Konrad-Zuse-Zentrum für Informationstechnik Berlin
Heilbronner Str. 10, D-10711 Berlin - Wilmersdorf

Frank Schmidt

Simultaneous Computation of the Lowest Eigenvalues and Eigenvectors of the Helmholtz Equation

Preprint TR 95-13 (November 1995)

SIMULTANEOUS COMPUTATION OF THE LOWEST EIGENVALUES AND EIGENVECTORS OF THE HELMHOLTZ-EQUATION

FRANK SCHMIDT

Contents

Abstract

This report collects a number of proposals to determine the lowest eigensolutions of the scalar Helmholtz equation. The basic routine of all discussed algorithms is the standard Rayleigh quotient minimization process. The minimization is performed in a direct multilevel manner, and a subspace iteration is used to determine simultaneously a couple of eigensolutions. As smoother the nonlinear Gauß-Seidel, the nonlinear conjugate gradient method and a preconditioned version of this method are compared with respect to their efficiency. The numerical examples are based on realistic 1D and 2D models of integrated optics components.

1 Introduction

The most prominent example of the modern optical information techniques is the use of optical fibers in telecomunication networks. Today, nearly all long distance networks and even local networks are realized using optical fibers, because they are suprerior to the conventional electrical transmission lines in nearly all technical parameters, where the most important ones are the low transmission loss (below 1 dB/km) and the large transmission bandwith (some dozen Gbit/s in the infrared region). This development has been started in the early 1970 th with the first low-loss optical fiber (the Corning Glass C^0-fiber). At the same time, the improvements of lasers and the developement of coherent optics had created a need for waveguide structures with which to build optical components and connect them to optical circuits. These optical waveguides should allow the planar fabrication of components and the integration into planar optical circuits which at the same time was already proven to be of great advantage in electronic circuits. These integrated optical circuits should combine the advantages of an effecient production process with the large transmission capabilities of an optical signal processing. But until today, most of all comercial used circuits are expensive hybrid assemblies, i. e. , combinations of pure optical (e. g. photodiode) and pure electronical components (e. g. transistor) based on different semiconductor materials. This depends on the initial fabrication difficulties caused by the very high technological demands. Nevertheless, after a period of the stabilization of the technology during the last decade, the idea of integrated optical circuits becomes more important again and some experts expect a breakthrough in the near future ('fiber to the home').
Apart from the technological progress, there is a need for better physical models and for faster and more reliable numerical simulation tools. An essential prerequisite for the design of integrated optical components like switches, modulators, lasers, couplers, gratings etc. is the knowledge of the eigensolutions of the optical field in some cross sections of the component. In many cases (e. g. directional coupler), the pure knowledge of the lowest few eigenmodes allows the design of the component. Within the rich variety of components, we find such with simple geometries, e. g. step-index fibers with a circular refractive index profile, and such with very complex geometries, e. g. some kinds of switches or branches. It is the aim of this paper to supply the component designer with effective algorithms to solve the arising multi-scale problems.
An essential aspect is that depending on the various working principles of the components to investigate, there are different accuracy demands on the eigenvalues and eigenvectors to be met by a numerical solution. Even the case of a component supporting some very closely neighboring modes (directional coupler) has to be solved with high accuracy, which means for the design tasks to an relative eigenvalue error of 10^{-6} to 10^{-8}.

2 General Problem

We investigate the Helmholtz eigenproblem

$$-\Delta u_i - \epsilon_D u_i = \lambda_i u_i, \quad i = 0, 1, \dots \tag{1}$$

in a convex domain $\Omega \subset \mathbf{R}^2$ equipped with homogeneous Dirichlet boundary conditions. The function $\epsilon_D \in \mathbf{C}(\Omega)$ is a piecewise continuous, bounded complex function representing the normalized dielectric function with $0 < \Re(\epsilon_D)$.

The adjoint problem

The complex conjugate version of (1) defines the adjoint problem with solutions w^*

$$-\Delta w_i^* - \overline{\epsilon_D} w_i^* = \overline{\lambda_i} w_i^*. \tag{2}$$

and (compare with (1))

$$w_i^* = \overline{u_i}. \tag{3}$$

Using the standard L^2 scalar product for complex functions,

$$(u, v) = \int_\Omega \overline{u} v \, dx$$

we obtain the Helmholtz equation in weak form

$$\begin{aligned} a(v, u) &= \lambda \cdot (v, u), \quad u, v \in H_0^1(\Omega), \\ \text{with} \quad a(v, u) &= (\partial_x v, \partial_x u) + (\partial_y v, \partial_y u) - (v, \epsilon_D u). \end{aligned} \tag{4}$$

and its adjoint version

$$\begin{aligned} a^*(v, w^*) &= \overline{\lambda} \cdot (v, w^*), \quad w^*, v \in H_0^1(\Omega), \\ \text{with} \quad a^*(v, w^*) &= (\partial_x v, \partial_x w^*) + (\partial_y v, \partial_y w^*) - (v, \overline{\epsilon_D} w^*). \end{aligned} \tag{5}$$

It is, due to this symmetry,

$$\begin{aligned} \overline{a^*(v, w^*)} &= (\partial_x w^*, \partial_x v) + (\partial_y w^*, \partial_y v) - (w^*, \epsilon_D v) \\ &= a(w^*, v). \end{aligned} \tag{6}$$

