{tt}\cdot(\mathbf{z}\times\mathbf{x}\psi_M^{(e)})]$$

$$-k_0^2\;(\mathbf{x}\psi_M^{(e)})\cdot(\mathbf{h}_g\psi_N^{(e)})$$

$$-\gamma^2\;(\mathbf{z}\times\mathbf{x}\psi_M^{(e)})\cdot[\overline{\overline{\kappa}}_{tt}\cdot(\mathbf{z}\times\mathbf{h}_g\psi_N^{(e)})]\Big\}\,ds$$

$$+\sum_{p=1}^{P^*}\sum_{M=1}^{Np}\sum_{N=1}^{Np}\sum_{g=1}^{G}\overset{(p)}{\Xi}_h^M\,\overset{(p)}{\Xi}_g^N\;\delta^p\int_{L_p^{pec}}2\Big\{$$

$$\nabla_t\cdot(\mathbf{h}_g\psi_N^{(p)})\;\{\mathbf{z}\times[\overline{\overline{\kappa}}_{tt}\cdot(\mathbf{z}\times\mathbf{x}\psi_M^{(p)})]\}\cdot\mathbf{n}\Big\}\,dl$$

$$+\sum_{q=1}^{Q^*}\sum_{M=1}^{Nq}\sum_{N=1}^{Nq}\sum_{g=1}^{G}\overset{(q)}{\Theta}_h^M\,\overset{(q)}{\Theta}_g^N\int_{L_q^{int}}2\Big\{$$

$$\nabla_t\cdot(\mathbf{h}_g\psi_N^{(p)})\;\{\mathbf{z}\times[(\overline{\overline{\kappa}}_{tt}^{(+)}-\overline{\overline{\kappa}}_{tt}^{(-)})\cdot(\mathbf{z}\times\mathbf{x}\psi_M^{(p)})]\}\cdot\mathbf{n}^{(+)}\Big\}\,dl=0$$

$$(h=1,\ \ldots,\ G)\qquad(6.4a)$$

$$\sum_{e=1}^{E}\sum_{M=1}^{Ne}\sum_{N=1}^{Ne}\sum_{g=1}^{G}\overset{(e)}{\Delta}_h^M\,\overset{(e)}{\Delta}_g^N\int_{S_e}2\Big\{$$

$$\kappa_{zz}\nabla_t\times(\mathbf{y}\psi_M^{(e)})\cdot\nabla_t\times(\mathbf{h}_g\psi_N^{(e)})$$

$$+\nabla_t\cdot(\mathbf{h}_g\psi_N^{(e)})\;\mathbf{z}\cdot\nabla_t\times[\overline{\overline{\kappa}}_{tt}\cdot(\mathbf{z}\times\mathbf{y}\psi_M^{(e)})]$$

$$- k_0^2 \, (\mathbf{y}\psi_M^{(e)})\cdot(\mathbf{h}_g\psi_N^{(e)})$$

$$- \gamma^2 \, (\mathbf{z}\times\mathbf{y}\psi_M^{(e)})\cdot[\overline{\overline{\kappa}}_{tt}\cdot(\mathbf{z}\times\mathbf{h}_g\psi_N^{(e)})]\Big\} \, ds$$

$$+ \sum_{p=1}^{P^*} \sum_{M=1}^{Np} \sum_{N=1}^{Np} \sum_{g=1}^{G} \Xi_h^{(p)M} \, \Xi_g^{(p)N} \, \delta^p \int_{L_p^{pec}} 2 \Big\{$$

$$\nabla_t\cdot(\mathbf{h}_g\psi_N^{(p)}) \; \{\mathbf{z}\times[\overline{\overline{\kappa}}_{tt}\cdot(\mathbf{z}\times\mathbf{y}\psi_M^{(p)})]\}\cdot\mathbf{n}\Big\} \, dl$$

$$+ \sum_{q=1}^{Q^*} \sum_{M=1}^{Nq} \sum_{N=1}^{Nq} \sum_{g=1}^{G} \Theta_h^{(q)M} \, \Theta_g^{(q)N} \int_{L_q^{int}} 2 \Big\{$$

$$\nabla_t\cdot(\mathbf{h}_g\psi_N^{(p)}) \; \{\mathbf{z}\times[(\overline{\overline{\kappa}}_{tt}^{(+)} - \overline{\overline{\kappa}}_{tt}^{(-)})\cdot(\mathbf{z}\times\mathbf{y}\psi_M^{(p)})]\}\cdot\mathbf{n}^{(+)}\Big\} \, dl = 0$$

$$(h = 1, \ldots, G) \qquad (6.4b)$$

6.4 The Matrix Eigenvalue Equation

The system of equations (6.4a) and (6.4b) can be expressed more succinctly as (6.5a) and (6.5b) respectively:

$$\sum_{g=1}^{G} \sum_{e=1}^{E} \sum_{M=1}^{Ne} \sum_{N=1}^{Ne} \Delta_h^{(e)M} \, \Delta_g^{(e)N} \Big\{ [\, S_{xx}^{(e)}(M,N) - \gamma^2 \, B_{xx}^{(e)}(M,N)]h_{x_g}$$

$$[\, S_{xy}^{(e)}(M,N) - \gamma^2 \, B_{xy}^{(e)}(M,N)]h_{y_g} \Big\}$$

$$+ \sum_{g=1}^{G} \sum_{p=1}^{P^*} \sum_{M=1}^{Np} \sum_{N=1}^{Np} \Xi_h^{(p)M} \, \Xi_g^{(p)N} \Big\{ W_{xx}^{(p)}(M,N) \, h_{x_g} + W_{xy}^{(p)}(M,N) \, h_{y_g} \Big\}$$

$$+ \sum_{g=1}^{G} \sum_{q=1}^{Q^*} \sum_{M=1}^{Nq} \sum_{N=1}^{Nq} \Theta_h^{(q)M} \, \Theta_g^{(q)N} \Big\{ T_{xx}^{(q)}(M,N) \, h_{x_g} + T_{xy}^{(q)}(M,N) \, h_{y_g} \Big\} = 0$$

$$(h = 1, \ldots, G) \qquad (6.5a)$$

$$\sum_{g=1}^{G} \sum_{e=1}^{E} \sum_{M=1}^{Ne} \sum_{N=1}^{Ne} \Delta_h^{(e)M} \, \Delta_g^{(e)N} \Big\{ [S_{yx}^{(e)}(M,N) - \gamma^2 \, B_{yx}^{(e)}(M,N)]h_{x_g}$$

$$[S_{yy}^{(e)}(M,N) - \gamma^2 \, B_{yy}^{(e)}(M,N)]h_{y_g} \Big\}$$

$$+ \sum_{g=1}^{G} \sum_{p=1}^{P^*} \sum_{M=1}^{Np} \sum_{N=1}^{Np} {}^{(p)}\Xi_h^M \, {}^{(p)}\Xi_g^N \left\{ W_{yx}^{(p)}(M,N)\ hx_g + W_{yy}^{(p)}(M,N)\ hy_g \right\}$$

$$+ \sum_{g=1}^{G} \sum_{q=1}^{Q^*} \sum_{M=1}^{Nq} \sum_{N=1}^{Nq} {}^{(q)}\Theta_h^M \, {}^{(q)}\Theta_g^N \left\{ T_{yx}^{(q)}(M,N)\ hx_g + T_{yy}^{(q)}(M,N)\ hy_g \right\} = 0$$

$$(h = 1, \ldots, G) \qquad (6.5b)$$

in (6.5)

$$S_{xx}^{(e)}(M,N) = \int_{S_e} \left\{ \kappa_{zz} \nabla_t \times (\mathbf{x}\psi_M^{(e)}) \cdot \nabla_t \times (\mathbf{x}\psi_N^{(e)}) - k_0^2\ (\mathbf{x}\psi_M^{(e)}) \cdot (\mathbf{x}\psi_N^{(e)}) \right.$$
$$\left. + \nabla_t \cdot (\mathbf{x}\psi_N^{(e)})\ \mathbf{z} \cdot \nabla_t \times [\bar{\bar{\kappa}}_{tt} \cdot (\mathbf{z} \times \mathbf{x}\psi_M^{(e)})] \right\} ds \qquad (6.6a)$$

$$S_{xy}^{(e)}(M,N) = \int_{S_e} \left\{ \kappa_{zz} \nabla_t \times (\mathbf{x}\psi_M^{(e)}) \cdot \nabla_t \times (\mathbf{y}\psi_N^{(e)}) - k_0^2\ (\mathbf{x}\psi_M^{(e)}) \cdot (\mathbf{y}\psi_N^{(e)}) \right.$$
$$\left. + \{\mathbf{z} \cdot \nabla_t \times [\bar{\bar{\kappa}}_{tt} \cdot (\mathbf{z} \times \mathbf{x}\psi_M^{(e)})]\}\ \nabla_t \cdot (\mathbf{y}\psi_N^{(e)}) \right\} ds \qquad (6.6b)$$

$$S_{yx}^{(e)}(M,N) = \int_{S_e} \left\{ \kappa_{zz} \nabla_t \times (\mathbf{y}\psi_M^{(e)}) \cdot \nabla_t \times (\mathbf{x}\psi_N^{(e)}) - k_0^2\ (\mathbf{y}\psi_M^{(e)}) \cdot (\mathbf{x}\psi_N^{(e)}) \right.$$
$$\left. + \{\mathbf{z} \cdot \nabla_t \times [\bar{\bar{\kappa}}_{tt} \cdot (\mathbf{z} \times \mathbf{y}\psi_M^{(e)})]\}\ \nabla_t \cdot (\mathbf{x}\psi_N^{(e)}) \right\} ds \qquad (6.6c)$$

$$S_{yy}^{(e)}(M,N) = \int_{S_e} \left\{ \kappa_{zz} \nabla_t \times (\mathbf{y}\psi_M^{(e)}) \cdot \nabla_t \times (\mathbf{y}\psi_N^{(e)}) - k_0^2\ (\mathbf{y}\psi_M^{(e)}) \cdot (\mathbf{y}\psi_N^{(e)}) \right.$$
$$\left. + \{\mathbf{z} \cdot \nabla_t \times [\bar{\bar{\kappa}}_{tt} \cdot (\mathbf{z} \times \mathbf{y}\psi_M^{(e)})]\}\ \nabla_t \cdot (\mathbf{y}\psi_N^{(e)}) \right\} ds \qquad (6.6d)$$

$$B_{xx}^{(e)}(M,N) = \int_{S_e} \left\{ - (\mathbf{z} \times \mathbf{x}\psi_M^{(e)}) \cdot [\bar{\bar{\kappa}}_{tt} \cdot (\mathbf{z} \times \mathbf{x}\psi_N^{(e)})] \right\} ds \qquad (6.7a)$$

$$B_{xy}^{(e)}(M,N) = \int_{S_e} \left\{ - (\mathbf{z} \times \mathbf{x}\psi_M^{(e)}) \cdot [\bar{\bar{\kappa}}_{tt} \cdot (\mathbf{z} \times \mathbf{y}\psi_N^{(e)})] \right\} ds \qquad (6.7b)$$

$$B_{yx}^{(e)}(M,N) = \int_{S_e} \left\{ - (\mathbf{z} \times \mathbf{y}\psi_M^{(e)}) \cdot [\bar{\bar{\kappa}}_{tt} \cdot (\mathbf{z} \times \mathbf{x}\psi_N^{(e)})] \right\} ds \qquad (6.7c)$$

$$B_{yy}^{(e)}(M,N) = \int_{S_e} \left\{ - (\mathbf{z} \times \mathbf{y}\psi_M^{(e)}) \cdot [\bar{\bar{\kappa}}_{tt} \cdot (\mathbf{z} \times \mathbf{y}\psi_N^{(e)})] \right\} ds \qquad (6.7d)$$

$$W_{xx}^{(p)}(M,N) = \int_{L_p^{pec}} \mathbf{n}\cdot\{\mathbf{z}\times[\bar{\bar{\kappa}}_{tt}\cdot(\mathbf{z}\times\mathbf{x}\psi_M^{(p)})]\}\ \nabla_t\cdot(\mathbf{x}\psi_N^{(p)})\ dl \tag{6.8a}$$

$$W_{xy}^{(p)}(M,N) = \int_{L_p^{pec}} \mathbf{n}\cdot\{\mathbf{z}\times[\bar{\bar{\kappa}}_{tt}\cdot(\mathbf{z}\times\mathbf{x}\psi_M^{(p)})]\}\ \nabla_t\cdot(\mathbf{y}\psi_N^{(p)})\ dl \tag{6.8b}$$

$$W_{yx}^{(p)}(M,N) = \int_{L_p^{pec}} \mathbf{n}\cdot\{\mathbf{z}\times[\bar{\bar{\kappa}}_{tt}\cdot(\mathbf{z}\times\mathbf{y}\psi_M^{(p)})]\}\ \nabla_t\cdot(\mathbf{x}\psi_N^{(p)})\ dl \tag{6.8c}$$

$$W_{yy}^{(p)}(M,N) = \int_{L_p^{pec}} \mathbf{n}\cdot\{\mathbf{z}\times[\bar{\bar{\kappa}}_{tt}\cdot(\mathbf{z}\times\mathbf{y}\psi_M^{(p)})]\}\ \nabla_t\cdot(\mathbf{y}\psi_N^{(p)})\ dl \tag{6.8d}$$

$$T_{xx}^{(q)}(M,N) = \int_{L_q^{int}} \mathbf{n}^{(+)}\cdot\{\mathbf{z}\times[(\bar{\bar{\kappa}}_{tt}^{(+)} - \bar{\bar{\kappa}}_{tt}^{(-)})\cdot(\mathbf{z}\times\mathbf{x}\psi_M^{(q)})]\}\nabla_t\cdot(\mathbf{x}\psi_N^{(q)})dl \tag{6.9a}$$

$$T_{xy}^{(q)}(M,N) = \int_{L_q^{int}} \mathbf{n}^{(+)}\cdot\{\mathbf{z}\times[(\bar{\bar{\kappa}}_{tt}^{(+)} - \bar{\bar{\kappa}}_{tt}^{(-)})\cdot(\mathbf{z}\times\mathbf{x}\psi_M^{(q)})]\}\nabla_t\cdot(\mathbf{y}\psi_N^{(q)})dl \tag{6.9b}$$

$$T_{yx}^{(q)}(M,N) = \int_{L_q^{int}} \mathbf{n}^{(+)}\cdot\{\mathbf{z}\times[(\bar{\bar{\kappa}}_{tt}^{(+)} - \bar{\bar{\kappa}}_{tt}^{(-)})\cdot(\mathbf{z}\times\mathbf{y}\psi_M^{(q)})]\}\nabla_t\cdot(\mathbf{x}\psi_N^{(q)})dl \tag{6.9c}$$

$$T_{yy}^{(q)}(M,N) = \int_{L_q^{int}} \mathbf{n}^{(+)}\cdot\{\mathbf{z}\times[(\bar{\bar{\kappa}}_{tt}^{(+)} - \bar{\bar{\kappa}}_{tt}^{(-)})\cdot(\mathbf{z}\times\mathbf{y}\psi_M^{(q)})]\}\nabla_t\cdot(\mathbf{y}\psi_N^{(q)})dl \tag{6.9d}$$

The systems of equations (6.5a) and (6.5b) can be combined to be expressed in matrix form as :

$$[A]\ \{h\} = \gamma^2\ [B]\ \{h\} \tag{6.10a}$$

or more explicitly as

$$\begin{bmatrix} [A_{xx}] & [A_{xy}] \\ [A_{yx}] & [A_{yy}] \end{bmatrix} \begin{Bmatrix} \{h_x\} \\ \{h_y\} \end{Bmatrix} = \gamma^2 \begin{bmatrix} [B_{xx}] & [B_{xy}] \\ [B_{yx}] & [B_{yy}] \end{bmatrix} \begin{Bmatrix} \{h_x\} \\ \{h_y\} \end{Bmatrix} \tag{6.10b}$$

where $\{h_x\}$ and $\{h_y\}$ are sub-vectors of $\{h\}$, $[A_{xx}]$, $[A_{xy}]$, $[A_{yx}]$, $[A_{yy}]$ and $[B_{xx}]$, $[B_{xy}]$, $[B_{yx}]$, $[B_{yy}]$ are sub-matrices (of order $G\times G$) of $[A]$ and $[B]$ respectively. The matrix $[A]$ is in fact the sum of three matrices:

$$[A] = [S] + [W] + [T] \tag{6.11a}$$

in sub-matrix form, there are relations:

$$[A_{xx}] = [S_{xx}] + [W_{xx}] + [T_{xx}] \tag{6.11b}$$

$$[A_{xy}] = [S_{xy}] + [W_{xy}] + [T_{xy}] \tag{6.11c}$$

$$[A_{yx}] = [S_{yx}] + [W_{yx}] + [T_{yx}] \tag{6.11d}$$

$$[A_{yy}] = [S_{yy}] + [W_{yy}] + [T_{yy}] \tag{6.11e}$$

where $[S_{xx}]$, $[S_{xy}]$, $[S_{yx}]$, $[S_{yy}]$, $[W_{xx}]$, $[W_{xy}]$, $[W_{yx}]$, $[W_{yy}]$, $[T_{xx}]$, $[T_{xy}]$, $[T_{yx}]$, $[T_{yy}]$ are the sub-matrices (of order $G \times G$) of [S], [W], and [T] respectively.

The matrix elements of the global matrices [S], [W], [T] and [B] can be obtained by simple summations of matrix elements of corresponding local element matrices $[S^{(e)}]$, $[B^{(e)}]$, $[W^{(p)}]$, $[T^{(q)}]$:

$$S_{xx}(g,h) = \sum_{e=1}^{E} \sum_{M=1}^{Ne} \sum_{N=1}^{Ne} \Delta_h^{(e)M} \Delta_g^{(e)N} S_{xx}^{(e)}(M,N) \tag{6.12a}$$

$$S_{xy}(g,h) = \sum_{e=1}^{E} \sum_{M=1}^{Ne} \sum_{N=1}^{Ne} \Delta_h^{(e)M} \Delta_g^{(e)N} S_{xy}^{(e)}(M,N) \tag{6.12b}$$

$$S_{yx}(g,h) = \sum_{e=1}^{E} \sum_{M=1}^{Ne} \sum_{N=1}^{Ne} \Delta_h^{(e)M} \Delta_g^{(e)N} S_{yx}^{(e)}(M,N) \tag{6.12c}$$

$$S_{yy}(g,h) = \sum_{e=1}^{E} \sum_{M=1}^{Ne} \sum_{N=1}^{Ne} \Delta_h^{(e)M} \Delta_g^{(e)N} S_{yy}^{(e)}(M,N) \tag{6.12d}$$

$$B_{xx}(g,h) = \sum_{e=1}^{E} \sum_{M=1}^{Ne} \sum_{N=1}^{Ne} \Delta_h^{(e)M} \Delta_g^{(e)N} B_{xx}^{(e)}(M,N) \tag{6.13a}$$

$$B_{xy}(g,h) = \sum_{e=1}^{E} \sum_{M=1}^{Ne} \sum_{N=1}^{Ne} \Delta_h^{(e)M} \Delta_g^{(e)N} B_{xy}^{(e)}(M,N) \tag{6.13b}$$

$$B_{yx}(g,h) = \sum_{e=1}^{E} \sum_{M=1}^{Ne} \sum_{N=1}^{Ne} \Delta_h^{(e)M} \Delta_g^{(e)N} B_{yx}^{(e)}(M,N) \tag{6.13c}$$

$$B_{yy}(g,h) = \sum_{e=1}^{E} \sum_{M=1}^{Ne} \sum_{N=1}^{Ne} \Delta_h^{(e)M} \Delta_g^{(e)N} B_{yy}^{(e)}(M,N) \tag{6.13d}$$

$$W_{xx}(g,h) = \sum_{p=1}^{P^*} \sum_{M=1}^{Np} \sum_{N=1}^{Np} {}^{(p)}\Xi_h^M \; {}^{(p)}\Xi_g^N \; W_{xx}^{(p)}(M,N) \qquad (6.14a)$$

$$W_{xy}(g,h) = \sum_{p=1}^{P^*} \sum_{M=1}^{Np} \sum_{N=1}^{Np} {}^{(p)}\Xi_h^M \; {}^{(p)}\Xi_g^N \; W_{xy}^{(p)}(M,N) \qquad (6.14b)$$

$$W_{yx}(g,h) = \sum_{p=1}^{P^*} \sum_{M=1}^{Np} \sum_{N=1}^{Np} {}^{(p)}\Xi_h^M \; {}^{(p)}\Xi_g^N \; W_{yx}^{(p)}(M,N) \qquad (6.14c)$$

$$W_{yy}(g,h) = \sum_{p=1}^{P^*} \sum_{M=1}^{Np} \sum_{N=1}^{Np} {}^{(p)}\Xi_h^M \; {}^{(p)}\Xi_g^N \; W_{yy}^{(p)}(M,N) \qquad (6.14d)$$

$$T_{xx}(g,h) = \sum_{q=1}^{Q^*} \sum_{M=1}^{Nq} \sum_{N=1}^{Nq} {}^{(q)}\Theta_h^M \; {}^{(q)}\Theta_g^N \; T_{xx}^{(q)}(M,N) \qquad (6.15a)$$

$$T_{xy}(g,h) = \sum_{q=1}^{Q^*} \sum_{M=1}^{Nq} \sum_{N=1}^{Nq} {}^{(q)}\Theta_h^M \; {}^{(q)}\Theta_g^N \; T_{xy}^{(q)}(M,N) \qquad (6.15b)$$

$$T_{yx}(g,h) = \sum_{q=1}^{Q^*} \sum_{M=1}^{Nq} \sum_{N=1}^{Nq} {}^{(q)}\Theta_h^M \; {}^{(q)}\Theta_g^N \; T_{yx}^{(q)}(M,N) \qquad (6.15c)$$

$$T_{yy}(g,h) = \sum_{q=1}^{Q^*} \sum_{M=1}^{Nq} \sum_{N=1}^{Nq} {}^{(q)}\Theta_h^M \; {}^{(q)}\Theta_g^N \; T_{yy}^{(q)}(M,N) \qquad (6.15d)$$

The matrix elements of all local element matrices can be finally expanded from (6.6) to (6.9) to the following expressions ready for programming:

$$S_{xx}^{(e)}(M,N) = \int_{S_e} \Big\{ \kappa_{zz} \frac{\partial}{\partial y} \psi_M^{(e)} \frac{\partial}{\partial y} \psi_N^{(e)} + \kappa_{yy} \frac{\partial}{\partial x} \psi_M^{(e)} \frac{\partial}{\partial x} \psi_N^{(e)} - k_0^2 \, \psi_M^{(e)} \psi_N^{(e)} - \kappa_{yx} \frac{\partial}{\partial y} \psi_M^{(e)} \frac{\partial}{\partial x} \psi_N^{(e)} \Big\} \, ds \qquad (6.16a)$$

$$S_{xy}^{(e)}(M,N) = \int_{S_e} \Big\{ - \kappa_{zz} \frac{\partial}{\partial y} \psi_M^{(e)} \frac{\partial}{\partial x} \psi_N^{(e)} + \kappa_{yy} \frac{\partial}{\partial x} \psi_M^{(e)} \frac{\partial}{\partial y} \psi_N^{(e)} - \kappa_{yx} \frac{\partial}{\partial y} \psi_M^{(e)} \frac{\partial}{\partial y} \psi_N^{(e)} \Big\} \, ds \qquad (6.16b)$$

$$S_{yx}^{(e)}(M,N) = \int_{S_e} \Big\{ - \kappa_{zz} \frac{\partial}{\partial x} \psi_M^{(e)} \frac{\partial}{\partial y} \psi_N^{(e)} + \kappa_{xx} \frac{\partial}{\partial y} \psi_M^{(e)} \frac{\partial}{\partial x} \psi_N^{(e)} - \kappa_{xy} \frac{\partial}{\partial x} \psi_M^{(e)} \frac{\partial}{\partial x} \psi_N^{(e)} \Big\} \, ds \qquad (6.16c)$$

$$S_{yy}^{(e)}(M,N) = \int_{S_e} \Big\{ \kappa_{zz} \frac{\partial}{\partial x} \psi_M^{(e)} \frac{\partial}{\partial x} \psi_N^{(e)} + \kappa_{xx} \frac{\partial}{\partial y} \psi_M^{(e)} \frac{\partial}{\partial y} \psi_N^{(e)} - k_0^2 \psi_M^{(e)} \psi_N^{(e)} - \kappa_{xy} \frac{\partial}{\partial x} \psi_M^{(e)} \frac{\partial}{\partial y} \psi_N^{(e)} \Big\} ds \quad (6.16d)$$

$$B_{xx}^{(e)}(M,N) = \int_{S_e} \kappa_{yy} \psi_M^{(e)} \psi_N^{(e)} ds \quad (6.17a)$$

$$B_{xy}^{(e)}(M,N) = -\int_{S_e} \kappa_{yx} \psi_M^{(e)} \psi_N^{(e)} ds \quad (6.17b)$$

$$B_{yx}^{(e)}(M,N) = -\int_{S_e} \kappa_{xy} \psi_M^{(e)} \psi_N^{(e)} ds \quad (6.17c)$$

$$B_{yy}^{(e)}(M,N) = \int_{S_e} \kappa_{xy} \psi_M^{(e)} \psi_N^{(e)} ds \quad (6.17d)$$

$$W_{xx}^{(p)}(M,N) = \int_{L_p^{pec}} \Big\{ - \kappa_{yy} (\mathbf{x}\cdot\mathbf{n}) \psi_M^{(p)} \frac{\partial}{\partial x} \psi_N^{(p)} + \kappa_{yx} (\mathbf{y}\cdot\mathbf{n}) \psi_M^{(p)} \frac{\partial}{\partial x} \psi_N^{(p)} \Big\} dl \quad (6.18a)$$

$$W_{xy}^{(p)}(M,N) = \int_{L_p^{pec}} \Big\{ - \kappa_{yy} (\mathbf{x}\cdot\mathbf{n}) \psi_M^{(p)} \frac{\partial}{\partial y} \psi_N^{(p)} + \kappa_{yx} (\mathbf{y}\cdot\mathbf{n}) \psi_M^{(p)} \frac{\partial}{\partial y} \psi_N^{(p)} \Big\} dl \quad (6.18b)$$

$$W_{yx}^{(p)}(M,N) = \int_{L_p^{pec}} \Big\{ - \kappa_{xx} (\mathbf{y}\cdot\mathbf{n}) \psi_M^{(p)} \frac{\partial}{\partial x} \psi_N^{(p)} + \kappa_{xy} (\mathbf{x}\cdot\mathbf{n}) \psi_M^{(p)} \frac{\partial}{\partial x} \psi_N^{(p)} \Big\} dl \quad (6.18c)$$

$$W_{yy}^{(p)}(M,N) = \int_{L_p^{pec}} \Big\{ - \kappa_{xx} (\mathbf{y}\cdot\mathbf{n}) \psi_M^{(p)} \frac{\partial}{\partial y} \psi_N^{(p)} + \kappa_{xy} (\mathbf{x}\cdot\mathbf{n}) \psi_M^{(p)} \frac{\partial}{\partial y} \psi_N^{(p)} \Big\} dl \quad (6.18d)$$

$$T_{xx}^{(q)}(M,N) = \int_{L_q^{int}} \Big\{ - (\kappa_{yy}^{(+)} - \kappa_{yy}^{(-)}) (\mathbf{x}\cdot\mathbf{n}^{(+)}) \psi_M^{(q)} \frac{\partial}{\partial x} \psi_N^{(q)} + (\kappa_{yx}^{(+)} - \kappa_{yx}^{(-)}) (\mathbf{y}\cdot\mathbf{n}^{(+)}) \psi_M^{(q)} \frac{\partial}{\partial x} \psi_N^{(q)} \Big\} dl \quad (6.19a)$$

$$T_{xy}^{(q)}(M,N) = \int_{L_q^{int}} \Big\{ -(\kappa_{yy}^{(+)} - \kappa_{yy}^{(-)})\,(\mathbf{x}\cdot\mathbf{n}^{(+)})\,\psi_M^{(q)}\frac{\partial}{\partial y}\psi_N^{(q)} + (\kappa_{yx}^{(+)} - \kappa_{yx}^{(-)})\,(\mathbf{y}\cdot\mathbf{n}^{(+)})\,\psi_M^{(q)}\frac{\partial}{\partial y}\psi_N^{(q)} \Big\}\, dl \quad (6.19b)$$

$$T_{yx}^{(q)}(M,N) = \int_{L_q^{int}} \Big\{ -(\kappa_{xx}^{(+)} - \kappa_{xx}^{(-)})\,(\mathbf{y}\cdot\mathbf{n}^{(+)})\,\psi_M^{(q)}\frac{\partial}{\partial x}\psi_N^{(q)} + (\kappa_{xy}^{(+)} - \kappa_{xy}^{(-)})\,(\mathbf{x}\cdot\mathbf{n}^{(+)})\,\psi_M^{(q)}\frac{\partial}{\partial x}\psi_N^{(q)} \Big\}\, dl \quad (6.19c)$$

$$T_{yy}^{(q)}(M,N) = \int_{L_q^{int}} \Big\{ -(\kappa_{xx}^{(+)} - \kappa_{xx}^{(-)})\,(\mathbf{y}\cdot\mathbf{n}^{(+)})\,\psi_M^{(q)}\frac{\partial}{\partial y}\psi_N^{(q)} + (\kappa_{xy}^{(+)} - \kappa_{xy}^{(-)})\,(\mathbf{x}\cdot\mathbf{n}^{(+)})\,\psi_M^{(q)}\frac{\partial}{\partial y}\psi_N^{(q)} \Big\}\, dl \quad (6.19d)$$

6.5 Properties of The Resultant Matrix Equation

In the resultant matrix eigenvalue equation (6.10a), the matrices [A] and [B] may, in general, be complex and non-symmetric. It is easy to see that the global matrices [S], [W], [T] and [B] have the same symmetries as their corresponding local element matrices $[S^{(e)}]$, $[W^{(p)}]$, $[T^{(q)}]$ and $[B^{(e)}]$.

If the waveguide is isotropic and inhomogeneous, then the matrix elements of local element matrices can be simplified as

$$S_{xx}^{(e)}(M,N) = \int_{S_e} \Big\{ \kappa\frac{\partial}{\partial y}\psi_M^{(e)}\frac{\partial}{\partial y}\psi_N^{(e)} + \kappa\frac{\partial}{\partial x}\psi_M^{(e)}\frac{\partial}{\partial x}\psi_N^{(e)} - k_0^2\,\psi_M^{(e)}\psi_N^{(e)} \Big\}\, ds \quad (6.20a)$$

$$S_{xy}^{(e)}(M,N) = \int_{S_e} \Big\{ -\kappa\frac{\partial}{\partial y}\psi_M^{(e)}\frac{\partial}{\partial x}\psi_N^{(e)} + \kappa\frac{\partial}{\partial x}\psi_M^{(e)}\frac{\partial}{\partial y}\psi_N^{(e)} \Big\}\, ds \quad (6.20b)$$

$$S_{yx}^{(e)}(M,N) = \int_{S_e} \Big\{ -\kappa\frac{\partial}{\partial x}\psi_M^{(e)}\frac{\partial}{\partial y}\psi_N^{(e)} + \kappa\frac{\partial}{\partial y}\psi_M^{(e)}\frac{\partial}{\partial x}\psi_N^{(e)} \Big\}\, ds \quad (6.20c)$$

$$S_{yy}^{(e)}(M,N) = \int_{S_e} \Big\{ \kappa\frac{\partial}{\partial x}\psi_M^{(e)}\frac{\partial}{\partial x}\psi_N^{(e)} + \kappa\frac{\partial}{\partial y}\psi_M^{(e)}\frac{\partial}{\partial y}\psi_N^{(e)} - k_0^2\,\psi_M^{(e)}\psi_N^{(e)} \Big\}\, ds \quad (6.20d)$$

$$B_{xx}^{(e)}(M,N) = \int_{S_e} \kappa \, \psi_M^{(e)} \, \psi_N^{(e)} \, ds \tag{6.21a}$$

$$B_{xy}^{(e)}(M,N) = 0 \tag{6.21b}$$

$$B_{yx}^{(e)}(M,N) = 0 \tag{6.21c}$$

$$B_{yy}^{(e)}(M,N) = \int_{S_e} \kappa \, \psi_M^{(e)} \, \psi_N^{(e)} \, ds \tag{6.21d}$$

$$W_{xx}^{(p)}(M,N) = 0 \tag{6.22a}$$

$$W_{xy}^{(p)}(M,N) = 0 \tag{6.22b}$$

$$W_{yx}^{(p)}(M,N) = 0 \tag{6.22c}$$

$$W_{yy}^{(p)}(M,N) = 0 \tag{6.22d}$$

$$T_{xx}^{(q)}(M,N) = \int_{L_q^{int}} \left\{ - (\kappa^{(+)} - \kappa^{(-)}) \, (\mathbf{x}\cdot\mathbf{n}^{(+)}) \, \psi_M^{(q)} \, \frac{\partial}{\partial x} \psi_N^{(q)} \right\} dl \tag{6.23a}$$

$$T_{xy}^{(q)}(M,N) = \int_{L_q^{int}} \left\{ - (\kappa^{(+)} - \kappa^{(-)}) \, (\mathbf{x}\cdot\mathbf{n}^{(+)}) \, \psi_M^{(q)} \, \frac{\partial}{\partial y} \psi_N^{(q)} \right\} dl \tag{6.23b}$$

$$T_{yx}^{(q)}(M,N) = \int_{L_q^{int}} \left\{ - (\kappa^{(+)} - \kappa^{(-)}) \, (\mathbf{y}\cdot\mathbf{n}^{(+)}) \, \psi_M^{(q)} \, \frac{\partial}{\partial x} \psi_N^{(q)} \right\} dl \tag{6.23c}$$

$$T_{yy}^{(q)}(M,N) = \int_{L_q^{int}} \left\{ - (\kappa^{(+)} - \kappa^{(-)}) \, (\mathbf{y}\cdot\mathbf{n}^{(+)}) \, \psi_M^{(q)} \, \frac{\partial}{\partial y} \psi_N^{(q)} \right\} dl \tag{6.23d}$$

It can be easily proven from (6.20) to (6.23) that for lossless isotropic inhomogeneous waveguide, the resultant matrices [A] and [B] are real, [B] is symmetric and positive definite, but [A] is asymmetric. The asymmetry of [A] evidence the possible presence of complex modes although the dielectric is lossless. It is interesting to note from (6.20), (6.22), and (6.23) that the source of this asymmetry and then, of the existence of complex modes, resides in the line integral part [T] (6.23) (because [A] = [S] + [W] + [T], [S] is symmetric and [W] is zero here). If we did not take into account the line integral term [T] then [A] would also be

symmetric in such a manner that the possible existence of complex modes could not be included. We can see that it is the existence of inhomogeneity in the dielectric which makes complex modes possible although it does not guarantee their presence.

