Pure Appl. Opt. 2 (1993) 211–233. Printed in the UK

Film mode matching: a versatile numerical method for vector mode field calculations in dielectric waveguides

Aasmund Sv Sudbø

Norwegian Telecom Research, PO Box 83, N-2007 Kjeller, Norway
and
Physics Department, University of Oslo, PO Box 1048, Blindern, N-0316 Oslo, Norway

Received 21 May 1992, in final form 17 September 1992

Abstract. The old film mode matching or transverse resonance method for calculating mode fields in dielectric waveguides is given a new formulation that makes it an efficient, accurate, and general numerical method, ready to exploit standard computer libraries for numerical linear algebra. Both the scalar and the vector field cases are treated for a rather general class of waveguide cross sections, where the method (like the method of lines) is more accurate and more efficient than sophisticated finite-element and finite-difference methods.

1. Introduction

The calculation of mode fields in dielectric waveguides is of fundamental importance both in optics and in microwave technology. Literally hundreds of papers have been published on the subject, and even more on the associated problem of finding mode indices or repetencies. The most challenging problem is the design of numerical methods whereby the mode field can be calculated accurately everywhere in the waveguide cross section, also near non-planar interfaces with large index of refraction differences. Specifically, there are fundamental analytical problems with the calculation of vector modes of rectangular waveguides, in that the mode field is singular at the corners of the waveguide core [1–3]. The singularity is a property of vector modes only, absent in the scalar approximation. Such corners appear in many model waveguides analysed for diode laser and integrated optics applications, and the associated field singularity shows the convergence of all known numerical methods for vector mode field calculations in such waveguides. This is a problem not usually discussed when numerical methods for mode field calculations are presented in the literature, probably because the results in [1–3] are absent from most standard references on optical waveguide theory.

Most simple methods for vector mode field calculations by design cannot yield mode fields that are accurate everywhere in the waveguide cross section. Finite-difference (FD) and finite-element (FE) methods with non-uniform meshes can yield accurate results, but at a considerable computer program development cost. One of the few simple methods that also can yield accurate results almost everywhere in the waveguide cross section is that of film mode matching (FMM). It is eminently suited for analysing a large class of model waveguides encountered in laser diode and integrated optics applications, namely those with a cross section consisting of a small number of homogeneous

0963-9659/93/030211+23 $07.50

rectangles. It is one of the oldest numerical techniques used for mode field calculations [4–7], and is widely known as the transverse resonance or equivalent network method [7–10]. The full vector modes can be calculated with the method, and semivectorial and scalar approximations emerge in a particularly transparent manner, as does also the widely used effective index approximation [6].

In its usual formulation [7, ch 11], [8, 9], the equivalent network method encounters severe numerical problems when the expansion of the mode field in film modes is carried to high order, because of the evanescent nature of the higher order film modes. Therefore, in recent work the ability of the method to produce good results with modest size computations has been emphasized [10]. When a proper formulation is used, however, there are no inherent numerical problems associated with the method, and the method is actually the efficient and general numerical method it is promised to be [8]. The development of such a formulation is the subject of this contribution. Incidentally, the equivalent network picture of the method actually provides some mental barriers for arriving at a numerically well behaved formulation, hence the term 'film mode matching' rather than the more widely known 'transverse resonance' is chosen for describing the method.

By way of example, we shall see that the FMM method yields results that are more accurate than those obtained with a fully fledged FE method [11], with order-of-magnitude smaller requirements for computer resources and program development. For the simple case of an arbitrarily shaped homogeneous core in a homogeneous cladding, methods that are as efficient and accurate as the FMM method are known [12, 13], but the FMM method is more straightforward, and applicable to a much wider class of waveguide geometries. The method of lines (MoL) [7, ch 6], [14, 15] is, however, about as efficient and straightforward as the FMM method, as discussed in section 5.

The strengths of the FMM method are most apparent when it is applied to vector mode calculations, but a natural course of presentation is to treat the scalar approximation first and then to generalize to the full vector treatment. For weak guides, however, almost any standard numerical method developed for solving problems involving the Laplacian operator works well for mode field calculations, and different applications favour different methods.

2. Background

Let us review the essential features of the film mode matching (FMM) method, with reference to the example of a semiconductor ridge waveguide shown in figure 1. We consider the waveguide cross section to be a sandwich of slices numbered $m = 1, 2, \ldots, M$. Each slice is considered to be cut from a film waveguide with film layers numbered $n = 1, 2, \ldots, N$. It is therefore meaningful to use the label m for the whole film waveguide that slice number m is cut from as well. We will thus say that slice m is cut from film m, and that there is layer n in film m. Figure 1 should make the terms 'slice' and 'layer' clear; the layers are perpendicular to the slices. Layer n in slice m has index of refraction $n^{(m,n)}$ and relative permittivity $\varepsilon^{(m,n)} = n^{(m,n)2}$. We choose our x-axis parallel to the film layers and perpendicular to the slice interfaces, our y-axis perpendicular to the film layers and parallel to the film slice interfaces, and our z-axis along the waveguide, parallel to the layers and slices. With reference to figure 1, we are of course free to choose our film orientation parallel or perpendicular to the guiding

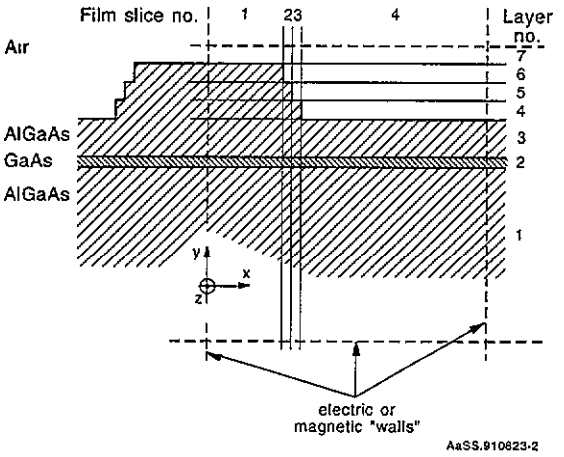

Figure 1. Typical semiconductor ridge waveguide cross section, modelled by a number of film guide slices. The sloped sides of the ridge have been approximated by staircases.

GaAs layer of the waveguide. In figure 1 there are $M = 4$ film slices, each with $N = 7$ layers. Symmetry planes (like the plane defined by $x = 0$ in figure 1) are planes where a mode field component or its gradient perpendicular to the plane is zero, allowing the calculation of mode fields to be restricted to one half (as in figure 1) or one quarter of the waveguide cross section.

The FMM method involves finding the TE and TM modes of each film (or just the scalar modes if the guide is weak), collecting the modes that have the same z-component k_z of the modal wavevector, and matching the field distributions at the slice interfaces by adjusting the modal amplitudes in each film. k_z will also be identified as the angular repetency (propagation constant) of the waveguide mode we seek. Only for certain values of k_z can a set of non-zero film mode amplitudes with matching distributions at the slice interfaces be found. Each set constitutes a waveguide mode.

The continuous spectrum of unbound film modes must somehow be discretized. Usually this discretization is achieved by the introduction of artificial boundaries, so-called walls, placed sufficiently far from the waveguide core that for the mode(s) of interest both the field components and their gradients are negligible. An alternative discretization of the continuous mode spectrum is developed in [10], but the artificial wall method is more straightforward, and the method in [10] does not have any numerical advantages over the artificial wall method. Since the influence of the walls can be easily checked by comparing results obtained with each of the two possible types of boundary conditions imposed at the wall ('electric' or 'magnetic', as discussed below), we shall use the method with artificial walls.

We shall limit our discussion to waveguides made of materials that are isotropic, non-magnetic and non-absorbing, in which case the film modes may be chosen so as to allow the film mode amplitudes to be real numbers. We shall then obtain a purely real eigenvalue problem, as opposed to the complex one of the equivalent network approach [8–10]. Generalization of the FMM method to treat absorbing or non-isotropic material is straightforward, but the real formulation must then be abandoned.

3. Theory

In this section we shall arrive at a (non-linear) eigenvalue equation for the angular repetency k_z of a waveguide mode. The derivation is quite similar to the one shown in detail in [8], and only deviations from the derivation in [8] will be emphasized. The main content of this section is the presentation of a suitable notation for the various quantities needed. The sub/superscript conventions used are summarized in table 1. The choice of x- and y-axes and sign of the exponent $j(\omega t - k_z z)$ are the same as in [8]. Again referring to figure 1, let the thickness of layer n be $d_y^{(n)}$ and the position of the interface between layer $n-1$ and n be $y^{(n)}$, so that

$$d_y^{(n)} = y^{(n+1)} - y^{(n)}. \tag{1a}$$

Similarly, let the thickness of film slice m be $d_x^{(m)}$ and the slice interface positions be at $x^{(m)}$, so that

$$d_x^{(m)} = x^{(m+1)} - x^{(m)}. \tag{1b}$$

The vacuum wavelength of the light is λ_0, and the corresponding angular repetency of the light is

$$k_0 = 2\pi/\lambda_0. \tag{2}$$

The relative permittivity in film m is

$$\varepsilon^{(m)}(y) = \varepsilon^{(m,n)} \qquad \text{for } y^{(n)} < y < y^{(n+1)}. \tag{3}$$

Table 1. Sub/superscript conventions. Superscripts refer to positions in the waveguide cross section; subscripts: (a) label components of a vector in space, (b) label the modes of a film.