For the discussion of some problems it is convenient to use the complex symmetric representation operator $A : H_0^1(\Omega) \to H_0^1(\Omega)$ defined by

$$a(v, u) = (v, Au) \quad v \in H_0^1(\Omega), \tag{7}$$

from which we obtain an abstract formulation of our problem

$$Au_i = \lambda_i u_i. \tag{8}$$

Orthogonality
Equations (4) and (5), specialized to test functions $v = w_j^*$ and $v = u_i$, respectively, yield the system

$$a(w_j^*, u_i) = \lambda_i (w_j^*, u_i) \tag{9}$$
$$a^*(u_i, w_j^*) = \overline{\lambda_j}(u_i, w_j^*). \tag{10}$$

The difference between the complex conjugate of (10) and (9) gives the usual orthogonality relation

$$(\lambda_i - \lambda_j)(u_i, \overline{u_j}) = 0.$$

If the eigensolutions u_i and u_j belong to different eigenvalues we find

$$(\overline{u_i}, u_j) = 0 \quad \text{if} \quad \lambda_i \neq \lambda_j.$$

Boundedness

We discuss the bounds of the Rayleigh quotient like expression

$$\lambda_R^c(u) = \frac{a(u,u)}{(u,u)} \tag{11}$$

for any possible choice $u \in H_0^1(\Omega)$ (the superscript c reminds that this is not the usual form of the generalized Rayleigh quotient which should use one factor u in (11) as the complex conjugate of the other one).
The real part of λ_R^c supplies a lower bound for the real part of all eigenvalues

$$\begin{aligned}
\Re(\lambda_R^c) &= \Re\left(\frac{a(u,u)}{(u,u)}\right) \\
&= \Re\left(\frac{(\nabla u, \nabla u) - (u, \epsilon_D u)}{(u,u)}\right) \\
&= \frac{(\nabla u, \nabla u) - (u, \Re(\epsilon_D) u)}{(u,u)} \\
&\geq -\frac{(u, \Re(\epsilon_D) u)}{(u,u)} \\
&\geq -\max(\Re(\epsilon_D)) \, . \qquad (12)
\end{aligned}$$

In the same way, we obtain

$$\Im(\lambda_R^c) = -\frac{(u, \Im(\epsilon_D)\, u)}{(u,u)}$$

and therefore

$$\min\left(\Im(\epsilon_D)\right) \leq -\Im(\lambda_R^c) \leq \max\left(\Im(\epsilon_D)\right) \, . \tag{13}$$

Orthogonality between residual and approximate eigensolution
An approximate eigensolution u_a may be given and we look for an related approximate eigenvalue λ_a. We define the condition such that the L^2-norm of the related residual r

$$Au_a - \lambda_a u_a = r \tag{14}$$

becomes a minimum $\|r\|^2 \to \min$

$$(r, Au_a) - \lambda_a(r, u_a) \;=\; (r, r) \quad \to \min\,.$$

A differentiation with respect to λ_a yields the orthogonality condition

$$(r, u_a) = 0\,, \tag{15}$$

which supplies (from (14))

$$(u_a, Au_a) - \lambda_a(u_a, u_a) \;=\; 0 \tag{16}$$

and therefore

$$\lambda_a \;=\; \frac{a(u_a, u_a)}{(u_a, u_a)}\,. \tag{17}$$

This shows that λ_a obtained in this way is exactly the Rayleigh quotient like expression λ_R^c defined in (11).

Generalized Rayleigh Quotient
The generalized Rayleigh quotient is defined as

$$\lambda_R(u) = \frac{a(\overline{u}, u)}{(\overline{u}, u)}. \tag{18}$$

It becomes stationary, if u is an eigensolution of the Helmholtz equation, i. e. , it is $\partial\lambda_R = 0$ for all functions $v \in H_0^1(\Omega)$

$$\begin{aligned} \partial\lambda_R(u; v) &= \left.\frac{d}{dt}\right|_{t=0} \lambda_R(u + tv) \\ &= \frac{2}{(\overline{u}, u)}\left(a(\overline{u}, v) - \lambda_R(\overline{u}, v)\right). \end{aligned}$$

For a discussion of the convergence rate we need the Taylor series expansion up to the term of second order

$$\lambda_R(u + tv) = \lambda_R(u) + \partial\lambda_R(u; v)t + \frac{1}{2}\partial^2\lambda_R(u; v)t^2 + \cdots\,, \tag{19}$$

where the second derivative is

$$\begin{aligned}
\partial^2\lambda_R(u;v) &= \partial\left(\partial\lambda_R(u;v)\right)\\
&= \left.\frac{d}{dt}\right|_{t=0} \partial\lambda_R(u+tv)\\
&= \frac{2}{(\overline{u},u)}(a(\overline{v},v)-\lambda_R(\overline{v},v)) - \frac{8(\overline{u},v)}{(\overline{u},u)^2}(a(\overline{u},v)-\lambda_R(\overline{u},v)).
\end{aligned}$$

Condition of the Eigenvalue Problem

The condition analysis follows the one given in [7]. We assume that an eigenpair u, λ of (8) is given and a small pertubation in form of an additional operator $t \cdot C, t \in \mathbf{R}$ changes the original problem to

$$(A+tC)\,u(t) = \lambda(t)\,u(t).$$

Differentiation at $t=0$ yields

$$A\dot{u} + Cu = \dot{\lambda}u + \lambda\dot{u}.$$

A scalar multiplication from left with the adjoint eigenvector of the unperturbed system gives for the left hand side

$$\begin{aligned}
(\overline{u}, A\dot{u}) + (\overline{u}, Cu) &= (A^*\overline{u}, \dot{u}) + (\overline{u}, Cu)\\
&= (\overline{\lambda}\overline{u}, \dot{u}) + (\overline{u}, Cu),
\end{aligned}$$

and for the right hand side

$$(\overline{u}, \dot{\lambda}u) + (\overline{u}, \lambda\dot{u}) = (\overline{\lambda}\overline{u}, \dot{u}) + \dot{\lambda}(\overline{u}, u)\,,$$

and finally we obtain

$$\dot{\lambda} = \partial\lambda(A;C) = \frac{(\overline{u}, Cu)}{(\overline{u}, u)}\,. \tag{20}$$