For isotropic homogeneous lossless waveguide, both [W] and [T] are zero matrices so that [A] and [B] are real symmetric. Therefore, no complex mode can exist.

In section 6.4, we have obtained a matrix eigenvalue equation of the canonical form:

$$[A]\ \{x\} = \lambda\ [B]\ \{x\} \tag{6.25}$$

where the eigenvalue λ is the square of propagation constant (real or complex). The matrices [A] and [B] are in general large and sparse. They are also non-symmetric (non-hermitian) in the presence of dielectric inhomogeneity and/or anisotropy. Furthermore, the matrix elements are complex (and so are the eigenvalues) in case of lossy dielectrics. For the lossless case, both matrices in (1) are real but the eigenvalues and eigenvectors can still possibly be complex conjugate (as in the case of complex modes in lossless guides).

6.6 Choice of Elements

6.6.1 Bilinear Quadrilateral Element

Quadrilateral is a particularly interesting element shape for our formulations. Quadrilateral elements provide for flexibility in geometric modelling that is comparable with triangles, and are used by many analysts in preference to triangles [115]. The bilinear shape functions in quadrilateral elements can improve the accuracy of the line integral terms in our formulation. Moreover, the use of quadrilateral elements, which are actually linear isoparametric elements, makes it easier to extend to (quadratic) isoparametric elements which are more suitable to follow arbitrary curves.

The domain of a straight-edged quadrilateral elements is defined by the locations of its four nodal points (x_a^e, y_a^e), $a = 1, \ldots, 4$ in the E^2-plane. We assume the nodal points are labeled in ascending order

corresponding to the counterclockwise direction (see Fig. 6.1). We seek a change of coordinates which maps the given quadrilateral into the biunit square, as depicted in Fig. 6.1. The biunit square is sometimes called the *parent domain* [115]. The coordinates of a point

$$\mathbf{u} = \begin{Bmatrix} u \\ v \end{Bmatrix} \tag{6.25}$$

in the biunit square are to be related to the coordinates of a point

$$\mathbf{x} = \begin{Bmatrix} x \\ y \end{Bmatrix} \tag{6.26}$$

in S_e by mappings of the form

$$x(u,v) = \sum_{a=1}^{4} \alpha_i(u,v)\, x_i^e \tag{6.27a}$$

$$y(u,v) = \sum_{a=1}^{4} \alpha_i(u,v)\, y_i^e \tag{6.27b}$$

The bilinear functions $\alpha_i(u,v)$ are expressed as

$$\alpha_i(u,v) = \frac{1}{4}(1 + u_i u)\,(1 + v_i v), \tag{6.28}$$

where u_i and v_i are defined in Fig. 6.1.

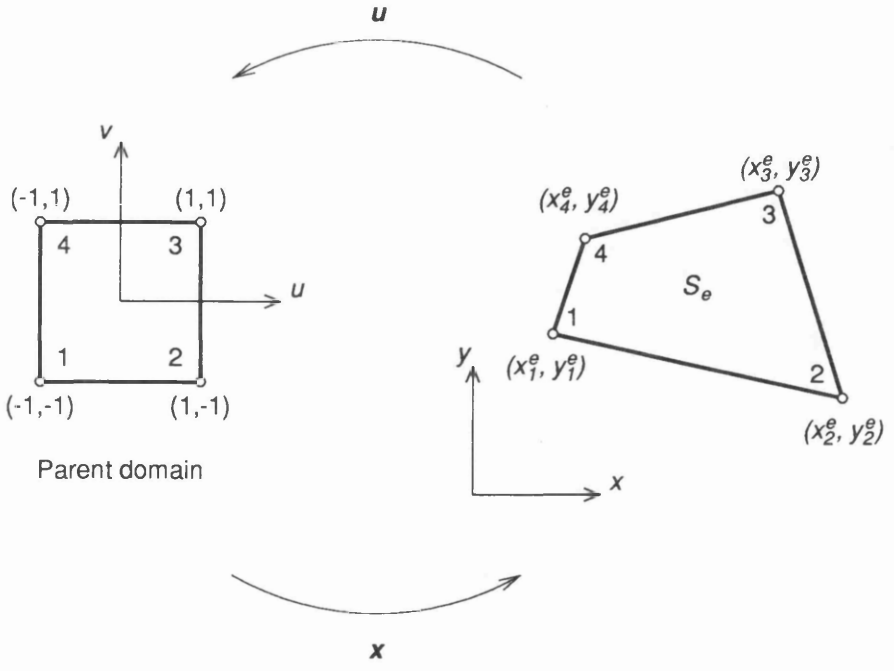


Fig. 6.1 Bilinear quadrilateral element domain and local numbering

The mapping (6.27) is bilinear, i.e., linear in u and v taken individually; straight lines in the u-v plane therefore map into straight lines in the x-y plane. The coordinate transformation (6.27) is stable as long as the determinant of its Jacobian matrix $[J] = \partial(x,y)/\partial(u,v)$ is greater than zero; This condition is equivalent to that all interior angles formed by two adjacent edges are less than 180° [116]; this is also the only condition to guarantee the smoothness of the trial functions on S_e (Note that N_a is always a smooth function of u and v).

The Jacobian matrix [J] of a coordinate transformation $\mathcal{T}$ expresses the local geometric properties of $\mathcal{T}$. Its magnitude det([J]) denotes the local area magnification, while its individual components show the relative twisting and stretching in the different coordinate directions.

It can be shown that the Jacobian matrix [J] reduces to constant ones if opposite sides of the quadrilateral in the x-y plane are of the same length [9]. This simplification occurs for parallelograms, geometric figures far more flexible than squares or rectangles.

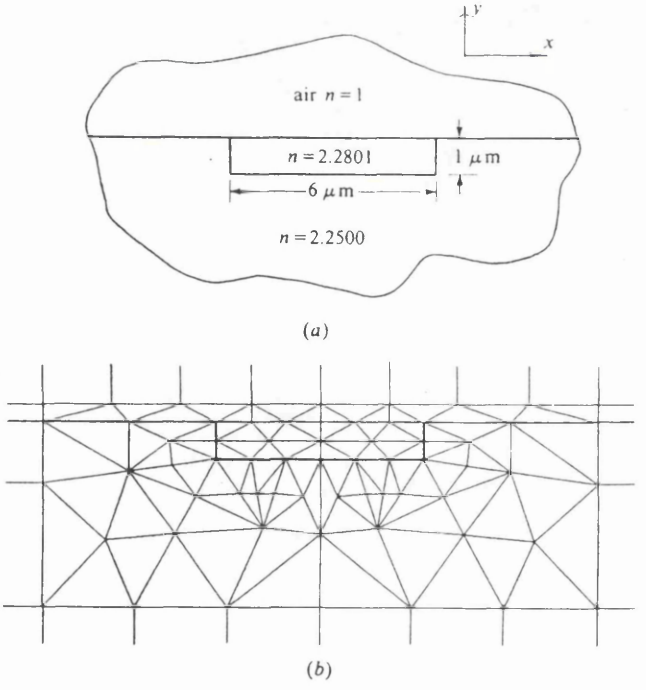


Fig. 6.2 An integrated optical channel waveguide.
(a) Basic geometry.
(b) A typical finite element mesh, showing orthodox elements bordered by infinite elements.

6.6.2 *Infinite Elements*

As we mentioned in Chapter 2, one difficulty concerning dielectric waveguides, particularly optical waveguides, is that they have open boundaries. Open-boundary problems may be solved without simple truncation by several techniques, all having in common the idea that the open region is subdivided into *interior* and *exterior* portions so that the interior part contains the structures and fields of principal interest. In this study, we choose to adopt the simple *infinite elements* [9], [63], [70]. The open regions of the problem space are bordered with elements extending to infinity, such as depicted in Fig. 6.2.

Every such border edge of a element is chosen parallel to one or other of the coordinate axes. Suppose this edge is parallel to the y-axis, so that the infinite element will be concerned with a finite range of y, say from y_1 to y_2 and an infinite range of x, say from x_1 to $+\infty$. A one dimensional trial function $U(x)$ is constructed for each of the relevant H-components so that it spans the nodes $y = y_1$ and $y = y_2$ (and any intermediate nodes, if high-order trial functions are being employed). Now within the infinite element an overall trial function

$$U(y)exp[-(x - x_1)A] \tag{6.29}$$

is used, where A, namely *decay factor*, is an additional undetermined parameter. The integrations necessary to establish the contribution of the infinite element will extend to $x = +\infty$, but of course with the presence of the exponential factors such infinite integrals present no particular difficulty. Each y-border infinite element associated with $x \longrightarrow \infty$ is given a similar treatment with the same decay factor to preserve continuity in the y-direction. Cases where the infinite problem region extends to $x = -\infty$ are dealt with merely by changing the sign in the exponential part of the trial function. Infinite elements corresponding to a x-border are also similarly dealt with employing

$$U(x)exp[-(y - y_1)B] \tag{6.30}$$

as the trial function while *corner* infinite elements are constructed with an appropriate variation

$$exp[-(x - x_1)A]\ exp[-(y - y_1)B] \tag{6.31}$$

The practice is to leave A and B undetermined so as to be optimized by the variational extremization process. The exponential form of the variation chosen to be 'built in' may be justified on the grounds that in simple cases where an analytic solution can be found for open dielectric waveguides, the fields do indeed die away in this fashion. An alternative way may be to use predetermined but plausible values of these parameters rather than go to the factors required to obtain their variational estimates.

6.6.3 Remarks

a) Shape of elements

It is often stated that triangular elements are responsible for the geometric flexibility of the two-dimensional finite element method. This is perhaps somewhat of an exaggeration as triangular shapes are not needed in practice. Most regions are conveniently discretized by arbitrary quadrilateral elements. It is also often stated that triangles enable modelling of particularly intricate geometries and that these shapes facilitate transition from coarsely meshed zones of a grid to finely meshed zones. This is, of course, true, but quadrilaterals are capable of doing the same thing at least to some degree. To illustrate this point, consider Fig. 6.3 in which a triangular zone is discretized into three quadrilaterals. Thus we see any triangular element could be replaced by quadrilaterals. Mesh generation for triangular regions via quadrilaterals may be handled similarly, see Fig. 6.4. Quadrilaterals may also be used to perform mesh transition as illustrated in Fig. 6.5.

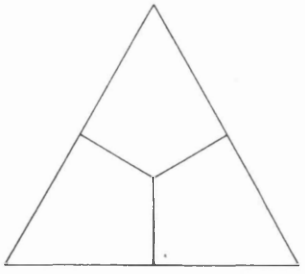


Fig. 6.3

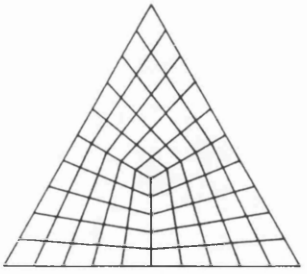


Fig. 6.4

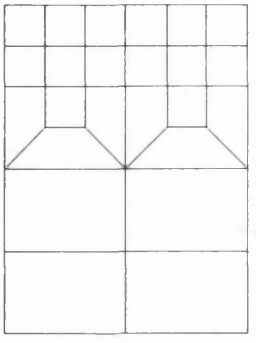


Fig. 6.5

b) Order of elements

The first order element with linear and bilinear polynomials are the simplest finite element forms. The motivation for introduction of higher order trial functions is obviously one of achieving a better approximation to the solution of the problem at hand.

Refinement of solution can be achieved either by successive creation of finer meshes of elements with the use of same trial functions or by introduction of successively higher order functions with a constant mesh subdivision.

From the practical point of view, obviously, the best choice will be the one achieving highest accuracy at least computational expense. It has been found that, sometimes, higher order approximations appear to be more cost effective (although the actual optimum is very problem dependent) [116]. However, for simplicity, we will always stick to using the lowest possible order trial functions.

6.7 Non-symmetric Sparse Matrix Solver

The single most time-consuming part of finite element computer programs is usually the solution of the resultant matrix equation. This is particularly evident in those cases needing the solution of eigenvalue problems. For large, sparse, real and symmetric (or complex hermitian) eigenvalue equations there are efficient commercial software available in such standard computer libraries as HARWELL [124], NAG [125], and IMSL [126] or special software developed within this group [62], [77]. However, for large, sparse, non-hermitian (real or complex) eigenvalue problems, there is no efficient commercial software available [86]-[88], [124]-[126].

Without an efficient sparse solver, one has to resort to the dense matrix algorithm QR (for real), or QZ (for real or complex) which are the only two available for complex non-hermitian problems [86]-[88], [124]-[126]. Using the dense solver, one can only treat a very limited size of problems at a great expense even with supercomputers [86]-[88] [133]-[135]. Roughly speaking, the cpu time and memory requirement of dense solvers are proportional to N^3 and N^2 respectively, where N denotes matrix order. As shown in [85], [86], it requires 27 MB memory and about

40 second cpu time for an complex problem of 508 unknowns (153 nodes) on a HITACHI S-810/10 supercomputer [85], and it requires about 30 seconds for a real problem of 578 unknowns on CRAY X-MP/48 supercomputer. It can be estimated that it is hardly possible to treat problems of more than 1000 unknowns on a top model supercomputer. More statistics are shown in section 7.6 in Chapter 7.

In order to take full advantage of the sparsity of matrix equation (6.5) and to solve it efficiently, a sparse matrix solver has been especially developed for our problem (6.10) or (6.25). The solver has been studied, programmed and tested by a colleague, at UCL, Zhu. We, of course, initiated the project to develop the solver and collaborated in tests and applications of the solver.

The general eigenvalue problem (6.25) is here solved using a subspace iteration algorithm [119]-[122] applied to non-symmetric matrices. We start with two sets of p right and left initial vectors $[X^{(0)}]$ and $[Y^{(0)}]$ of a length n ($p << n$, the order of matrices [A] and [B]), and an appropriately selected shift η, which may be complex. Two sets of trial vectors are simultaneously iterated using:

$$([A] - \eta\,[B])\,[X^{(s+1)}] = [B]\,[X^{(s)}] \tag{6.32a}$$

$$[Y^{(s+1)}]^T\,([A] - \eta\,[B]) = [Y^{(s)}]^T\,[B] \tag{6.32b}$$

where $[X^{(s)}]$ and $[Y^{(s)}]$ are $n \times p$ matrices presenting p vectors of length n. With adequate normalization of the iteration vectors after each iteration, a reduction of the order of the problem is performed using the transformations:

$$[A^{(s+1)}] = [Y^{(s+1)}]^T\,([A] - \eta\,[B])\,[X^{(s+1)}] \tag{6.33a}$$

$$[B^{(s+1)}] = [Y^{(s+1)}]^T\,[B]\,[X^{(s+1)}] \tag{6.33b}$$

Solution is completed by solving the (now dense) eigenvalue problems of much smaller order p:

$$[A^{(s+1)}]\,[U^{(s+1)}] = [B^{(s+1)}]\,[U^{(s+1)}]\,[\chi^{(s+1)}] \tag{6.34a}$$

$$[V^{(s+1)}]^T\ [A^{(s+1)}] = [\chi^{(s+1)}]\ [V^{(s+1)}]^T\ [B^{(s+1)}] \tag{6.34b}$$

and recovering the right and left eigenvectors:

$$[X] = [X^{(s+1)}]\ [U^{(s+1)}] \tag{6.35a}$$

$$[Y] = [Y^{(s+1)}]\ [V^{(s+1)}] \tag{6.35b}$$

These are now available to restart the iterations if necessary. Convergence can be tested by comparing the normalized difference with a given tolerance TOL:

$$\left| \frac{\lambda_i^{(s+1)} - \lambda_i^{(s)}}{\lambda_i^{(s+1)}} \right| \leq \text{TOL} \tag{6.36}$$

The current estimate to the ith eigenvalue ($i \leq p$) λ_i^{s+1} is given by

$$\lambda_i^{(s+1)} = \chi_i^{(s+1)} + \eta \tag{6.37}$$

where $\chi_i^{(s+1)}$ is the ith element of the diagonal matrix $[\chi^{(s+1)}]$ of (6.34).

Setting $p = 1$ reduces this procedure to inverse iteration and in this case successive updating of shift η can be used to accelerate convergence.

The crucial step regarding time consumption in the above procedure is the solution of the linear systems (6.32a) and (6.32b). We have implemented these using the package *ME28* for complex matrices from the Harwell Library of Subroutines [123]-[124]. This results in a very efficient solution, taking full advantage of the sparsity of the matrices and allowing the use of the same L-U decomposition of the matrices for both systems and repeatedly through the iterations.

A more compact form of this procedure can (and has been) achieved using only a single set of eigenvectors (*e.g.* the right eigenvectors). This is still valid but convergence has been found to require a few more (although faster) iterations. In this case [Y] = [X] is used in (3a). The procedure now consists only of (6.32a), (6.33a), (6.34a) and (6.35a). Our choice, though, has been for the use of both eigenvectors.

CHAPTER 7

COMPUTATIONAL RESULTS

7.1 Introduction

In the preceding chapters 5-6 we have derived and implemented the finite element variational formulation, and obtained the matrix eigenvalue equation which is readily for computer language coding.

In this chapter, we will demonstrate the strength of the new method presented in chapters 5-6 by a series of examples covering all four categories of inhomogeneous waveguides — isotropic lossless, anisotropic lossless, isotropic lossy, and anisotropic lossy waveguides. Considering optical waveguides, several open-bounded waveguiding structures are included with the use of infinite elements.

The finite element formulation has been coded in FORTRAN 77 language and the software has been tested on a variety of computers from workstations to supercomputers, and the results show good consistency. For some examples, both the dense solver F02BJF from NAG library and the sparse solver SGECS developed at UCL during the later stage of the study are used. Statistics of the sparse matrix eigenequation solver are also presented.

Depending on situation, the problem size may be referred to by one of the four parameters:

a) N_p : number of nodal points;
b) N_e : number of elements;
c) N_m : matrix order, where $N_m = 2N_p$;
d) N_u : number of unknowns, where N_u equals to N_m minus known boundary values.

When referring to questions of computing time or convergence, choosing N_u rather than N_p is more sensible, because computing cost depends explicitly on N_u whereas it may depend on $2N_p$, $3N_p$, $4N_p$, or $6N_p$ for different methods. When referring to computing memory requirement, the matrix order N_m is suitable.

7.2 Description of the FORTRAN Program

We begin by briefly looking at some computational aspects of this study. A software package, with the name FEMDI, has been developed based on the afore presented derivations. The FEMDI system mainly includes five parts — controller, preprocessor, matrix eigenvalue equation generator, matrix eigenvalue equation solver, and postprocessor. The FEMDI is coded in the standard FORTRAN 77 language, and its basic structure is shown in Fig. 7.1.

The controller, namely CASP1, controls

(a) input options: whether to get a finite element mesh data from a preprocessed data file or from the mesh generator subroutine directly;
(b) loop parameters when dispersion characteristics required;
(c) the order of subspace, tolerance of solutions for the sparse matrix solver;
(d) eigenvalue and eigenvector shift for the sparse matrix solver;
(e) output options: the number of modes for which their field distributions are required to be output for further postprocessing.

The preprocessor, namely MESHGE, is a mainly a mesh generator which produces a set of data including waveguide geometric structure, refractive index profile, and boundary-value conditions. The mesh generator itself could be an independent program just to produce the mesh data and output the data into a file, and the data file may be used later as the input file of FEMDI.

The matrix eigenvalue equation generator, namely CASP2, is the most important part of the program. Its purpose is to translate a physical dielectric waveguide problem from its finite element mesh data form into a digital-matrix form, which will be solved later as a general eigenvalue equation.

The matrix eigenvalue equation solver, namely SGECS, efficiently produces both eigenvalue and eigenvector solutions for the large, sparse, non-symmetric, complex matrix eigenvalue equation. The SGECS is specially developed for FEMDI within this group. The algorithm of the sparse matrix

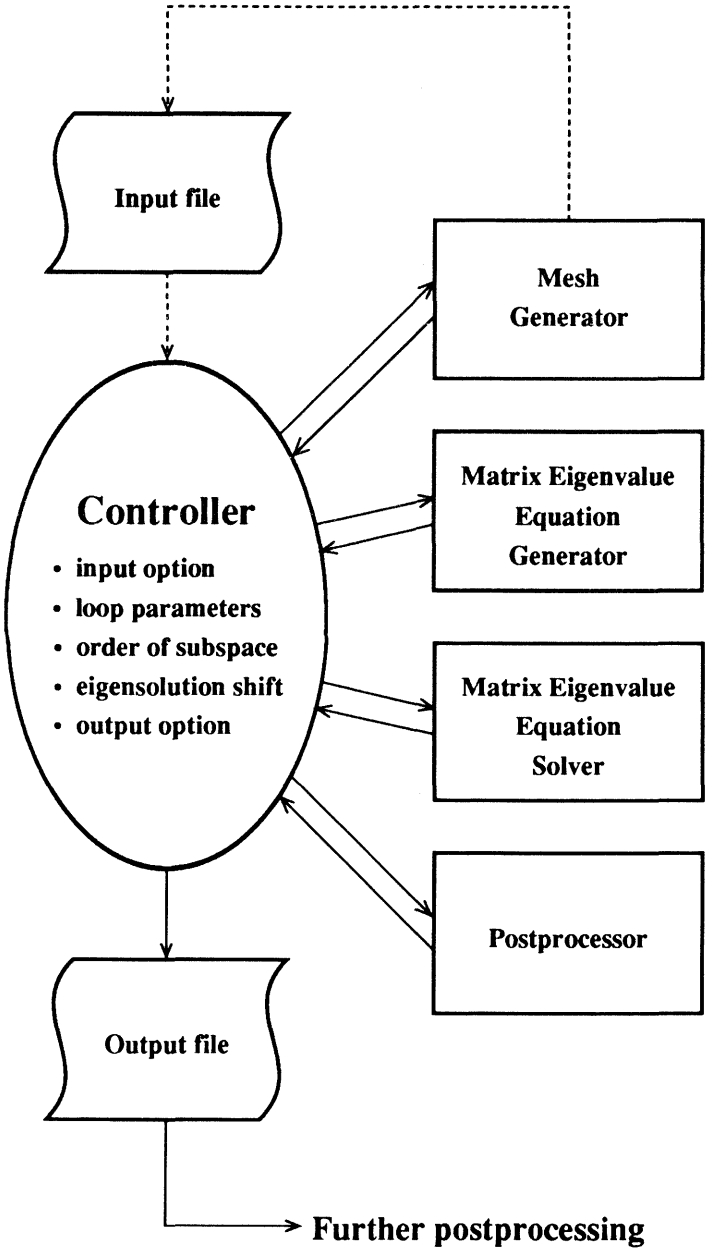


Fig. 7.1 Basic structure of the FEMDI software

solver is introduced in [136] and also described in section 6.7.

The postprocessor, with the name VALVEC, use eigenvalue solutions to workout

(a) propagation constants and normalized propagation constants;
(b) two-dimensional field distribution by filling in the missed known boundary values into the eigenvector solutions;
(c) one-dimensional field distribution on a given straight line within the waveguide cross section.

From the output data of FEMDI we can do further postprocessing such as 3-D, 2-D (contours), and 1-D plotting of field distribution as well as dispersion curve drawing via the use of standard library software or PC software.

As the requirement of this study, the main purpose of developing FEMDI is to provide numerical examples to verify the validity of the new variational finite element formulation. The computers used for this study are those at University College Computer Centre (UCCC) and University of London Computer Centre (ULCC) (see Appendix B), the access to these computers is through the local area network. Because there is not a fixed computing environment, we do not intend to develop a very friendly software at this stage. However, FEMDI has got the frame work of professional software, and so it will not be very difficult to develop it into a friendly package if a computer system is decided.

The FEMDI has been run on a variety of computers from workstations to supercomputers. Appendix B lists the computers (with such basic specifications as quoted speed, memory capacity, and operating system, etc.) used in the numerical computation for this study

7.3 Isotropic Lossless Waveguides

7.3.1 Dielectric-Slab-Loaded Metallic Rectangular Waveguide

As the first example, we consider a dielectric-slab-loaded metallic rectangular waveguide (inset of Fig. 7.1) which is one of few types of inhomogeneous waveguide structures for which analytical solutions exist. This kind of waveguide has been used as a basic test for new methods [81],

[84], [87].

Fig. 7.2 shows the dispersion curves of the lowest five modes in an dielectric-slab-loaded metallic rectangular waveguide. The results were obtained taking advantage of symmetry with a mesh of 18×6 rectangular elements (133 nodes) over half of the cross section of the guide. Excellent agreement with the analytical results can be observed despite the relatively coarse mesh used.

Fig. 7.3 shows the relative error, $|e|$, in the finite element solutions of propagation constant for the fundamental LSE_{10} and the first higher LSM_{11} modes in the waveguide as a function of the number of unknowns, N_u, at $k_0 b = 3$. Ten meshes, $2L \times L$ ($L = 2, 3, ..., 11$), of first order square elements are chosen in the computation.

The relative error e is defined by

$$e = (\beta - \bar{\beta})/\bar{\beta} \tag{7.1}$$

where β and $\bar{\beta}$ are the computed and exact values [15], respectively.

As expected in a variational approach, the relative error e monotonically decreases with the increase of matrix order and so with the number of elements and the number of nodes. It is also found that the finite element solution is a lower bound to the true solution.

We have confirmed not only from the eigenvalues but also from eigenvectors that spurious solutions do not appear. Fig. 7.3 shows the profiles of H_x and H_y components for the lowest 10 modes in a dielectric-slab-loaded waveguide. The profiles are obtained directly from the eigenvectors of the finite element solutions using in the case NAG subroutine F02BJF. The field profiles match very well with the exact solutions. The calculated H_y components of LSE_{10}, LSE_{20}, LSE_{30} modes and the H_x components of LSM_{11}, LSM_{21}, LSM_{31} modes, which are exactly zero in the analytical solutions, are at least 10 orders of magnitude smaller than their corresponding dominant components; well below the tolerance of any practical application. Besides, with an adjustable tolerance parameter which is available in the sparse solver, these errors can be reduced to whatever one wants subject to the limitation of computer resources.

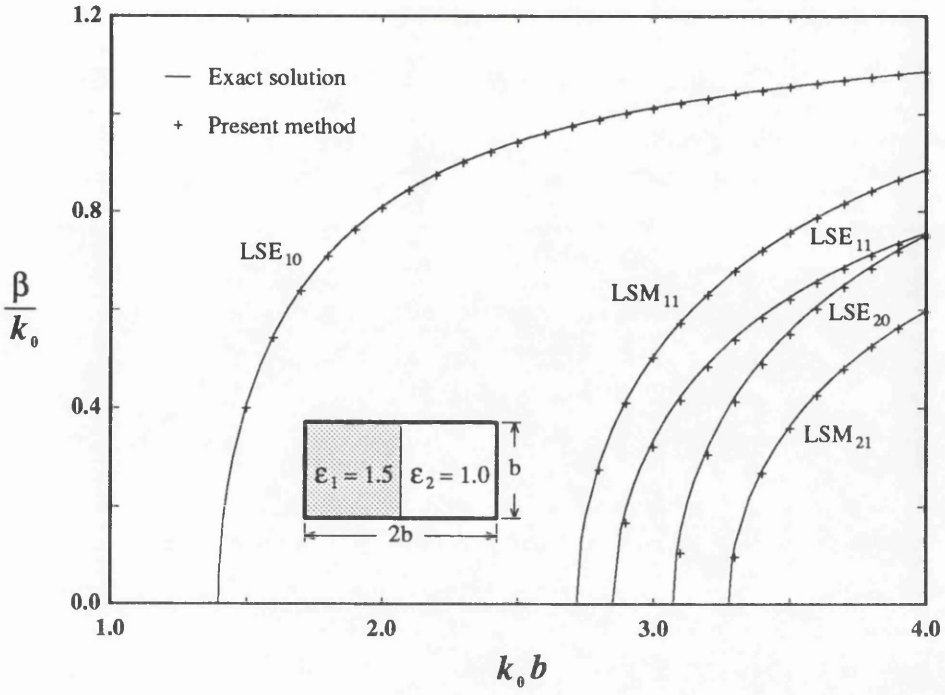


Fig. 7.2 Dispersion characteristics of the lowest five modes in an dielectric-slab-loaded rectangular waveguide (inset).

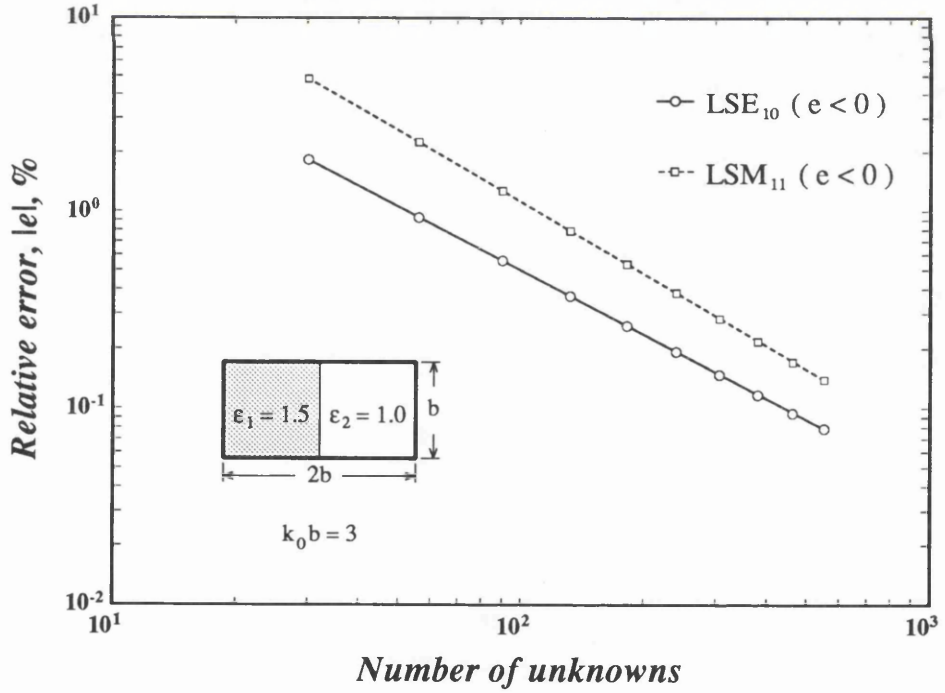
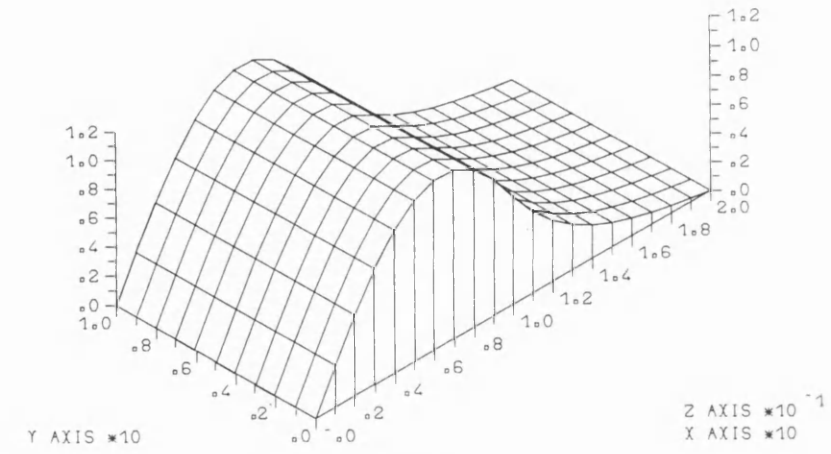
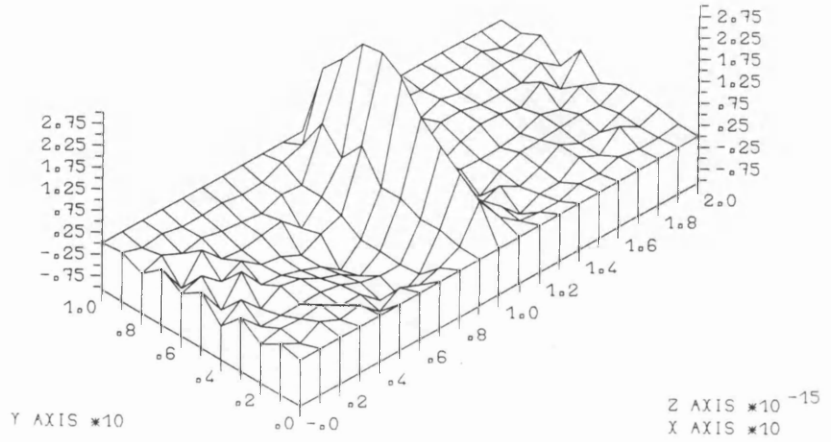


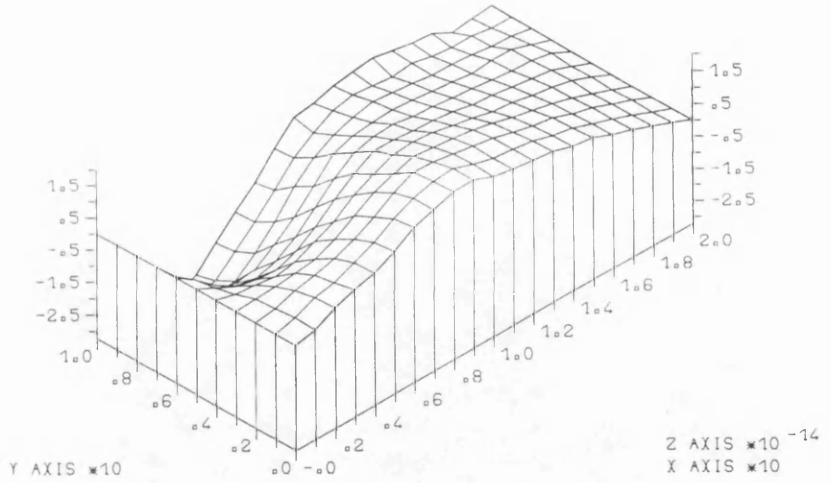
Fig. 7.3 Convergence of the finite element solutions ($k_0 b$ = 3).