	Range of values	Meaning
Superscript		
m	$1, 2, \ldots, M$	film (slice) pointer
M	—	number of film slices
n	$1, 2, \ldots, N$	layer pointer
N	—	number of layers in films
l	—	left or bottom side indicator
r	—	right or top side indicator
Subscript		
0	—	used in $k_0 = \omega/c$
f	—	used in film mode repetency k_f
x	—	x-direction indicator
y	—	y-direction indicator
z	—	z-direction indicator
a	—	antisymmetric indicator
s	—	symmetric indicator
p	h or e	polarization (TE or TM)
h	—	TE polarization indicator
e	—	TM polarization indicator
k	$1, 2, \ldots$	film mode pointer
l	$1, 2, \ldots$	film mode pointer

Each film has a set of film modes with mode functions $\varphi_k^{(m)}(y)$ such that in layer n of film m

$$\ddot{\varphi}_k^{(m)} + (\varepsilon^{(m,n)}k_0^2 - k_k^{(m)2})\varphi_k^{(m)} = 0. \tag{4}$$

where $\ddot{\varphi}_k^{(m)}$ is the second derivative of $\varphi_k^{(m)}$ with respect to y and $k_k^{(m)}$ is the angular repetency of mode k in film m. In the appendix the properties of the piecewise trigonometric/hyperbolic mode functions satisfying (4) are listed, and the usual numerical procedure for obtaining the eigenvalues $k_k^{(m)2}$ of (4) is outlined. For weak guides, the mode functions $\varphi_k^{(m)}(y)$ are the actual mode fields. Vector modes are of two types, TE and TM, depending on polarization. The field components of TE mode k can be expressed in terms of a mode function $\varphi_k^{(m)}(y)$ satisfying (4) that is continuous with continuous derivative $\dot{\varphi}_k^{(m)}(y)$, whereas the field components of TM mode k can be expressed in terms of a continuous function $\psi_k^{(m)}(y)$ also satisfying (4), but with $\dot{\psi}_k^{(m)}(y)/\varepsilon^{(m)}(y)$ continuous. Table A1 in the appendix lists the various field components expressed in terms of φ and ψ and their derivatives.

3.1. Scalar modes

In the scalar case the mode field in film m is expanded in film mode functions:

$$\varphi(x, y) = \sum_{k=1}^{\infty} u_k^{(m)}(x)\varphi_k^{(m)}(y). \tag{5}$$

As usual, the common factor $\exp[j(\omega t - k_z z)]$ has been omitted from the mode field expression (5). The mode amplitude $u_k^{(m)}(x)$ in slice m satisfies an equation like (4):

$$\ddot{u}_k^{(m)}(x) + (k_k^{(m)2} - k_z^2)u_k^{(m)}(x) = 0. \tag{6}$$

As in the appendix we obtain

$$u_k^{(m)}(x) = h(u_{sk}^{(m,l)}, u_{ak}^{(m,l)}/k_{xk}^{(m)}, k_{xk}^{(m)}(x - x^{(m)})) \tag{7a}$$

$$\dot{u}_k^{(m)}(x) = h(u_{ak}^{(m,l)}, -k_{xk}^{(m,l)}u_{sk}^{(m)}, k_{xk}^{(m)}(x - x^{(m)})) \tag{7b}$$

where

$$h(u, v, \eta) = u\cos(\eta) + v\sin(\eta) \tag{7c}$$

$$k_{xk}^{(m)} = (k_k^{(m)2} - k_z^2)^{1/2} \tag{8}$$

the amplitude of mode k in film m at the interface between film slices $m - 1$ and m is

$$u_{sk}^{(m,l)} = u_k^{(m)}(x^{(m)} + 0) \tag{9a}$$

and its x-derivative is

$$u_{ak}^{(m,l)} = \dot{u}_k^{(m)}(x^{(m)} + 0). \tag{9b}$$

The corresponding quantities at the interface at the other side of slice m, towards slice $m + 1$, are

$$u_{sk}^{(m,r)} = u_k^{(m)}(x^{(m+1)} - 0) \tag{10a}$$

$$u_{ak}^{(m,r)} = \dot{u}_k^{(m)}(x^{(m+1)} - 0). \tag{10b}$$

The essence of the film mode matching method for computing waveguide mode fields is that (referring to figure 1) in each homogeneous rectangular region of the waveguide cross section the mode field is a weighted sum of products of two trigonometric/hyperbolic functions of the form (7), one for the x direction and one for the y direction.

The relations between the u analogous to (A9) are

$$u_{ak}^{(m,l)} = -t_k^{(m)} u_{sk}^{(m,l)} + s_k^{(m)} u_{sk}^{(m,r)} \tag{11a}$$

$$u_{ak}^{(m,r)} = -s_k^{(m)} u_{sk}^{(m,l)} + t_k^{(m)} u_{sk}^{(m,r)} \tag{11b}$$

where

$$t_k^{(m)} = k_{xk}^{(m)} / \tan(k_{xk}^{(m)} d_x^{(m)}) \tag{11c}$$

$$s_k^{(m)} = k_{xk}^{(m)} / \sin(k_{xk}^{(m)} d_x^{(m)}). \tag{11d}$$

The FMM method will encounter numerical problems if any one of the $t_k^{(m)}$ vanishes or diverges for a value of k_z/k_0 very close to one of the waveguide mode indices. For most film modes this is not a problem, since $k_{xk}^{(m)}$ is imaginary for all but a few low order film modes. It may be necessary to adjust the position of the artificial boundary at $y = y^{(1)}$ or $y = y^{(N+1)}$ (or the layer thicknesses $d_x^{(m)}$) slightly to avoid such accidental numerical problems with the few film modes for which $k_{xk}^{(m)}$ is real.

Let us introduce matrix notation, where $\boldsymbol{u}$ is a column vector with components u_k, and $\mathbf{M}$ is a matrix with elements M_{kl}. We define $u_{sk}^{(m,l)}$ to be elements of a column vector $\boldsymbol{u}_s^{(m,l)}$, and similarly for the other film mode amplitudes in (9) and (10). If we then introduce diagonal matrices $\mathbf{T}^{(m)}$ and $\mathbf{S}^{(m)}$ with diagonal elements

$$T_{kk}^{(m)} = t_k^{(m)} = k_{xk}^{(m)} / \tan(k_{xk}^{(m)} d_x^{(m)}) \tag{12a}$$

$$S_{kk}^{(m)} = s_k^{(m)} = k_{xk}^{(m)} / \sin(k_{xk}^{(m)} d_x^{(m)}) \tag{12b}$$

(11) may be written

$$\boldsymbol{u}_a^{(m,l)} = -\mathbf{T}^{(m)} \boldsymbol{u}_s^{(m,l)} + \mathbf{S}^{(m)} \boldsymbol{u}_s^{(m,r)} \tag{13a}$$

$$\boldsymbol{u}_a^{(m,r)} = -\mathbf{S}^{(m)} \boldsymbol{u}_s^{(m,l)} + \mathbf{T}^{(m)} \boldsymbol{u}_s^{(m,r)}. \tag{13b}$$

Introducing the inner product notation

$$\langle \varphi | \psi \rangle = (y^{(N+1)} - y^{(1)})^{-1} \int_{y^{(1)}}^{y^{(N+1)}} \varphi(y) \psi(y) \, \mathrm{d}y \tag{14}$$

we may define an overlap integral or coupling matrix between modes of film m and film m'

$$O_{kl}^{(m,m')} = \langle \varphi_k^{(m)} | \varphi_l^{(m')} \rangle. \tag{15}$$

$O_{kl}^{(m,m')}$ are the elements of a matrix $\mathbf{O}^{(m,m')}$ whose transpose is

$$\mathbf{O}^{(m,m')T} = \mathbf{O}^{(m',m)}. \tag{16}$$

When the orthogonal functions $\varphi_k^{(m)}$ are normalized, $\mathbf{O}^{(m,m)}$ is equal to the identity matrix $\mathbf{I}$, and each matrix $\mathbf{O}^{(m,m')}$ is a unitary matrix:

$$\mathbf{O}^{(m,m')} \mathbf{O}^{(m',m)} = \mathbf{O}^{(m',m)} \mathbf{O}^{(m,m')} = \mathbf{I}. \tag{17}$$

The orthogonality relation (17) does not hold exactly when the matrix $\mathbf{O}^{(m,m')}$ is truncated to finite dimensions for practical calculations. Equation (17) must be well

approximated at each interface between neighbouring film slices before the FMM method converges. When this requirement is satisfied, the 'relative convergence' problem [7, ch 10] often associated with the method is not encountered.

Furthermore, with orthonormal film mode functions $\varphi_k^{(m)}$, continuity of the mode field $\varphi(x, y)$ and its normal derivative across the interface between slices m and $m+1$ implies that

$$u_{sk}^{(m,r)} = \sum_{l=1}^{\infty} O_{kl}^{(m,m+1)} u_{sl}^{(m+1,l)} \tag{18a}$$

$$u_{ak}^{(m,r)} = \sum_{l=1}^{\infty} O_{kl}^{(m,m+1)} u_{al}^{(m+1,l)} \tag{18b}$$

or in matrix notation

$$\boldsymbol{u}_s^{(m,r)} = \mathbf{O}^{(m,m+1)} \boldsymbol{u}_s^{(m+1,l)} \tag{18c}$$

$$\boldsymbol{u}_a^{(m,r)} = \mathbf{O}^{(m,m+1)} \boldsymbol{u}_a^{(m+1,l)}. \tag{18d}$$

Combining (13) and (18) to eliminate $\boldsymbol{u}_a^{(m,l)}$, $\boldsymbol{u}_a^{(m,r)}$ and $\boldsymbol{u}_s^{(m,r)}$ we obtain

$$(\mathbf{O}^{(m,m-1)}\mathbf{T}^{(m-1)}\mathbf{O}^{(m-1,m)} + \mathbf{T}^{(m)})\boldsymbol{u}_s^{(m,l)} = \mathbf{O}^{(m,m-1)}\mathbf{S}^{(m-1)}\boldsymbol{u}_s^{(m-1,l)} + \mathbf{S}^{(m)}\mathbf{O}^{(m,m+1)}\boldsymbol{u}_s^{(m+1,l)} \tag{19a}$$

for $2 < m < M$, and

$$(\mathbf{O}^{(2,1)}\mathbf{T}^{(1)}\mathbf{O}^{(1,2)} + \mathbf{T}^{(2)})\boldsymbol{u}_s^{(2,l)} = \mathbf{S}^{(2)}\mathbf{O}^{(2,3)}\boldsymbol{u}_s^{(3,l)} \tag{19b}$$

$$(\mathbf{O}^{(M,M-1)}\mathbf{T}^{(M-1)}\mathbf{O}^{(M-1,M)} + \mathbf{T}^{(M)})\boldsymbol{u}_s^{(M,l)} = \mathbf{O}^{(M,M-1)}\mathbf{S}^{(M-1)}\boldsymbol{u}_s^{(M-1,l)}. \tag{19c}$$

If $M = 2$, we obtain the simple

$$(\mathbf{O}^{(1,2)}\mathbf{T}^{(2)} + \mathbf{T}^{(1)}\mathbf{O}^{(1,2)})\boldsymbol{u}_s^{(2,l)} = 0. \tag{19d}$$