As long as the Taylor series expansion holds, it is

$$\lambda(t) = \lambda + \partial\lambda(A;C)t + \cdots\,,$$

and the deviation $\Delta\lambda$ of the perturbed eigenvalue from the original one is in first order

$$\Delta\lambda = \frac{(\overline{u}, Cu)}{(\overline{u}, u)}\,. \tag{21}$$

This result can further be approximated with

$$
\begin{aligned}
|\Delta\lambda| &= \left|\frac{(\overline{u}, Cu)}{(\overline{u}, u)}\right| \\
&\leq \frac{\|u\|^2 \, \|C\|}{|(\overline{u}, u)|} \\
&= \kappa \|C\| \,. \qquad (22)
\end{aligned}
$$

Here, κ denotes the condition number

$$
\kappa = \frac{\|u\|^2}{|(\overline{u}, u)|} = \frac{(u, u)}{|(\overline{u}, u)|},
$$

which is 1 if the eigenvector and its adjoint are parallel. For normal matrices A, the result (22) is not only of first order but exact, see e. g. [5], chapter 2.

Residual based error estimation

From a different point of view we can use the same consideration to estimate the iteration error. Assume that an approximate eigenpair u_a, λ_a is given, which does not fulfills (8) exactly, but leads to a residual r

$$
Au_a - \lambda_a u_a = r \,.
$$

Now we can define an operator which causes a weak perturbation of the original equation with a normalized vector $\|u_a\| = 1$

$$
\begin{aligned}
Au_a - r &= \lambda_a u_a \\
Au_a - r\,(u_a, u_a) &= \lambda_a u_a \\
Au_a - Cu_a &= \lambda_a u_a \,,
\end{aligned}
$$

where the linear operator $C : H_0^1(\Omega) \to H_0^1(\Omega)$ is defined for fixed vectors u_a and r as

$$
Cv = r(u_a, v) \,,
$$

with $\|C\| = \sup\{\|Cv\| : v \in H_0^1, \|v\| \leq 1\} = \|r\|$. This means we consider the given approximated eigenpair as the exact solution of a nearby partial differential equation. Now, using (21) we can estimate the eigenvalue error in first order to

$$
\begin{aligned}
\Delta\lambda &= \frac{(\overline{u}, Cu)}{(\overline{u}, u)} \\
&= \frac{(\overline{u}, r(u, u)u)}{(\overline{u}, u)} \\
&= \frac{(\overline{u}, r)}{(\overline{u}, u)}. \qquad (23)
\end{aligned}
$$

The numerator of the right hand side of the last equation can be rewritten as

$$\begin{aligned}(\overline{u}, r) &= (\overline{u}, Au_a) - \lambda_a(\overline{u}, u_a) \\ &= a(\overline{u}, u_a) - \lambda_a(\overline{u}, u_a) \\ &=: r(\overline{u})\,.\end{aligned}$$

Unfortunately, the exact solution $\overline{u}$ is not available, so that we have go back to (22) to obtain an applicable residual based estimate

$$|\Delta\lambda| \leq \kappa\|C\| = \kappa\|r\|. \tag{24}$$

3 Self-Adjoint Problem

Error analysis for the self-adjoint problem

In the case of a purely real dielectric function ϵ_D the operator A becomes self-adjoint and hence we can represent any approximate eigenfunction u_a (which is assumed to be normalized) by a spectral decomposition using the orthonormal eigenfunctions e_i

$$u_a = \sum_i \alpha_i e_i\,, \quad (u_a, u_a) = 1\,.$$

Here we use a numbering of the modes, which assignes the mode under investigation to λ_0, its next neighbor λ_1 etc. The spectral decomposition of the residual becomes

$$\begin{aligned} r &= (A - \lambda_R)u_a \\ &= \sum_i \alpha_i(\lambda_i - \lambda_R)e_i \\ &= \alpha_0(\lambda_0 - \lambda_R)e_0 + (\lambda_1 - \lambda_R)\left(\alpha_1 e_1 + \sum_{i\geq 2}\frac{\lambda_i - \lambda_R}{\lambda_1 - \lambda_R}\alpha_i e_i\right), \end{aligned}$$

where in the last equation all factors

$$q_i = \frac{\lambda_i - \lambda_R}{\lambda_1 - \lambda_R} \geq 1 \quad \text{for} \quad i > 1\,.$$

Therefore we can represent the true eigensolution, which belongs to the λ_0, with the help of a smoothing operator $S : L^2(\Omega) \to H_0^1(\Omega)$ by

$$\begin{aligned} u_a - \frac{1}{\lambda_1 - \lambda_R}Sr &= u_0 \\ &= \left(\alpha_0 - \alpha_0\frac{\lambda_0 - \lambda_R}{\lambda_1 - \lambda_R}\right)e_0 \\ &= \alpha_0\left(\frac{\lambda_1 - \lambda_0}{\lambda_1 - \lambda_R}\right)e_0, \end{aligned} \tag{25}$$

where (25) defines S. A comparison between the spectral decomposition of u_a and u_0 shows that the operator S has the following properties

- it leaves the spectral component which belongs to λ_0 and the nearest neighbour, λ_1, unchanged $Se_0 = e_0$, $Se_1 = e_1$
- it dampes all higher spectral components, $Se_i = (1/q_i)e_i, i > 1$.