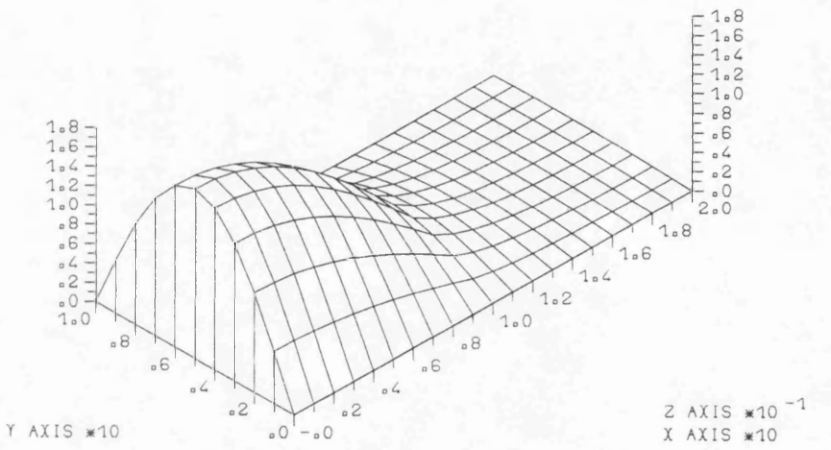
(a) H_x of LSE_{10}



(b) H_y of LSE_{10}

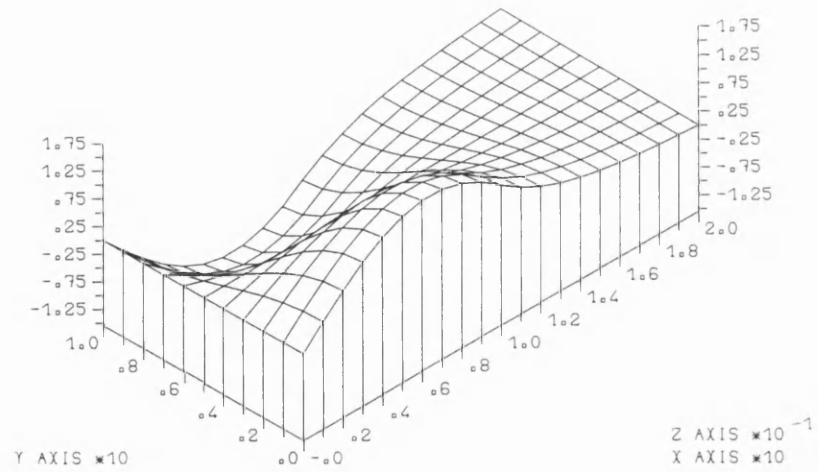


(c) H_x of LSM_{11}

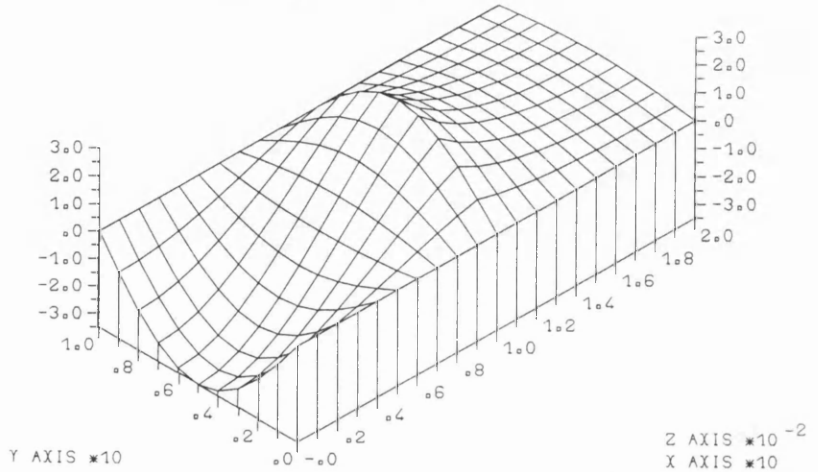


(d) H_y of LSM_{11}

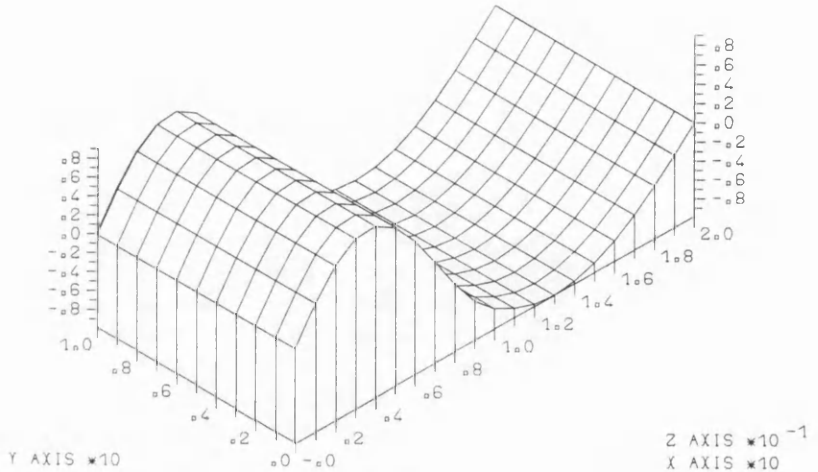
Fig. 7.4(a)-(d) Field profiles of LSE_{10} mode and LSM_{11} mode



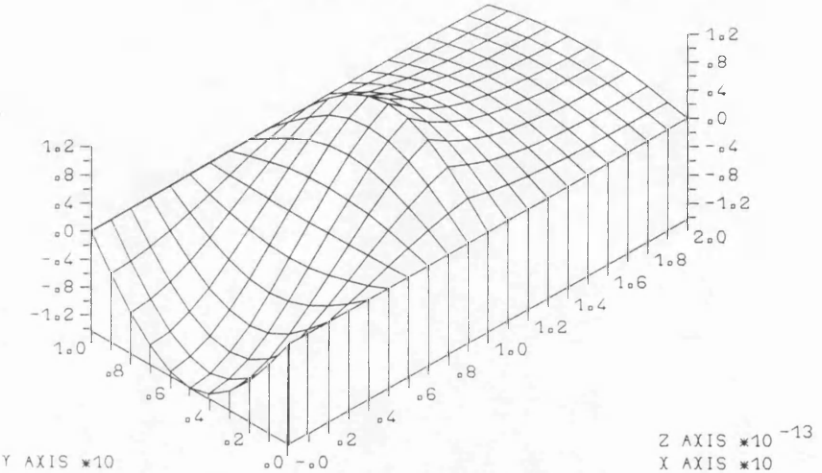
(e) H_x of LSE_{11}



(f) H_y of LSE_{11}

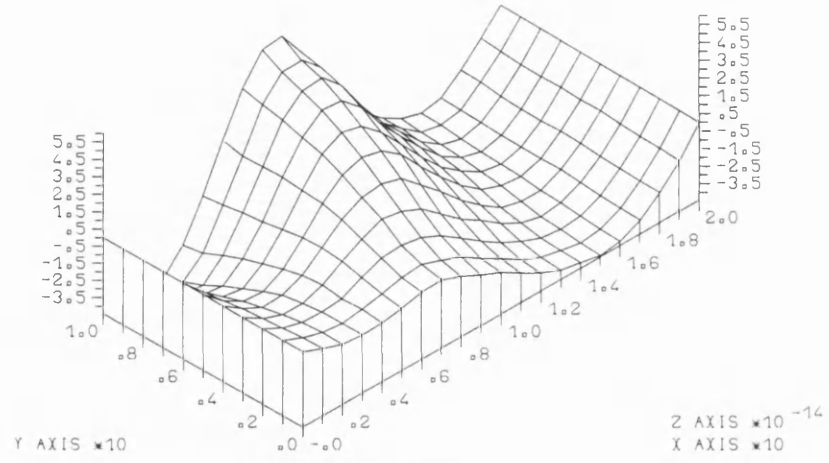


(g) H_x of LSE_{20}

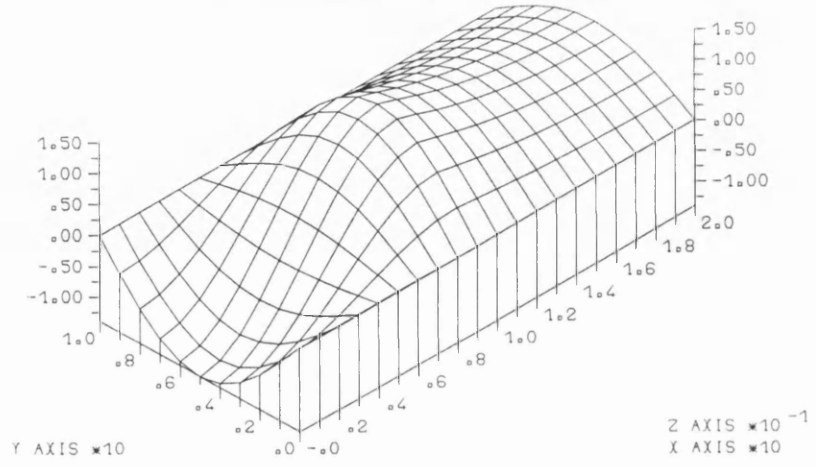


(h) H_y of LSE_{20}

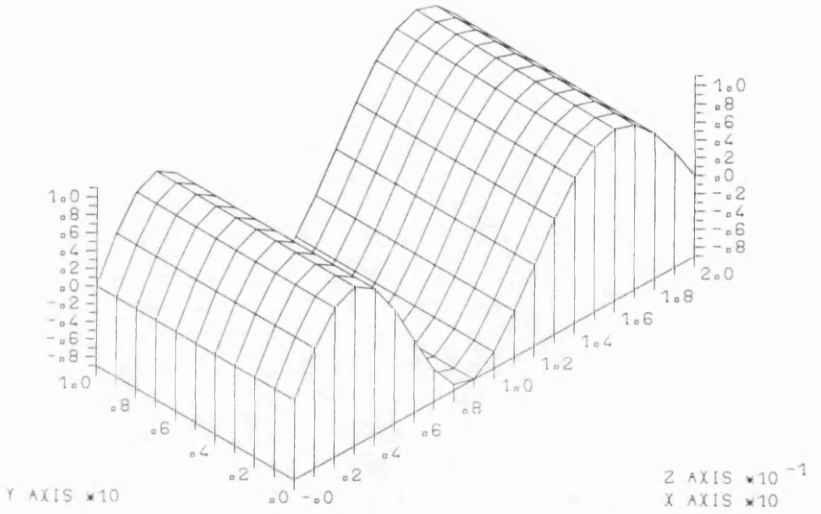
Fig. 7.4(e)-(h) Field profiles of LSE_{11} mode and LSE_{20} mode



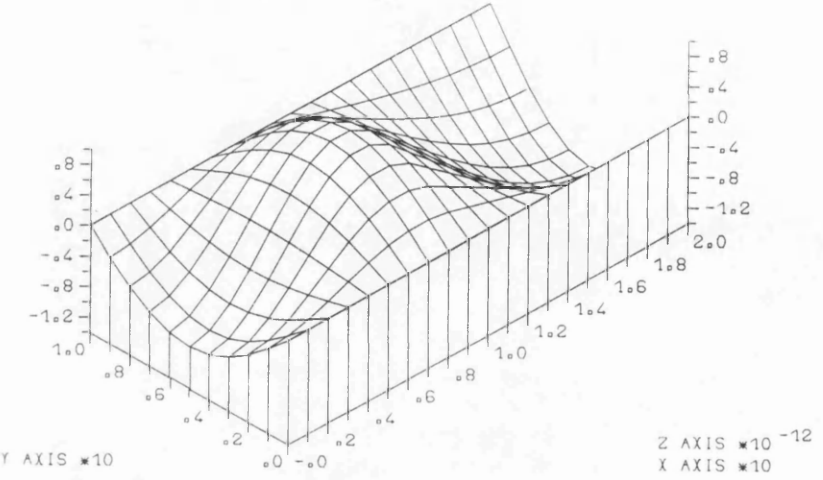
(i) H_x of LSM_{21}



(j) H_y of LSM_{21}

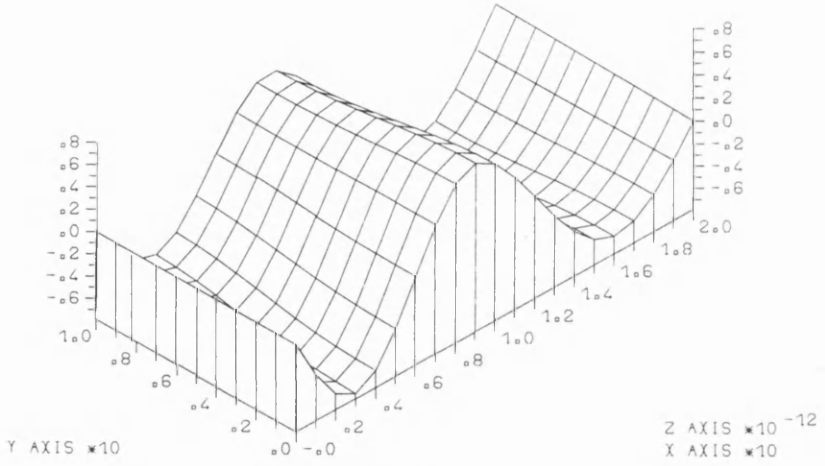


(k) H_x of LSE_{30}

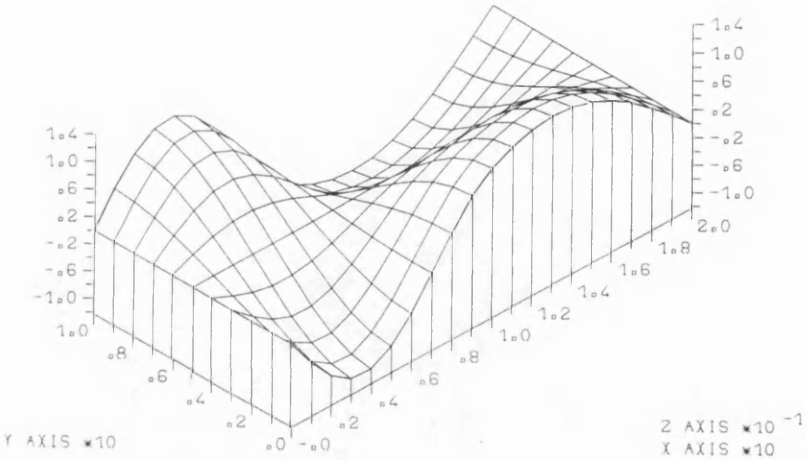


(l) H_y of LSE_{30}

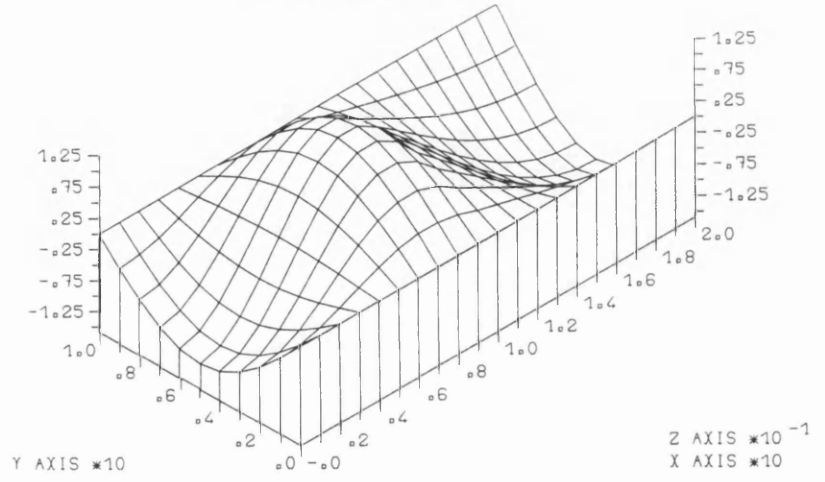
Fig. 7.4(i)-(l) Field profiles of LSM_{21} mode and LSE_{30} mode



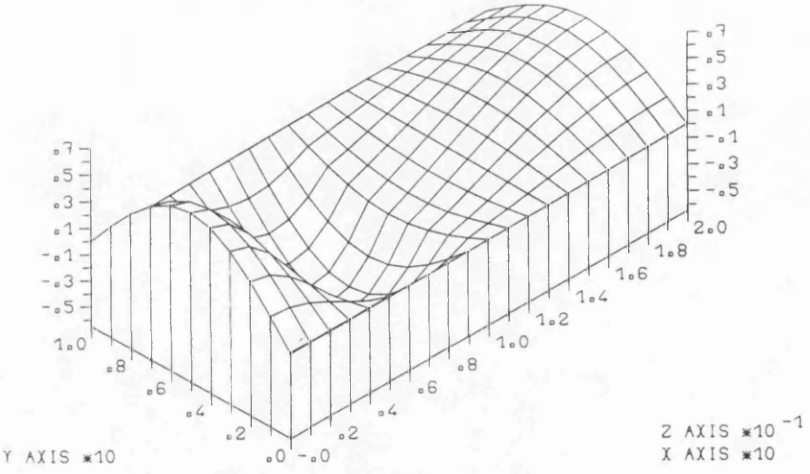
(m) H_x of LSM_{31}



(o) H_x of LSE_{21}

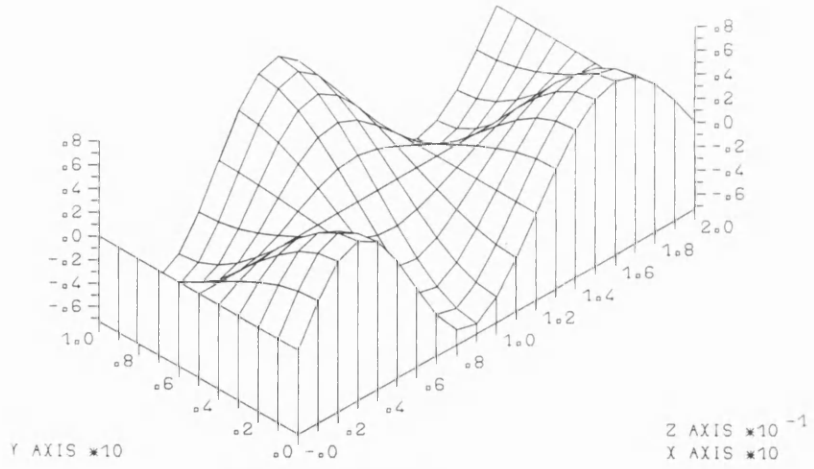


(n) H_y of LSM_{31}

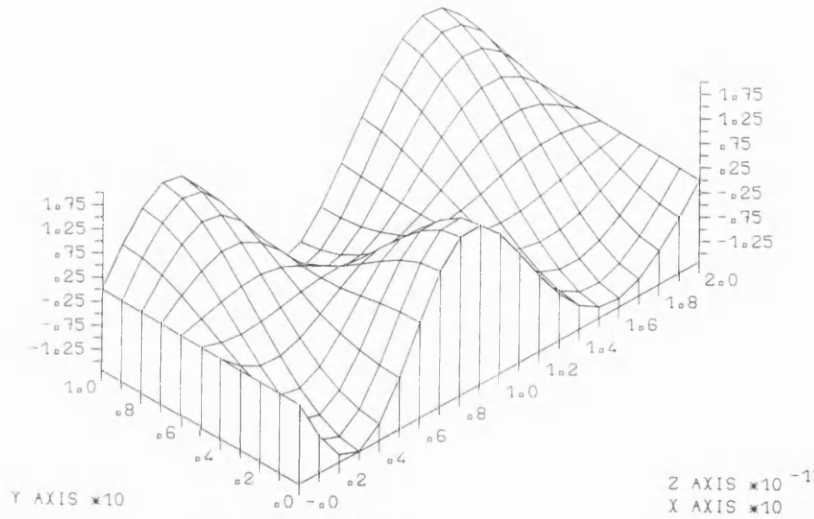


(p) H_y of LSE_{21}

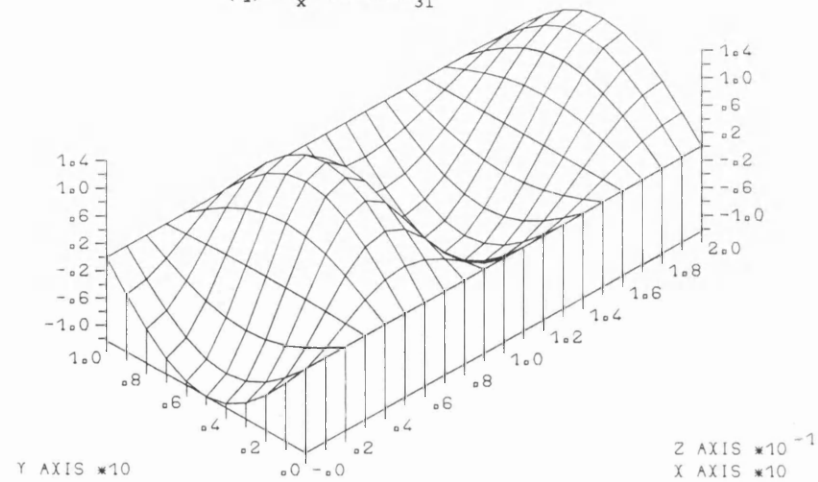
Fig. 7.4(m)-(p) Field profiles of LSM_{31} mode and LSE_{21} mode



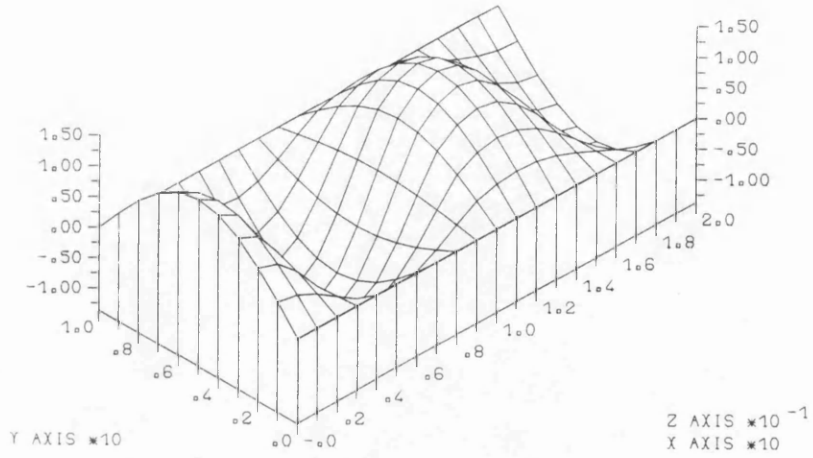
(q) H_x of LSE_{31}



(s) H_x of LSM_{41}



(r) H_y of LSE_{31}



(t) H_y of LSM_{41}

Fig. 7.4(q)-(t) Field profiles of LSE_{31} mode and LSM_{41} mode

7.3.2 *Complex Modes in Shielded Image Waveguides*

Complex waves are modes existing in lossless, inhomogeneous waveguides having complex propagation constants [19]. They always occur in pairs with the propagation constant of one being the complex conjugate of the other. Although they carry no real power, the influence of their presence has been recognized in the analysis of waveguide discontinuities. It has been observed that complex modes have to be included in the field expansion used in mode matching procedures for analysis of waveguide discontinuities [18], [127], and that their omission may lead to serious errors.

Fig. 7.5 shows dispersion characteristics of the lowest six modes in a lossless dielectric image waveguide ($\varepsilon = 9$), shielded with a conventional rectangular *Ku*-band housing (15.799×7.899 mm^2, 12.4-18 GHz). For simplicity, the normalized phase and attenuation constant are plotted in the same diagram in the opposite direction. Dispersion curves of normalized propagation constants for the frequency range 12-18 GHz

show close agreement with those presented by Strube and Arndt [18]. The dotted lines indicate complex waves with $\gamma_{cw} = \pm \alpha_{cw} \pm j\beta_{cw}$.

Taking advantage of symmetry, half of the cross section of the guide is divided into a mesh of 31×27 = 837 non-uniform rectangular elements (896 nodes). Both matrices are in this case real, the numbers of unknowns are 1674 for PEC and 1676 for PMC symmetry walls respectively.

Fig. 7.6 shows the effect of varying the permittivity of the dielectric insert on the normalized propagation constant $\gamma/k_0 = (\alpha/k_0, j\beta/k_0)$ at frequency 14 GHz. In Fig. 7.6, $H_{mn\square}$ indicate H_{mn} (TE_{mn}) modes in homogeneous metallic rectangular waveguide. One can see that even for low permittivity, a complex mode exists at this frequency. It is also interesting to note that as the permittivity increases, complex modes appear intermittently.

For curiosity, Figs. 7.7 and 7.8 show, for the first time, the profiles and contours of the H_x and H_y components of the complex mode in Fig. 7.5 at f = 14 GHz, respectively.

The examples also show the completeness of the solutions which do not miss complex modes — an essential part of guided wave spectrum. To our knowledge, these are the first finite element solutions of complex modes.

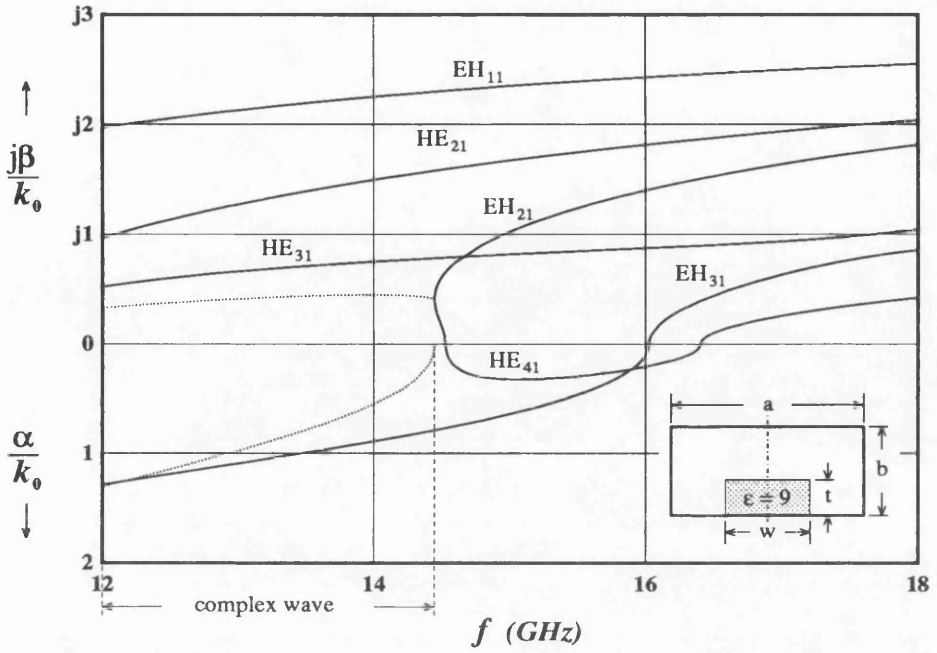


Fig. 7.5 Propagation constant $\gamma/k_0 = (\alpha/k_0, \beta/k_0)$ versus frequency f of a shielded image waveguide in a conventional *Ku*-band housing with $\varepsilon = 9$, a=15.799 mm, b=7.899 mm, w=3.45 mm, t=3.2 mm.

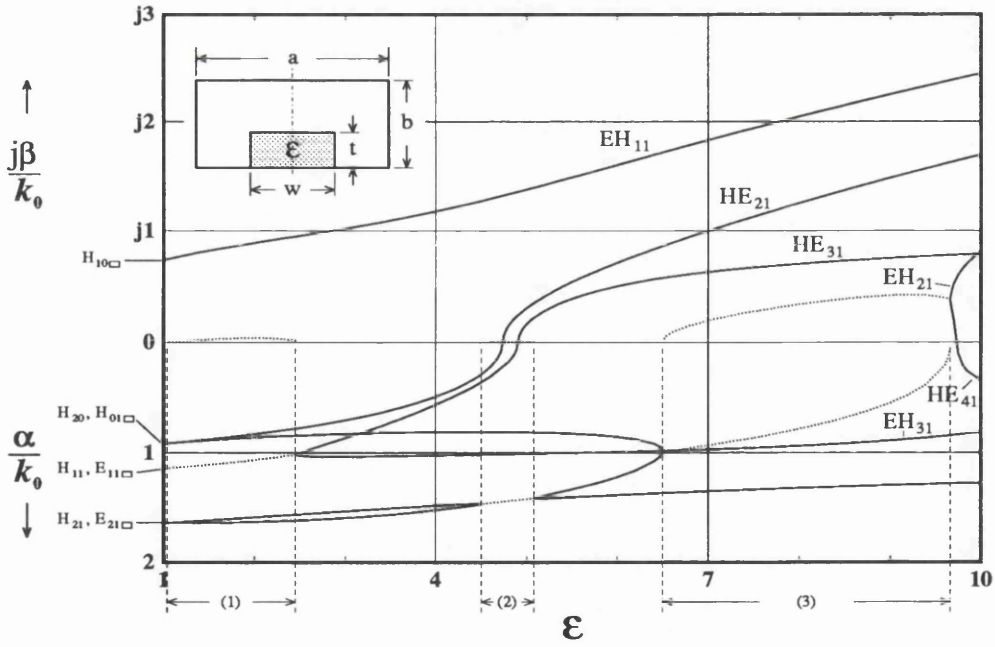


Fig. 7.6 Propagation constant $\gamma/k_0 = (\alpha/k_0, \beta/k_0)$ versus relative permittivity ε of a shield image waveguide with f = 14 GHz, a = 15.799 mm, b = 7.899 mm, w = 3.45 mm, t = 3.2 mm.

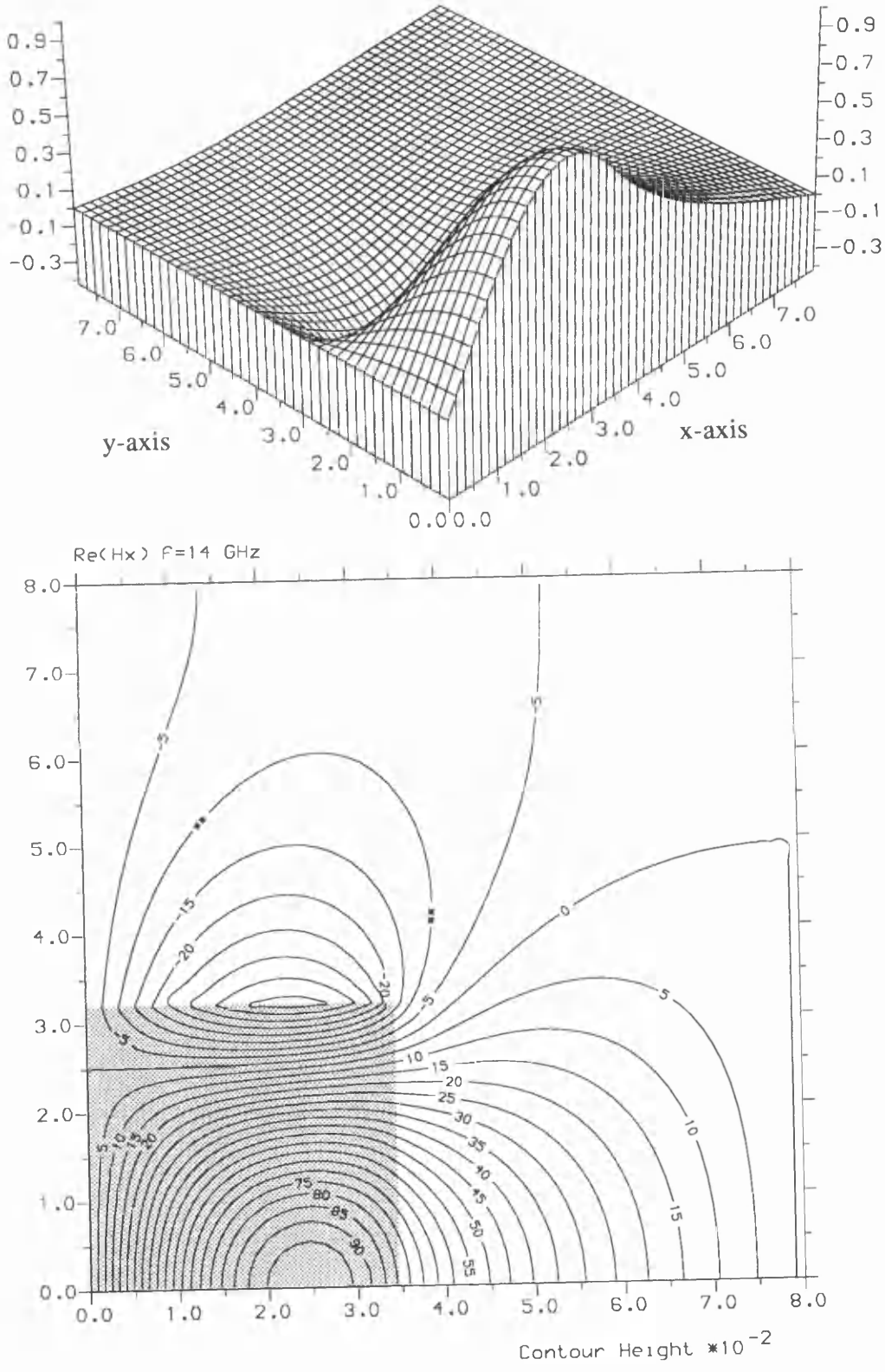


Fig. 7.7a Profile of real part of H_x component at f = 14 GHz.