Equation (19) holds if the boundary conditions are $\boldsymbol{u}_s^{(1,l)} = \boldsymbol{u}_s^{(M,r)} = 0$. If the boundary condition for $x = x^{(1)}$ is $\boldsymbol{u}_a^{(1,l)} = 0$, $\mathbf{T}^{(1)}$ in (19*b*) and (19*d*) must be replaced by a diagonal matrix $\mathbf{T}^{(1)*}$ whose diagonal elements are

$$T_{kk}^{(1)*} = -k_{xk}^{(1)2}/t_k^{(1)} = -k_{xk}^{(1)} \tan(k_{xk}^{(1)} d_x^{(1)}) \tag{20a}$$

and if the boundary condition for $x = x^{(M+1)}$ is $\boldsymbol{u}_a^{(M,r)} = 0$, $\mathbf{T}^{(M)}$ in (19*c*) and (19*d*) must be replaced by a diagonal matrix $\mathbf{T}^{(M)*}$ whose diagonal elements are

$$T_{kk}^{(M)*} = -k_{xk}^{(M)2}/t_k^{(M)} = -k_{xk}^{(M)} \tan(k_{xk}^{(M)} d_x^{(M)}). \tag{20b}$$

Equation (19) constitutes a non-linear eigenvalue problem for the angular repetency k_z of the waveguide mode. When this eigenvalue problem is solved iteratively, the coupling matrices $\mathbf{O}^{(m,m+1)}$ need to be calculated only once, since they are independent of k_z. Only the diagonal matrices $\mathbf{T}^{(m)}$ and $\mathbf{S}^{(m)}$ depend on k_z and must be recalculated for each step in the iteration. When the matrices are truncated to the upper left K by K corner, corresponding to the K lowest-order film modes being considered in each film, (19) is a $K(M-1)$ by $K(M-1)$ eigenvalue problem of the general form

$$\mathbf{M}(k_z)\boldsymbol{u} = 0, \tag{21}$$

where the matrix $\mathbf{M}$ has the band structure shown in figure 2(a).

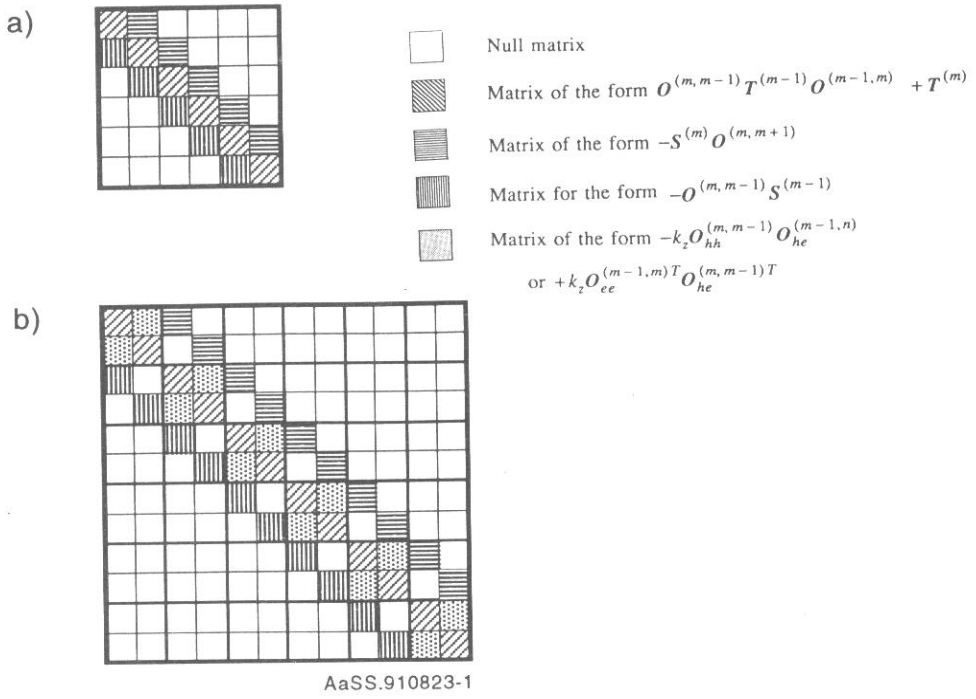

Figure 2. (a) Structure of the matrix $\mathbf{M}$ of the eigenvalue problem (21) of the mode matching method in the scalar field case, equation (19). Each square is an L by L matrix, where L is the number of film modes considered in each film. The total dimension of the matrix is $L(M-1)$, where M is the number of film slices used to model the waveguide cross section. (b) As (a), for the vector field case, equation (33). The total dimension of the matrix is $2L(M-1)$, where L is the number of TE or TM modes. In (a) and (b), the number of films is $M = 7$. Note the band structure of the matrices.

We want a stable and efficient numerical method for solving the non-linear eigenvalue problem (21) that also works when the dimension of $\mathbf{M}$ is large, i.e., if we have a large number of film modes and/or many slices in the model waveguide cross section. Such a method is described below.

Take any vector $\boldsymbol{v}$ with a non-negligible component along the eigenvector $\boldsymbol{u}$ of (21) and solve the equation

$$\mathbf{M}(k_z)\boldsymbol{u}' = \boldsymbol{v} \tag{22}$$

for $\boldsymbol{u}'$, using a standard linear equations solver. As k_z approaches an eigenvalue, the solution $\boldsymbol{u}'$ diverges. So does the output from the linear equations solver, but if it is normalized, it approaches the corresponding eigenvector [16]. A procedure for solving a non-linear eigenvalue problem like (21) (regardless of the dimension of the problem) is therefore first to plot as a function of k_z the norm of the output $\boldsymbol{u}'$ from a linear equations solver applied to (22), in an interval where an eigenvalue for k_z is believed to exist. The search can then be localized to k_z intervals where the norm of $\boldsymbol{u}'$ is much larger than the norm of $\boldsymbol{v}$. In such an interval, consider the function $f(k_z) = 1/u'_p$, where u'_p is the p component of $\boldsymbol{u}'$. We assume that p is chosen so that the p component of the eigenvector $\boldsymbol{u}$ of (21) is non-negligible. Sufficiently close to the eigenvalue, $f(k_z)$ is

a continuous function of a single variable with a zero crossing at the eigenvalue, if the eigenvalue is isolated and non-degenerate. (Exactly degenerate eigenvalues only arise in symmetric geometries that ought to be analysed in their symmetric halves or quarters.) Standard simple numerical procedures may then be used to locate the zero accurately. In the above procedure, the film mode amplitudes $\boldsymbol{u}$, i.e. the waveguide mode field, are obtained together with the angular repetency k_z of the waveguide mode. The procedure must be repeated for each waveguide mode, if more than one mode is needed.

A numerically useful property of (19) is the fact that for higher order film modes the $k_{xk}^{(m)}$ are imaginary with large absolute value, so that unless the slice thickness $d_x^{(m)}$ is small,

$$S_{kk}^{(m)} \approx 2|k_{xk}^{(m)}| \exp(-|k_{xk}^{(m)}| d_x^{(m)}) \tag{23}$$

is negligible if the mode order k is sufficiently large. Hence the right-hand sides of (19) vanish for higher order modes, and (19) uncouple into separate sets of equations for each slice. Thus, instead of being a nuisance, the evanescent nature of the higher order film modes is numerically helpful. The dimensions of the matrices that need to be handled directly may therefore be reduced considerably from the value $L(M-1)$ of (19) as it stands.

3.2. Vector modes

The expansion (5) for scalar waveguide mode fields may be generalized to vector mode fields, with a slight complication: the mode field components of TE and TM film modes in the x and z directions depend on the direction of propagation of the modes. In the equivalent network formulation of the FMM method, the focus is on waves corresponding to $+k_x$ and $-k_x$ in table A1, whereas we are interested in the symmetric and antisymmetric combination of these waves. This choice of variables is necessary to avoid the numerical problems encountered when evanescent film modes are included in the equivalent network formulation, and leads to a real formulation. The vector generalization of (5) is

$$F(x, y) = \sum_{p=e,h} \sum_{k=1}^{\infty} [u_{pk}^{(m)}(x) F_{spk}^{(m)}(y) + \dot{u}_{pk}^{(m)}(x) F_{apk}^{(m)}(y)] \tag{24}$$

where F represents any one of the electric (E_x, E_y, E_z) or magnetic $(cB_x = \mu_0 cH_x$, $cB_y = \mu_0 cH_y$, or $cB_z = \mu_0 cH_z)$ mode field components in slice m. $F_{spk}^{(m)}$ are the corresponding symmetric film mode components in film m with TE $(p = h)$ and TM $(p = e)$ polarization, and $F_{apk}^{(m)}$ the antisymmetric ones, derived from table A1 via

$$F_{pk}^{(m)}(x, y, z, t) = [F_{spk}^{(m)}(y) - \mathrm{j}k_x F_{apk}^{(m)}(y)] \exp[\mathrm{j}(\omega t - k_{xpk}^{(m)}(x - x^{(m)}) - k_z z)] \tag{25a}$$

and listed in table 2. The amplitudes $u_{pk}^{(m)}(x)$ of the symmetric film mode field components are real and satisfy equations like (6) and (7*a*), whereas the amplitudes of the antisymmetric film mode film components are the x derivatives $\dot{u}_{pk}^{(m)}(x)$ of the amplitudes $u_{pk}^{(m)}(x)$, satisfying (7*b*). Note that the x and y components of the fields $F_{spk}^{(m)}(y)$ and $F_{apk}^{(m)}(y)$ are real, whereas the z components are purely imaginary. $(k_0/k_{pk}^{(m)})^2 u_{spk}^{(m,l)}$ and $u_{apk}^{(m,l)}$ are the amplitudes of the symmetric and the antisymmetric film mode field components $F_{spk}^{(m)}(y)$ and $F_{apk}^{(m)}(y)$, respectively, in film m at $x = x^{(m)} + 0$, so that

$$u_{pk}^{(m)}(x) = h((k_0/k_{pk}^{(m)})^2 u_{spk}^{(m,l)}, u_{apk}^{(m,l)}/k_{xpk}^{(m)}, k_{xpk}^{(m)}(x - x^{(m)})) \tag{25b}$$

Table 2. Symmetric and antisymmetric film mode components. This is derived from table A1 in the appendix, via $F = [F_s(y) - \mathrm{j}k_x F_a(y)]\exp[\mathrm{j}(\omega t - k_x x - k_z z)]$, where $\omega = ck_0$ is the angular frequency, c is the speed, and k_0 is the angular repetency of the light in vacuum. The properties of the mode functions $\varphi(y)$ and $\psi(y)$ are discussed in the appendix. The mode has wavevector $k_f = k_x\hat{x} + k_z\hat{z}$, and the angular repetency of the mode is $k_f = |k_f| = (k_x^2 + k_z^2)^{1/2}$. k_f is determined from the film mode equation (A1) for TE modes), and if we choose a value for k_z, k_x is given by $k_x = (k_f^2 - k_z^2)^{1/2}$. An overall normalization factor with the dimension of electric field has been omitted.