With these properties we have $\|Sr\| \leq \|r\|$ for all r and $\|S\| = 1$. Now it follows, using the inverse triangle inequality

$$
\begin{aligned}
(u_0 - u_a, u_0 - u_a) &= \frac{(Sr, Sr)}{(\lambda_1 - \lambda_R)^2} && (26)\\
&\geq \|u_a\|^2 - \|u_0\|^2 && (27)\\
\|u_0\|^2 &\geq \|u_a\|^2 - \frac{(Sr, Sr)}{(\lambda_1 - \lambda_R)^2} && (28)\\
&\geq 1 - \frac{\|r\|^2}{(\lambda_1 - \lambda_R)^2}. && (29)
\end{aligned}
$$

From (15), (23), (25) and(29) we obtain the desired estimate concerning the eigenvalue error

$$
\begin{aligned}
|\Delta\lambda| &= \frac{|(u, r)|}{(u, u)}\\
&= \frac{1}{(u, u)} \left| (u_a, r) - \frac{(Sr, r)}{\lambda_1 - \lambda_R} \right|\\
&\leq \frac{1}{(u, u)} \frac{(r, r)}{|\lambda_1 - \lambda_R|}\\
&\leq \frac{\|r\|^2}{|\lambda_1 - \lambda_R|} \frac{1}{\left(1 - \frac{\|r\|^2}{(\lambda_1 - \lambda_R)^2}\right)}. && (30)
\end{aligned}
$$

Further we get an estimate for the error of the approximated eigensolution

$$
\|u_0 - u_a\| \leq \frac{\|r\|}{|\lambda_1 - \lambda_R|}. \tag{31}
$$

Both (30) as well as (31) contain the spectral distance $|\lambda_1 - \lambda_R|$ as significant factor. Usually, only an estimate for the eigenvalue of the nearest neighbour is available, e. g. via (24), which is always applicable, such that we have a deviation $\Delta\lambda_1$ from λ_1 to take into account. A small deviation $\Delta\lambda_1 \ll |\lambda_1 - \lambda_R|$ causes an influence on the critical term like

$$
\begin{aligned}
\frac{1}{|\lambda_1 - \lambda_R| \pm \Delta\lambda_1} &= \frac{1}{|\lambda_1 - \lambda_R|} \frac{1}{1 \pm \frac{\Delta\lambda_1}{|\lambda_1 - \lambda_R|}}\\
&\approx \frac{1}{|\lambda_1 - \lambda_R|} \left(1 \mp \frac{\Delta\lambda_1}{|\lambda_1 - \lambda_R|}\right),
\end{aligned}
$$

which transforms the uncertainty $\Delta\lambda_1$ of λ_1 into a realistic estimation

$$|\Delta\lambda|_{\Delta\lambda_1\neq 0} \approx |\Delta\lambda|_{\Delta\lambda_1=0}\left(1+\frac{|\Delta\lambda_1|}{|\lambda_1-\lambda_R|}\right).$$

Minimax definition of the eigenvalues
The minimax defintion of the k-th eigenvalue [6], vol. 2, chap. 7, supplies a useful frame to construct a numerical algorithm. The smallest eigenvalue is characterized by the smallest Rayleigh quotient

$$\lambda(u) = \min_{u\in H_0^1} \frac{a(u,u)}{(u,u)}.$$

The k-th eigenvalue is determined, without the knowledge of the $k-1$ preceding eigenvalues, by

$$\begin{aligned} \lambda_k(u_k) &= \max_{v_i}\min_{u\in V}\frac{a(u,u)}{(u,u)} \\ \text{subject to the constraints}\quad (u,v_i) &= 0, \end{aligned}$$

where $v_i, i = 1\ldots k-1$, is any set of $k-1$ linearly independent functions from the space of admissible functions $V = H_0^1$.

4 Discussion of the Rayleigh-Quotient-Minimization

From these definitions we get the discrete problem through restriction onto $V_h \subset H_0^1$

$$\text{Discretization:}\quad \lambda(u_h) = \min_{v_h\in V_h}\frac{a(v_h,v_h)}{(v_h,v_h)},\quad V_h\subset H_0^1.$$

As the subspace V_h lies in H_0^1, the minimum property of the Rayleigh quotient is passed to the discrete formulation. If the numerical agorithm to detemine the lowest eigenvalue and eigenvector is based on this property, one obtains in a straight forward way the Rayleigh quotient-minimization procedure proposed by Faddejew and Faddejewa [1] in 1963

$$R(u_h+tv_h) = \min_{t\in R}\frac{a(u_h+tv_h,u_h+tv_h)}{(u_h+tv_h,u_h+tv_h)},$$

i. e. , the Rayleigh quotient is one dimensional minimized in direction of the function v_h. The use of all nodal basis functions results in the nonlinear Gauß-Seidel-minimization, the use of global, mutually orthogonal basis functions leads to the nonlinear cg-minimization (Bradwick, Fletcher 1966 [4], Polak 1971 [3]).