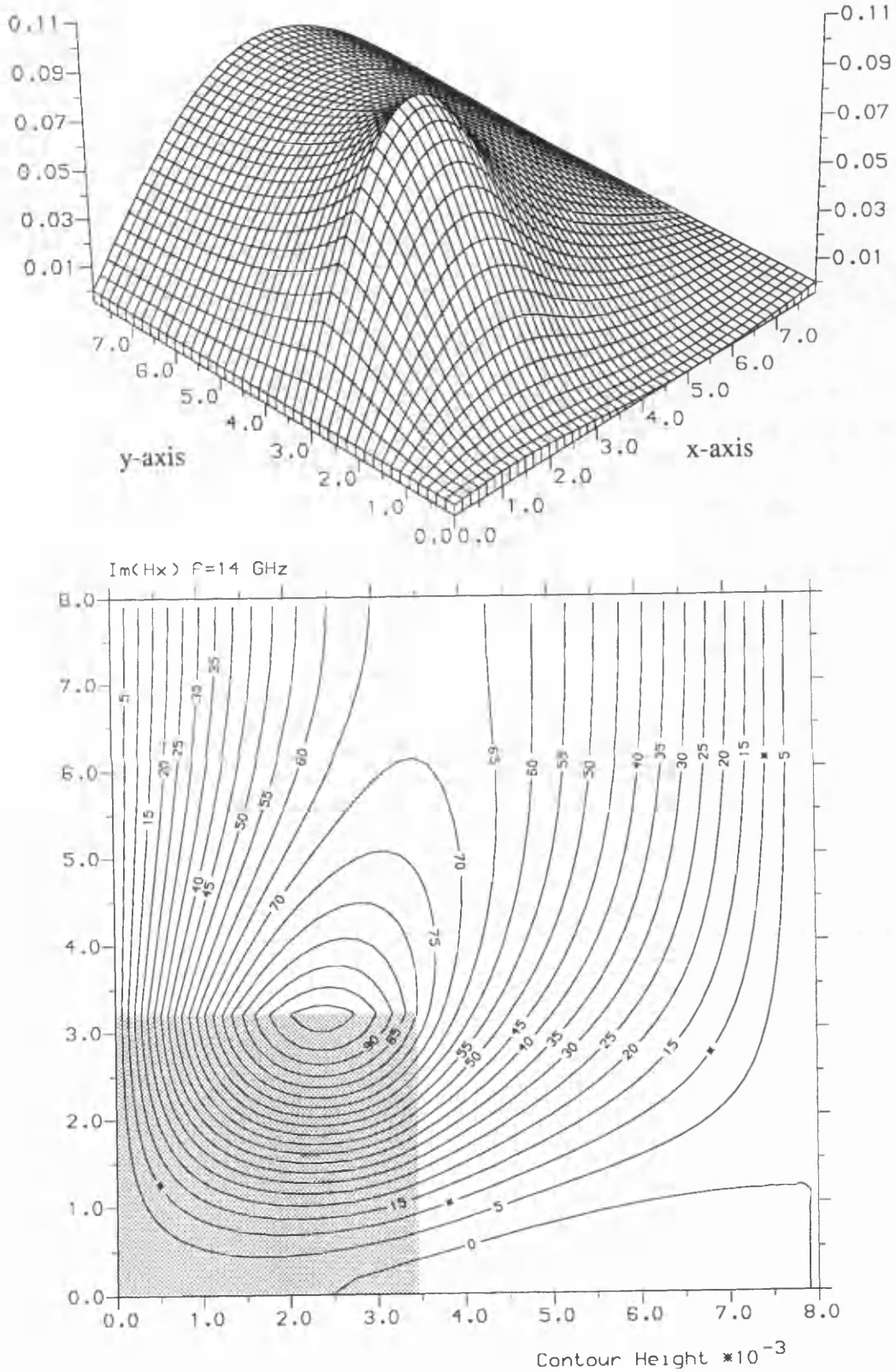


Fig. 7.7b Profile of imaginary part of H_x component at f = 14 GHz.

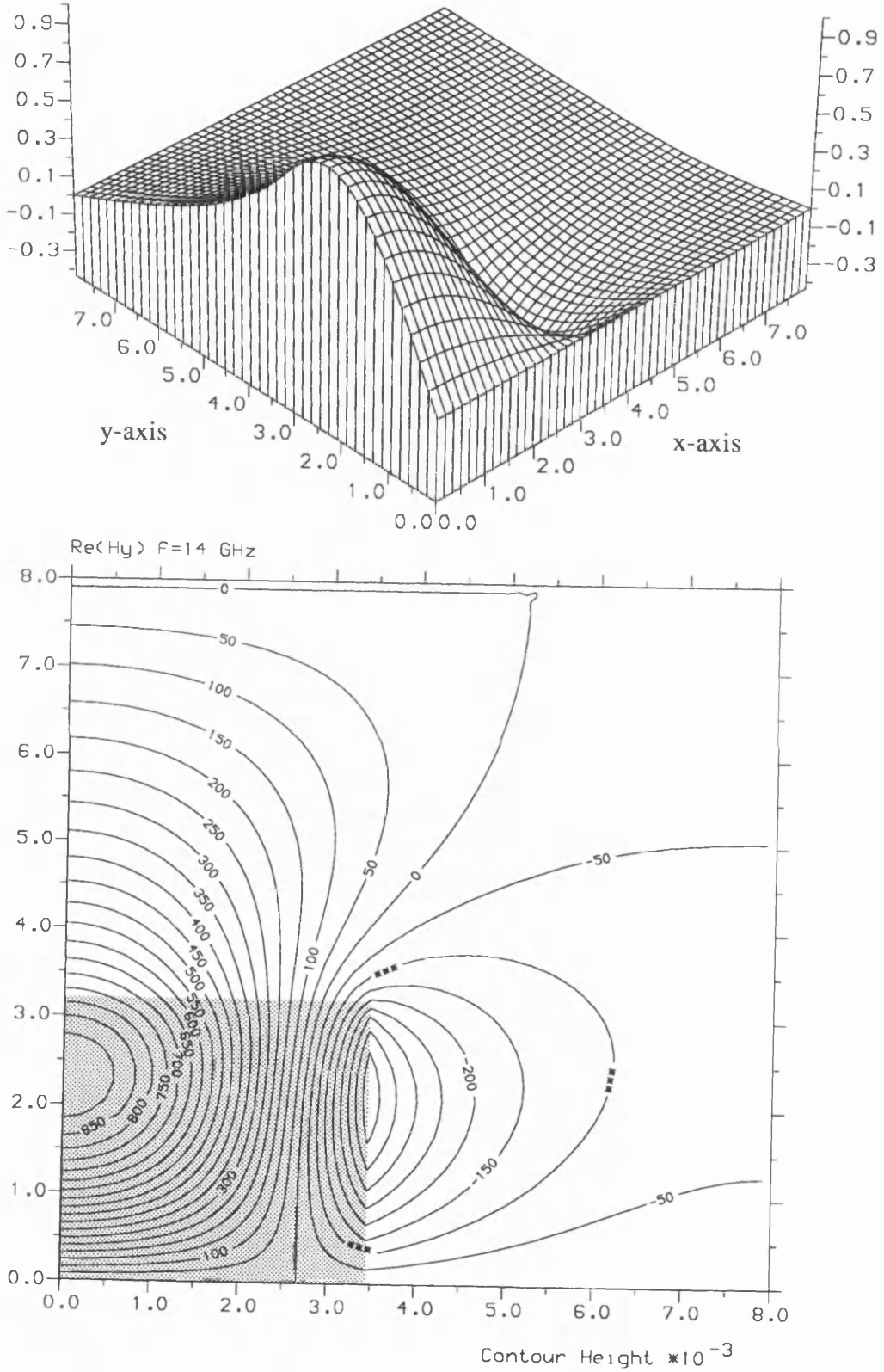


Fig. 7.8a Profile of real part of H_y component at f = 14 GHz.

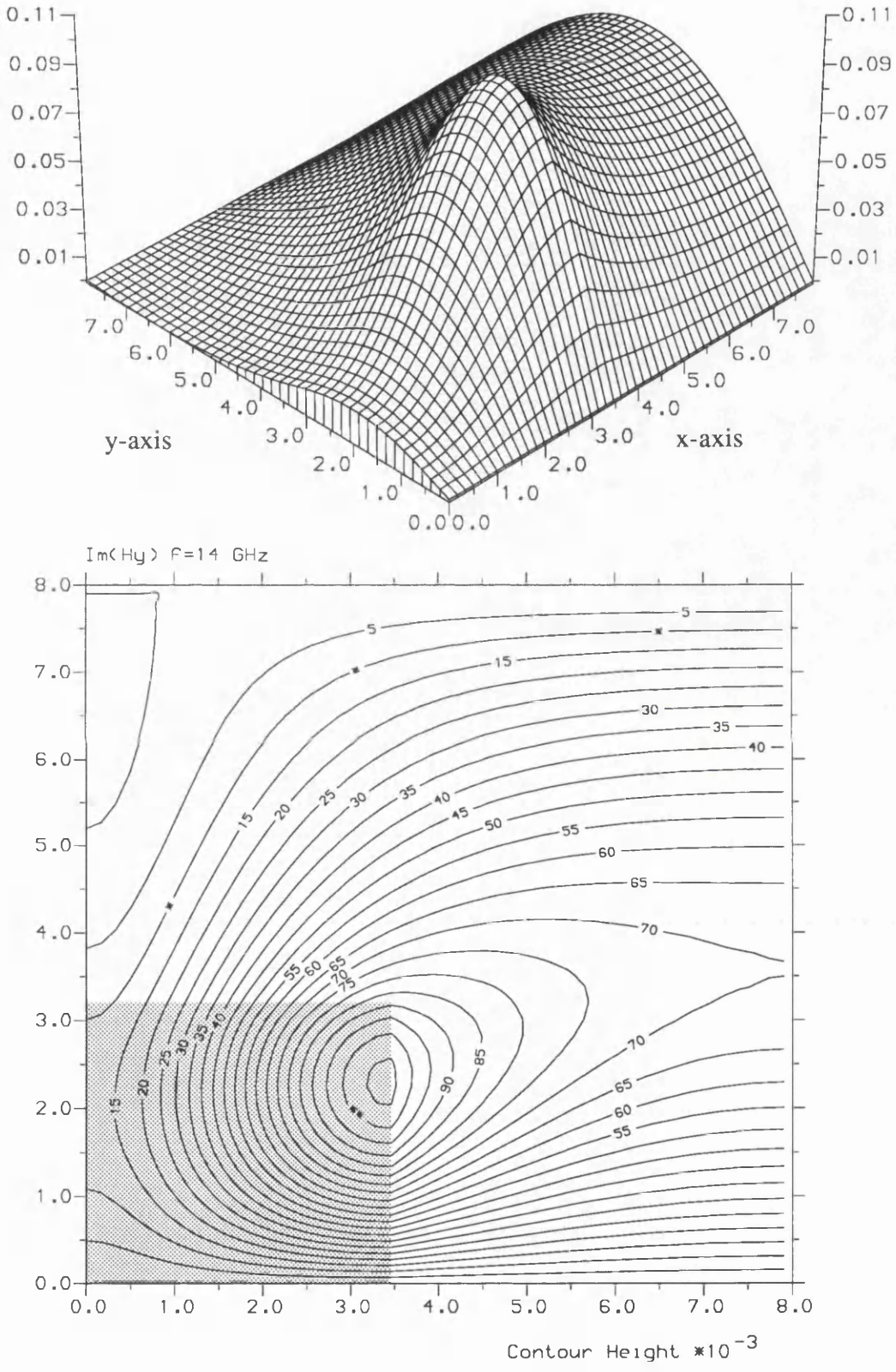


Fig. 7.8b Profile of imaginary part of H_y component at f = 14 GHz.

7.3.3 *Rectangular Dielectric Waveguide*

Fig. 7.9 shows the dispersion characteristics for the E^x_{mn} and E^y_{mn} modes of an isotropic rectangular dielectric waveguide of height t and width w buried in a medium with a refractive index n_2 of value 1.0; the refractive index of the core n_1 is 1.5. The dispersion curves are drawn in terms of the normalized index b and normalized frequency v, which are defined by:

$$b = ((\beta/k_0)^2 - n_2^2)/(n_1^2 - n_2^2) \tag{7.2}$$

$$v = k_0 t \sqrt{n_1^2 - n_2^2}\,/\pi \tag{7.3}$$

We compare our solutions with the results of Goell [16], showing excellent agreement even at low frequency. Goell's solution is derived from cylindrical harmonic analysis, and has often been used as a standard for comparison in literature [81], [83], [86]. However, a finite element solution is more versatile than a cylindrical harmonic analysis.

Unlike the crude simple truncation at a certain distance with artificial conductor walls enclosing the dielectric core [81], [86], we have used infinite elements in this example to extend a fixed finite element area of dimension (w+w)×(t+w). This improves the solution substantially in the lower frequency range (in this example $v < 0.5$). The decay factor can be easily optimized by looking for one that minimizes the eigenvalue solution as our formulation is a variational one. Fig. 7.10 shows the variation of the optimum decay factor versus frequency.

In this example, only one-quarter of the cross-section has been divided in 456 quadrilateral elements utilizing the inherent symmetry of the different modes. The CPU time is about 35 seconds for each point on a SUN SPARC 2 workstation, the memory requirement is less than 3 Mbytes.

Fig. 7.11 shows the contours of the H_x and H_y components of the lowest four modes at $v = 1.5$ in one quarter of the waveguide. The shaded area indicates the position of the dielectric core. The contours agree well with the mode designations and profiles [16] in both shape and magnitude for all four modes.

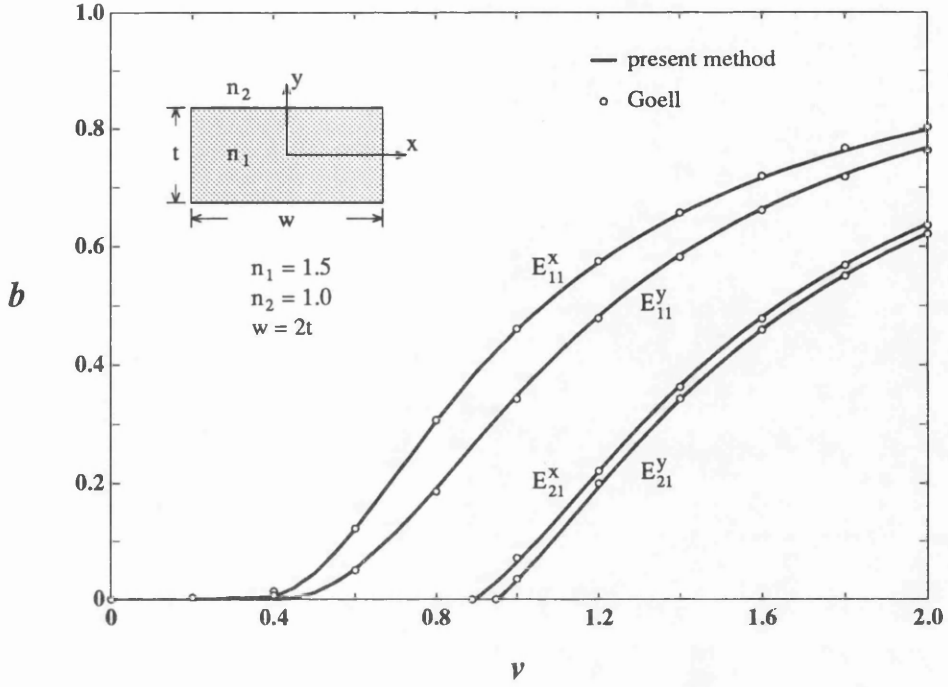


Fig. 7.9 Dispersion characteristics of the lowest four modes in an isotropic rectangular dielectric waveguide.

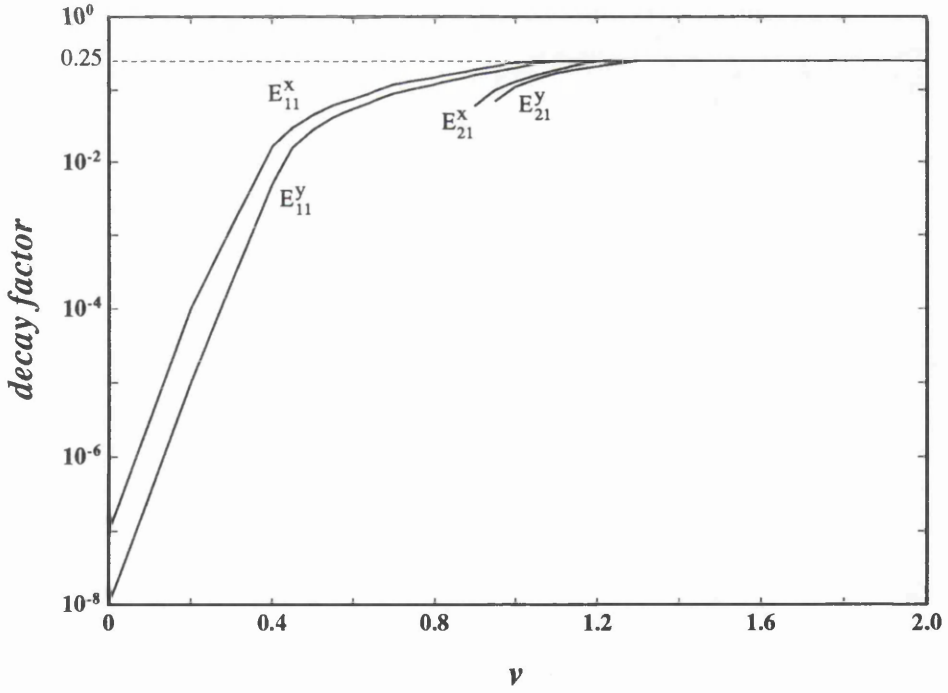


Fig. 7.10 Variation of optimum decay factors with normalized frequency.

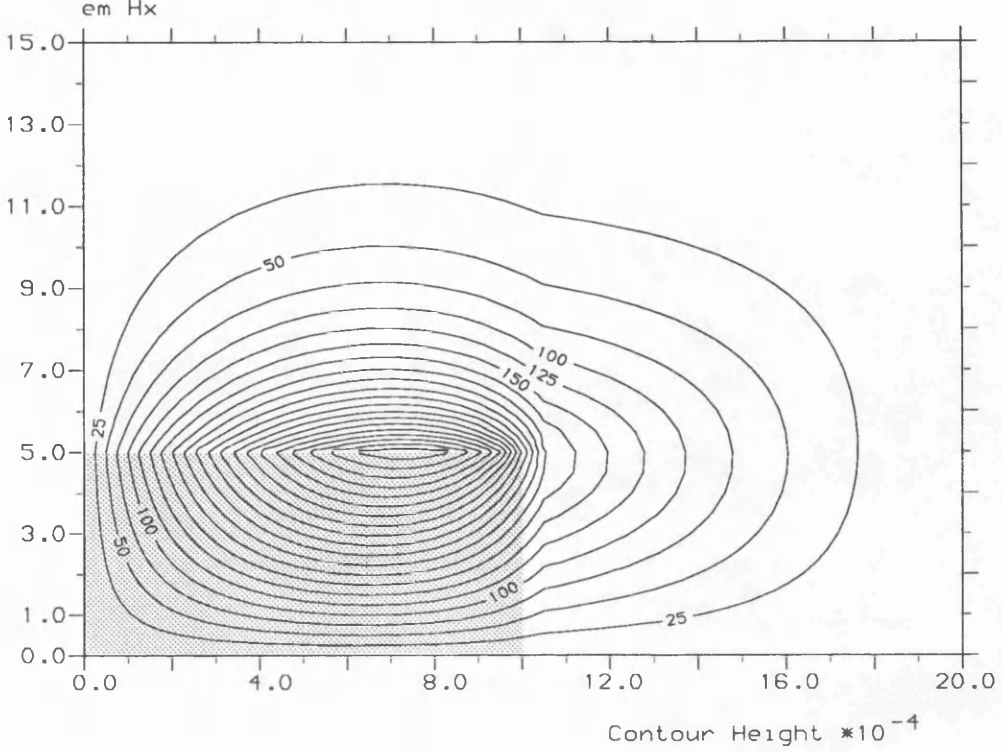


Fig. 7.11a Contour of H_x of E^x_{11} mode at v = 1.5.

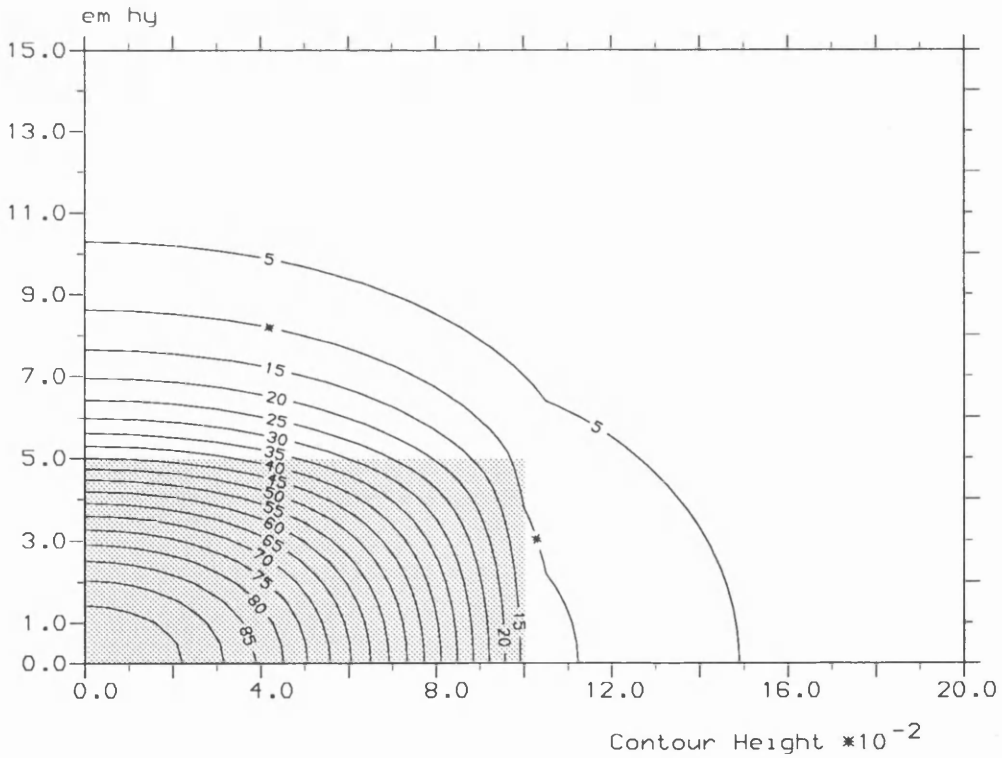


Fig. 7.11b Contour of H_y of E^x_{11} mode at v = 1.5.

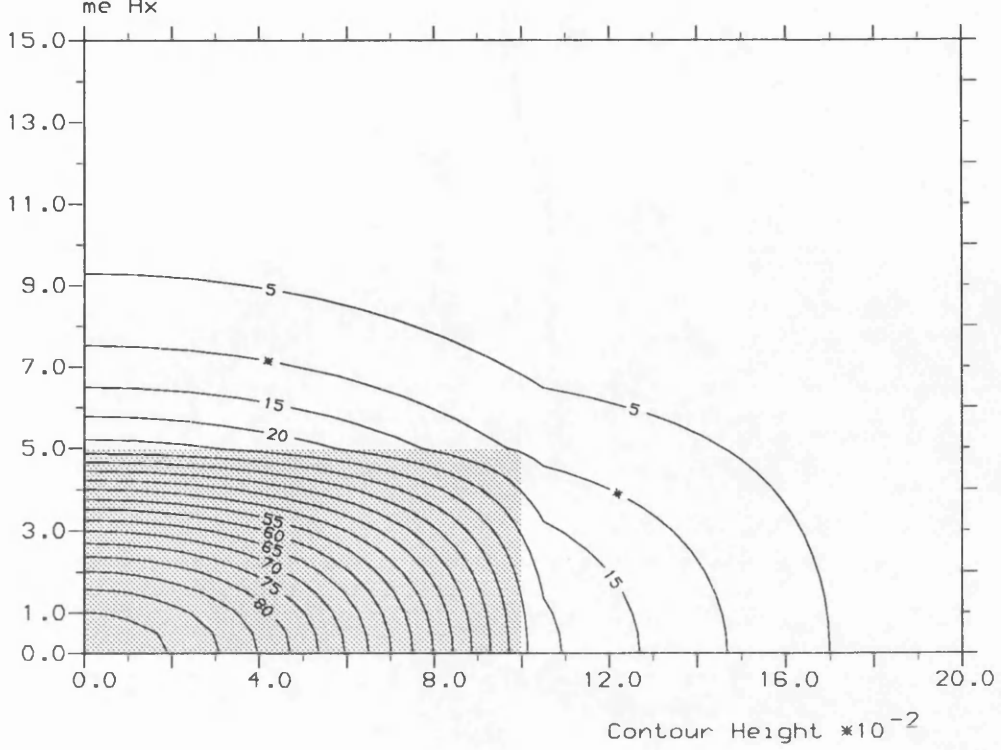


Fig. 7.11c Contours of H_x of E^y_{11} mode at v = 1.5.

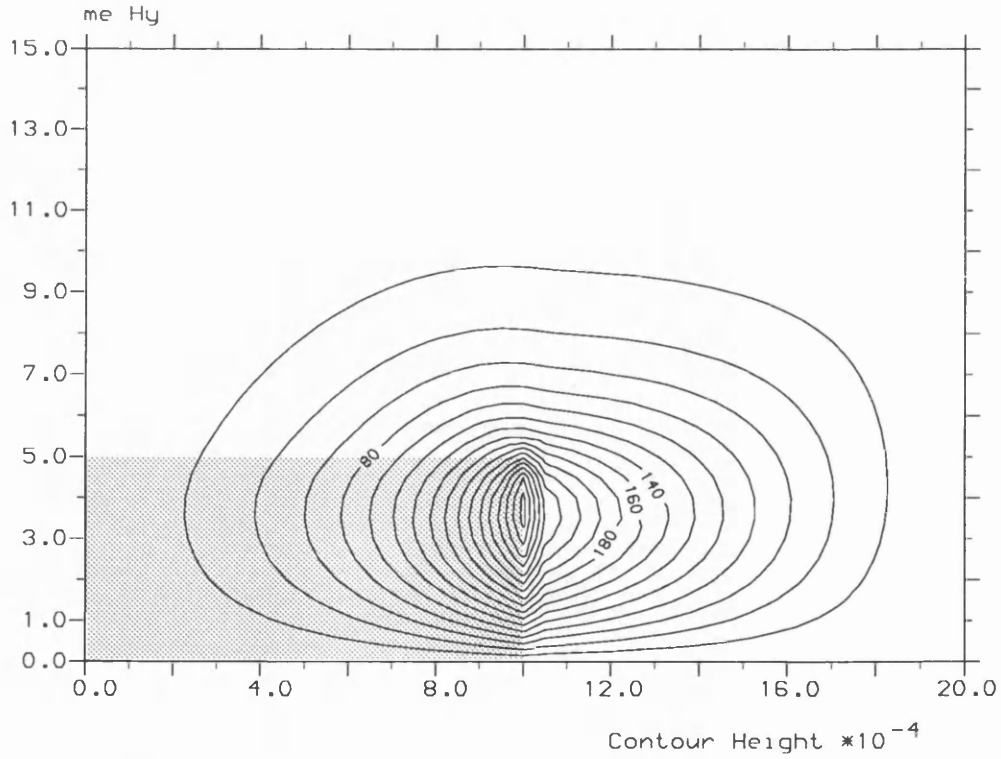


Fig. 7.11d Contours of H_y of E^y_{11} mode at v = 1.5.

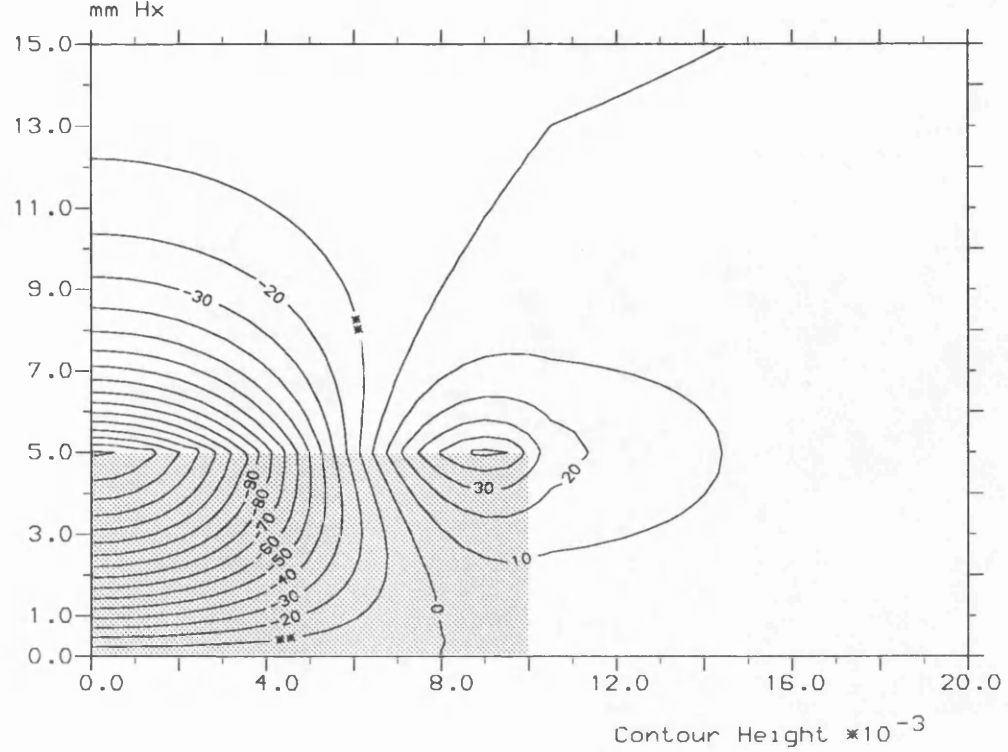


Fig. 7.11e Contours of H_x of E^x_{21} mode at v = 1.5.

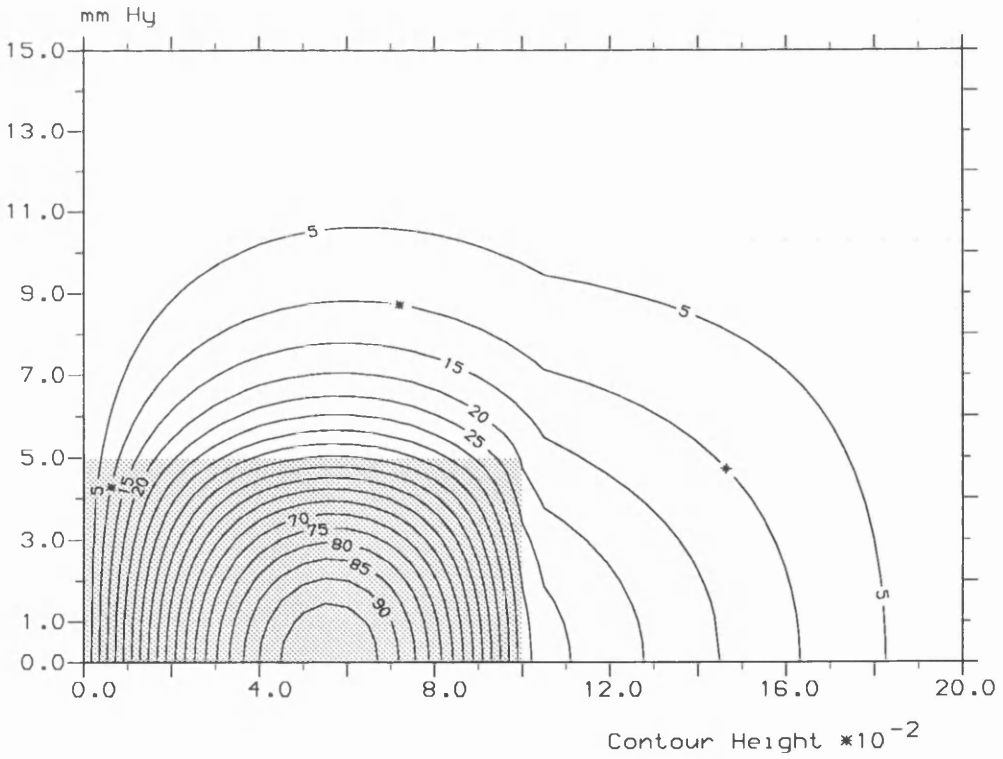


Fig. 7.11f Contours of H_y of E^x_{21} mode at v = 1.5.

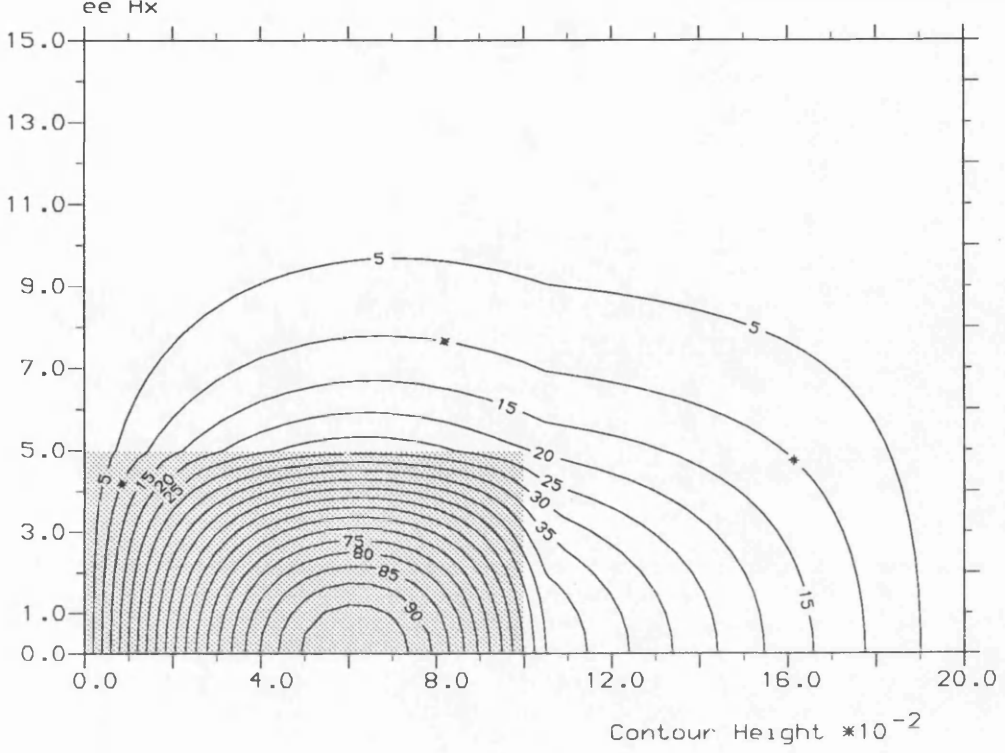


Fig. 7.11g Contours of H_x of E^y_{21} mode at v = 1.5.