	TE		TM	
F	F_{sh}	F_{ah}	F_{se}	F_{ae}
E_x	$(k_z/k_0)\varphi(y)$	0	0	$-\dot{\psi}(y)/(k_0^2\varepsilon(y))$
E_y	0	0	$(-k_f^2/k_0^2)\psi(y)/\varepsilon(y)$	0
E_z	0	$-\mathrm{j}\varphi(y)/k_0$	$(\mathrm{j}k_z/k_0^2)\dot{\psi}(y)/\varepsilon(y)$	0
cB_x	0	$\dot{\varphi}(y)/k_0^2$	$(k_z/k_0)\psi(y)$	0
cB_y	$(k_f^2/k_0^2)\varphi(y)$	0	0	0
cB_z	$(-\mathrm{j}k_z/k_0^2)\dot{\varphi}(y)$	0	0	$-\mathrm{j}\psi(y)/k_0$

where the function h is defined in (7*c*). The corresponding amplitudes at $x = x^{(m+1)} - 0$ are $(k_0/k_{pk}^{(m)})^2 u_{spk}^{(m,r)}$ and $u_{apk}^{(m,r)}$. (The factor $(k_0/k_{pk}^{(m)})^2$ is discussed below.) Equations (11) and (12) are then applicable also to the vector film mode amplitudes, and (13) takes the form

$$\boldsymbol{u}_{ap}^{(m,l)} = -\mathbf{T}_p^{(m)}\boldsymbol{u}_{sp}^{(m,l)} + \mathbf{S}_p^{(m)}\boldsymbol{u}_{sp}^{(m,r)} \tag{26a}$$

$$\boldsymbol{u}_{ap}^{(m,r)} = -\mathbf{S}_p^{(m)}\boldsymbol{u}_{sp}^{(m,l)} + \mathbf{T}_p^{(m)}\boldsymbol{u}_{sp}^{(m,r)} \tag{26b}$$

for $p = h$ (TE modes) and $p = e$ (TM modes). In analogy with (12) the diagonal matrices $\mathbf{T}_p^{(m)}$ and $\mathbf{S}_p^{(m)}$ have the diagonal elements

$$T_{pkk}^{(m)} = (k_0/k_{pk}^{(m)})^2 k_{xpk}^{(m)}/\tan(k_{xpk}^{(m)} d_x^{(m)}) \tag{27a}$$

$$S_{pkk}^{(m)} = (k_0/k_{pk}^{(m)})^2 k_{xpk}^{(m)}/\sin(k_{xpk}^{(m)} d_x^{(m)}) \tag{27b}$$

where

$$k_{xpk}^{(m)} = (k_{pk}^{(m)2} - k_z^2)^{1/2} \tag{28}$$

and $k_{pk}^{(m)}$ are the angular repetencies of the film modes. The film mode amplitudes in films m and $m+1$ are coupled at the interface at $x = x^{(m+1)}$ in a generalization of (18):

$$\boldsymbol{u}_{sh}^{(m,r)} = \mathbf{O}_{hh}^{(m,m+1)}\boldsymbol{u}_{sh}^{(m+1,l)} \tag{29a}$$

$$\boldsymbol{u}_{ah}^{(m,r)} = \mathbf{O}_{hh}^{(m,m+1)}\boldsymbol{u}_{ah}^{(m+1,l)} - k_z\mathbf{O}_{he}^{(m,m+1)}\boldsymbol{u}_{se}^{(m+1,l)} \tag{29b}$$

$$\boldsymbol{u}_{se}^{(m,r)} = \mathbf{O}_{ee}^{(m,m+1)}\boldsymbol{u}_{se}^{(m+1,l)} \tag{29c}$$

$$\boldsymbol{u}_{ae}^{(m,r)} = \mathbf{O}_{ee}^{(m+1,m)T}\boldsymbol{u}_{ae}^{(m+1,l)} + k_z\mathbf{O}_{he}^{(m+1,m)T}\boldsymbol{u}_{sh}^{(m+1,l)} \tag{29d}$$

where the coupling matrices $\mathbf{O}_{hh}^{(m,m')}$, $\mathbf{O}_{he}^{(m,m')}$ and $\mathbf{O}_{ee}^{(m,m')}$ have matrix elements

$$O_{hhkl}^{(m,m')} = \langle\varphi_k^{(m)}|\varphi_l^{(m')}\rangle \tag{30a}$$

$$O_{eekl}^{(m,m')} = \langle\psi_k^{(m)}|\psi_l^{(m')}/\varepsilon^{(m')}\rangle \tag{30b}$$

$$O_{hekl}^{(m,m')} = \langle\dot{\varphi}_k^{(m)}|\psi_l^{(m')}/\varepsilon^{(m')}\rangle/k_{hk}^{(m)2} + \langle\varphi_k^{(m)}|\dot{\psi}_l^{(m')}/\varepsilon^{(m')}\rangle/k_{el}^{(m')2} \tag{30c}$$

and the φ and the ψ are orthonormal:

$$\langle \varphi_k^{(m)} | \varphi_l^{(m)} \rangle = \langle \psi_k^{(m)} | \psi_l^{(m)} / \varepsilon^{(m)} \rangle = \delta_{kl}. \tag{31}$$

When we include the factor $(k_0/k_{pk}^{(m)})^2$ in (25*b*) we can retain the simple normalization (31), we obtain expressions for $O_{hhkl}^{(m,m')}$ and $O_{eekl}^{(m,m')}$ that are independent of the film mode angular repetencies $k_{pk}^{(m)}$, and the similarity between the scalar and the vector case is emphasized. On the other hand, the presence of the factor $(k_0/k_{pk}^{(m)})^2$ in (25*b*) is then a manifestation of the essential difference between scalar and vector fields, namely that the scalar field is, and the vector field is not, continuous across the slice interfaces.

The matrix $\mathbf{O}_{hh}^{(m,m')}$ is identical to the matrix $\mathbf{O}^{(m,m')}$ of the scalar case, and satisfies (16) and (17). The matrix $\mathbf{O}_{ee}^{(m,m')}$ also satisfies (17), i.e., the inverse of $\mathbf{O}_{ee}^{(m,m')}$ is $\mathbf{O}_{ee}^{(m',m)}$, but $\mathbf{O}_{ee}^{(m',m)}$ is not the transpose of $\mathbf{O}_{ee}^{(m,m')}$. The TE–TM cross coupling matrix $\mathbf{O}_{he}^{(m,m')}$ vanishes for $m' = m$, and satisfies

$$\mathbf{O}_{he}^{(m,m')}\mathbf{O}_{ee}^{(m',m)} + \mathbf{O}_{hh}^{(m,m')}\mathbf{O}_{he}^{(m',m)} = 0. \tag{32}$$

The vector mode generalization of (19) is then obtained from (26) and (29):

$$\begin{aligned}(\mathbf{O}_{hh}^{(m,m-1)}\mathbf{T}_h^{(m-1)}\mathbf{O}_{hh}^{(m-1,m)} + \mathbf{T}_h^{(m)})\boldsymbol{u}_{sh}^{(m,l)} \\ = \mathbf{O}_{hh}^{(m,m-1)}\mathbf{S}_h^{(m-1)}\boldsymbol{u}_{sh}^{(m-1,l)} + \mathbf{S}_h^{(m)}\mathbf{O}_{hh}^{(m,m+1)}\boldsymbol{u}_{sh}^{(m+1,l)} \\ - k_z\mathbf{O}_{hh}^{(m,m-1)}\mathbf{O}_{he}^{(m-1,m)}\boldsymbol{u}_{se}^{(m,l)}\end{aligned} \tag{33a}$$

$$\begin{aligned}(\mathbf{O}_{ee}^{(m-1,m)T}\mathbf{T}_e^{(m-1)}\mathbf{O}_{ee}^{(m-1,m)} + \mathbf{T}_e^{(m)})\boldsymbol{u}_{se}^{(m,l)} \\ = \mathbf{O}_{ee}^{(m-1,m)T}\mathbf{S}_e^{(m-1)}\boldsymbol{u}_{se}^{(m-1,l)} + \mathbf{S}_e^{(m)}\mathbf{O}_{ee}^{(m,m+1)}\boldsymbol{u}_{se}^{(m+1,l)} \\ + k_z\mathbf{O}_{ee}^{(m-1,m)T}\mathbf{O}_{he}^{(m,m-1)T}\boldsymbol{u}_{sh}^{(m,l)}\end{aligned} \tag{33b}$$

for $2 < m < M$. Like in (19) for $m = 2$ and $m = M$, $\boldsymbol{u}_{sp}^{(1,l)}$ and $\boldsymbol{u}_{sp}^{(M+1,l)}$, respectively, must be set to zero in (33). The band structure of the eigenvalue problem matrix represented by (33) is illustrated in figure 2(b). As in the scalar case, we have two possible boundary conditions at $x = x^{(1)}$ and $x = x^{(M+1)}$: The vanishing of E_y, E_z and B_x (so-called electric wall), corresponding to the vanishing of $\dot{u}_{hk}^{(m)}(x)$ and $u_{ek}^{(m)}(x)$, or the vanishing of E_x, B_y and B_z (so-called magnetic wall), corresponding to the vanishing of $u_{hk}^{(m)}(x)$ and $\dot{u}_{ek}^{(m)}(x)$. Like in (19), $\mathbf{T}_p^{(1)}$ or $\mathbf{T}_p^{(M)}$ in (33) must be modified according to (20) (with the additional factor $(k_0/k_{pk}^{(m)})^2$ as in (27)) if the boundary conditions are $\boldsymbol{u}_{ap}^{(1,l)} = 0$ or $\boldsymbol{u}_{ap}^{(M,r)} = 0$ instead of $\boldsymbol{u}_{sp}^{(1,l)} = 0$ or $\boldsymbol{u}_{sp}^{(M,r)} = 0$, respectively.