A multigrid formulation (McCormick 1992, [2]) is obtained, if additional nodal basis functions $v_H \in V_H$ of a coarser space $V_H \subset V_h$ as search directions are added

$$
\begin{aligned}
R(u_h + tv_H) &= \min_{t\in R} \frac{a(u_h+tv_H, u_h+tv_H)}{(u_h+tv_H, u_h+tv_H)} \\
&= \min_{t\in R} \frac{a(u_h,u_h) + 2t\cdot a(u_h,v_H) + t^2 a(v_H,v_H)}{(u_h,u_h)+2t\cdot(u_h,v_H)+t^2(v_H,v_H)}.
\end{aligned}
$$

With the help of the mapping π_h, which converts the nodal value representation of a given vector $\underline{v_h}$ into the related finite element function v_h with basis functions Ψ_i,

$$\pi_h : \mathbf{R}^N \to L^2, \quad \pi_h \underline{v_h} = v_h = \sum_i \underline{v_h}(i)\Psi_i$$

one obtains an algorithmically direct accessible form of the representation of fine grid quantities on the coarse mesh

$$
\begin{aligned}
a(u_h, v_H) &= \langle A_h \underline{u_h}, \pi_h^{-1} v_H \rangle \\
&= \langle A_h \underline{u_h}, \pi_h^{-1}\pi_H\pi_H^{-1} v_H \rangle \\
&= \langle (\pi_h^{-1}\pi_H)^* A_h \underline{u_h}, \underline{v_H} \rangle.
\end{aligned}
$$

Altogether, this gives the restrictions

$$A_h\underline{u_h} \to I_h^H \cdot (A_h \cdot \underline{u_h}), \quad M_h\underline{u_h} \to I_h^H \cdot (M_h \cdot \underline{u_h}),$$

where I_h^H means the usual restriction operator. Through the additional restriction one has to realize approximately twice the effort to perform one V-cycle, in comparison with the linear procedure, if numerator and denominator are updated recursively.

Algorithmic realization of the minimax-principle

So far we have concentrated on the minimization of the discrete Rayleigh quotient, i. e. , we have approximated the lowest eigenvector. In order to approximate the next few eigenvectors, we apply the minimax-principle in the following way. Assume, a space of $k-1$ linear independend approximations $S_{k-1} = (s_1, s_2, \ldots, s_{k-1})$ is given. Now we carry out one step of the minimax procedure in modified form

$$
\begin{aligned}
\lambda_k(u_k) &= \max_{s_i \in S_{k-1}} \min_{u\in V_h} \frac{a(u,u)}{(u,u)} \\
\text{subject to the constraints} \quad (u, s_i) &= 0, \quad i = 1\ldots k-1\,.
\end{aligned}
$$

Here we have restricted the whole space over which to maximize to the space S_{k-1}. If we have found a function u, which fulfills the minimum condition, we have to obey the maximum condition for all $k-1$ linear independend vectors

$v \in S_{k-1}$. With $\alpha_i \in \mathbf{R}, i = 1 \ldots k$, the task to determine the maximum can be formulated as

$$\begin{aligned} \lambda_k(s_k) &= \max_{\alpha_1,\ldots,\alpha_k} \frac{a(\sum_{i=1}^k \alpha_i s_i, \sum_{i=1}^k \alpha_i s_i)}{(\sum_{i=1}^k \alpha_i s_i, \sum_{i=1}^k \alpha_i s_i)} \\ &= \max_{\underline{\alpha} \in \mathbf{R}^k} \frac{a_{Ritz}(\underline{\alpha}, \underline{\alpha})}{(\underline{\alpha}, \underline{\alpha})_{Ritz}} . \end{aligned}$$

This shows, that the determination of the maximum required in the minimax-principle is the same as to solve a new eigenproblem of the dimension k. This Ritz-projection step and the different proposals for a multi-level minimization of the Rayleigh quotient are the basis of the algorithmic proposals in in the following two sections. In the following we consider only the case of the self-adjoint eigenproblem. The more general symmetric complex eigenproblem will be the topic of future work.

The basic procedure of all proposed algorithms is the one-dimensional minimization along a given function p. This procedure, lineMin, is described in Fig. 1 using a pseudo code notation. It determines the scale-factor T, and updates the vectors Ax, Mx and the numbers xAx, xMx. Note, that the procedure itself does not know the current solution vector approximation. All information necessary is contained within these vectors and numbers.

In order to test the behavior of the Rayleigh quotient minimization in the context of typical applications from integrated optics, we investigated at first 1D models two different classes of problems. The first class concerns single and multimode waveguides, which are characterized by a clear separation of the lowest modes. The second class concerns coupler structures being in fact multimode waveguides too, but posses low eigenvalues, which are very close to each other. All iterations have been performed with uniform refinement.

4.1 Nonlinear Gauß-Seidel

Fig. 2 demonstrates the application of the line search procedure lineMin in the classic way, resulting in the standard node based minimization routine [1].

The algorithmic realization of the multigrid procedure is explained in Fig. 3. Note, that the smoother, which is in this case the nonlinear Gauß-Seidel-smoother can be exchanged by a cg-based smoother discussed later.

Fig. (4) and (5) show the typical iteration behavior using an example from the first problem class (slab waveguide). As expected, the number of iterations remains asymptotically constant and the convergence rate is excellent.