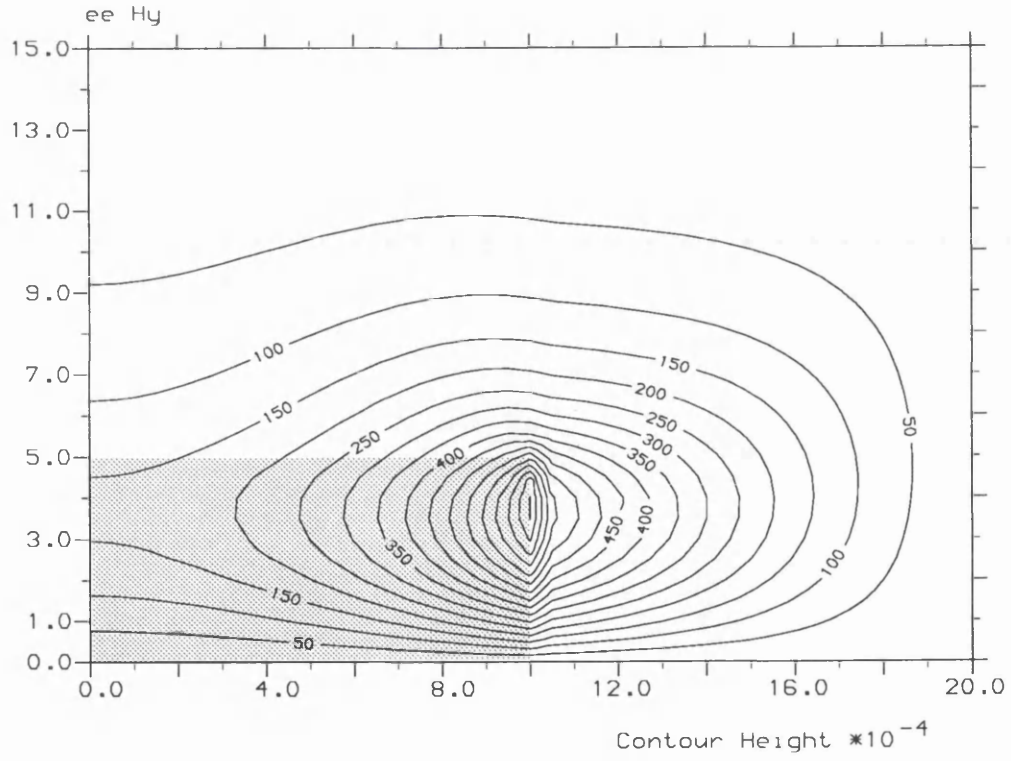


Fig. 7.11h Contours of H_y of E^y_{21} mode at v = 1.5.

7.3.4 Rib Waveguide

In Fig. 7.12 we compare our solutions of GaAs/GaAlAs rib waveguide structure (inset) with the results by *the spectral index method* and *the modified weighted index method* [128]. The curve for the dominant H^y_{11} mode is drawn in terms of the normalized index $b = ((\beta/k_0)^2 - n_3^2)/(n_2^2 - n_3^2)$ and the layer depth D with n_1= 1.0 (air), n_2= 3.40 (GaAs), n_3= 3.44 (GaAlAs), $2W$ = 3 μm, $H + D$ = 1 μm, λ = 1.15 μm. The normalized index b is very sensitive to the value β, therefore it is preferable to use b rather than β for comparison purposes.

A mesh of quadrilateral elements with 992 nodal points on half of the rib guide cross-section is used, the symmetry plane is placed at x = 0, the finite-to-infinite element boundaries are placed at x = 2.0 μm, y = –1.0 μm, and y = 1.2 μm. The sparse matrix solver is used. On a SUN SPARC 2 workstation, the cpu time is about 35 seconds for one layer depth D.

The rib waveguide is considered as an open structure, hence infinite elements are used with optimized decay factors. Using infinite elements gives an accurate solution even when the finite-to-infinite element boundary is placed very close to the guide. Fig. 7.13 illustrates the field distribution with a close finite-to-infinite element boundary, showing good agreement with the corresponding 'true' field in Fig. 7.14 obtained with far finite-to-infinite element boundary (x = 0, x = 3.0 μm, y = –2.5 μm, and y = 1.5 μm).

Figs. 7.14 and 7.15 show the profiles and contours of the dominant components of H^y_{11}, H^x_{21} modes, showing good agreement with those corresponding modes in [128]-[130].

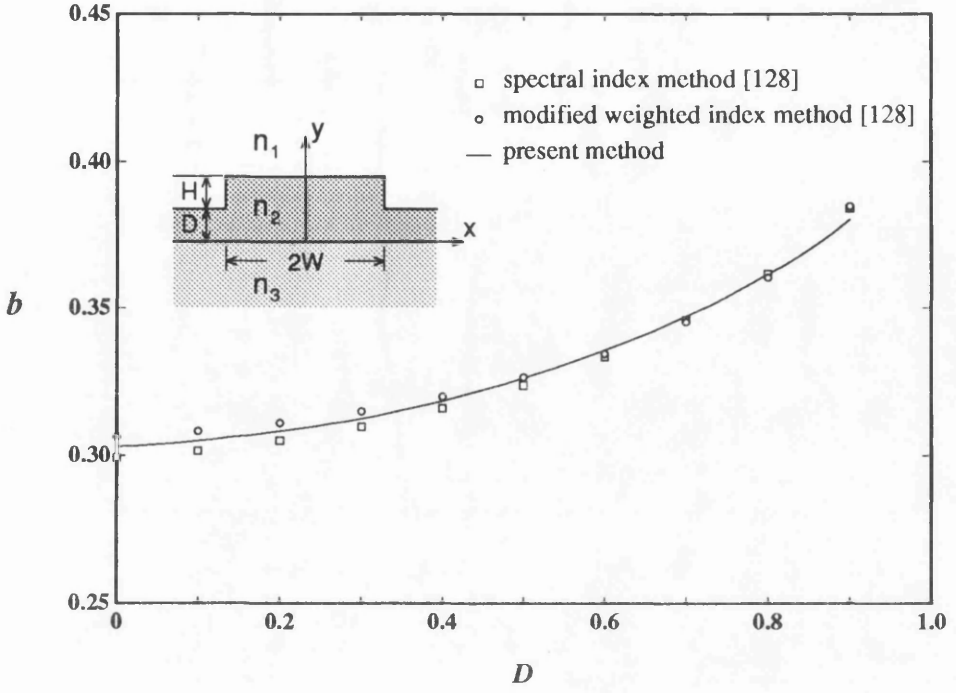


Fig. 7.12 Normalized indices b of H_{11}^{y} mode vs the layer depth D with n_1= 1.0 (air), n_2= 3.44 (GaAs), n_3= 3.40 ($Ga_{0.9}Al_{0.1}As$), $2W$ = 3 μm, $H + D$ = 1 μm, λ = 1.15 μm.

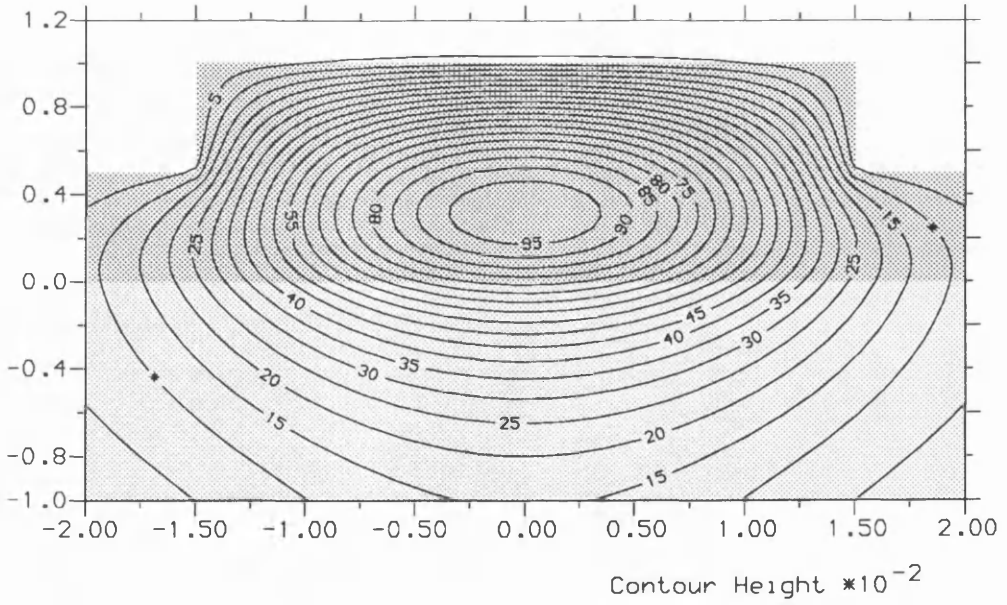


Fig. 7.13 Contours of the dominant component of fundamental H_{11}^{y} mode in a rib waveguide with a close finite-to-infinite boundary (x=0, x=2.0, y=-1.0, y=1.2) with D=0.5.

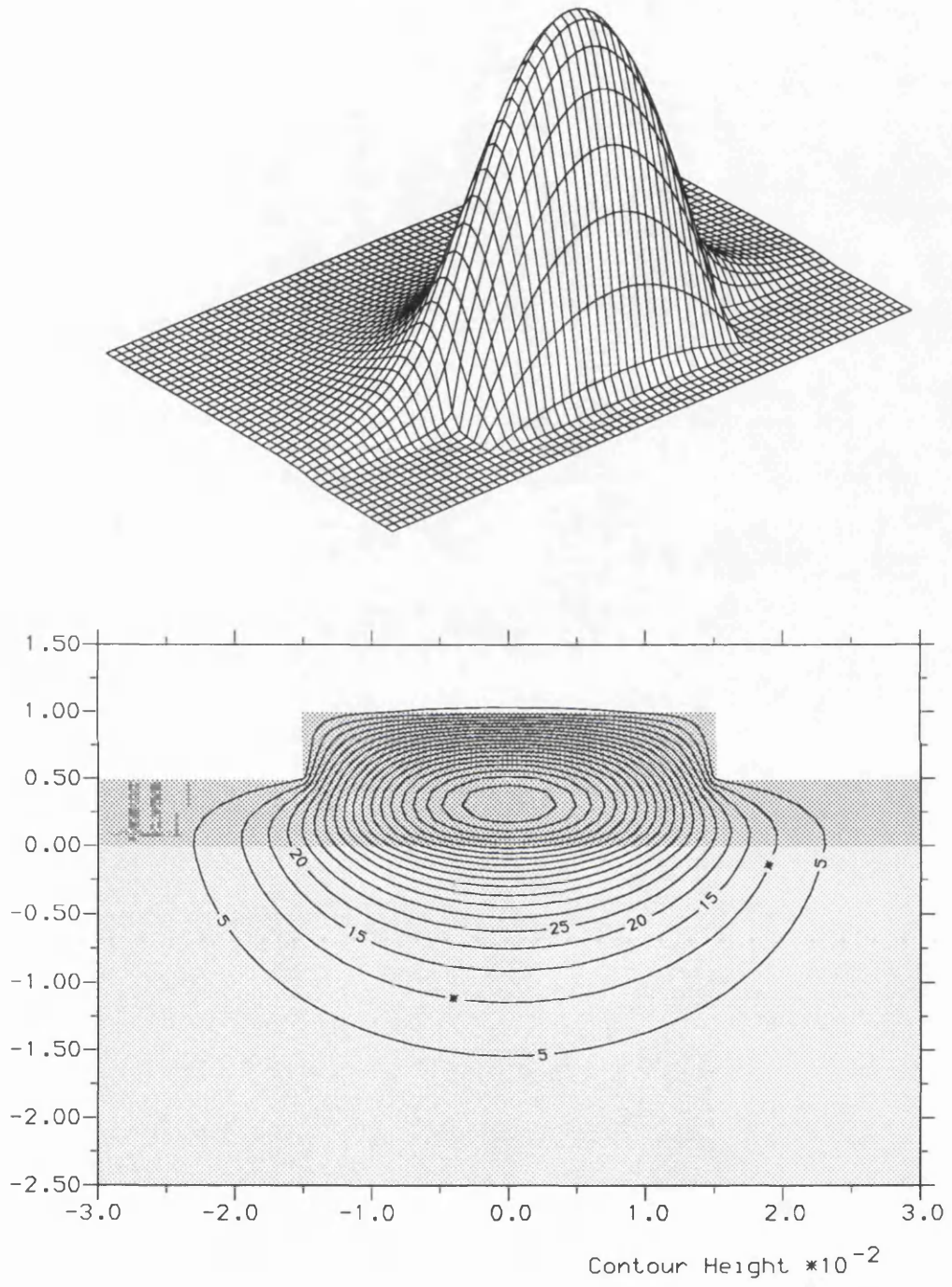


Fig. 7.14 Profile of the dominant component H_y of H_{11}^y mode in a rib guide with λ = 1.15 μm, n_1 = 1.0 (air), n_2 = 3.44 (GaAs), n_3= 3.40 ($Ga_{0.9}Al_{0.1}As$), $2W$ = 3 μm, $H = D$ = 0.5 μm.

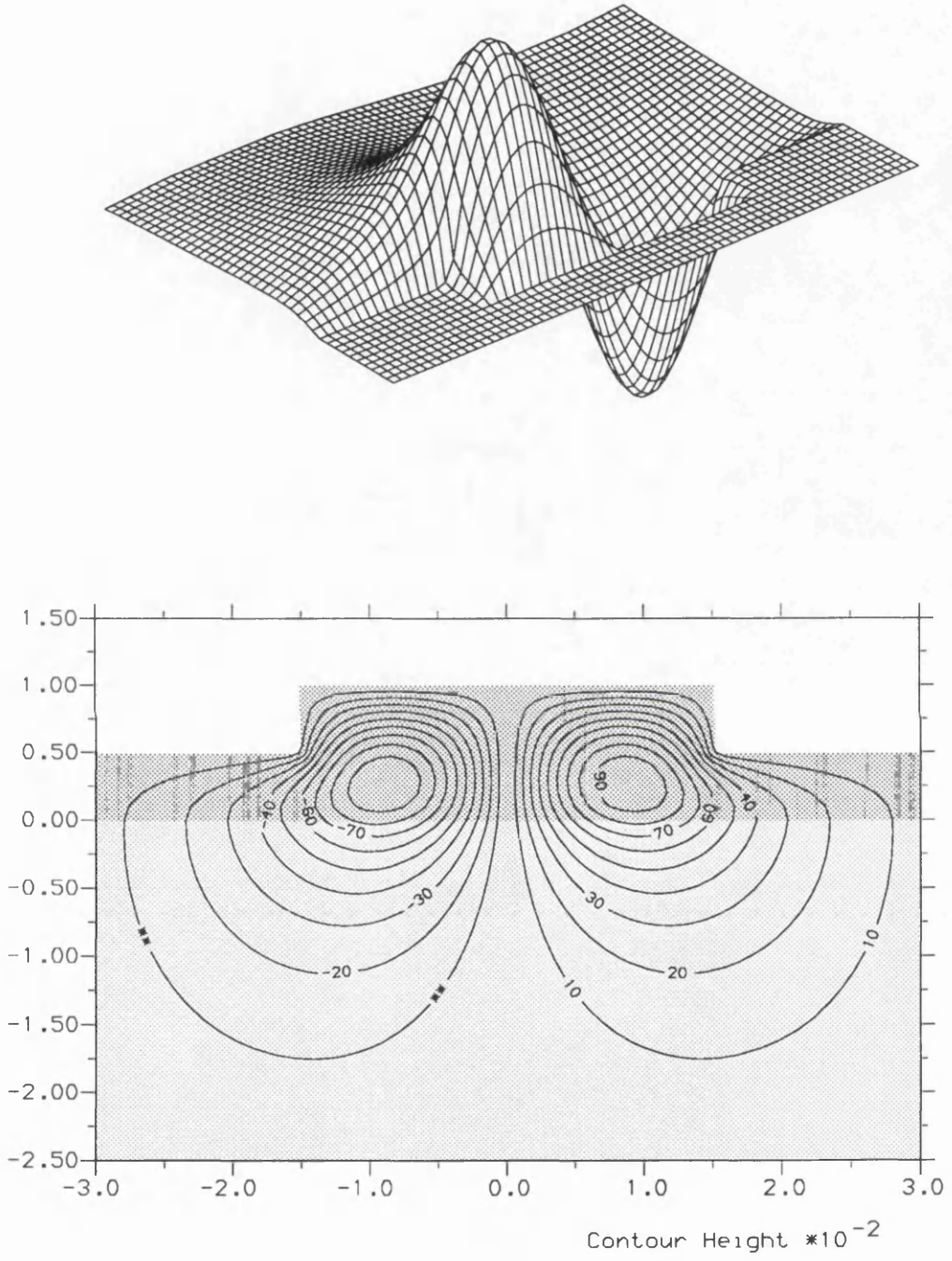


Fig. 7.15 Profile of the dominant component H_x of H_{21}^x mode in a rib guide with λ = 1.15 μm, n_1 = 1.0 (air), n_2 = 3.44 (GaAs), n_3= 3.40 ($Ga_{0.9}Al_{0.1}As$), $2W$ = 3 μm, H = D = 0.5 μm.

7.4 Anisotropic Lossless Waveguide

7.4.1 Rectangular Dielectric Waveguide

Fig. 7.16 shows the dispersion characteristics for the E^x_{mn} and E^y_{mn} modes of an anisotropic rectangular dielectric waveguide of height t, width $W = 2t$, core permittivity $n_x^2 = n_z^2 = 2.31$, $n_y^2 = 2.19$, and cladding permittivity $n_2^2 = 2.05$. Our results agree excellently with the results of Ohtaka [131]. Ohtaka's results are obtained by a variational method with cylindrical-harmonic-function expansion, and have been used frequently as a standard for comparison of results of anisotropic dielectric waveguide [81], [86], [88]. Similarly with the example of isotropic rectangular dielectric waveguide in section 7.3.3, the use of infinite elements greatly improves the accuracy of solution at the lower frequency range ($k_0 t < 3.5$), giving better results than [81], [86], [88]. It is worth mentioning that Svedin [88] uses infinite elements, but he does not get the good agreement in the lower frequency range.

The finite element area adopted and the mesh used in this example are the same of example in section 7.2.3. Also in this anisotropic case, no spurious solutions appear.

7.4.2 Channel Waveguide

Fig. 7.17 shows the dispersion characteristics for the lowest four H^x_{mn} modes of an anisotropic $LiNbO_3$ channel waveguide. The structure and parameters of the channel waveguide are shown in the inset of Fig. 7.17. Our results agree well for all four modes with Rahman and Davies [69] and for the dominant mode with Vandenbulcke and Lagasse [55], and Koshiba *et al.* [67].

Utilizing the inherent symmetry of the different modes, only half of the cross-section of the waveguide are divided into a mesh of 306 nodes with infinite elements beyond x = 0.8W and y= -2t and 0.5t. The CPU time is about 30 seconds per point on a SUN SPARC 2 workstation.

Fig. 7.18 shows the contours of the dominant magnetic field components of the lowest four modes at $k_0 t = 20$, showing clearly the field distribution of each mode.

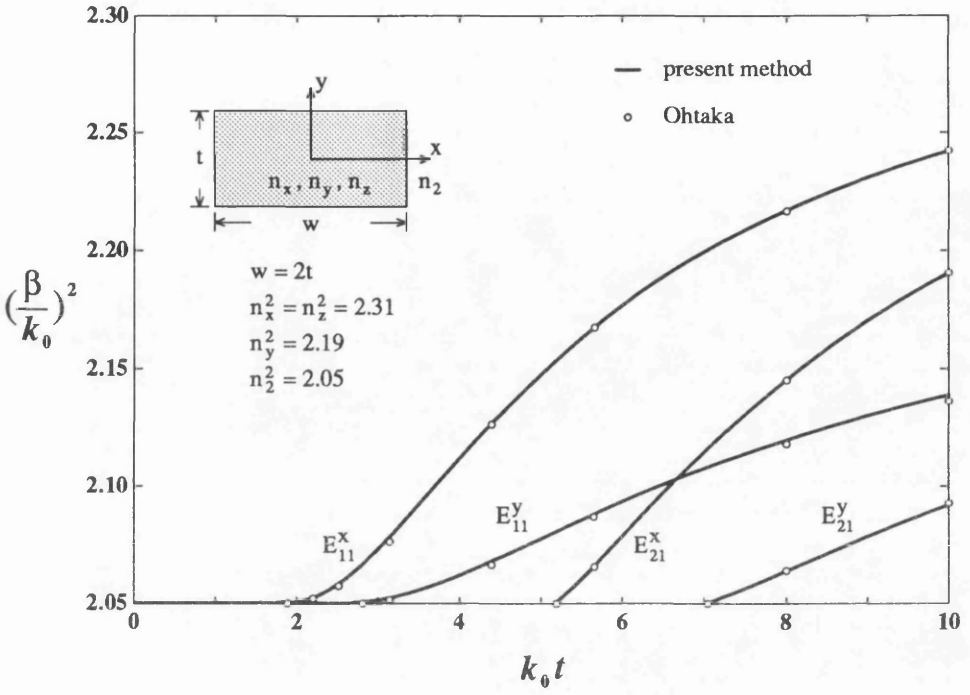


Fig. 7.16 Dispersion characteristics for the first four modes of an anisotropic rectangular dielectric waveguide.

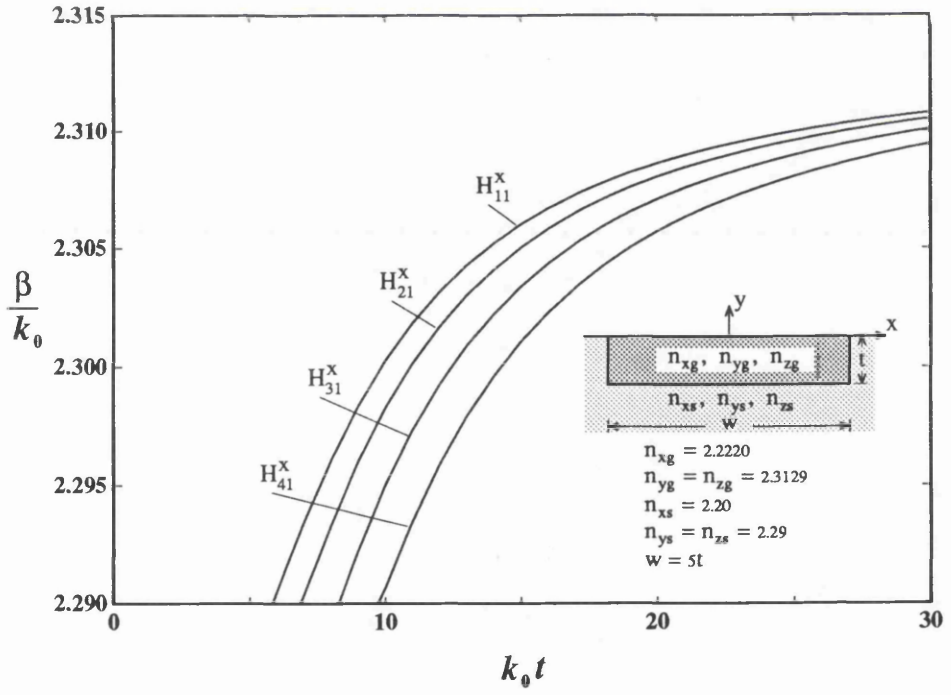


Fig. 7.17 Dispersion characteristics for the lowest four H_{mn}^x modes

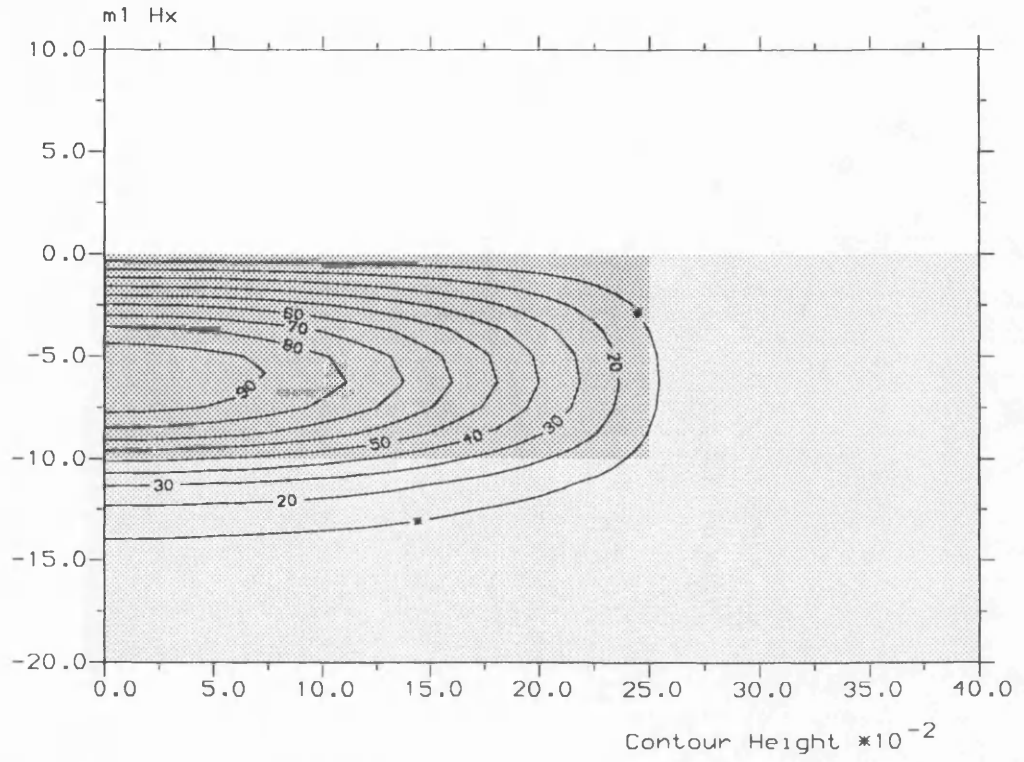


Fig. 7.18a Contours of dominant Hx component of H_{11}^{x} mode $k_0 t$ = 20.

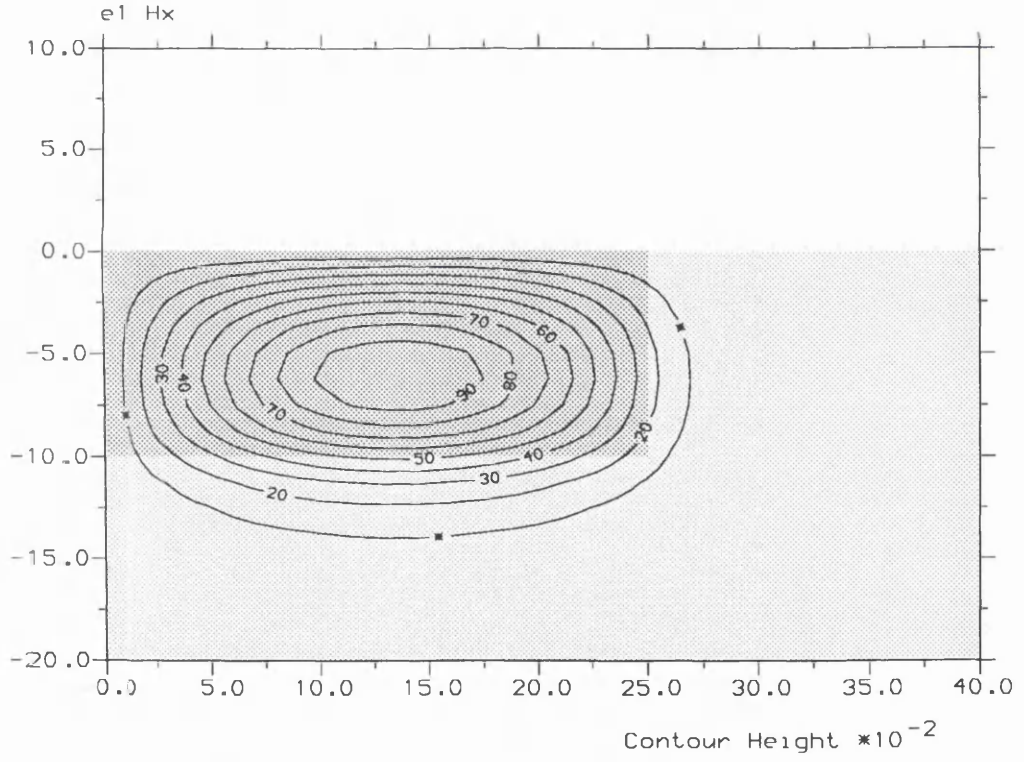


Fig. 7.18b Contours of dominant Hx component of H_{21}^{x} mode $k_0 t$ = 20.

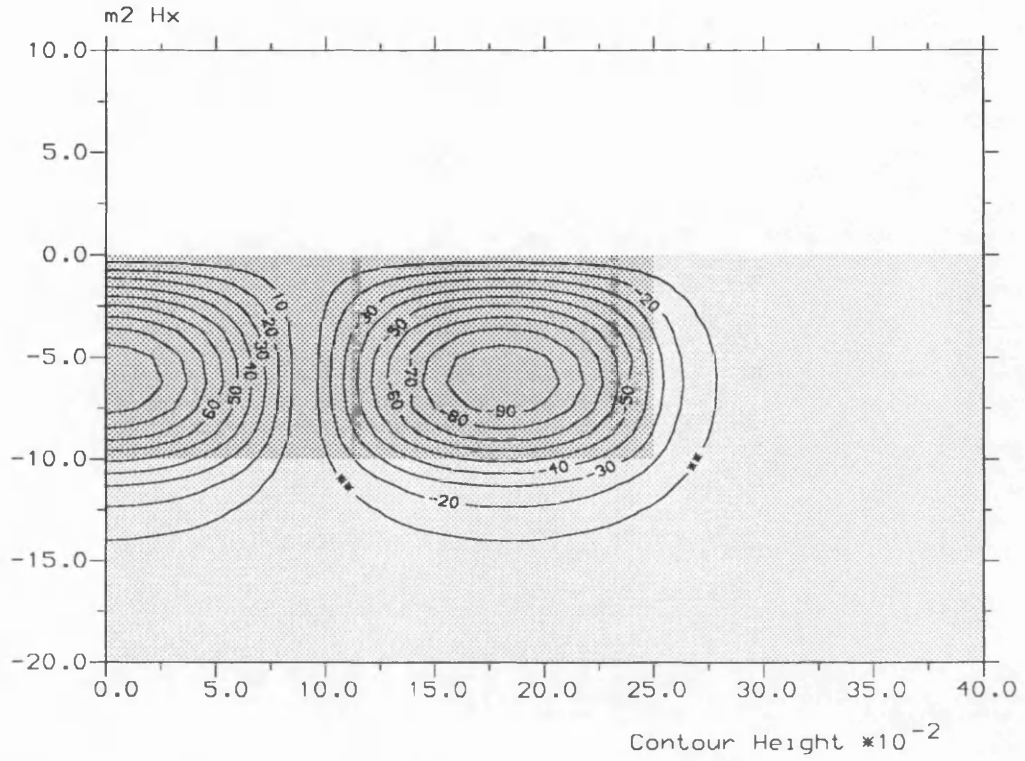


Fig. 7.18c Contours of dominant H_x component of H_{31}^x mode $k_0 t = 20$.

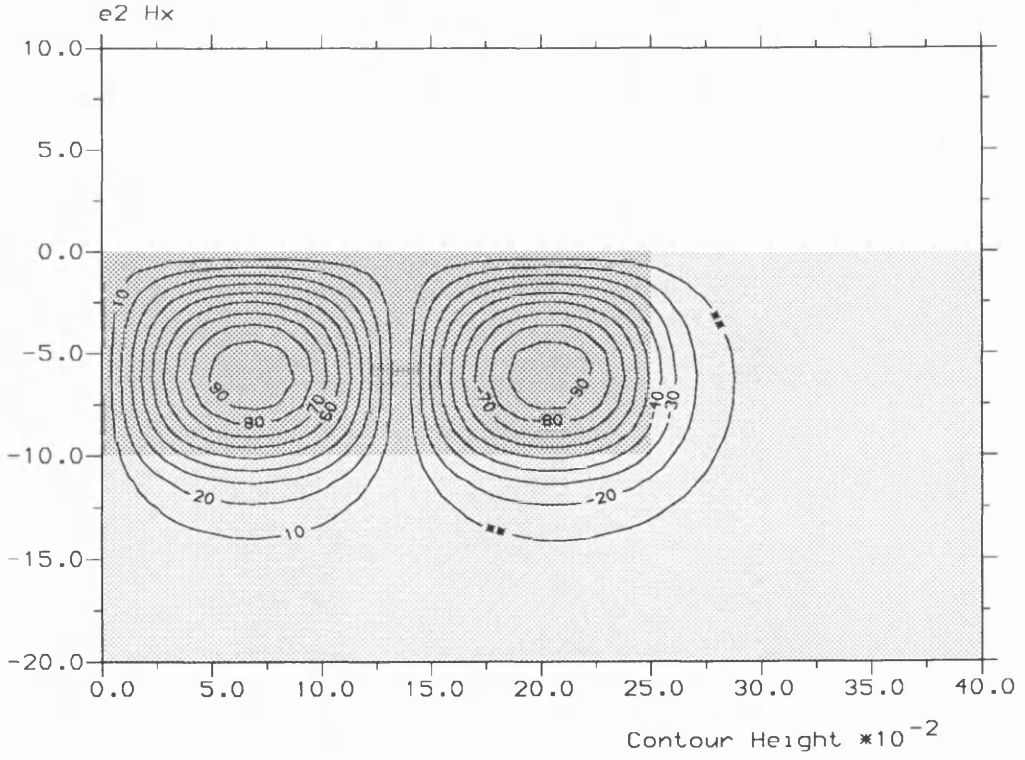


Fig. 7.18d Contours of dominant H_x component of H_{41}^x mode $k_0 t = 20$.