The overlap integral or coupling matrices in (30) can be identified among the **P**, **Q**, **R** and **S** matrices in [8], and the derivation of (29) closely parallels corresponding derivations in [8] and [9]. The coupling matrices in (30) and the film mode amplitudes $u_{pk}^{(m)}(x)$ in (24) (and the amplitudes and derivatives in (26) and (29)) are all real, whereas in the equivalent network formulation [8], the corresponding 'voltages' and 'currents' are complex. Thus we have a formalism that is as simple as that of the equivalent network, involves only real matrices, and can be used as a general numerical method for calculating mode fields in dielectric waveguides.

Most of what was said about numerical computations in the scalar case (19) also applies to the vector case (33). The dimension of the eigenvalue problem matrix is doubled, since both TE and TM film modes must be considered. In addition to the unitary overlap integral matrix $\mathbf{O}_{hh}^{(m,m+1)}$ of the scalar case, a total of three other overlap integral matrices $\mathbf{O}_{ee}^{(m,m+1)}$, $\mathbf{O}_{he}^{(m,m+1)}$ and $\mathbf{O}_{he}^{(m+1,m)}$ must be calculated at each film slice interface.

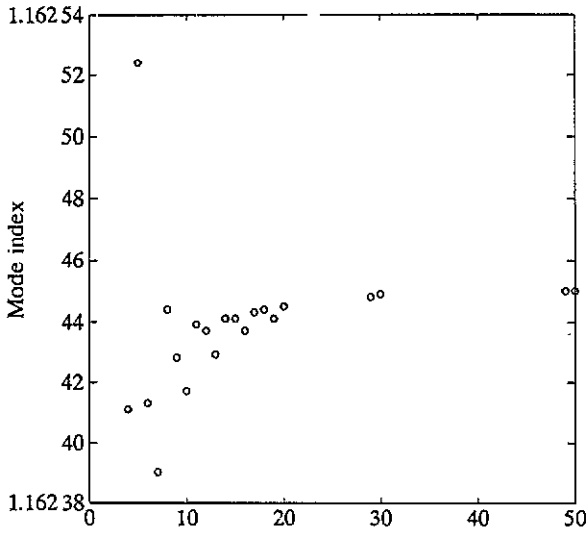

Figure 3. Mode index calculated for the fundamental mode of the model square waveguide described in the text, as a function of number of TE/TM film mode pairs included in each film in the calculation.

If the TE–TM mode coupling at the interfaces is neglected, only the first matrix is needed, and (33) decouple into a TE and a TM case for the whole waveguide. The TE case (33*a*) and the scalar case (19*a*) then look identical. The TM case (33*b*) looks slightly different, since $\mathbf{O}_{ee}^{(m,m+1)}$ is not unitary like $\mathbf{O}_{hh}^{(m,m+1)}$. What makes the TE case different from the scalar case is the factor $(k_0/k_{pk}^{(m)})^2$ that appears in the definitions (25*b*) of $\boldsymbol{u}_p^{(m)}(x)$ and (27) of $\mathbf{T}_p^{(m)}$ and $\mathbf{S}_p^{(m)}$.

4. Results

For illustration purposes we consider a square single mode waveguide with an index of refraction in the core significantly different from that in the cladding. This is a kind of guide where we expect pronounced differences between the results obtained with the various approaches, scalar, semivectorial and vectorial. (In practical structures like semiconductor ridge waveguides, the differences are expected to be much smaller, because the field is confined mainly to regions with small variations in the refractive index.) To show how the method tackles the singular behaviour of the vector mode field at the corners of the waveguide core we consider a core small enough that the mode field extends into the cladding and is non-negligible near the corners of the square waveguide core.

The waveguide chosen for this purpose is one with a core $\lambda/2$ on the side, where λ is the wavelength of the light, and with an index of refraction of 1.5 in the core and 1.0 in the cladding. Artificial electric or magnetic walls were placed at $y = \pm 0.76\lambda$ and at $x = \pm 3.01\lambda$. The program for calculating the effective indices and field components of

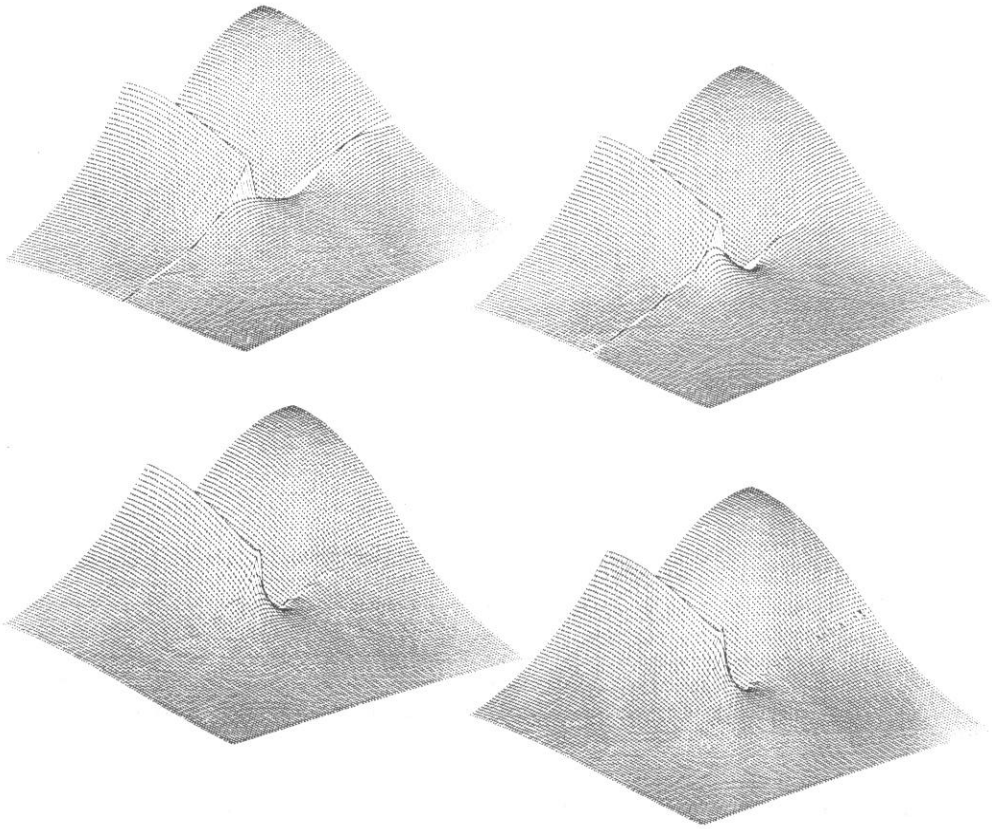

Figure 4. Surface plot of the distribution of the major transverse electric component (the y component) of the fundamental mode field of the model square waveguide described in the text, calculated with the film mode matching method for one quadrant of the waveguide cross section. The region bounded by $x = 0$, $x = 0.5\lambda$, $y = 0$, and $y = 0.5\lambda$ is shown. The four plots correspond to 10, 20, 40 and 80 TE/TM film mode pairs included in the waveguide mode field expansion. The top corner of each plot is at the centre of the guide, the x-axis runs to the right from the top and the y-axis runs to the left. The field is y-polarized.

the waveguide modes was written in the programming language Matlab†, and the calculations were performed on a personal computer with an 80486 processor.

The properties of the method for vector field calculations are clearly displayed in figures 3 and 4. All the results and the graphics presented in figures 3 and 4 were obtained in a few minutes of processing time on the personal computer. As the order of approximation is increased (i.e. more film modes are included in the expansion of the waveguide mode field), the mode index k_z/k_0 rapidly converges to six digits, as shown in figure 3. Figure 4 shows the distribution of the major transverse component of the electric field of the fundamental mode, calculated for one quadrant of the square waveguide, for different orders of approximation. The numerically troublesome continuity requirement on the field near the corner of the core/cladding interface is well satisfied, even in low-order approximations. For the highest order approximation the cusp associated

† '*Matlab*' is a registered trade mark of The MathWorks, Inc., 21 Eliot Street, South Natick, MA 01760, USA.

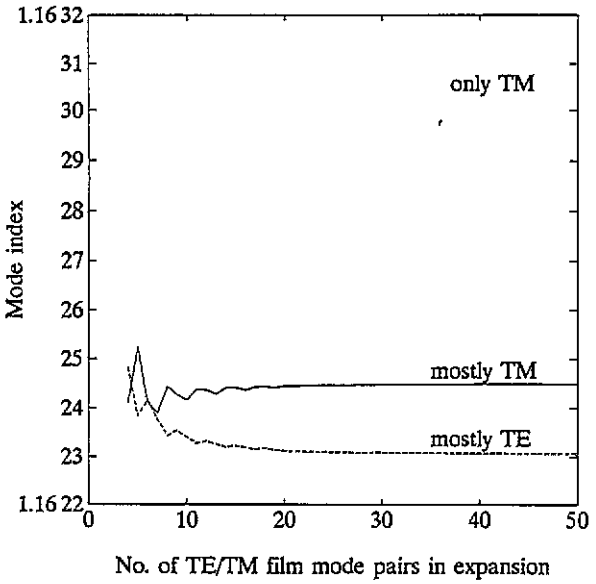

Figure 5. Mode index calculated for the fundamental mode of the model square waveguide described in the text, as a function of the number L of TM film modes included in the calculation. Expansions based on TM modes only (top trace), dominated by TM modes (centre) and dominated by TE modes (bottom) are shown. As in figure 3, results were calculated for all L values between 4 and 20, and for L equal to 29, 30, 49 and 50.

with the divergence of the field at this corner is apparent. The cusp is very clearly displayed, also in lower order approximations, in the other transverse E-field component (not shown). The cusp is more clearly displayed in the results presented in [3], where both transverse E-field components are shown, and where more film modes were used in the calculation, so that the spatial resolution was four times better than in figure 4.