Convergence in case of modes with close eigenvalues

In some applications from the second class of problems, the nonlinear Gauß-

function $(Ax, Mx, xAx, xMx, T) = \mathbf{lineMin}(A, M, Ax, Mx, xAx, xMx, p)$

$$\begin{array}{rclcrcl} \beta & = & (Ax)'p; & \quad & \gamma & = & p'Ap; \\ b & = & (Mx)'p; & \quad & c & = & p'Mp; \end{array}$$

$$\begin{array}{rcl} q & = & b \cdot \gamma - \beta \cdot c; \\ r & = & c \cdot xAx - \gamma \cdot xMx; \\ s & = & xMx \cdot \beta - b \cdot xAx; \end{array}$$

$$t_{1,2} = \frac{r}{2q} \cdot \left(1 \pm \sqrt{1 - \frac{4qs}{r^2}} \right);$$

$$\begin{array}{rclcrcl} Ax_{1,2} & = & Ax + t_{1,2}A\,p; & \quad & Mx_{1,2} & = & Mx + t_{1,2}M\,p; \\ xAx_{1,2} & = & xAx + 2t_{1,2}\beta + t_{1,2}^2\gamma; & \quad & xMx_{1,2} & = & xMx + 2t_{1,2}b + t_{1,2}^2 c; \end{array}$$

$$\begin{array}{rcl} RQ_{1,2} & = & \dfrac{xAx_{1,2}}{xMx_{1,2}}; \\ i & = & i(\min\limits_{1,2} RQ_i); \\ Ax & = & Ax_i; \\ xAx & = & xAx_i \\ Mx & = & Mx_i; \\ xMx & = & xMx_i; \\ T & = & t_i; \end{array}$$

end;

FIG. 1. *Line search procedure*

```
function   (Ax, Mx, xAx, xMx, d) = NLGS(A, M, Ax, Mx, xAx, xMx)

   d = 0;

   for i = 1 : N   // all nodes
                         p                   = 0;
                         p(i)                = 1;
             (Ax, Mx, xAx, xMx, t) = lineMin(A, M, Ax, Mx, xAx, xMx, p);
                         d                   = d + t · p;
   end;

end;
```

FIG. 2. *Nonlinear Gauß-Seidel minimization*

Seidel-minimization becomes qualitatively slower than before. In order to analyze this behavior, let us consider a linearized minimization procedure. If we are close to the exact eigenvalue, we can expand the Rayleigh quotient into a Taylor series up to the terms of second order (see (19)) and the requirement

$$\frac{\partial \lambda_R(u+tv)}{\partial t} = 0$$

is approximately fulfilled by

$$t = -\frac{\partial \lambda_R(u;v)}{\partial^2 \lambda_R(u;v)} .$$

In general, we have a situation, where $\partial\lambda_R(u;v)$ is close to zero for all v and $|\partial^2\lambda_R(u;v)| \gg |\partial\lambda_R(u;v)|$. In this case, the numerator $\partial\lambda_R(u;v)$ dominates the nodal correction. The spectral decomposition of the eigenvector approximation u with the help of the orthonormal set of eigenfunctions e_i gives the representation

$$\begin{aligned} \partial\lambda_R(u;v) &= \frac{2}{(u,u)}\left(a(u,v) - \lambda_R(u,v)\right) \\ &= \frac{2}{(u,u)}\left(\sum_i \alpha_i(\lambda_i - \lambda_R)(e_i,v)\right) . \end{aligned}$$

This shows that the nodal correction depends on the search-direction v and on the spectral distance between the Rayleigh quotient and the neighbouring modes. If $\Delta\lambda$ denotes the spectral distance between the neighbouring modes, we have

$$\text{if} \quad \Delta\lambda \to 0 \quad \text{then} \quad t \to 0 \quad .$$

function $(Ax, Mx, xAx, xMx, d) =$ **RQMinMG**$(l, A_l, M_l, Ax_l, Mx_l, xAx, xMx)$

$$
\begin{aligned}
d_l &= 0; \\
dNew_l &= 0;
\end{aligned}
$$

`// initialize`

if $\quad l = 1$ `// exact solution on the coarsest grid`

$$\lambda_l(x_1) = \min_{v \in V_1} \frac{x_1' A_1 \, x_1}{x_1' M_1 \, x_1}$$

else

$$
\begin{aligned}
& && \texttt{//presmooth} \\
(Ax_l, Mx_l, xAx, xMx, dNew_l) &= \mathbf{NLGS}(A_l, M_l, Ax_l, Mx_l, xAx, xMx); \\
d_l &= d_l + dNew_l \\
& && \texttt{//restriction} \\
A_{l-1} &\leftarrow A_l; \\
M_{l-1} &\leftarrow M_l; \\
Ax_{l-1} &\leftarrow Ax_l; \\
Mx_{l-1} &\leftarrow Mx_l; \\
(Ax_{l-1}, Mx_{l-1}, xAx, xMx, d_{l-1}) &= \mathbf{RQMinMG}(l-1, A_{l-1}, M_{l-1}, \\
& \quad\; Ax_{l-1}, Mx_{l-1}, xAx, xMx); \\
& && \texttt{//prolongation} \\
d_{l-1} &\rightarrow dNew_l; \\
d_l &= d_l + dNew_l; \\
Ax_l &= Ax_l + A_l \cdot dNew_l; \\
Mx_l &= Mx_l + M \cdot dNew_l; \\
& && \texttt{//postsmooth} \\
(Ax_l, Mx_l, xAx, xMx, dNew_l) &= \mathbf{NLGS}(A_l, M_l, Ax_l, Mx_l, xAx, xMx); \\
d_l &= d_l + dNew_l;
\end{aligned}
$$

end;