7.5 Isotropic Lossy Waveguides

7.5.1 Dielectric-Loaded Metallic Rectangular Waveguide

Fig. 7.19 shows the relative error of the finite element solutions, of the propagation constant for the fundamental TE_{10} and the first higher order TE_{01} modes in a rectangular metallic waveguide filled with lossy homogeneous isotropic dielectric of relative permittivity $\varepsilon = 1.5 - j1.5$, as a function of the number of unknowns (usually less than or equal to $2N_p$). Six meshes, 4×2, 12×6, 28×14, 56×28, 64×32, and 100×50 first order square elements, are chosen in the numerical computation. The statistics of CPU and memory requirement for this example are shown in Figs. 7.26 and 7.27 in section 7.6.

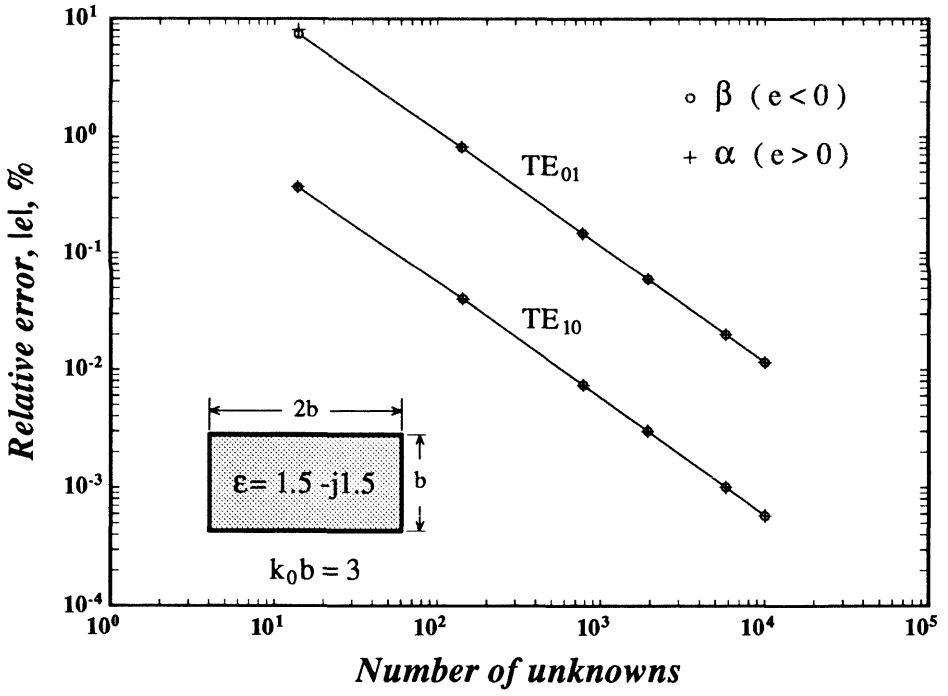


Fig. 7.19 The relative error in the finite element solutions of the propagation constant for the fundamental TE_{10} mode and higher order TE_{01} mode in the lossy dielectric-loaded metallic rectangular waveguide (inset) as a function of the number of unknowns, using square first order elements.

The relative error *e* is defined by

$$e = \begin{cases} (\alpha - \bar{\alpha})/\bar{\alpha} & \text{for attenuation constant} \\ (\beta - \bar{\beta})/\bar{\beta} & \text{for phase constant} \end{cases} \tag{7.4}$$

where (α , β) and $(\bar{\alpha} , \beta)$ are the finite element and exact solutions, respectively. The exact solutions are

$$\gamma = \alpha + j\beta = k_0 \left[\left(\frac{m\,\pi}{k_0 a}\right)^2 + \left(\frac{n\,\pi}{k_0 b}\right)^2 - \varepsilon' + j\varepsilon'' \right]^{\frac{1}{2}} \tag{7.5}$$

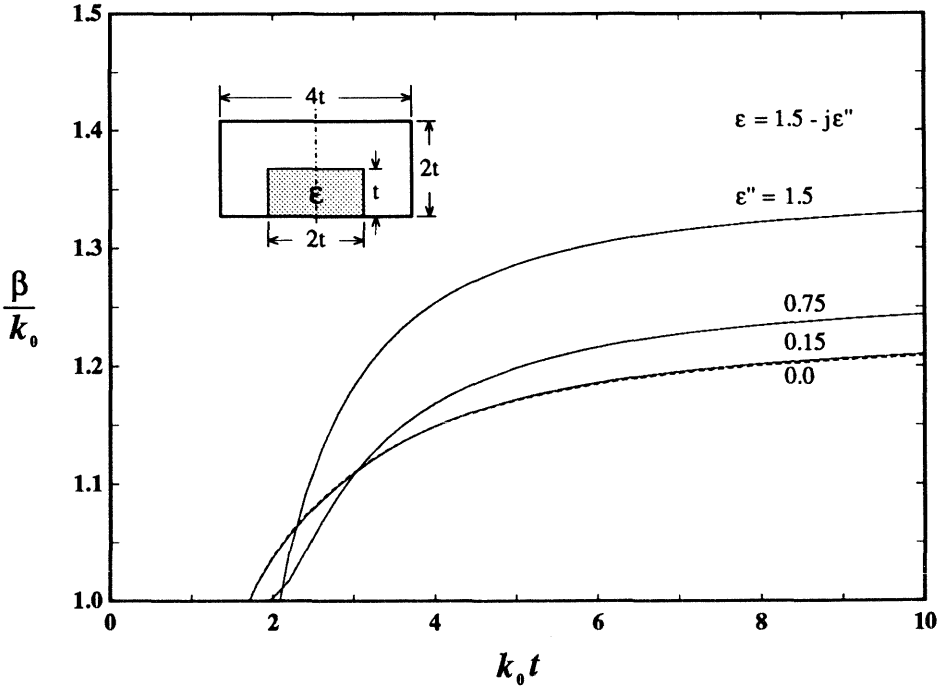
where, m and n stand for the mode indices for the x and y directions, respectively.

It is easily seen from Fig. 7.19 that the relative error decreases as the number of unknowns increases. Also it is interesting to note that the directions of convergence are opposite between the real and imaginary parts of the propagation constant; i.e., $e > 0$ for α whereas $e < 0$ for β.

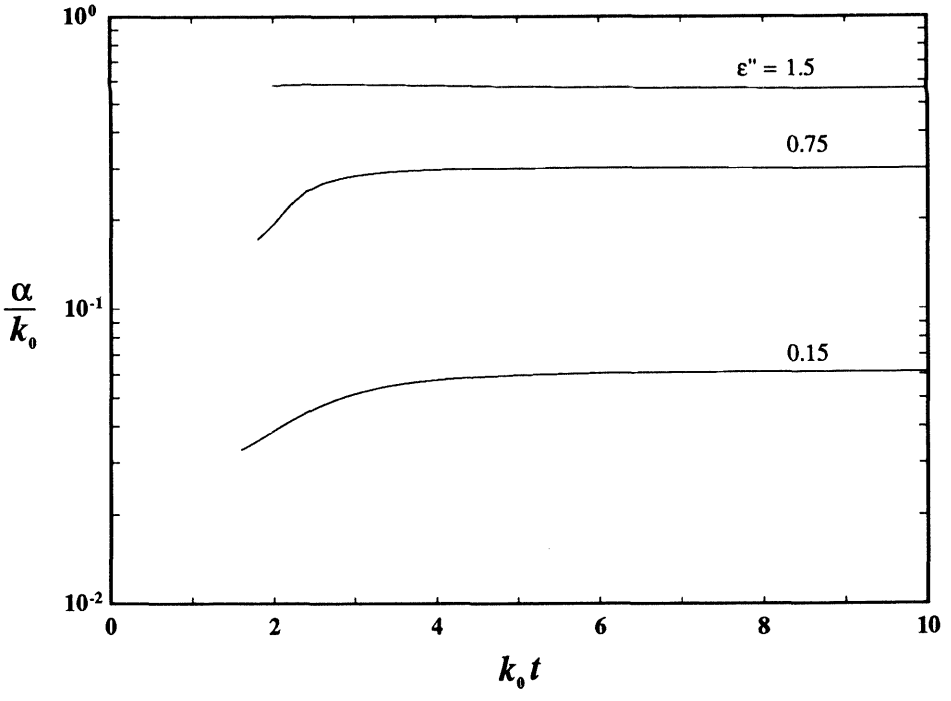
7.5.2 Shielded Image Waveguide

Fig. 7.20 shows the dispersion characteristics in the slow-wave region for the E^y_{11} mode of a lossy isotropic image waveguide, taking the imaginary part of relative permittivity, ε'', as a parameter. As it can be seen from Fig. 7.20a, the phase constant β for $\varepsilon'' = 0.15$ is very close to that for $\varepsilon'' = 0$, i.e., the lossless case.

Our results show the same trend as those of Hayata *et al.* [85]. For the α/k_0 curves both results are very close, whereas for the β/k_0 curves they differ considerably especially towards the lower frequency range. Hayata *et al.*'s three β/k_0 curves ($\varepsilon'' = 0.15$, 0.75, and 1.5) start from about $k_0 t = 3.4$, 3.9, 4.2, and end at about $\beta/k_0 = 1.17$, 1.20, 1.28, respectively, while our corresponding curves start from about $k_0 t = 1.71$, 1.95, 2.09, and end at about $\beta/k_0 = 1.210$, 1.243, and 1.330, respectively. It can be explained that our results are more accurate. One reason is that our solution always provides a lower bound to the true β solution and our results for this example are greater than those of Hayata *et al.* throughout the range in Fig. 7.20a. The other reason is that Hayata *et al.* use a very coarse mesh of only 81 nodes (for which the memory requirement is 7.6 MB !), and we use a much more finer mesh of 625 nodes (for which the memory requirement is less than 2.5 MB , the CPU time is about 17 seconds for each point on a SUN SPARC 2 workstation).



(a) Normalized phase constant.



(b) Normalized attenuation constant.

Fig. 7.20 Dispersion characteristics of E_{11}^{y} mode in shielded image waveguide with isotropic lossy dielectrics

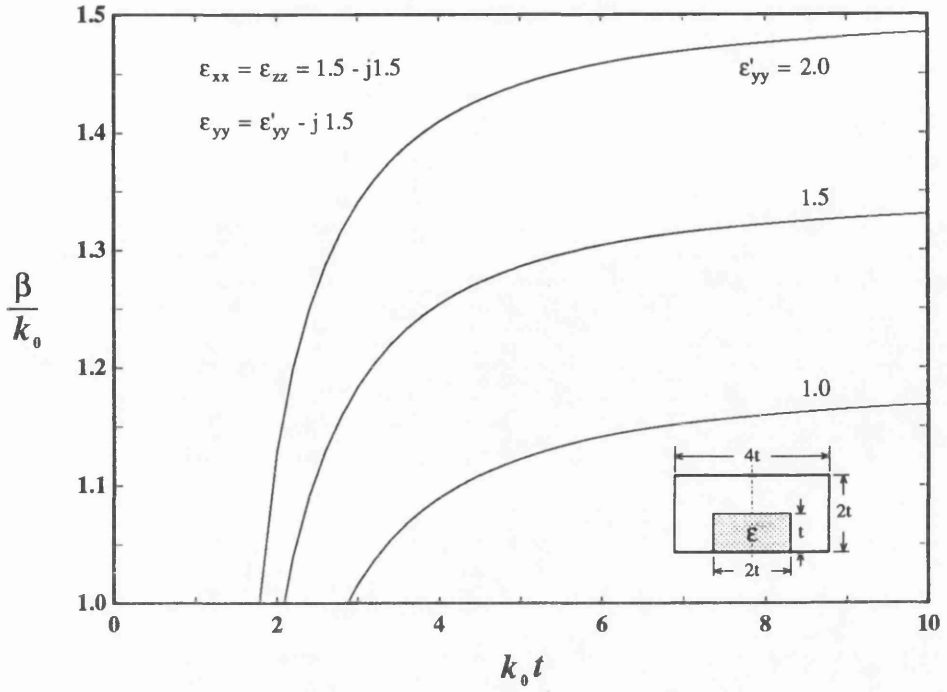
Figs. 7.21 and 7.22 show the dispersion characteristics in the slow wave region for the E^y_{11} mode of a lossy anisotropic image waveguide shown in the insets. The real part of ε_{yy}, ε'_{yy}, is chosen as a parameter in Fig. 7.21 ("dielectric anisotropy"), whereas the imaginary part of ε_{yy}, ε''_{yy}, is chosen in Fig. 7.22 ("conductivity anisotropy"). Comparison between Figs. 7.21(a) and 7.22(a) clearly shows that a similar effect is seen in the phase behaviour from the two types of anisotropy. On the contrary, from the comparison between Figs. 7.21(b) and 7.22(b), the opposite effect is found in the attenuation behaviour from the two types of anisotropy. That is, the attenuation becomes smaller as ε'_{yy} increases while it becomes larger as ε''_{yy} increases.

Similar to Fig. 7.20, the results in Figs. 7.21 and 7.22 show the same trend as those of Hayata *et al.* [85] but they differ in value considerably for the β/k_0 curves.

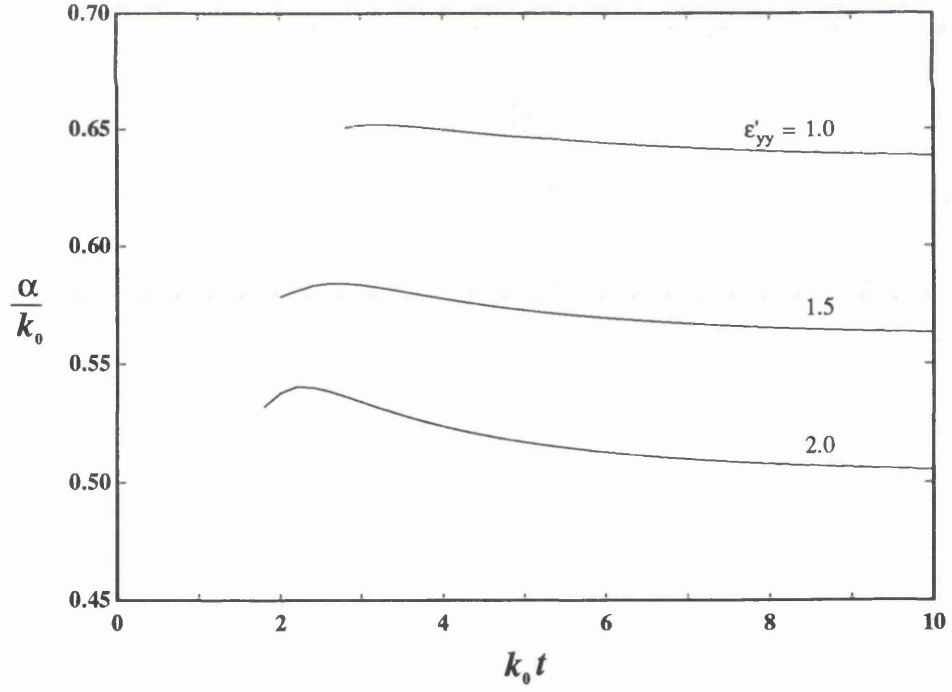
For curves in Fig. 7.21(a), the three β/k_0 curves (for ε'_{yy} = 2.0, 1.5, and 1.0) in [85] start from about $k_0 t$ = 3.6, 4.2, 5.7 and end at about β/k_0 = 1.44, 1.28, 1.12, respectively, while our corresponding curves start from about $k_0 t$ = 1.79, 2.09, 2.87, and end at about β/k_0 = 1.485, 1.330, and 1.168 respectively.

For curves in Fig. 7.22(a), the three β/k_0 curves (for ε''_{yy} = 2.0, 1.5, and 1.0) in [85] start from about $k_0 t$ = 3.9, 4.2, 4.5 and end at about β/k_0 = 1.36, 1.28, 1.23, respectively, while our corresponding curves start from about $k_0 t$ = 1.96, 2.09, 2.22, and end at about β/k_0 = 1.399, 1.330, and 1.270, respectively.

For the same reason as in the example in section 7.5.2, our results are much more accurate and economical than those of Hayata *et al.* [85]. A mesh of 625 nodes (24×24 quadrilateral elements) is used. The CPU time is about 17 seconds for each frequency and memory requirement is less than 2.5 MB on a SUN SPARC 2 workstation (with quoted speed at about 4.2 MFLOP, see Appendix B). This compares very favourably with Hayata's mesh of only 81 nodes needing 7.6 MB of memory and about 10 second CPU time for each frequency on a HITACHI S-810/10 supercomputer (with quoted speed at about 800 MFLOPS [97]).

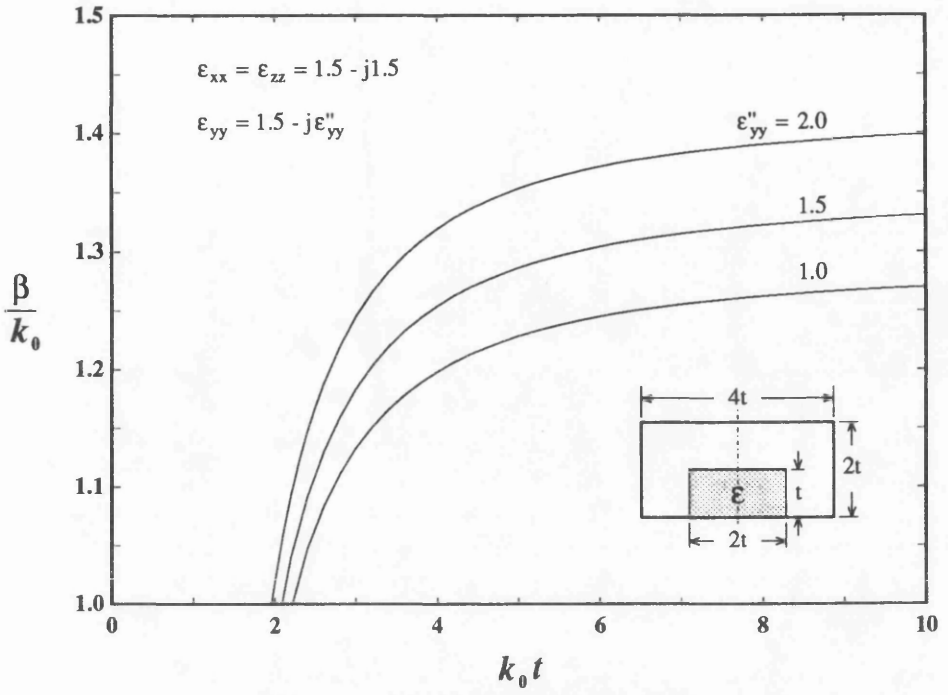


(a) Normalized phase constant.

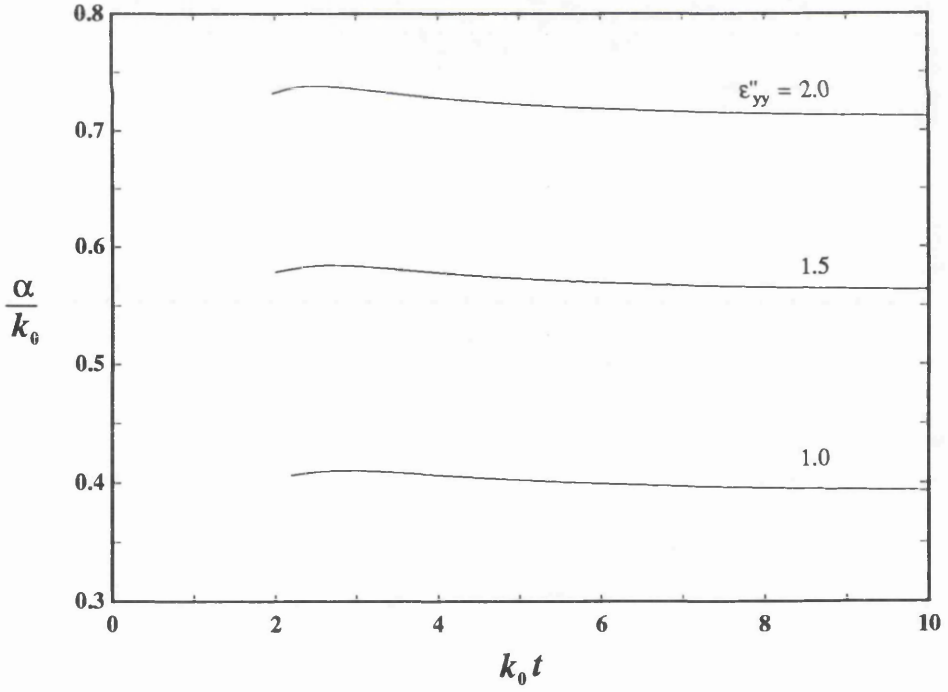


(b) Normalized attenuation constant.

Fig. 7.21 Dispersion characteristics of E^y_{11} mode in shielded image waveguide with anisotropic lossy dielectrics.



(a) Normalized phase constant.



(b) Normalized attenuation constant.

Fig. 7.22 Dispersion characteristics of E^y_{11} mode in shielded image waveguide with isotropic lossy dielectrics

7.7 Statistics of the Matrix Eigenequation Solvers

In this section, we show the statistics of CPU time and memory requirements of the new sparse matrix solver for several examples of use in connection with our new finite element formulation. For comparison purpose we have used the dense matrix algorithm QZ, the only algorithm available for this type of problem in standard computer libraries. In particular, we used the real number routine F02BJF for lossless waveguide problems and the complex number routine F02GJF for lossy waveguide problems from the NAG library [125].

Figs. 7.23 and 7.24 show the measured CPU time and estimated memory requirement using the dense real matrix solver F02BJF subroutine for the dielectric-slab-loaded metallic rectangular waveguide solution in section 7.2.1. The computation was performed on an Amdahl 5890 computer (see Appendix B). A curve fitting algorithm is used to approximate the measured CPU time t_{dr} by the function in (7.6), shown by the dotted line in Fig. 7.23.

$$t_{dr} \approx 3.288 \cdot 10^{-6} \cdot N_u^{3.054} \quad \text{(seconds)} \tag{7.6}$$

The memory requirement m_{dr} for a dense real matrix problem is estimated by (7.7)

$$m_{dr} \approx 0.096\ N_p^2 + 0.132\ N_p + 200 \quad \text{(kilobytes)} \tag{7.7}$$

where the assumption of 8 bytes for a DOUBLE PRECISION data, 4 bytes for a REAL data, and 2 bytes for an INTEGER data is adopted (see Appendix B).

Formulae (7.6) and (7.7) agree to the well known fact of CPU time and memory requirement being proportional to N^3 and N^2 respectively for large matrix order N when using a dense matrix solver.

Figs. 7.25 and 7.26 show the statistics of CPU time and memory requirement for the example of complex modes shown in section 7.3.2 using the sparse matrix solver with a subspace of order 6. For comparison the corresponding complex dense matrix routine F02GJF from the NAG library is used solving the same problem.

For this example, the measured CPU time t_{s6} and the memory m_{s6} for the sparse matrix solver can be fitted and estimated by (7.8) and (7.9),

respectively

$$t_{s6} \approx 9.596 \cdot 10^{-4}\ N_u^{1.6095} + 12.3 \quad \text{(seconds)} \tag{7.8}$$

$$m_{s6} \approx 4.496\ N_p + 250 \quad \text{(kilobytes)} \tag{7.9}$$

The memory requirement for the dense complex matrix solver F02GJF, m_{dc} can be estimated by (7.10)

$$m_{dc} \approx 0.192\ N_p^2 + 0.164\ N_p + 200 \quad \text{(kilobytes)} \tag{7.10}$$

It is apparent from the figures that the dense matrix solver rapidly becomes impractical (even with supercomputers) with increasing matrix order. For the same example using the sparse matrix solver, both the CPU time and memory are drastically reduced. For this example a subspace of order 6 is used. If only one mode was desired, the CPU time could be reduced even further.

Figs. 7.27 and 7.28 show the statistics of CPU time and memory requirement for the calculation of a lossy waveguide problem shown in section 7.5.1. In this example, the subspace of order 1 is chosen. The corresponding fitting function for the measured CPU time t_{s1} and the estimated function for the memory requirement m_{s1} are expressed in (7.11) and (7.12), respectively

$$t_{s1} \approx 4.3047 \cdot 10^{-4}\ N_u^{1.4855} + 0.2 \quad \text{(seconds)} \tag{7.11}$$

$$m_{s1} \approx 3.694\ N_p + 250 \quad \text{(kilobytes)} \tag{7.12}$$

Fig. 7.27 shows three computing times: (a) the total CPU time which includes mesh generating, matrix element calculation and assembling, and solution of the eigenvalue equation; (b) the matrix solver CPU time which only includes the CPU time used for solving the matrix eigenvalue equation; (c) the CPU time per iteration in the sparse matrix solver. Comparing the total time and the matrix solver time clearly shows that the most time consum ing part of the finite element solution is solving the matrix eigenvalue equation.

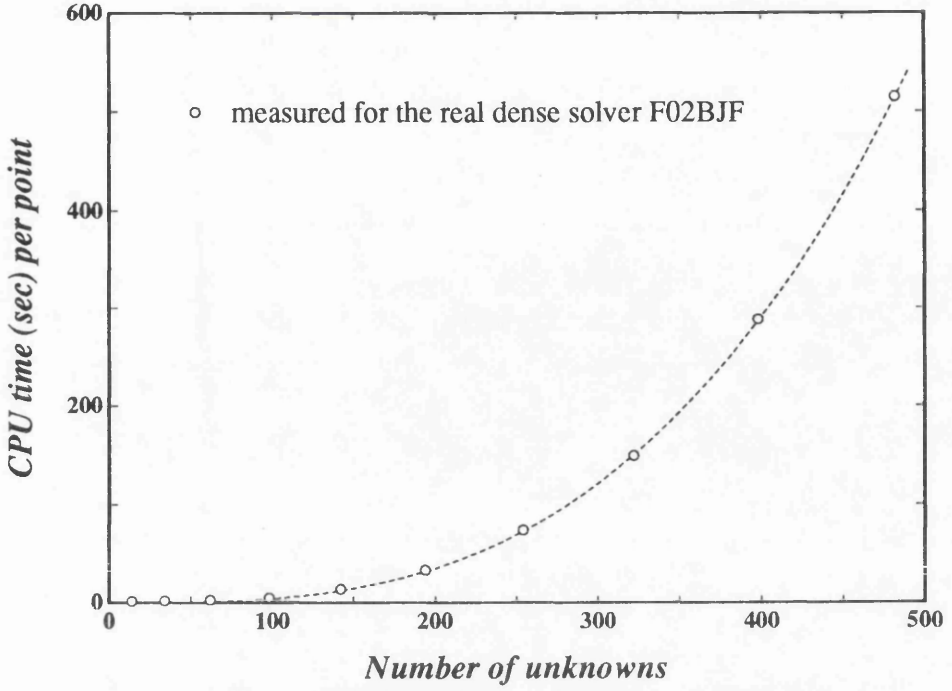


Fig. 7.23 Measured CPU time for the real dense NAG library routine F02BJF using Amdahl 5890 computer

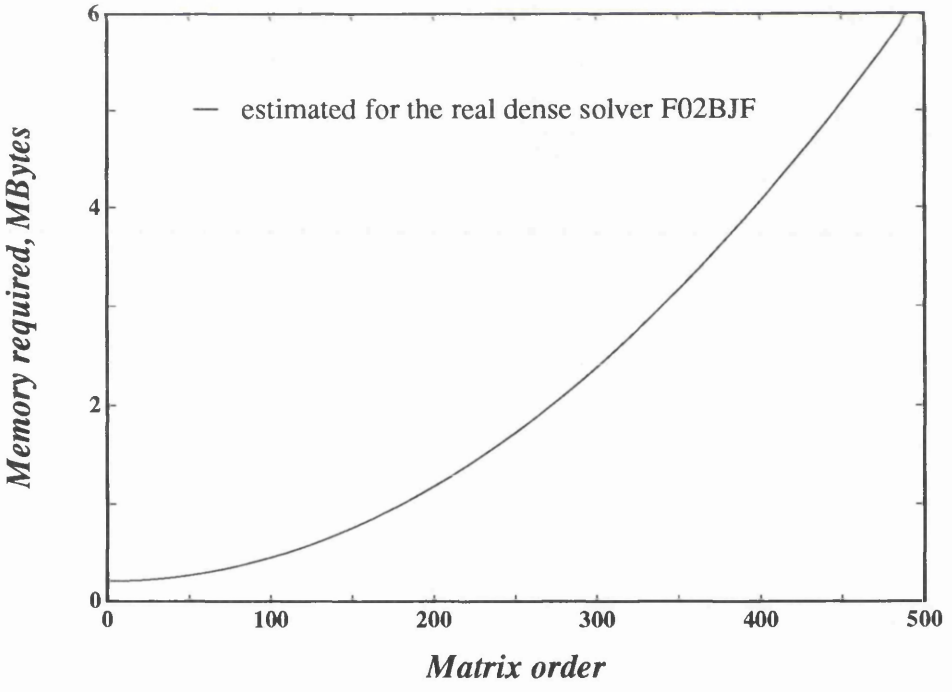


Fig. 7.24 Estimated memory requirement for the real dense NAG library routine F02BJF

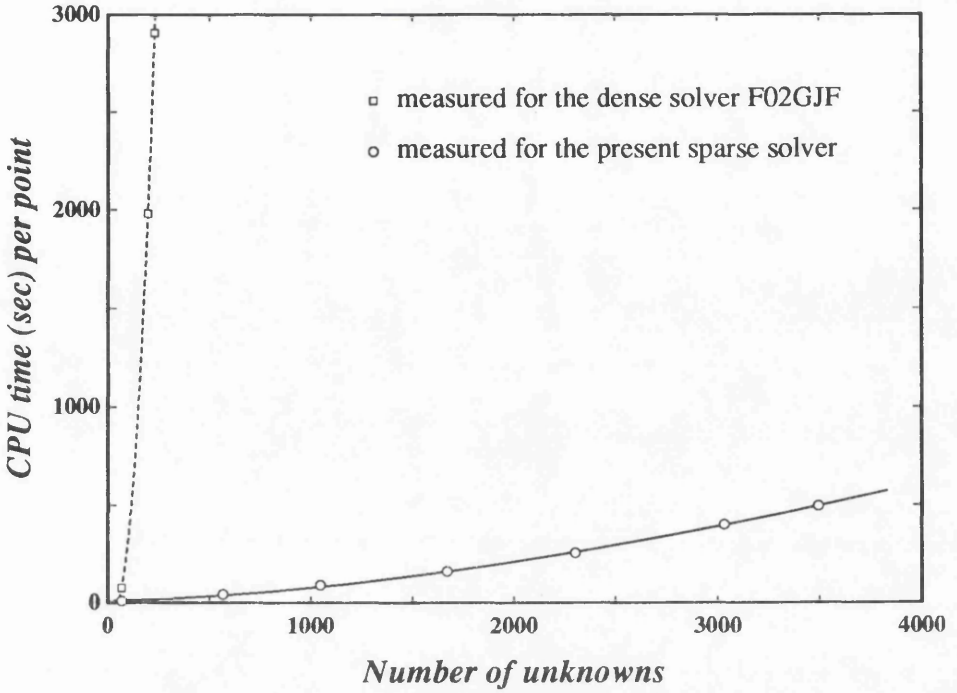


Fig. 7.25 CPU time for the sparse solver and its NAG equivalent complex dense routine F02GJF on a SUN SPARCstation 2.

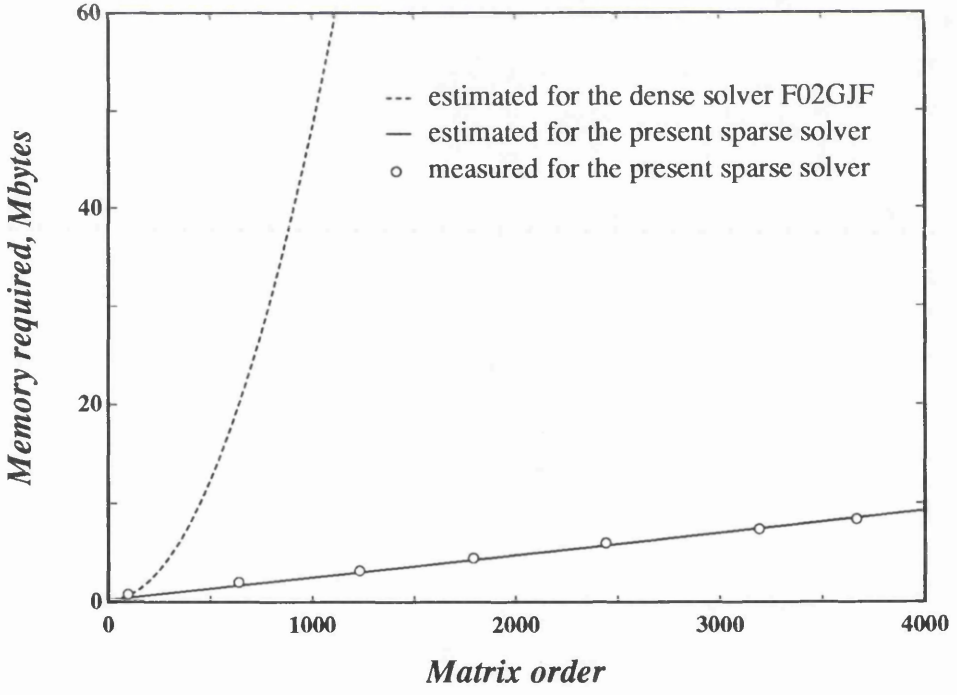


Fig. 7.26 Memory requirement of the sparse solver with a subspace of order 6 and its NAG equivalent complex dense subroutine F02GJF.