There are small ripples in the calculated mode field near the (artificial) interface between the two films defined by the mode matching method. The ripples are a numerical artifact introduced by the singular behaviour of the field at the corner of the waveguide core. Only the two diverging transverse components of the electric mode field exhibit these ripples; the longitudinal electric and the three magnetic field components (that are continuous) calculated all look smooth.

Figure 5 shows the calculated mode index k_z/k_0 of the fundamental mode as a function of number of film modes included in the calculation, for both x- and y-polarized full vector modes, and for the best semivectorial approximation. Figure 6 shows distributions of the major (x or y) component of the transverse electric mode field. The scalar field approximation is shown together with x-polarized (i.e. dominated by TE film modes), and y-polarized (i.e. dominated by TM film modes) fields, for semivectorial and vectorial expansions. 50 film modes (or TE/TM mode pairs) were included in each film in the calculation.

The full vector mode expansion where TM film modes dominate and the expansion where TE modes dominate yield very similar results, and the semivectorial approximation based on TM film modes only is hardly distinguishable from its full vector mode counterpart. From figure 5 we see that the convergence properties of the above three expansions are quite similar, with the semivectorial approximation introducing an error

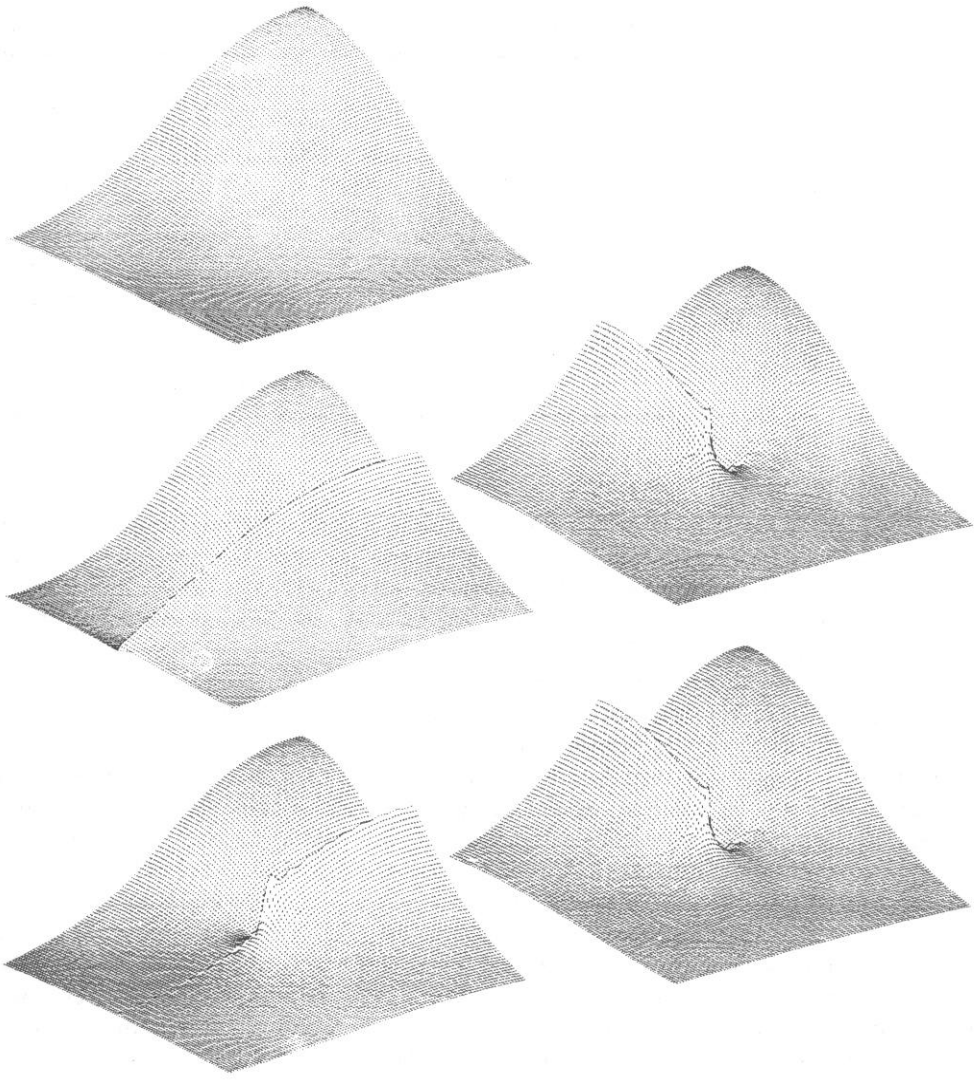

Figure 6. Distribution of the major transverse electric component of the fundamental mode field of the model square waveguide described in the text, calculated with the film mode matching method for one quadrant of the waveguide cross section. The scalar field approximation (top), two semivectorial approximations (centre), and the two possible polarizations of the vector field (bottom) are shown, with TE film mode dominated cases to the left and TM dominated cases to the right. Axes are as in figure 4.

of about 6×10^{-4} in the calculated mode index in this particular example. (The difference between x- and y-polarized cases is mainly due to the non-square shape of the artificial enclosure, and vanishes for a square enclosure, or if the artificial walls are moved further from the waveguide core.) The semivectorial approximation based on the TE film modes only is, however, linearly polarized and almost as inadequate as the scalar approximation. Both approximations yield mode field distributions that at best can be considered qualitatively correct.

Table 3. Comparison between the results in [17] and those obtained with the FMM method.

Computational window height (μm)	Type of artificial wall	Mode index		
		TE	TM	TE (asym.)
20 [17]		3.255 166 33	3.241 078 91	3.192 067 93
6.0001	e	3.256 107 51	3.241 911 29	3.193 357 41
	m	3.256 114 95	3.241 911 17	3.193 382 22
4.0001	e	3.256 109 15	3.241 915 22	3.193 321 59
	m	3.256 105 81	3.241 907 23	3.193 394 35
2.0001	e	3.255 172 65	3.243 143 89	3.191 000 17
	m	3.256 998 72	3.240 587 88	3.195 366 88

The square waveguide described above is useful for illustration purposes, but it is not very practical. A recently published calculation on a diode laser amplifier structure provides such a practical example [11]. The waveguide analysed consists of two parallel slab waveguides on top of each other, both 1.5 μm wide, the bottom one 0.2 μm thick with an index of refraction of 3.4, the top one 0.15 μm thick with an index of refraction of 3.53, spaced 0.1 μm apart in a medium with an index of refraction of 3.17. The structure was analysed in [11] with an FE method using a non-uniform grid, even though the structure is perfect for the application of the much simpler FMM method described above.

In table 1 results obtained for the mode index k_z/k_0 with the FMM method are compared with the FE method results in the first line of table VI of [11]. Three computational windows were considered, each 20 μm wide as in [11], with heights of 2.0001 μm, 4.0001 μm and 6.0001 μm, as compared to the 20 μm of [11]. The waveguide core was offset 0.000 15 μm vertically from the centre of the computational window. Results were calculated both for electric and magnetic wall boundary conditions on the top and bottom sides of the computational window, for a wavelength of 1.55 μm. To obtain the same spatial resolution in the solutions, 25 TE/TM film mode pairs were used in the mode field expansion with the 2.0001 μm high computational window, 50 mode pairs with 4.0001 μm, and 75 mode pairs with 6.0001 μm.

Table 3 indicates that for the fundamental TE and TM modes the spatial resolution considered yields results for the mode index that are accurate to between five and six decimal places. Calculations with 100 mode pairs for the 2.0001 μm high computational window confirm this conclusion. The results for the fundamental modes do not seem to be affected by the artificial walls when they are more than about 2 μm from the core, while the results for the lowest order asymmetric mode are affected in the fifth decimal place. Artificial walls only 1 μm from the core affect the results in the third decimal place. The results in [11] are about 0.001 smaller than those obtained here and in a calculation with the method of lines [17].

5. Discussion

Film mode matching is a well established method for calculating approximations to mode fields in dielectric waveguides. The present work establishes its efficiency as an

accurate numerical method. As mentioned already, the formulation discussed here is in terms of real vectors and matrices, instead of the complex ones obtained if the equivalent network analogy [8–10] is used. Since the real formulation is obtained without an accompanying doubling of the dimension of the matrices and vectors, we gain numerical efficiency by abandoning the equivalent network analogy. Furthermore, a computer program for mode field calculations based on the real formulation is as straightforward as one based on the equivalent network formulation. The 'mesh' geometry of the model waveguide cross section considered here is, however, more general than the geometry usually considered with the equivalent network. Last, but not least, no numerical problems associated with the evanescent nature of the higher-order film modes are encountered. This last feature of the method is a result of the particular choice of unknowns $u_{spk}^{(m,l)}$ in the eigenvalue problem (33), not that this choice leads to a real formulation.

Figures 3 and 4 and table 3 lead one to believe that film mode matching is probably one of the most accurate and efficient methods available for the calculation of vector mode fields in dielectric waveguides. The analytic problems at the corners of the waveguide core appear as oscillations in the calculated mode fields, localized near the interfaces between film slices defined by the method. The oscillations are well behaved and well understood from completely analogous behaviour of the standard Fourier expansion of a discontinuous function of a single variable. For aesthetic purposes, they can be eliminated by extrapolation everywhere except in the immediate vicinity of the analytic singularity. When the results obtained with the FMM method are compared with the results obtained with a well developed FE method (table 3) the accuracy of the much simpler and faster FMM method is striking.

Detailed accuracy of the kind seen in figure 6 for the computed mode fields provided by the FMM method can only be matched by an FE method with an impractically fine mesh. The same can be said about finite-difference methods. The results in table 3 for the mode indices are accurate to better than a few parts in 10^5, the accuracy claimed in [11]. The differences between the FMM results and the FE results [11] are actually one order of magnitude larger. One possible explanation for the differences is the omission of the penalty parameter extrapolation in [11]. The sign of the differences is consistent with this explanation. The conclusion is clear: use the FMM instead of any FE method whenever waveguide geometry allows.