FIG. 3. *Multigrid procedure using the nonlinear Gauß-Seidel smoother*

iterations

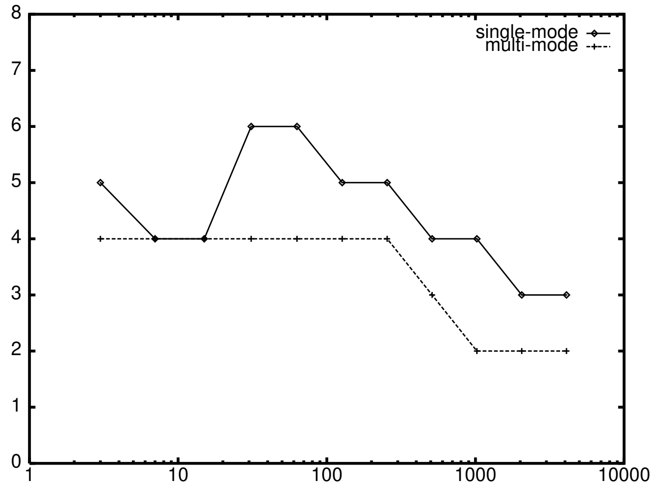

FIG. 4. *Number of iterations vs. number of nodes (nonlinear Gauß-Seidel minimization)*

conv.rate

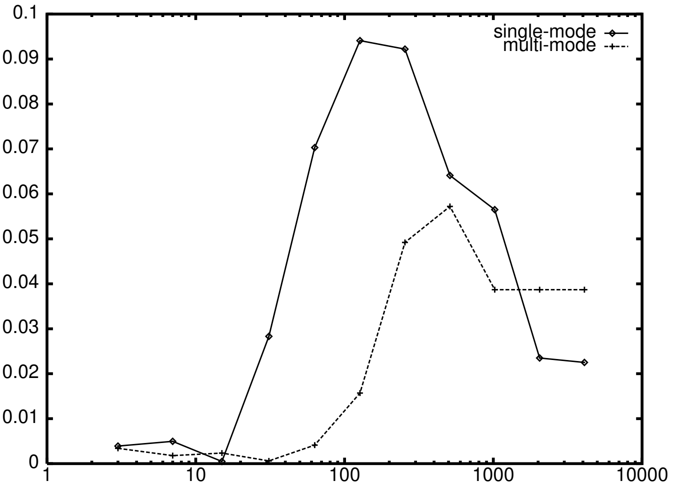

FIG. 5. *Convergence rate vs. number of nodes nonlinear Gauß-Seidel minimization)*

The convergence rate diminishes when the neighboring eigenvalues becomes closer. The refractive index distribution of an test-problem characterized by very close neighboring modes is given in Fig. 7. The coefficients given here do not belong to any real problem, because the coefficient-jumb is much larger than in real problems, but the model is well suited to test the iteration behavior of the different algorithms. The related iteration numbers are shown in Fig. iter1D. The nonlinear Gauß-Seidel procedure discussed so far needs on the second level a very large number of iterations, which makes the method for practical purposes to slow.

4.2 Nonlinear CG

In order to have an alternative minimization procedure, we consider the nonlinear conjugate-gradient based minimization [4] and [3].

function $(Ax, Mx, xAx, xMx, d) = \mathbf{NLCG}(A, M, Ax, Mx, xAx, xMx)$

$$RQ = \frac{xAx}{xMx};$$

$$d = 0$$

for $i = 1 : cgSteps$

$$g_1 = \frac{2}{xMx}(A\,x - RQ\,Mx);$$

if $i == 1$

$p = -g_1;$

else

$$\epsilon_1 = \frac{(g_1' - g_2')g_1}{g_2' g_2};$$

$p = -g_1 + \epsilon_1 p_1;$

end;

$(Ax, Mx, xAx, xMx, t) = \mathbf{lineMin}(A, M, Ax, Mx, xAx, xMx, p);$

$d = d + t \cdot p;$

$p_1 = p;$

$g_2 = g_1;$

$RQ = \frac{xAx}{xMx};$

end;

FIG. 6. *Nonlinear conjugate-gradient minimization (Polak-Ribiere)*

The nonlinear cg-minimization in its multigrid version, with 3 smoothing steps at

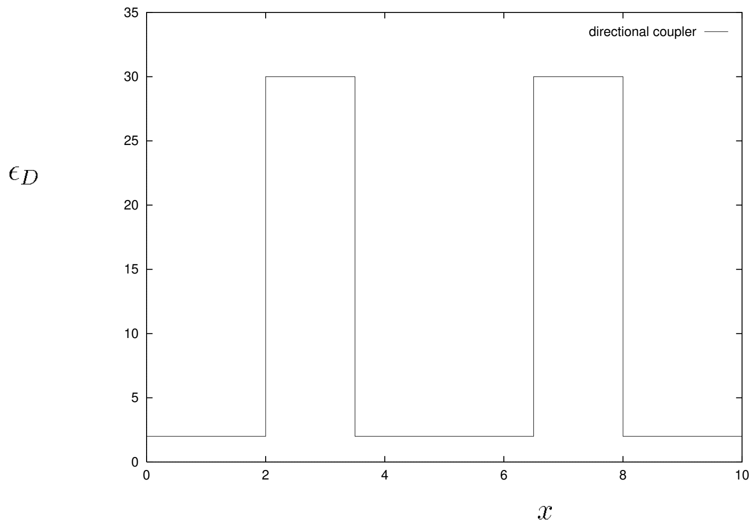

FIG. 7. *Refractive index distribution for the 1D model problem*

each level, applied to our model problem supplies a more stable iteration (Fig.8) than the nodal based minimization. Moreover, the number of iterations needed to meet the desired accuracy (a relative error of 10^{-7} in the eigenvalue) is lower.