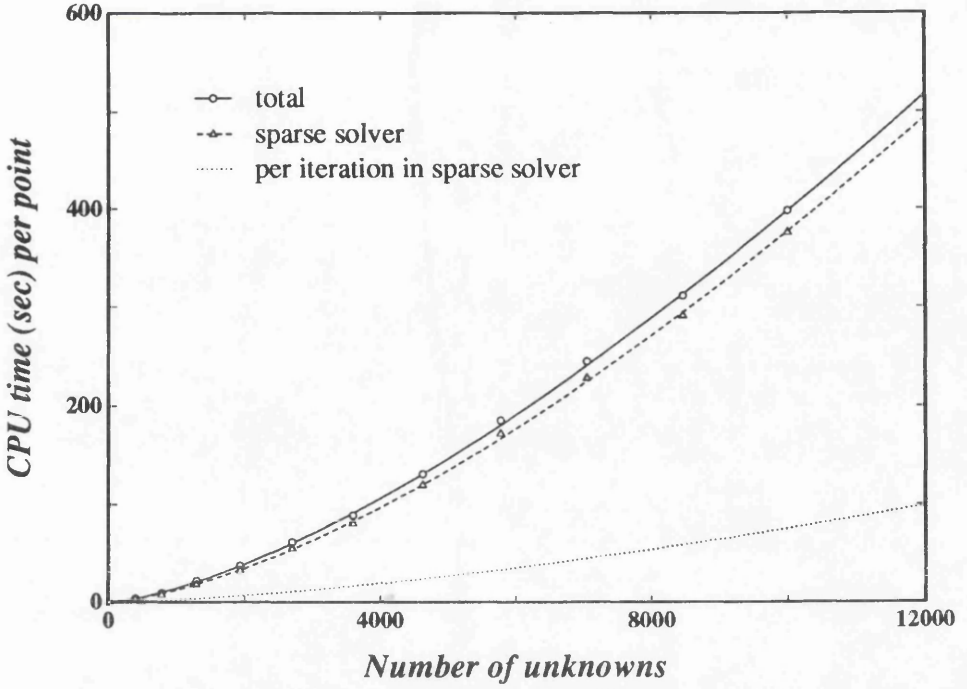


Fig. 7.27 CPU time for a lossy waveguide using the sparse solver on a SUN SPARCstation 2.

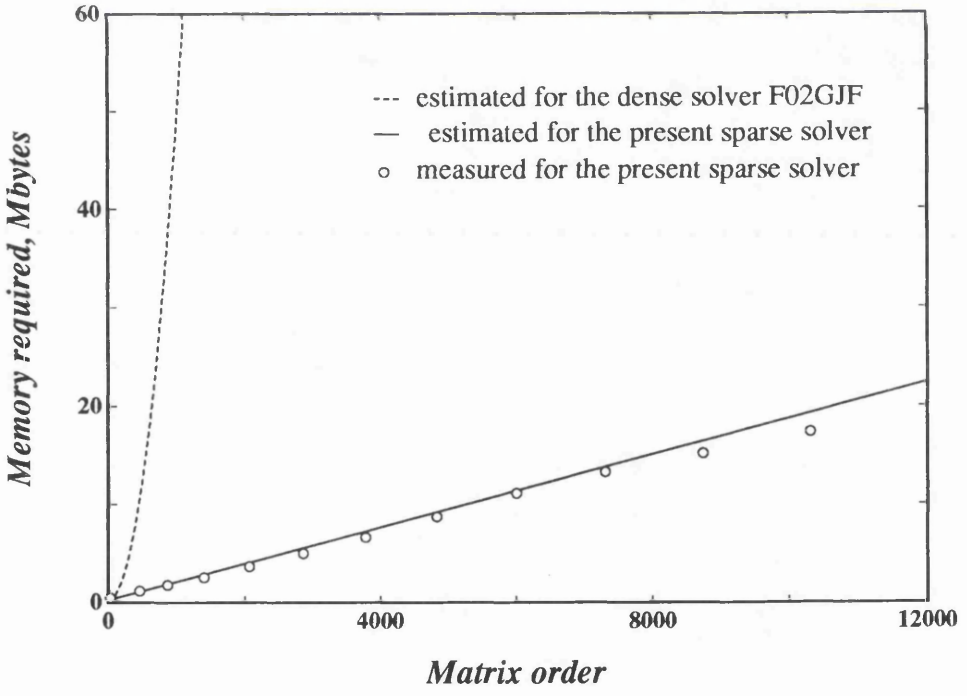


Fig. 7.28 Memory requirement of the sparse solver with a subspace of order 1 and its NAG equivalent complex dense subroutine F02GJF.

7.8 Two More Interesting Examples

7.8.1 A Lossy Optical Waveguide

In this subsection, we show the finite element solutions of a practical lossy optical waveguide structure provided by the French company *Alcatel* [132].

Fig. 7.29 shows the waveguide structure and material parameters. The losses of the P^+InP buffer layer and the N^+InP substrate need to be taken into account. In addition, the loss of the metal cladding of the waveguide has also been included, as metals are highly absorbing at the optical frequency range.

In this example, a mesh of 1760 nodal points with infinite elements on the cross-section of waveguide is used. The finite-to-infinite element border is placed at x = -2.5 μm, x = 2.5 μm, y = -1.2 μm, and y = 1.96 μm. The working wavelength λ is 1.52 μm. With the sparse solver, the computations were run on an IBM RISC 6000 workstation (see Appendix B). The cpu time is about 113 seconds for one calculation with the subspace of order 1.

Table 7.1 shows the finite element solutions of the propagation constants of the two lowest modes of the GaInAsP/InP optical waveguide shown in Fig. 7.29 in comparison with Alcatel's own calculations [132] by the *effective index method* [1]. The relative errors of the phase constants between our results and those of Alcatel are very small. The effective index method can only approximate certain types of structures, but the finite element method is much more versatile and can treat more complicated problems.

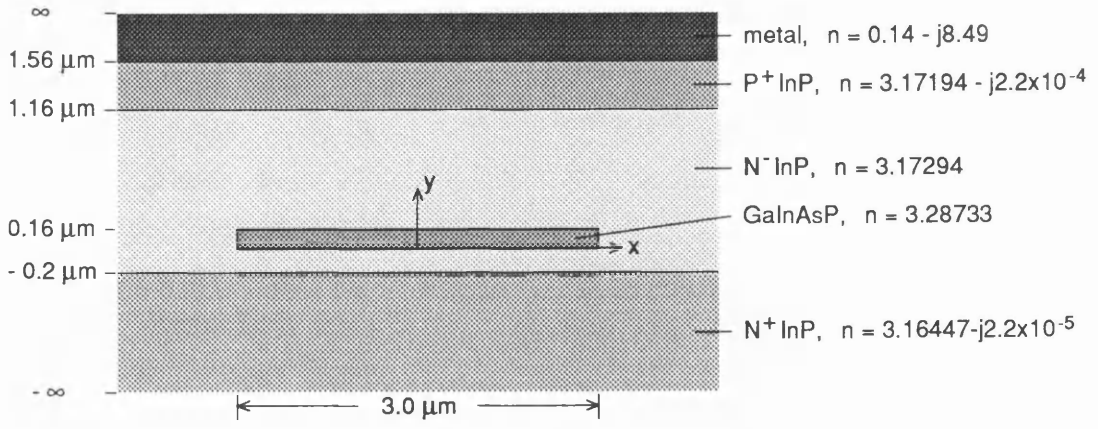


Fig. 7.29 A metal-clad buried GaInAsP/InP optical waveguide

Table 7.1 Normalized propagation constant $\gamma/k_0 = \alpha/k_0 + j\beta/k_0$

	Alcatel's calculation	Present method	relative error of β
TE mode	6.30E-6+j3.175304	3.16E-5+j3.175835	1.67E-4
TM mode	1.54E-5+j3.173879	1.29E-5+j3.173762	-3.68E-5

7.8.2 Variations of Complex Modes in lossy Waveguide

In section 7.3.2, we have shown the complex modes in lossless waveguide. For curiosity, it may be interesting to see what might happen if some small losses are introduced in the lossless waveguide supporting complex modes. Figs. 7.30 and 7.31 show the variations of complex modes and their derivative modes in lossy waveguides.

In connection with Fig. 7.5, Fig. 7.30 plots the dispersion curves of normalized propagation constant $\gamma/k_0 = (\alpha/k_0, \beta/k_0)$ against frequency f of an image waveguide of the same geometry with the waveguide in Fig. 7.5 at the frequencies around the complex mode bifurcation. Fig. 7.31 plots the dispersion curves of the square of the propagation constant, namely, the eigenvalue, on the complex plane. Fig. 7.31 gives a different view of the behaviour of complex modes and may also be helpful to understanding the variations shown in Fig. 7.30.

In Fig. 7.30, the single-dot-dashed lines indicate the complex modes (14.3 - 14.4065 GHz) and their derivative modes in the image guide with lossless material $\varepsilon = 9.0$; in Fig. 7.31, the dotted and dashed lines indicate the pair of eigenvalues representing the complex modes (between 14.0 - 14.4065 GHz), where they are exactly complex conjugate, and their derivative modes. In both Figs. 7.30 and 7.31, the two-dot-dashed and three-dot-dashed lines indicate respectively the two perturbed complex modes and their derivative modes in the image guide with a small loss material $\varepsilon = 9.0 - j\ 0.01$.

In connection with Fig. 7.5, Fig. 7.32 show the variations of propagation constant against loss (ε'') for the HE_{21} mode and the perturbed complex modes at $f = 12$ GHz, where $\varepsilon = 9.0 - j\varepsilon''$.

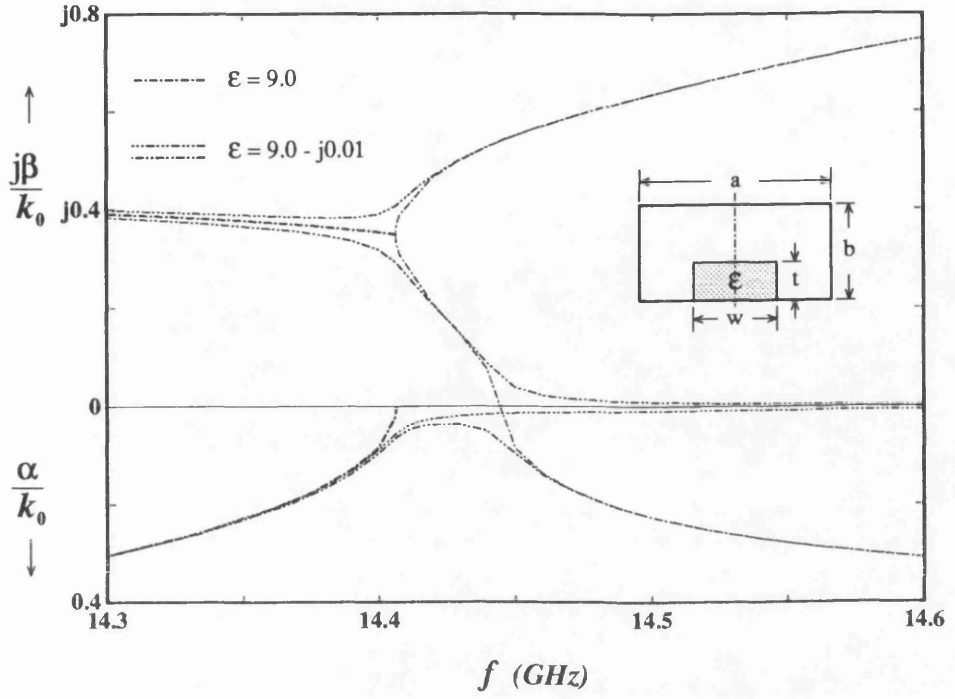


Fig. 7.30 Normalized propagation constant $\gamma/k_0 = (\alpha/k_0,\ \beta/k_0)$ versus frequency f.

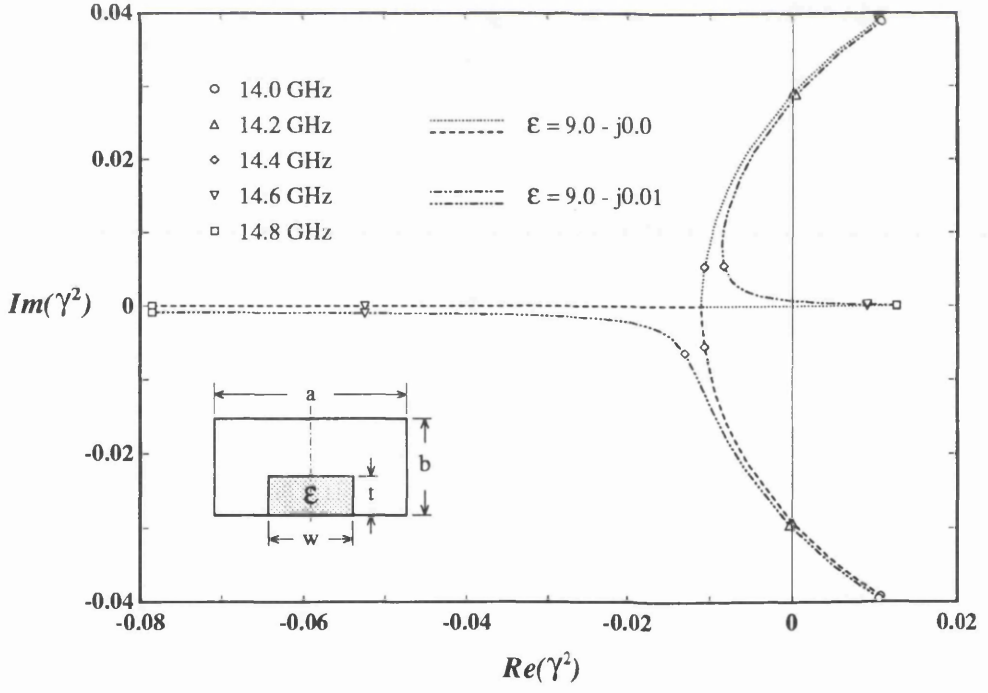
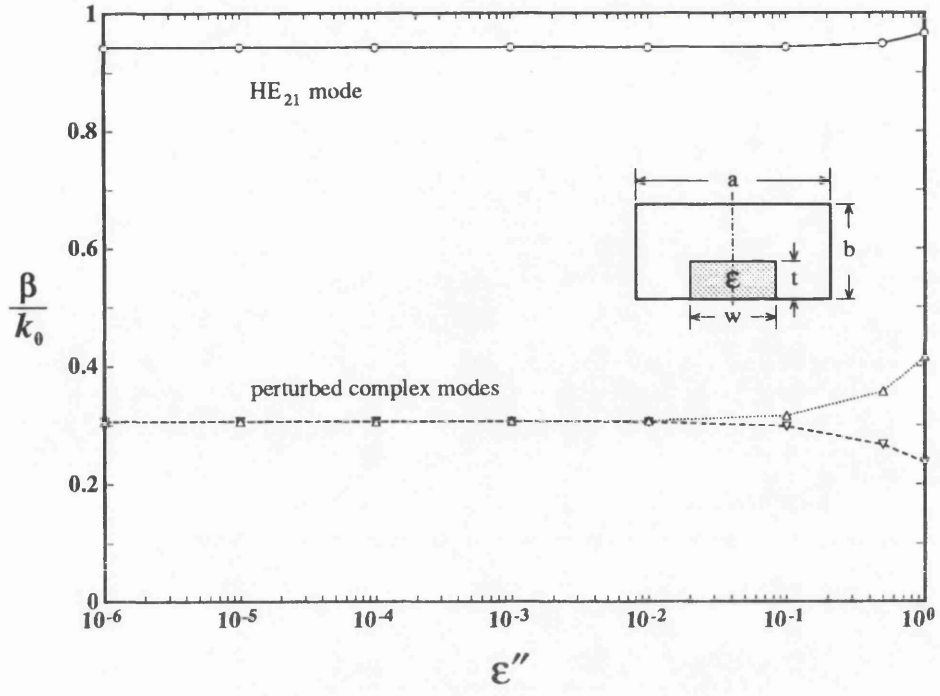
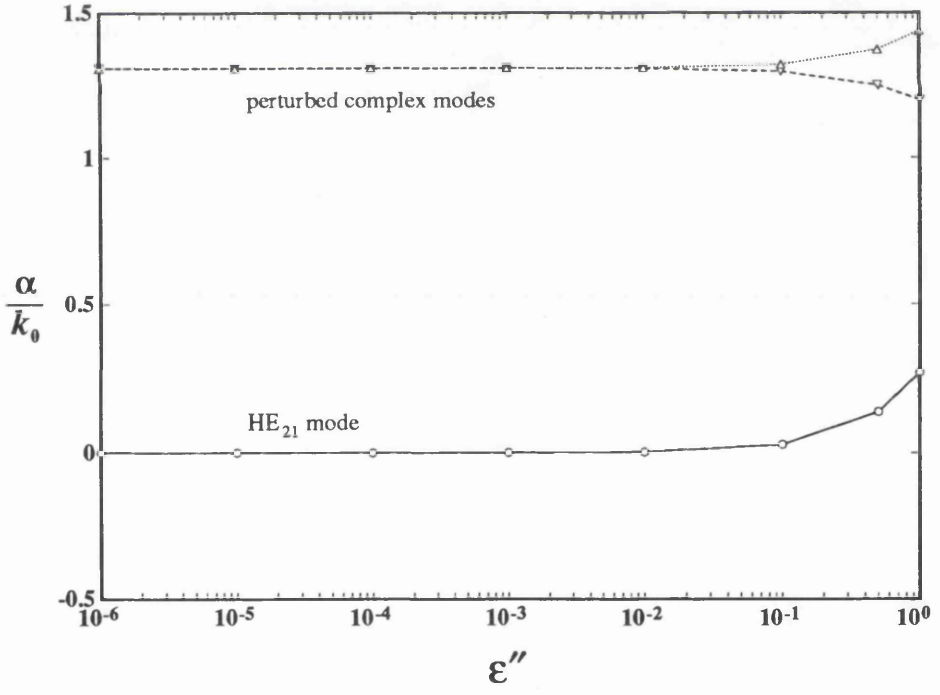


Fig. 7.31 The eigenvalues γ^2 plotted in the complex plane.



(a) Normalized phase constant.



(b) Normalized attenuation constant.

Fig. 7.32 Variations propagation constant versus loss (ε'') of the HE_{21} mode and the perturbed complex modes at f = 12 GHz.

7.9 Remarks

The computational results of all the examples are very satisfactory. No spurious solutions appear in any of the examples which cover all categories of dielectric waveguide problem — lossless isotropic, lossless anisotropic, lossy isotropic, lossy anisotropic dielectric waveguides including both closed and open structures. Furthermore, complex modes in lossless waveguide are successfully analyzed, showing the completeness of the solutions from the new formulation. Finite element solutions of complex modes have not been achieved by other finite element formulations.

Statistics of the sparse matrix eigenequation solver show its high efficiency. With this solver, one can efficiently solve problems with more than 10 thousand unknowns just on a medium-sized workstation.

The standard FORTRAN 77 software of the formulation has been run on a variety of computers from workstations to supercomputers. The computational results show good consistency

The final two examples of a practical lossy optical waveguide and the variation of complex modes in lossy waveguide provide further proofs of the applicability and effectiveness of the new finite element formulation.

CHAPTER 8

CONCLUSION

8.1 Introduction

In the final chapter, we will briefly summarize some important points of view and major achievements obtained during this study. In order to keep the clarity and completeness of this chapter, some of the statements appeared in previous chapter are restated in this chapter.

8.2 The Criteria of Judgments

In judging the appropriateness of a general finite element formulation for dielectric waveguide problems, the following 8 criteria may be adopted.

1) The formulation should be robust and capable of including as many waveguide features, such as arbitrarily shaped cross-section, inhomogeneity, anisotropy, and significant loss (or gain), as possible.
2) The formulation should not include undesirable non-physical spurious modes, and also should not miss any physical solutions such as complex modes.
3) The resultant matrix equation of the formulation should be well-conditioned (for instance, it should not breakdown due to the choice of shape function or mesh).
4) The electromagnetic fields are preferred to be represented in terms of only magnetic field **H**, because the magnetic field is continuous everywhere, and no special treatment is needed to enforce normal component of electric field at dielectric interfaces.
5) The solution should be direct for complex propagation constant in terms of specified real frequency, so as to be more efficient and reliable.
6) If possible, the formulation variables should be represented by only two field components, the least number necessary to represent a general problem, thus minimizing the unknowns.

7) The formulation should lead to a canonical eigenvalue equation which can be solved efficiently.

8) The resultant matrices of the formulation should be highly sparse able to utilize a fast and efficient matrix equation solver. This is of decisive importance for large problems, even on a supercomputer.

Criterion 1 refers to the problem coverage; criteria 2 and 3 refer to robustness and applicability; criteria 4 and 5 refer to simplicity and user friendliness; criteria 5 to 8 refer to ability to treat large problems.

8.3 Origin of Spurious Modes

Strictly speaking, the classical definitions (the two curl equations or the double-curl equation with associated tangential or normal boundary conditions) of the source-free boundary-value problems are not complete, they may only be adequate for traditional analytical solutions. For an approximate solution, the classical definition can introduce non-physical spurious solutions. The insufficient definitions of a source-free boundary-value problem are the inherent origin of spurious modes.

8.4 Elimination of Spurious Modes

For approximate solutions of Maxwell source-free boundary-value problems, the classical definitions are not sufficient. One of the following complete definitions should be adopted. They are sufficient to eliminate the spurious modes.

E-H-definition:

The equations

$$\nabla \times \mathbf{E} = -j\,\omega\,\mu_0\bar{\bar{\mu}}\cdot\mathbf{H}$$

$$\nabla \times \mathbf{H} = j\,\omega\,\varepsilon_0\bar{\bar{\varepsilon}}\cdot\mathbf{E}$$

$$\nabla \cdot (\varepsilon_0\bar{\bar{\varepsilon}}\cdot\mathbf{E}) = 0$$

$$\nabla \cdot (\mu_0\bar{\bar{\mu}}\cdot\mathbf{H}) = 0$$

The interface conditions

$$\mathbf{n} \times (\mathbf{E}_a - \mathbf{E}_b) = \mathbf{0}$$

$$\mathbf{n} \times (\mathbf{H}_a - \mathbf{H}_b) = \mathbf{0}$$

$$\mathbf{n} \cdot (\bar{\bar{\varepsilon}}_a \cdot \mathbf{E}_a - \bar{\bar{\varepsilon}}_b \cdot \mathbf{E}_b) = 0$$

$$\mathbf{n} \cdot (\bar{\bar{\mu}}_a \cdot \mathbf{H}_a - \bar{\bar{\mu}}_b \cdot \mathbf{H}_b) = 0$$

The boundary conditions

$$\mathbf{n} \times \mathbf{E} = \mathbf{0} \qquad \text{(on PEC)}$$

$$\mathbf{n} \times \mathbf{H} = \mathbf{0} \qquad \text{(on PMC)}$$

$$\mathbf{n} \cdot \bar{\bar{\mu}} \cdot \mathbf{H} = 0 \qquad \text{(on PEC)}$$

$$\mathbf{n} \cdot \bar{\bar{\varepsilon}} \cdot \mathbf{E} = 0 \qquad \text{(on PMC)} \qquad \blacksquare$$

H definition:

The equations

$$\nabla \times (\bar{\bar{\varepsilon}}^{-1} \cdot \nabla \times \mathbf{H}) - \omega^2 \varepsilon_0 \mu_0 \bar{\bar{\mu}} \cdot \mathbf{H} = \mathbf{0}$$

$$\nabla \cdot (\mu_0 \bar{\bar{\mu}} \cdot \mathbf{H}) = 0$$

The interface conditions

$$\mathbf{n} \times (\mathbf{H}_a - \mathbf{H}_b) = \mathbf{0}$$

$$\mathbf{n} \cdot (\bar{\bar{\mu}}_a \cdot \mathbf{H}_a - \bar{\bar{\mu}}_b \cdot \mathbf{H}_b) = 0$$

$$\mathbf{n} \cdot (\nabla \times \mathbf{H}_a - \nabla \times \mathbf{H}_b) = 0$$

$$\mathbf{n} \times (\bar{\bar{\varepsilon}}_a^{-1} \cdot \nabla \times \mathbf{H}_a - \bar{\bar{\varepsilon}}_b^{-1} \cdot \nabla \times \mathbf{H}_b) = \mathbf{0}$$

The boundary conditions

$$\mathbf{n} \cdot \bar{\bar{\mu}} \cdot \mathbf{H} = 0 \qquad \text{(on PEC)}$$

$$\mathbf{n} \times \mathbf{H} = \mathbf{0} \qquad \text{(on PMC)}$$

$$\mathbf{n} \times (\bar{\bar{\varepsilon}}^{-1} \cdot \nabla \times \mathbf{H}) = \mathbf{0} \qquad \text{(on PEC)}$$

$$\mathbf{n} \cdot \nabla \times \mathbf{H} = 0 \qquad \text{(on PMC)} \qquad \blacksquare$$

E definition:

The equations

$$\nabla \times (\bar{\bar{\mu}}^{-1} \cdot \nabla \times \mathbf{E}) - \omega^2 \varepsilon_0 \mu_0 \bar{\bar{\varepsilon}} \cdot \mathbf{E} = \mathbf{0}$$

$$\nabla \cdot (\varepsilon_0 \bar{\bar{\varepsilon}} \cdot \mathbf{E}) = 0$$

The interface conditions

$$\mathbf{n} \times (\mathbf{E}_a - \mathbf{E}_b) = \mathbf{0}$$

$$\mathbf{n} \cdot (\bar{\bar{\varepsilon}}_a \cdot \mathbf{E}_a - \bar{\bar{\varepsilon}}_b \cdot \mathbf{E}_b) = 0$$

$$\mathbf{n} \cdot (\nabla \times \mathbf{E}_a - \nabla \times \mathbf{E}_b) = 0$$

$$\mathbf{n} \times (\bar{\bar{\mu}}_a^{-1} \cdot \nabla \times \mathbf{E}_a - \bar{\bar{\mu}}_b^{-1} \cdot \nabla \times \mathbf{E}_b) = \mathbf{0}$$

The boundary conditions

$$\mathbf{n} \times \mathbf{E} = \mathbf{0} \qquad \text{(on PEC)}$$

$$\mathbf{n} \cdot \bar{\bar{\varepsilon}} \cdot \mathbf{E} = 0 \qquad \text{(on PMC)}$$

$$\mathbf{n} \cdot \nabla \times \mathbf{E} = 0 \qquad \text{(on PEC)}$$

$$\mathbf{n} \times (\bar{\bar{\mu}}^{-1} \cdot \nabla \times \mathbf{E}) = \mathbf{0} \qquad \text{(on PMC)} \qquad \blacksquare$$

Remarks:

For *only* **H** approximation under the **H**-definition or only **E** approximation under the **E**-definition, the interface and boundary conditions with the curl operator may not have all to be imposed, only the tangential and normal conditions of the field are essential and have to be imposed. ■

8.5 The Variational Finite Element Formulation

In this study, a variational finite element formulation for the full-wave analysis of microwave and optical waveguide problems with arbitrary cross section and inhomogeneous, transverse-anisotropic, and lossy dielectrics have been derived so as to eliminate the spurious modes in finite element solutions of dielectric waveguide problems.

The computational results of implementing the formulation are very satisfactory. No spurious solution appear in any of the examples. The coverage of examples is quite wide. The examples cover lossless isotropic, lossless anisotropic, lossy isotropic, lossy anisotropic dielectrics as well as both closed and open structures. Furthermore, solutions for complex modes in lossless waveguide are successfully achieved, showing the completeness of the solutions of the method.

Summarizing the study and computational results, this formulation has the following features:

1) it can treat a wide range of dielectric waveguide problems with arbitrarily-shaped cross section, inhomogeneity, transverse-anisotropy, and loss (or gain);

2) it totally eliminates troublesome non-physical spurious solutions which ordinarily appear interspersed with the correct results in many other vectorial finite element solutions;
3) it allows direct solution for propagation constant at a specified frequency (rather than for the frequency at a specified propagation constant as in the usual approaches);
4) the numerical efficiency of solution is maximized since this formulation uses only two magnetic field components, this being achieved without losing the matrix sparsity which only depends on the topology of the mesh used. This property is of decisive importance for solving large-size problems;
5) it provides the capability to compute complex modes in lossless waveguide, showing the completeness of the solutions;

8.6 The Sparse Matrix Eigenequation Solver

An efficient matrix solver for large, sparse, non-symmetric (real or complex) matrix eigenequations has been especially developed by Zhu for this work [136].

Statistics of the sparse matrix eigenequation solver show its high efficiency and drastically drop of computing time and memory requirement comparing to the only available standard library subroutines.

The use of a 'shift' in the sparse solver makes it possible to concentrate on a particular (dominant or higher) mode with the minimum computing cost to achieve best results. With this solver, one can efficiently solve problems with more than 10 thousand complex unknowns on a medium-sized workstation.

The sparse matrix solver, developed for this study, is apparently the unique efficient solver to solve large, sparse, non-symmetric, and complex matrix eigenequation. No other solver with similar functions is available or has been reported.

8.7 Concluding Remarks

There is a pressing need for robust and numerically efficient computer simulation techniques for analysis and design of optical and microwave waveguides with arbitrary dielectric profile and arbitrary cross-section.

Facilities are needed for considering loss (or gain), arbitrary polarization and without troublesome spurious modes. The objectives of this study are to develop a method to satisfy the need.

In this study, all the original objectives have been achieved. A effective finite element method for dielectric waveguide problems with arbitrary cross-section, inhomogeneity, transversely-anisotropy, and loss (or gain) has been developed. Our method shows considerable advances over the state-of-the-art methods. Using this method, one can establish more realistic models of real waveguides. This method provides a powerful tool to analyze more complicated waveguiding structures in optoelectronics and microwaves.

This study provides the first finite element solutions of complex modes in lossless waveguides and the variations of complex modes in lossy waveguides, and the first finite element solutions of (large) practical lossy waveguide problems. This method fills in the gap of adequate treatment of lossy waveguide.

Our finite element method together with our unique efficient sparse matrix solver provides probably the only approach in the world for some complicated waveguide problems.

APPENDIX A

ADJOINTNESSES OF THE DIFFERENTIAL OPERATORS

This appendix gives the details of investigating the adjointnesses of the differential operators in Eq. (5.1) (in chapter 5). To begin with, we rewrite Eq. (5.1) as follows:

$$\mathbb{A}_1 \mathbf{H}_t + \mathbb{A}_2 \mathbf{H}_t + \mathbb{A}_3 \mathbf{H}_t + \gamma^2 \mathbb{B} \mathbf{H}_t = \mathbf{0} \tag{A.1}$$

where each of the individual differential operators are expressed as:

$$\mathbb{A}_1 \_ = \nabla_t \times (\kappa_{zz} \nabla_t \times \_) \tag{A.2a}$$

$$\mathbb{A}_2 \_ = - \mathbf{z} \times [\bar{\bar{\kappa}}_{tt} \cdot \nabla_t \times (\mathbf{z} \nabla \cdot \_)] \tag{A.2b}$$

$$\mathbb{A}_3 \_ = - \omega^2 \mu_0 \varepsilon_0 \_ \tag{A.2c}$$

$$\mathbb{B} \_ = \mathbf{z} \times [\bar{\bar{\kappa}}_{tt} \cdot (\mathbf{z} \times \_)] \tag{A.2d}$$

The adjoint operator of $\mathbb{L}$ is an operator $\mathbb{L}^a$ such that

$$\langle \mathbf{H}_t^a, \mathbb{L} \mathbf{H}_t \rangle = \langle \mathbb{L}^a \mathbf{H}_t^a, \mathbf{H}_t \rangle + \text{b.t.} \tag{A.3}$$

where 'b.t.' stands for boundary terms, $\langle \cdot , \cdot \rangle$ is a inner product.

An operator $\mathbb{L}$ is a self-adjoint operator if $\mathbb{L}^a = \mathbb{L}$.

After defining a real inner product for our problem as follows:

$$\langle \mathbf{A}_t, \mathbf{B}_t \rangle = \int_S \mathbf{A}_t \cdot \mathbf{B}_t \, ds \tag{A.4}$$

where S is the cross-section of the waveguide, we investigate each term of Eq. (A.1) individually.