From the comparison of figure 6 of scalar and semivectorial approximations with the full vector results, it is evident that the field of modes with a polarization such that they can be expanded in TM film modes only are as accurate as their full vector counterparts, for all practical purposes. Similar statements cannot be made about the semivectorial approximation based on TE modes only or about the scalar approximation. These approximations have problems that one encounters whenever one is interested in the actual magnitude of the mode field near non-planar interfaces with large index of refraction differences. (A practical example is a DFB laser with a grating at the interface between the semiconductor and an oxide cover.) That an expansion based on TM modes only is better than one based on TE modes only is reasonable, since the discontinuity of the mode field near the corners of the core is explicitly present in the TM modes only.

A few pros and cons of the FMM method are detailed below.

Pros: 1. The method is straightforward. Writing a computing program implementing the method is easy, especially if a programming language incorporating matrix algebra is used. This is by far the most important feature of the method.

2. The method can be used both as a quick method of approximation and as a fully fledged accurate numerical method for mode field calculations. In the formulation discussed above, no numerical stability problems pop up when the number of film mode basis functions is increased. In particular, there are no problems associated with the evanescent nature of the higher-order modes, as discussed already.
3. The method handles both scalar and vector fields, formally with almost equal ease. Semivectorial approximations emerge with the same ease.
4. The method is much more efficient than finite-element, finite-difference or orthogonal-function-expansion methods, for waveguides composed of homogeneous regions with well defined interfaces. The reason is that it is a kind of boundary integral method, i.e. the mode field in the entire waveguide cross section does not have to be considered in the eigenvalue problem to be solved, only the field along the boundaries represented by the interfaces between the film slices defined by the method.

The method works best for waveguide cross sections consisting of a small number of homogeneous rectangles. Because the description of this geometry is simple, it is the preferred geometry for many model waveguide calculations for practical applications. The rectangles have corners where the index of refraction is different inside and outside the corners. As discussed above, such corners represent an analytical problem for the calculation of vector mode fields, in that the fields diverge at the corner [1–3]. In the vector case it is therefore more important than in the scalar case to have numerically efficient methods, if there are corners in the waveguide cross section. The mode matching method provides numerical efficiency that allows the field divergence at the corner to be represented numerically, as shown in figures 4 and 6, and even clearer in [3].

Cons: 1. The eigenvalue problem (33) that one eventually arrives at is non-linear, and each separate eigenvalue and corresponding eigenvector must be found by iteration. Many other numerical methods for mode field calculations, like finite-element, finite-difference or orthogonal-function-expansion methods, yield eigenvalue problems that are linear. Linear problems are faster than non-linear ones to solve numerically, especially if many eigenvalues and eigenvectors are needed.
2. In vector mode calculations, the analytical problem associated with the presence of sharp right-angle corners in the waveguide cross section can be viewed as a problem with the FMM method, not with the waveguide itself. The waveguide may well have a cross section that is better modelled by rectangles with rounded corners or by graded index distributions than the sharp corners with index steps required by the FMM method. Many real waveguide structures, however, actually have cross sections with numerically troublesome corners, i.e., well defined interfaces with a local radius of curvature much smaller than a wavelength.
3. When the geometry of the waveguide cross section deviates from the rectangular (e.g. the slanted sides of the ridge waveguide of figure 1), corners that are not present in the real guide are introduced in the model guide by the method (see figure 1). In light of what has been said above

about the trouble of corners in vector mode calculations, these artificial corners are clearly undesirable.

4. The number of basis functions needed for a given accuracy increases when the distance from the core to the artificial boundaries of the film guides increases. This is a problem for modes near cut-off, because their mode fields extend far into the cladding, so that the artificial walls must be placed far from the core. The distance to the walls can be minimized, however, by comparison of mode fields and repetencies calculated for each of the two possible boundary conditions at the artificial boundaries.
5. A troublesome problem with the method has been the 'relative convergence' phenomenon [7, ch 10]. This problem is not encountered with the purely dielectric waveguides usually considered for integrated optics, unless the waveguide structure is such that the approximate orthogonality of the overlap integral matrices $\mathbf{O}^{(m,m+1)}$ can only be satisfied with impractically large dimensions of the matrices.

Finally, let us consider a viable alternative, the method of lines (MoL). It is a rather close relative of the FMM method, as discussed in [7, ch 6]. For the class of waveguide geometries considered here, both methods are quite straightforward and computationally efficient. The expansion of the mode field in film modes is roughly equivalent to an equidistant sampling of the field in the direction perpendicular to the film, in the high order limit. Thus the number of modes considered in the FMM method corresponds to the number of lines in the MoL, and the computational efficiency of the FMM method as presented here is about the same as that of the MoL with equidistant lines. They both yield non-linear eigenvalue problems of the form (21), of comparable dimensions. The dimension of the eigenvalue problem can be reduced by non-equidistant sampling of the field, e.g. near field singularities, where high spatial resolution is needed, and in the cladding far from the core, where low resolution is needed. With the FMM method a non-equidistant sampling can be achieved with a change of basis, e.g. by expanding the functions $\varphi_k^{(m)}$ and $\psi_k^{(m)}$ in other function sets, whereas for the MoL a formulation with non-equidistant lines is more straightforward and is already developed [15]. In this respect the MoL has an advantage over the FMM method. Otherwise, there is no reason to expect one method in general to be preferable over the other. Most of what is said above about the advantages of the FMM method over FE and FD methods is equally applicable to the MoL.

6. Conclusion

The old film mode matching method for calculating mode fields in rectangular dielectric waveguides has been re-examined. A numerically stable formulation that is optimized for exploiting standard computer libraries for numerical linear algebra has been developed. The method is straightforward, numerically efficient, accurate, and applicable to scalar and vector field calculations with almost equal ease. The method does not circumvent the fundamental analytic problem with vector fields in rectangular waveguides, but its numerical efficiency is comparable to that of the method of lines, and permits a brute-force treatment of the problem. For waveguide geometries where it is applicable, the method is much simpler, faster and more accurate than finite-element and finite-difference methods.

Appendix. Film modes

In this section the most important quantities describing film modes are listed together with relations useful when a computer program for waveguide mode field calculations based on the film mode matching method is written.

We shall consider real solutions of the film mode equation

$$\ddot{\varphi} + (\varepsilon(y)k_0^2 - k_f^2)\varphi = 0 \tag{A1}$$

where the film is made up of homogeneous layers, so that $\varepsilon(y)$ is a piecewise constant function given by

$$\varepsilon(y) = \varepsilon^{(n)} \qquad \text{for } y^{(n)} < y < y^{(n+1)} \tag{A2}$$

The film layer interfaces are at positions $y^{(n)}$, $n = 2, 3, \ldots, N$. $\ddot{\varphi}(y)$ is the second derivative $\mathrm{d}^2\varphi/\mathrm{d}y^2$ of the mode function $\varphi(y)$, and k_f is the angular repetency of the film mode. Let us for a moment limit our attention to scalar modes, so that the mode field $\varphi(y)$ and its derivative $\dot{\varphi}(y)$ are both continuous functions. The boundary conditions we consider are that either $\varphi(y)$ or $\dot{\varphi}(y)$ is zero for $y = y^{(1)}$ and $y = y^{(N+1)}$.

In layer n of thickness $d^{(n)}$ between $y^{(n)}$ and $y^{(n+1)}$ $\varphi(y)$ has the form

$$\varphi(y) = \varphi_s^{(n,l)} \cos[k_y^{(n)}(y - y^{(n)})] + (\varphi_a^{(n,l)}/k_y^{(n)}) \sin[k_y^{(n)}(y - y^{(n)})] \tag{A3}$$

where

$$k_y^{(n)} = (\varepsilon^{(n)}k_0^2 - k_f^2)^{1/2} \tag{A4}$$

$k_y^{(n)}$ is either real or purely imaginary, so that $\varphi(y)$ is real if $\varphi_s^{(n,l)}$ and $\varphi_a^{(n,l)}$ are real. $\varphi_a^{(n,l)}$ is the function value at the bottom (equivalently: left-hand) side of layer n,

$$\varphi_s^{(n,l)} = \varphi(y^{(n)} + 0) \tag{A5a}$$

and $\varphi_a^{(n,l)}$ is the y-derivative at the bottom side of the layer,

$$\varphi_a^{(n,l)} = \dot{\varphi}(y^{(n)} + 0). \tag{A5b}$$

The use of $\varphi_a^{(n,l)}$ as a variable instead of $\varphi_a^{(n,l)}/k_y^{(n)}$ has the advantage that (A3) is well behaved for vanishing $k_y^{(n)}$.

Let us also introduce the corresponding quantities at the top (equivalently: right-hand) side of the layer:

$$\varphi_s^{(n,r)} = \varphi(y^{(n+1)} - 0) \tag{A5c}$$

$$\varphi_a^{(n,r)} = \dot{\varphi}(y^{(n+1)} - 0). \tag{A5d}$$

Alternative expressions for $\varphi(y)$ and $\dot{\varphi}(y)$ that are particularly useful when $k_y^{(n)}$ is imaginary are

$$\varphi(y) = \big(\varphi_s^{(n,l)} \sin[k_y^{(n)}(y^{(n+1)} - y)] + \varphi_s^{(n,r)} \sin[k_y^{(n)}(y - y^{(n)})]\big)/\sin(k_y^{(n)}d^{(n)}) \tag{A6a}$$

$$\dot{\varphi}(y) = \big(\varphi_a^{(n,l)} \sin[k_y^{(n)}(y^{(n+1)} - y)] + \varphi_a^{(n,r)} \sin[k_y^{(n)}(y - y^{(n)})]\big)/\sin(k_y^{(n)}d^{(n)}) \tag{A6b}$$

where

$$d^{(n)} = y^{(n+1)} - y^{(n)}. \tag{A6c}$$

Dropping the layer label n and the subscript y on $k_y^{(n)}$ we have a number of ways of relating the quantities in (A5):

$$\varphi_s^{(r)} = \varphi_s^{(l)} \cos(kd) + (\varphi_a^{(l)}/k) \sin(kd) \tag{A7a}$$

$$\varphi_a^{(r)} = \varphi_a^{(l)} \cos(kd) - \varphi_s^{(l)} k \sin(kd) \tag{A7b}$$

$$\varphi_s^{(l)} = \varphi_s^{(r)} \cos(kd) - (\varphi_a^{(r)}/k) \sin(kd) \tag{A8a}$$

$$\varphi_a^{(l)} = \varphi_a^{(r)} \cos(kd) + \varphi_s^{(r)} k \sin(kd) \tag{A8b}$$

$$\varphi_a^{(l)} = k[-\varphi_s^{(l)}/\tan(kd) + \varphi_s^{(r)}/\sin(kd)] \tag{A9a}$$

$$\varphi_a^{(r)} = k[\varphi_s^{(r)}/\tan(kd) - \varphi_s^{(l)}/\sin(kd)] \tag{A9b}$$

$$\varphi_s^{(l)} = [\varphi_a^{(l)}/\tan(kd) - \varphi_a^{(r)}/\sin(kd)]/k \tag{A10a}$$

$$\varphi_s^{(r)} = [-\varphi_a^{(r)}/\tan(kd) + \varphi_a^{(l)}/\sin(kd)]/k. \tag{A10b}$$