The 1st term:

$$\langle \mathbf{H}_t^a , \mathbb{A}_1 \mathbf{H}_t \rangle = \langle \mathbf{H}_t^a , \nabla_t \times (\kappa_{zz} \nabla_t \times \mathbf{H}_t) \rangle$$

$$= \langle \nabla_t \times \mathbf{H}_t^a \, , \, \kappa_{zz} \nabla_t \times \mathbf{H}_t \rangle + C_1 = \langle \kappa_{zz} \nabla_t \times \mathbf{H}_t^a \, , \, \nabla_t \times \mathbf{H}_t \rangle + C_1$$

$$= \langle \nabla_t \times (\kappa_{zz} \nabla_t \times \mathbf{H}_t^a) \, , \, \mathbf{H}_t \rangle + C_1 + C_2$$

$$= \langle \mathbb{A}_1 \mathbf{H}_t^a \, , \, \mathbf{H}_t \rangle + \text{b.t.} \qquad \text{(A.5)}$$

where

$$C_1 = \int_C [(\kappa_{zz} \nabla_t \times \mathbf{H}_t) \times \mathbf{H}_t^a] \cdot \mathbf{n} \, dl$$

$$= j \omega \varepsilon_0 \int_C (\mathbf{E}_z \times \mathbf{H}_t^a) \cdot \mathbf{n} \, dl$$

$$= j \omega \varepsilon_0 \int_C (\underline{\mathbf{H}_t^a \times \mathbf{n}}) \cdot \mathbf{E}_z \, dl \longrightarrow 0 \quad \text{on PMC}$$

$$= j \omega \varepsilon_0 \int_C (\mathbf{n} \times \underline{\mathbf{E}_z}) \cdot \mathbf{H}_t^a \, dl \longrightarrow 0 \quad \text{on PEC}$$

$$= 0 \quad \text{(on PEC and PMC)} \qquad \text{(A.6)}$$

and

$$C_2 = \int_C [\mathbf{H}_t \times (\kappa_{zz} \nabla_t \times \mathbf{H}_t^a)] \cdot \mathbf{n} \, dl$$

$$= j \omega \varepsilon_0 \int_C (\mathbf{H}_t \times \mathbf{E}_z^a) \cdot \mathbf{n} \, dl$$

$$= j \omega \varepsilon_0 \int_C (\underline{\mathbf{n} \times \mathbf{H}_t}) \cdot \mathbf{E}_z^a \, dl \longrightarrow 0 \quad \text{on PMC}$$

$$= j \omega \varepsilon_0 \int_C (\underline{\mathbf{E}_z^a \times \mathbf{n}}) \cdot \mathbf{H}_t \, dl \longrightarrow 0 \quad \text{on PEC}$$

$$= 0 \quad \text{(on PEC and PMC)} \qquad \text{(A.7)}$$

The 2nd term:

$$\langle \mathbf{H}_t^a \, , \, \mathbb{A}_2 \mathbf{H}_t \rangle = \langle \mathbf{H}_t^a \, , \, - \mathbf{z} \times [\bar{\bar{\kappa}}_{tt} \cdot \nabla_t \times (\mathbf{z} \nabla_t \cdot \mathbf{H}_t)] \rangle$$

$$= \langle \mathbf{z} \times \mathbf{H}_t^a \, , \, \bar{\bar{\kappa}}_{tt} \cdot \nabla_t \times (\mathbf{z} \nabla_t \cdot \mathbf{H}_t) \rangle$$

$$= \langle \bar{\bar{\kappa}}_{tt}^{\,T} \cdot (\mathbf{z} \times \mathbf{H}_t^a) \, , \, \nabla_t \times (\mathbf{z} \nabla_t \cdot \mathbf{H}_t) \rangle$$

$$= \langle \nabla_t \times [\bar{\bar{\kappa}}_{tt}^{\,T} \cdot (\mathbf{z} \times \mathbf{H}_t^a)] \, , \, \mathbf{z} \nabla_t \cdot \mathbf{H}_t \rangle + C_3$$

$$= \langle \mathbf{z} \cdot \nabla_t \times [\bar{\bar{\kappa}}_{tt}^{\,T} \cdot (\mathbf{z} \times \mathbf{H}_t^a)] \, , \, \nabla_t \cdot \mathbf{H}_t \rangle + C_3$$

$$= <- \nabla_t \{ \mathbf{z} \cdot \nabla_t \times [\overline{\overline{\kappa}}_{tt}^{T} \cdot (\mathbf{z} \times \mathbf{H}_t^a)] \} , \mathbf{H}_t > + C_3 + C_4$$

$$\neq <\mathbb{A}_2 \mathbf{H}_t^a, \mathbf{H}_t > + \text{b.t.} \qquad \text{(A.8)}$$

where

$$C_3 = \int_C (\nabla_t \cdot \mathbf{H}_t) \{ \mathbf{z} \times [\overline{\overline{\kappa}}_{tt}^{T} \cdot (\mathbf{z} \times \mathbf{H}_t^a)] \} \cdot \mathbf{n} \, dl$$

$$= 0 \quad \text{(on PMC)} \qquad \text{(A.9)}$$

and

$$C_4 = \int_C \{ \mathbf{z} \cdot \nabla_t \times [\overline{\overline{\kappa}}_{tt}^{T} \cdot (\mathbf{z} \times \mathbf{H}_t^a)] \} (\mathbf{H}_t^a \cdot \mathbf{n}) \, dl$$

$$= 0 \quad \text{(on PEC)} \qquad \text{(A.10)}$$

The 3rd term:

$$<\mathbf{H}_t^a , \mathbb{A}_3 \mathbf{H}_t > = <\mathbf{H}_t^a , - \omega^2 \mu_0 \varepsilon_0 \mathbf{H}_t > = <- \omega^2 \mu_0 \varepsilon_0 \mathbf{H}_t^a , \mathbf{H}_t >$$

$$= <\mathbb{A}_3 \mathbf{H}_t^a , \mathbf{H}_t > \qquad \text{(A.11)}$$

The 4th term:

$$<\mathbf{H}_t^a , \mathbb{B}_1 \mathbf{H}_t > = <\mathbf{H}_t^a , \mathbf{z} \times [\overline{\overline{\kappa}}_{tt} \cdot (\mathbf{z} \times \mathbf{H}_t)] >$$

$$= - <\mathbf{z} \times \mathbf{H}_t^a , \overline{\overline{\kappa}}_{tt} \cdot (\mathbf{z} \times \mathbf{H}_t) >$$

$$= - <\overline{\overline{\kappa}}_{tt}^{T} \cdot (\mathbf{z} \times \mathbf{H}_t^a) , \mathbf{z} \times \mathbf{H}_t >$$

$$= <\mathbf{z} \times [\overline{\overline{\kappa}}_{tt}^{T} \cdot (\mathbf{z} \times \mathbf{H}_t^a)] , \mathbf{H}_t >$$

$$= <\mathbb{B}_1 \mathbf{H}_t^a , \mathbf{H}_t > \quad (\text{if } \overline{\overline{\kappa}}_{tt}^{T} = \overline{\overline{\kappa}}_{tt}) \qquad \text{(A.12)}$$

Because the permittivity tensor is assumed symmetric (see Eq. (3.1) in chapter 3), we know from (A.5), (A.8), (A.11) and (A.12) that operators, $\mathbb{A}_1$, $\mathbb{A}_3$, and $\mathbb{B}$ are self-adjoint, and only $\mathbb{A}_2$ is non-self-adjoint.

APPENDIX B

LIST OF COMPUTERS USED FOR THIS STUDY

During the last three years of my Ph.D. study, there were two replacements of computers at University College Computer Centre (UCCC) and two replacement of computers at University of London Computer Centre (ULCC). The frequent replacements caused some disruptions to the research, but, on the other hand, the author has had the chance to experience a variety of computers. The computers used for the study are listed in the table below.

computer model **location** **period available to the author**	**dependence** ***** most * least	**operating system** **quoted speed** **memory capacity**
GEC EUCLID 4000 UCCC 12/1988 to 06/1990	*	OS4000 — 1.8 Mbytes
PYRAMID 98x UCCC 05/1990 to 09/1991	*	OSx — —
SUN SPARC 2 workstation UCCC 12/1990 to 09/1991	*****	SunOS 4.2 MFLOPS 56 Mbytes
IBM RISC 6000-540 workstation UCCC 07/1991 to now	***	AIX 13.7 MFLOPS 128 Mbytes
AMDAHL 5890-300 large computer ULCC 12/1989 to 09/1991	****	MVS/XA — 32 Mbytes
CRAY X-MP/28 supercomputer ULCC 12/1990 to 09/1991	**	UNICOS < 500 MFLOPS 64 Mbytes
CONVEX C210 supercomputer ULCC 08/1991 to now	*	ConvexOS < 500 MFLOPS 256 Mbytes

Notes:
(1) MFLOPS : mega floating-point operation per second.
(2) The OSx, SunOS, AIX, UNICOS, and ConvexOS are all UNIX-like operating systems.
(3) CRAY gives 64-bit precision for REAL and INTEGER, and 128-bit precision for DOUBLE PRECISION. While other machines usually give 32-bit precision for REAL, 16-bit or 32-bit precision for INTEGER, and 64-bit precision for DOUBLE PRECISION.
(4) CONVEX C210 will soon be replaced by a CONVEX C3840 supercomputer with the memory capacity of 1 gigabytes.

REFERENCES

[1] ***Nishihara H., Haruna M, Suhara T.***, *Optical Integrated Circuits.* New York: McGraw-Hill, 1989.

[2] ***Alferness R.C.*** (ed.), Special Section on Integrated Optics and Optoelectronics, *Proc. IEEE*, Vol. 75, No. 11, pp. 1472-1524, 1987.

[3] ***Wolf H.F***, *Handbook of Fibre Optics: Theory and Applications*, London: Granada, 1979.

[4] ***Wada O., Sanada T., Kuno M., Fujii T.***, "Very low threshold current ridge-waveguide AlGaAs/GaAs single-quantum-well lasers," *Electron. Lett.*, Vol. 21, No. 22, pp. 1025-1026, Oct. 1985.

[5] ***Garrett B., Glew R.W.***, "Low-threshold, high-power zero-order lateral-mode DQW-SCH metal-clad ridge waveguide (AlGa)As/GaAs lasers," *Electron. Lett.*, Vol. 23, No. 8, pp. 373-374, Apr. 1987.

[6] ***Koch T.B.***, *Computation of Wave Propagation in Integrated Optical Devices*, Ph.D. Thesis, University College London, Dec. 1989.

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

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

[9] ***Silvester P.P., Ferrari R.L.***, *Finite Elements for Electrical Engineers,* 2nd. ed. Cambridge, England: Cambridge University Press, 1990.

[10] ***Desai C.S., Abel J.F.***, *Introduction to the Finite Element Method*, New York: Van Nostrand Reinhold, 1972.

[11] A Guide to Engineering Workstations, *IEEE Spectrum*, Apr. 1991.

[12] ***Rodgers P.***, "En route to Europe's teraflops," *Physics World*, Vol. 4, No. 2, pp. 13-14, Feb. 1991

[13] ***Wang R., Denerdash N.A.***, "A Combined Vector Potential-Scaler Potential Method for FE Computation of 3D Magnetic Fields in Electrical Devices with Iron Cores," *Digest of the Fourth Biennial IEEE Conference on Electromagnetic Field Computation*, CA-03,

Toronto, 22-24 Oct. 1990.

[14] A Special Guide to Data Communications — High-Speed Networks and Interconnection, *IEEE Spectrum*, Aug. 1991.

[15] ***Collin R.E.***, *Field Theory of Guided Waves*. New York: McGraw-Hill, 1960.

[16] ***Goell J.E.***, "A circular-harmonic computer analysis of rectangular dielectric waveguides," *Bell Syst. Tech. J.*, Vol. 48, pp. 2133-2160, Sept. 1969.

[17] ***Solbach K., Wolff I***, "The electromagnetic fields and the phase constants of dielectric image lines," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-26, No. 4, pp. 266-274, Apr. 1974.

[18] ***Strube J, Arndt F,*** "Rigorous hybrid-mode analysis of the transition from rectangular waveguide to shielded dielectric image guide," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-33, No. 5, pp. 391-400, May 1985.

[19] ***Clarricoats P.J.B., Slinn K.R.,*** "Complex modes of propagation in dielectric loaded circular waveguide," *Electron. Lett.*, Vol. 1, No. 5, pp. 145-146, Jul. 1965.

[20] ***Davies J.B.***, "Review of methods for numerical solution of the hollow-waveguide problem," *Proc. IEE,* Vol. 119, No.1, pp. 33-37, Jan. 1972.

[21] ***Ng F.L.***, "Tabulation of methods for the numerical solution of the hollow waveguide problem," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-22, No. 3, pp. 322-329, Mar. 1974.

[22] ***Saad S.M.***, "Review of numerical methods for the analysis of arbitrarily-shaped microwave and optical dielectric waveguides," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-33, No. 10, pp. 894-899, Oct. 1985.

[23] ***Koshiba M., Hayata K., Suzuki M.***, "Finite-element method analysis of microwave and optical waveguides - Trends in countermeasures to spurious solutions," *Electronics and Communication in Japan, Part 2*, Vol. 70, No. 9, pp. 96-108, 1987.

[24] ***Rahman B.M.A., Fernandez F.A., Davies J.B.***, "Review of finite element methods for microwave and optical waveguides," *Proc. IEEE*, to be published.

[25] ***Arlett P.L., Bahrani A.K., Zienkiewcz***, "Applications of finite elements to the solution of Helmholtz's equation," *Proc. IEE*, Vol. 115, No. 12, pp. 1762-1766, Dec. 1968.

[26] ***Silvester P.***, "A general high-order finite-element waveguide analysis program," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-17, No. 4, pp. 204-210, Apr. 1969.

[27] ***Daly P.***, "Polar geometry waveguides by finite-element methods," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-22, No. 3, pp. 202-209, Mar. 1974.

[28] ***Daly P.***, "Finite element approach to propagation in elliptical and parabolic waveguides," *International Journal for Numerical Methods in Engineering*, Vol. 20, pp. 681-688, 1984.

[29] ***Koshiba M., Hayata K., Suzuki M.***, "Approximate scalar finite-element analysis of anisotropic optical waveguides with off-diagonal elements in a permittivity tensor," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-32, No. 6, pp. 587-593, Jun. 1984.

[30] ***Chiang K.S.***, "Finite-element analysis of optical fibres with iterative treatment of the infinite 2-D space," *Opt. Quantum Electron.*, Vol. 17, No. 6, pp. 381-391, Nov. 1985.

[31] ***Wu R.B., Chen C.H.***, "A scalar variational conformal mapping technique for weakly guiding dielectric waveguides," *IEEE J. Quantum Electron.*, Vol. QE-22, No. 5, pp. 603-609, May 1986.

[32] ***Koshiba M., Hayata K., Suzuki M.***, "Approximate scalar finite-element analysis of anisotropic optical waveguides," *Electron. Lett.*, Vol. 18, No. 10, pp. 411-413, May 1982.

[33] ***Mustacich R.V.***, "Scalar finite element analysis of electrooptic modulation in diffused channel waveguides and poled waveguides in polymer thin films," *Appl. Opt.*, Vol. 27, No. 17, pp. 3732-3737, Sept. 1988.

[34] ***Berk A.D.***, "Variational principles for electromagnetic resonators and waveguides," *IRE Trans. Antennas Propagat.*, Vol. AP-4, pp. 104-111, Apr. 1956.

[35] ***Morishita K., Kumagai N.***, "Unified approach to the derivation of variational expression for electromagnetic fields," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-25, No. 1, pp. 34-40, Jan. 1977.

[36] ***Morishita K., Kumagai N.***, "Systematic derivation of variational expressions for electromagnetic and/or acoustic waves," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-26, No. 9, pp. 684-689, Sept. 1978.

[37] ***Jeng G., Wexler A.***, "Self-adjoint variational formulation of problems having non-self-adjoint operators," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-26, No. 2, pp. 91-94, Feb. 1978.

[38] ***Chen C.H., Lien C.D.***, "The variational principle for non-self-adjoint electromagnetic problems," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-28, No. 8, pp. 878-886, Aug. 1980.

[39] ***Ahmed S., Daly P.***, "Finite-element methods for inhomogeneous waveguides," *Proc. IEE*, Vol. 116, No. 10, pp. 1661-1664, Oct. 1969.

[40] ***Csendes Z.J., Silvester P.***, "Numerical solution of dielectric loaded waveguides: I- Finite-element analysis," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-18, No. 12, pp. 1124-1131, Dec. 1970.

[41] ***Daly P.***, "Hybrid-mode analysis of microstrip by finite-element methods," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-19, No. 1, pp. 19-25, Jan. 1971.

[42] ***Corr D.G., Davies J.B.***, "Computer analysis of the fundamental and higher order modes in single and coupled microstrip," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-20, No. 10, pp. 669-678, Oct. 1972.

[43] ***Aubourg M., Villotte J.P., Godon F., Garault Y.***, "Finite element analysis of lossy waveguides - Application to microstrip lines on semiconductor substrate," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-31, No. 4, pp. 326-331, Apr. 1983.

[44] ***Tzuang C.-K., Itoh T.***, "Finite-element analysis of slow-wave Schottky contact printed lines," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-34, No. 12, pp. 1483-1489, Dec. 1986.

[45] ***Eswarappa, G.I. Costache, Hoefer W.J.R.***, "Finlines in rectangular and circular waveguide housings including substrate mounting and bending effects - finite element analysis," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-37, No. 2, pp. 299-306, Feb. 1989.

[46] ***Gibson A.A.P., Helszajn J.***, "Finite element solution of longitudinally magnetized elliptical gyromagnetic waveguides,"

IEEE Trans. Microwave Theory Tech., Vol. MTT-37, No. 6, pp. 999-1005, Jun. 1989.

[47] ***Yeh C., Dong S.B., Oliver W.***, "Arbitrarily shaped inhomogeneous optical fiber or integrated optical waveguides," *J. Appl. Phys.*, Vol. 46, No. 5, pp. 2125-2129, May 1975.

[48] ***Bird T.S.***, "Propagation and radiation characteristics of rib waveguide," *Electron. Lett.*, Vol. 13, No. 14, pp. 401-403, Jul. 1977.

[49] ***Okamoto K., Okoshi T.***, "Vectorial wave analysis of inhomogeneous optical fibres using finite element method," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-26, No. 2, pp. 109-114, Feb. 1978.

[50] ***Yeh C., Ha K., Dong S.B., Brown W.P.***, "Single-mode optical waveguides," *Appl. Opt.*, Vol. 18, pp. 1490-1504, May 1979.

[51] ***Ikeuchi M., Sawami H., Niki H.***, "Analysis of open-type dielectric waveguides by the finite-element iterative method," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-29, No. 3, pp. 234-239, Mar. 1981.

[52] ***Mabaya N., Lagasse P.E., Vandenbulcke P.***, "Finite element analysis of optical waveguides," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-29, No. 6, pp. 600-605, Jun. 1981.

[53] ***Oyamada K., Okoshi T.***, "Two-dimensional finite-element calculation of propagation characteristics of axially nonsymmetrical optical fibres," *Radio Sci.*, Vol. 17, No. 1, pp. 109-116, Jan. 1982.

[54] ***Welt D., Webb J.***, "Finite-element analysis of dielectric waveguides with curved boundaries," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-33, No. 7, pp. 576-585, Jul. 1985.

[55] ***Vandenbulcke P., Lagasse P.E.***, "Eigenmode analysis of anisotropic optical fibres or integrated optical waveguides," *Electron. Lett.*, Vol. 12, No. 5, pp. 120-122, Mar. 1976.

[56] ***McAulay A.D.***, "Variational finite-element solution for dissipative waveguides and transportation application," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-25, No. 5, pp. 382-392, May 1977.

[57] ***English W.J.***, "Vector variational solutions of inhomogeneously loaded cylindrical waveguide structures," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-19, No. 1, pp. 9-18, Jan. 1971.

[58] ***English W.J., Young F.J.***, "An *E* vector variational formulation of

the Maxwell equations for cylindrical waveguide problems," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-19, No. 1, pp. 40-46, Jan. 1971.

[59] ***Konrad A.***, "Vector variational formulation of electromagnetic fields in anisotropic media," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-24, No. 9, pp. 553-559, Sept. 1976.

[60] ***Konrad A.***, "Higher-order triangular finite elements for electromagnetic waves in anisotropic media," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-25, No. 5, pp. 353-360, May 1977.

[61] ***Katz J.***, "Novel solution of 2-D waveguides using the finite element method," *Appl. Opt.*, Vol. 21, No. 15, pp. 2747-2750, Aug. 1982.

[62] ***Davies J.B., Fernandez F.A., Philippou G.Y.***, "Finite element analysis of all modes in cavities with circular symmetry," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-30, No. 11, pp. 1975-1980, Nov. 1982.

[63] ***Rahman B.M.A., Davies J.B.***, "Finite-element analysis of optical and microwave waveguide problems," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-32, No. 1, pp. 20-28, Jan. 1984.

[64] ***Hara M., Wada T., Fukasawa T., Kikuchi F.***, "A three dimensional analysis of RF electromagnetic fields by the finite element method," *IEEE Trans. Magnetics*, Vol. MAG-19, No. 6, pp. 2417-2420, Nov. 1983.

[65] ***Rahman B.M.A., Davies J.B.***, "Penalty function improvement of waveguide solution by finite elements," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-32, No. 8, pp. 922-928, Aug. 1984.

[66] ***Hano M.***, "Finite-element analysis of dielectric-loaded waveguides," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-32, No. 10, pp. 1275-1279, Oct. 1984.

[67] ***Koshiba M., Hayata K., Suzuki M.***, "Improved finite-element formulation in terms of the magnetic field vector for dielectric waveguides," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-33, No. 3, pp. 227-233, Mar. 1985.

[68] ***Koshiba M., Hayata K., Suzuki M.***, "Finite-element formulation in terms of the electric-field vector for electromagnetic waveguide problems," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-33,

No. 10, pp. 900-905, Oct. 1985.

[69] **Rahman B.M.A., Davies J.B.**, "Finite-element solution of integrated optical waveguides," *J. Lightwave Technol.*, Vol. LT-2, No. 5, pp. 682-688, Oct. 1984.

[70] **Rahman B.M.A., Davies J.B.**, "Vector-H finite element solution of GaAs/GaAlAs rib waveguides," *IEE Proc., Pt. J*, Vol. 132, No. 6, pp. 349-353, Dec. 1985.

[71] **Koshiba M., Hayata K., Suzuki M.**, "Vector *E*-field finite element analysis of dielectric optical waveguides," *Appl. Opt.*, Vol. 25, No. 1, pp. 10-11, Jan. 1986.

[72] **Koshiba M., Hayata K., Suzuki M.**, "Finite-element solution of anisotropic waveguides with arbitrary tensor permittivity," *J. Lightwave Technol.*, Vol. LT-4, No. 2, pp. 121-126, Feb. 1986.

[73] **Hayata K., Eguchi M., Koshiba M., Suzuki M.**, "Vectorial wave analysis of stress-applied polarization-maintaining optical fibers by the finite-element method," *J. Lightwave Technol.*, Vol. LT-4, No. 2, pp. 133-139, Feb. 1986.

[74] **Hayata K., Koshiba M., Suzuki M.**, "Lateral mode analysis of buried heterostructure diode lasers by the finite-element method," *IEEE J. Quantum Electron.*, Vol. QE-22, No. 6, pp. 781-788, Jun. 1986.

[75] **Hayata K., Eguchi M., Koshiba M., Suzuki M.**, "Vectorial wave analysis of side-tunnel type polarization-maintaining optical fibers by variational finite elements," *J. Lightwave Technol.*, Vol. LT-4, No. 8, pp. 1090-1096, Aug. 1986.

[76] **Young T.P.**, "Design of integrated optical circuits using finite elements," *IEE Proc. Pt.A*, Vol. 135, pp. 135-144, Mar. 1988.

[77] **Fernandez F.A.**, *Finite Element Analysis of Rotationally Symmetric Electromagnetic Cavities*, Ph.D. Thesis, University College London, Sept. 1981.

[78] **Cvetkovic S.R., Davies J.B.**, "Self-adjoint vector variational formulation for lossy anisotropic dielectric waveguide," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-34, No.1, pp. 129-134, Jan. 1986.

[79] **Cvetkovic S.R.**, *Finite Element Analysis of Lossy Dielectric Waveguides Based on Variational Principles*, Ph.D. thesis,

University College London, Jul. 1987.

[80] ***Kobelansky A.J., Webb J.P.***, "Eliminating spurious modes in finite-element waveguide problems by using divergence-free fields," *Electron. Lett.* Vol. 22, No. 11, pp. 569-570, May 1986.

[81] ***Hayata K., Koshiba M., Eguchi M., Suzuki M.***, "Vectorial finite-element method without any spurious solution for dielectric waveguiding problems using transverse magnetic-field component," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-34, No. 11, pp. 1120-1124, Nov. 1986.

[82] ***Su C.-C.***, "Origin of spurious modes in the analysis of optical fibre using the finite-element or finite-difference technique," *Electron. Lett.*, Vol. 21, No. 19, pp. 858-860, Sept. 1985.

[83] ***Su C.-C.***, "A combined method for dielectric waveguides using the finite-element technique and the surface integral equations method," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-34, No. 11, pp. 1140-1146, Nov. 1986.

[84] ***Angkaew T., Matsuhara M., Kumagai N.***, "Finite-element analysis of waveguide modes: A novel approach that eliminates spurious modes," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-35, No. 2, pp. 117-123, Feb. 1987.

[85] ***Hayata K., Miura K., Koshiba M.***, "Finite-element formulation for lossy waveguides," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-36, No. 2, pp. 268-276, Feb. 1988.

[86] ***Chew W.C., Nasir M.A.***, "A variational analysis of anisotropic, inhomogeneous dielectric waveguides," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-37, No. 4, pp. 661-668, Apr. 1989.

[87] ***Svedin J.A.M.***, "A numerically efficient finite-element formulation for the general waveguide problem without spurious modes," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-37, No. 11, pp. 1708-1715, Nov. 1989

[88] ***Svedin J.A.M.***, "A modified finite-element method for dielectric waveguides using an asymptotically correct approximation on infinite elements," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-39, No. 2, pp. 258-266, Feb. 1991.

[89] ***Bardi I, Biro O.***, "An efficient finite-element formulation without

spurious modes for anisotropic waveguides," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-39, No. 7, pp. 1133-1139, Jul. 1991.

[90] ***Bossavit A., Mayergoyz I.***, "Edge-elements for scattering problems," *IEEE Trans. Magnetics*, Vol. MAG-25, No. 4, pp. 2816-2821, July 1989.

[91] ***Lee J.F., Sun D.K., Cendes Z.J.***, "Tangential vector finite element for electromagnetic field computation," *Digest of the Fourth Biennial IEEE Conference on Electromagnetic Field Computation*, DB-04, Toronto, 22-24 Oct. 1990.

[92] ***Lee J.F., Sun D.K., Cendes Z.J.***, "Full-wave analysis of dielectric waveguides using tangential vector finite elements," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-39, No. 8, pp. 1262-1271, Aug. 1991.

[93] ***Silvester P., Lowther D.A., Carpenter C.J.***, "Exterior finite element for 2-dimensional field problems with open boundaries," *Proc. IEE*, Vol. 124, No. 12, pp. 1267-1270, Dec. 1977.

[94] ***Cendes Z.J.***, "A note on the finite-element solution of exterior-field problems," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-24, No. 7, pp. 468-473, Jul. 1976.

[95] ***Emson C.R.I***, "Methods for the solution of open-boundary electro-magnetic field problems," *IEE Proc., Pt. A*, Vol. 135, No. 3, pp. 151-158, Mar. 1988.

[96] ***McDougall M.J., Webb J.P.***, "Infinite elements for analysis of open dielectric waveguides," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-37, No. 11, pp. 1724-1731, Nov. 1989.

[97] ***Dongarra J.J,*** "Performance of various computers using standard linear equation software," *Supercomputing Review*, Jan. 1990.

[98] ***Van Bladel J.***, *Electromagnetic Fields*. New York: McGraw-Hill, 1964.

[99] ***Stakgold I.***, *Boundary Value Problems of Mathematical Physics*, Vol. II. New York: Macmillan, 1968.

[100] ***Lynch D.R., Paulsen K.D.***, "Origin of vector parasites in numerical Maxwell solutions," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-39, No. 8, pp. 383-394, Aug. 1991.

[101] ***Paulsen K.D., Lynch D.R.,*** "Elimination of vector parasites in finite element Maxwell solutions," *IEEE Trans. Microwave Theory*

Tech., Vol. MTT-39, No. 8, pp. 395-404, Aug. 1991.

[102] ***Ferrari R.L., Maile G.L.***, "Three-dimensional finite element method for solving electromagnetic field problems," *Electron. Lett.*, Vol. 14, pp. 467-468, 1978.

[103] ***Konrad A.***, "A direct three-dimensional finite element method for the solution of electromagnetic fields in cavities," *IEEE Trans. Magnetics*, Vol. MAG-21, pp. 2276-2279, 1985.

[104] ***Crowley C. W., Silvester P.P., Hurwitz,*** "Covariant projection elements for 3D vector field problems," *IEEE Trans. Magnetics*, Vol. MAG-24, No. 1, pp. 397-400, Jan. 1988.

[105] ***Pinchuk A.R., Crowley C.W., Silvester P.P.***, " Spurious solutions to vector diffusion and wave field problems," *IEEE Trans. Magnetics*, Vol. MAG-24, No. 1, pp. 158-161, Jan. 1988.

[106] ***Wong S.H., Cendes Z.J.,*** "Combined finite element-modal solution of three-dimensional eddy current problems," *IEEE Trans. Magnetics*, Vol. MAG-24, No. 6, pp. 2685-2687, Nov. 1988.

[107] ***Wong S.H., Cendes Z.J.,*** "Numerically stable finite element methods for Galerkin solution of eddy current problems," *IEEE Trans. Magnetics*, Vol. MAG-25, No. 4, pp. 3019-3021, Jul. 1989.

[108] ***Whiteman J.R.*** (ed.), *The Mathematics of Finite Elements and Applications*. London: Academic Press, 1973

[109] ***Mikhlin S.C.***, *Variational Methods in Mathematical Physics*, New York: Macmillian, 1964.

[110] ***Cairo L., Kahan T.***, *Variational Techniques in Electromagnetism.* New York: Gordon and Breach, 1965.

[111] ***Aubin J.P.***, *Applied Functional Analysis*. New York: John Wiley & Sons, 1979.

[112] ***Kendall P.C.***, private communication, Dec. 1989.

[113] ***Schechter R.S.***, *The Variational Method in Engineering*. New York: McGraw-Hill, 1967.

[114] ***Schweig E., Bridges W.B.***, "Computer analysis of dielectric waveguides: A finite difference method," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-32, No. 5, pp. 531-541, May 1984.

[115] ***Hughes T.J.R.***, *The Finite Element Method - Linear Static and Dynamic Finite Element Analysis*. Prentice-Hall, 1987.

[116] **Zienkiewicz O.C., Morgan K.**, *Finite Elements and Approximation.* New York: John Wiley & Sons, 1983.

[117] **Noor A.K., Pilkey W.D.** (ed), *State-of-the-Art Surveys on Finite Element Technology.* New York: The American Society of Mechanical Engineers, 1983.

[118] **Kardestuncer H., Norrie D.H.** (ed.), *Finite Element Handbook.* New York: McGraw-Hill, 1987.

[119] **Bathe K.J., Wilson E.L.**, *Numerical methods in finite element analysis.* New Jersey: Prentice Hall, 1976.

[120] **Clint M., and Jennings A.**, "A simultaneous iteration method for the unsymmetric eigenvalue problem," *J. Inst. Maths Applics.*, Vol. 8, pp. 111-121, 1971.

[121] **Jennings A., Stewart W.J.** , "Simultaneous iteration for partial eigensolution of real matrices," *J. Inst. Maths Applics.*, Vol. 15, pp. 351-361, 1975.

[122] **Dong S.B.**, "A block-Stodola eigensolution technique for large algebraic systems with non-symmetrical matrices," *International Journal for Numerical Methods in Engineering*, Vol. 11, pp. 247-267, 1977.

[123] **Duff I.S.**, "ME28: A sparse unsymmetric linear equation solver for complex equations," *ACM Transaction on Mathematical Software*, Vol. 7, No. 4, pp. 505-511, Dec. 1981.

[124] *Harwell Subroutine Library*, Harwell Laboratory, Oxford, England.

[125] *NAG Fortran Library*, Numerical Algorithms Group Ltd., Oxford, England.

[126] *IMSL Fortran Math/Library*, IMSL Inc., Texas, U.S.A.

[127] **Mrozowski M,** *Waves in Shielded Lossless Isotropic Waveguiding Structures*, Ph.D. thesis, Technical University of Gdansk, Poland, Dec. 1989.

[128] **Robson P.N., Kendall P.C.** (ed), *Rib Waveguide Theory by the Spectral Index Method.* Taunton, Somerset, England: Research Studies Press, John Wiley & Sons, 1990.

[129] **Stern M.S.**, "Semivectorial polarised finite difference method for optical waveguides with arbitrary index profiles," *IEE Proc.*, *Pt. J*, Vol. 135, No. 1, pp. 56-63, Feb. 1988.

[130] ***Stern M.S.***, "Semivectorial polarised ***H*** field solutions for dielectric waveguides with arbitrary index profiles," *IEE Proc., Pt. J*, Vol. 135, No. 5, pp. 333-338, Oct. 1988.

[131] ***Ohtaka M***, "Analysis of the guided modes in the anisotropic dielectric rectangular waveguides," (in Japanese) *Trans. Inst. Electron. Commun. Eng. Japan*, Vol. J64-C, pp. 674-681, Oct. 1981.

[132] ***Laboratoires de Marcussis, Centre de Recherche de la CGE, Alcatel***, (France), private communications, 1991.

[133] ***Fernandez F.A., Lu Y.***, "A variational finite element formulation for dielectric waveguides in terms of transverse magnetic fields," *Digest of the Fourth Biennial IEEE Conference on Electromagnetic Field Computation*, PA-03, Toronto, 22-24 Oct. 1990.

[134] ***Fernandez F.A., Lu Y.***, "Variational finite element analysis of dielectric waveguides with no spurious solutions," *Electron. Lett.*, Vol. 26, No. 25, pp. 2125-2126, 1990.

[135] ***Fernandez F.A., Lu Y.***, "Finite element analysis of complex modes in lossless waveguides," *Proc. of SBMO 91 International Microwave Conference*, Rio de Janeiro, 22-25 Jul. 1991, pp. 318-323.

[136] ***Fernandez F.A., Davies J.B., Zhu S., Lu Y.***, "Sparse matrix eigenvalue solver for finite element solution of dielectric waveguides," *Electron. Lett.*, Vol. 27, No. 26, pp. 1824-1826, Sept. 1991.

[137] ***Fernandez F.A., Lu Y.***, "A variational finite element formulation for dielectric waveguides in terms of transverse magnetic fields," *IEEE Trans. Magnetics*, to be published.

[138] ***Lu Y., Fernandez F.A., Zhu S., Davies J.B.***, "A new variational finite element analysis of microwave and optical waveguides without spurious solutions," (to present at *International Conference on Computation in Electromagnetics*, London, 26-27 Nov. 1991.)