Within a homogeneous layer $\varphi(y)$ can also be written

$$\varphi(y) = h(\varphi_s^{(l)}, \varphi_a^{(l)}/k, k(y - y^{(l)})) \tag{A11a}$$

and its derivative

$$\dot{\varphi}(y) = h(\varphi_a^{(l)}, -k\varphi_s^{(l)}, k(y - y^{(l)})) \tag{A11b}$$

where $y^{(l)}$ is the position of the bottom of the layer and the trigonometric/hyperbolic function h is a function of three variables

$$h(u, v, \eta) = u \cos(\eta) + v \sin(\eta) \tag{A12a}$$

satisfying

$$\partial h/\partial \eta = h(v, -u, \eta). \tag{A12b}$$

Of interest for overlap integral computations are integrals of the type

$$O = \int_0^d \varphi_1(y)\varphi_2(y)\, dy \tag{A13}$$

where

$$\varphi_k(y) = h(\varphi_{sk}^{(l)}, \varphi_{ak}^{(l)}/k_k, k_k y) \tag{A14}$$

for $k = 1$ and 2. Generalizing (A5) we define

$$\varphi_{sk}^{(r)} = \varphi_k(d) = h(\varphi_{sk}^{(l)}, \varphi_{ak}^{(l)}/k_k, k_k d) \tag{A15a}$$

$$\varphi_{ak}^{(r)} = \dot{\varphi}_k(d) = h(\varphi_{ak}^{(l)}, -k_k \varphi_{sk}^{(l)}, k_k d) \tag{A15b}$$

and obtain

$$\begin{aligned} O &= \int_0^d \varphi_1(y)\varphi_2(y)\, dy \\ &= (\varphi_{s1}^{(r)}\varphi_{a2}^{(r)} - \varphi_{a1}^{(r)}\varphi_{s2}^{(r)} - \varphi_{s1}^{(l)}\varphi_{a2}^{(l)} + \varphi_{a1}^{(l)}\varphi_{s2}^{(l)})/(k_1^2 - k_2^2). \end{aligned} \tag{A16}$$

The expression is numerically inaccurate when $|k_1^2 - k_2^2|d^2 \ll 1$, in which case either a power series expansion of (A15) or a re-expression of (A16) in terms of $k_1 - k_2$

and $k_1 + k_2$ are useful. The special case $k_1 = k_2 = k$ is needed for mode function normalization

$$\int_0^d \varphi(y)^2 \, dy = \frac{1}{2k^2} \varphi_s^{(l)} \varphi_a^{(l)} - \varphi_s^{(r)} \varphi_a^{(r)} + \frac{d}{2} [\varphi_s^{(r)2} + (\varphi_a^{(r)}/k)^2]$$

$$= \frac{1}{2k^2} (\varphi_s^{(l)} \varphi_a^{(l)} - \varphi_s^{(r)} \varphi_a^{(r)}) + \frac{d}{2} [\varphi_s^{(l)2} + (\varphi_a^{(l)}/k)^2]. \tag{A17}$$

A numerical solution of (A1) (so-called shooting method) involves the iterated use of (A7) starting from the bottom and using the bottom side boundary condition, and (A8) from the top, for a given k_f. The continuity requirement on $\varphi(y)$ and $\dot{\varphi}(y)$ imply that

$$\varphi_s^{(n,r)} = \varphi_s^{(n+1,l)} \qquad \text{and} \qquad \varphi_a^{(n,r)} = \varphi_a^{(n+1,l)} \tag{A18}$$

which must be exploited together with (A7) and (A8). A particular layer interface position $y^{(n')}$ is chosen where a field mismatch, expressed as

$$\Delta^{(n')} = \varphi_s^{(n',r)} \varphi_a^{(n'+1,l)} - \varphi_a^{(n',r)} \varphi_s^{(n'+1,l)} \tag{A19}$$

is computed. In (A19) $\varphi_s^{(n',r)}$ and $\varphi_a^{(n',r)}$ are calculated by iterating (A7) and (A18) from $n = 1$, whereas $\varphi_s^{(n'+1,l)}$ and $\varphi_a^{(n'+1,l)}$ are calculated by iterating (A8) and (A18) from $n = N$. The zeros of $\Delta^{(n')}$ as a function of k_f then yield the film modes. For numerical reasons n' should not be at a boundary, $n' = 1$ or $n' = N + 1$.

The vector modes of the film are of two kinds, TE or TM, depending on polarization. The TE mode field components can be expressed in terms of a scalar mode field function of the type $\varphi(y)$ discussed above, as shown in table A1. The TM mode field components can be expressed in terms of a scalar mode field function $\psi(y)$ satisfying (A1) everywhere except at the layer interfaces, where $\dot{\psi}(y)/\varepsilon(y)$ instead of $\dot{\psi}(y)$ is required to be continuous. Thus for $\psi(y)$ we may define quantities $\psi_s^{(n,l)}$, $\psi_a^{(n,l)}$, $\psi_s^{(n,r)}$ and $\psi_a^{(n,r)}$ in analogy with (A5), satisfying relations like (A7)–(A10). For TM modes, however, (A18) takes the form

$$\psi_s^{(n,r)} = \psi_s^{(n+1,l)} \qquad \text{and} \qquad \psi_a^{(n,r)}/\varepsilon^{(n)} = \psi_a^{(n+1,l)}/\varepsilon^{(n+1)} \tag{A20}$$

Table A1. Film mode field components. All components are to be multiplied by an arbitrary normalization constant with dimension of electric field, and by $\exp[\mathrm{j}(\omega t - k_x x - k_z z)]$, where $\omega = ck_0$ is the angular frequency, c is the speed and k_0 is the angular repetency of the light in vacuum. The properties of the dimensionless mode functions $\varphi(y)$ and $\psi(y)$ are discussed in the appendix. The mode has wavevector $\boldsymbol{k}_f = k_x \hat{\boldsymbol{x}} + k_z \hat{\boldsymbol{z}}$ and the angular repetency of the mode is $k_f = |\boldsymbol{k}_f| = (k_x^2 + k_z^2)^{1/2}$. k_f is determined from the film mode equation ((A1) for TE modes) and, if we choose a value for k_z, k_x is given by $k_x = (k_f^2 - k_z^2)^{1/2}$.

F	TE	TM
E_x	$(+k_z/k_0)\varphi(y)$	$(+\mathrm{j}k_x/k_0^2)\dot{\psi}(y)/\varepsilon(y)$
E_y	0	$(-k_f^2/k_0^2)\psi(y)/\varepsilon(y)$
E_z	$(-k_x/k_0)\varphi(y)$	$(+\mathrm{j}k_z/k_0^2)\dot{\psi}(y)/\varepsilon(y)$
cB_x	$(-\mathrm{j}k_x/k_0^2)\dot{\varphi}(y)$	$(+k_z/k_0)\psi(y)$
cB_y	$(+k_f^2/k_0^2)\varphi(y)$	0
cB_z	$(-\mathrm{j}k_z/k_0^2)\dot{\varphi}(y)$	$(-k_x/k_0)\psi(y)$

and (A19) takes the form

$$\Delta^{(n')} = \varepsilon^{(n')}\psi_s^{(n',r)}\psi_a^{(n'+1,l)} - \varepsilon^{(n'+1)}\psi_s^{(n'+1,l)}\psi_a^{(n',r)}. \tag{A21}$$

The TM mode field components are also listed in table A1.

As to the boundary conditions, if $\varphi(y) = 0$ at a boundary, then $\dot{\psi}(y) = 0$ at the same boundary, corresponding to vanishing E_x, E_z and B_y (so-called electric wall), and if $\dot{\varphi}(y) = 0$ at the boundary, $\psi(y) = 0$ at the same boundary, corresponding to vanishing E_y, B_x and B_z (so-called magnetic wall).

References

[1] Bach Andersen J and Solodukhov V V 1978 *IEEE Trans. Antenn. Prop.* **AP-26** 598–602
[2] van Bladel J 1991 *Singular Electromagnetic Fields and Sources* (Oxford: Clarendon) Ch 4
[3] Sudbø A S 1992 *J. Lightwave Technol.* **10** 418–19
[4] Schlosser W and Unger H G 1966 *Advances in Microwaves* (New York: Academic) pp 319–87
[5] Solbach K and Wolff I 1978 *IEEE Trans. Microwave Theory Techn.* **MTT-26** 266–74
[6] Mittra R, Hou Y-L and Jannejad V 1980 *IEEE Trans. Microwave Theory Techn.* **MTT-28** 36–43
[7] Itoh T (ed) 1988 *Numerical Techniques for Microwave and Millimeter-wave Passive Structures* (New York: Wiley)
[8] Peng S T and Oliner A A 1981 *IEEE Trans. Microwave Theory Techn.* **MTT-29** 843–55
[9] Koshiba M and Suzuki M 1986 *J. Lightwave Technol.* **4** 656–64
[10] Dagli N 1990 *IEEE J. Quantum Electron.* **26** 98–108
[11] Young T 1992 *J. Lightwave Technol.* **10** 626–33
[12] Kishi N and Okoshi T 1989 *IEEE Trans. Microwave Theory Techn.* **4** 89–102
[13] Vassallo C 1990 *J. Lightwave Technol.* **8** 1723–9
[14] Liskovets O A 1965 *Differentsial'nye Uravneniya* **1** 1662–78
[15] Diestel H 1984 *IEEE J Quantum Electron.* **QE-20** 1288–93
[16] Golub G H and Van Loan C F 1989 *Matrix Computations* 2nd edn (Baltimore, MD: Johns Hopkins University Press) p 383
[17] Pregla R 1992 Private communication