4.3 Preconditioned Nonlinear CG

Preconditioning of conjugate gradient methods has been proved to be of great advantage not only for linear systems but also for the Rayleigh quotient minimization [9]. As we have the nonlinear cg and the nonlinear Gauß-Seidel available, we construct a preconditioned cg combining both methods in generalization of the linear preconditioning.
This combination is represented in Fig. 9. The nodal based minimization is imbedded within the conjugate-gradient frame like it is the usual way to perform a preconditioned cg-routine. In contrast to the way of preconditioning described in [9], which uses a linear system solution, we propose a complete nonlinear method.
In this way, the structure of the cg-method is maintained.

5 Subspace Iterations

From the analysis of the iteration behavior it became clear that the convergence rate diminishes when the eigenvalues occur in clusters. This remains true independent of the iteration procedure used to minimize the Rayleigh quotient. A common idea to overcome this difficulty is to embed the single-vector iteration into a subspace iteration such that the closely neighboring modes are iterated

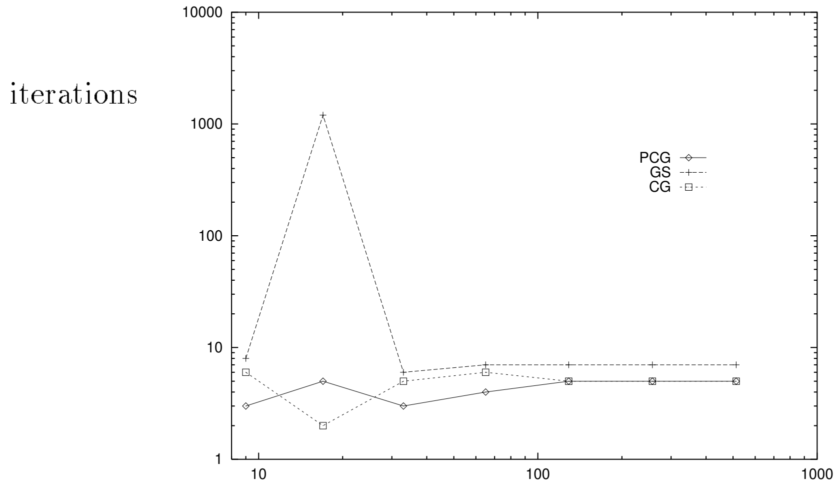

FIG. 8. *Number of iterations vs. number of nodes*

simultaneously. Within this section we we want to discuss possible realizations of such subspace iterations based on the minimization routines (nonlinear cg, nonlinear Gauß-Seidel, preconditioned nonlinear cg).

5.1 Simultaneous Nonlinear CG

Fig. 10 demonstrates the structure of cg-based simultaneous multigrid Rayleigh Quotient minimization procedure. The only difference with respect to the standard procedure Fig. 6 lies in the use of the orthogonalization algorithm orthoL^2, which makes the current search direction orthogonal to all preceeding $col - 1$ ones. In order to complete the discussed minimax-approach, a Ritz-step has to be added after each minimization run.

5.2 Simultaneous Nonlinear Gauß-Seidel

In contrast to the nonlinear cg, it becomes impossible to orthogonalize the single search direction of the nodal based nonlinear Gauß-Seidel with respect to the preceeding vectors of the subspace S, if the basis vectors of the subspace was already mutually orthogonal. Instead of using search directions with numbers of degrees of freedom larger than the dimension of the subspace, which would make the orthogonalization possible, we propose an orthogonalization *after* a the complete run over each level.

function $(Ax, Mx, xAx, xMx, d) = \mathbf{NLPCG}(A, M, Ax, Mx, xAx, xMx)$

$d = 0;$

for $\quad i = 1 : cgSteps$

$$RQ = \frac{xAx}{xMx};$$

$$g_1 = \frac{2}{xMx}(A\,x - RQ\,Mx);$$

$$\begin{array}{rclrcl} Ax_c & = & Ax; & Mx_c & = & Mx; \\ xAx_x & = & xAx; & xMx_x & = & xMx; \end{array} \quad \texttt{// dummies}$$

$$(Ax_c, Mx_c, xAx_c, xMx_c, pg_1) \quad = \quad \mathbf{NLGS}(A, M, Ax_c, Mx_c, xAx_c, xMx_c);$$

if $i == 1$

$\quad p = -pg_1;$

else

$$\epsilon_1 = \frac{(g_1' - g_2')pg_1}{g_2' pg_2};$$

$\quad p = -pg_1 + \epsilon_1 p_1;$

end;

$$(Ax, Mx, xAx, xMx, t) = \mathbf{lineMin}(A, M, Ax, Mx, xAx, xMx, p);$$

$$\begin{array}{rcl} d & = & d + t \cdot p; \\ p_1 & = & p; \\ g_2 & = & g_1; \\ pg_2 & = & pg_1; \end{array}$$

end;

end;

FIG. 9. *Preconditioned nonlinear conjugate-gradient minimization*

```
function   (Ax, Mx, xAx, xMx, d) = NLCGn(col, S, A, M, Ax, Mx, xAx, xMx)

      // It is assumed that x is L^2 -orthogonal to all n
      // column-vectors of the subspace S

RQ  =  xAx / xMx;


   for     i = 1 : cgSteps
           g_1 = 2/(xMx) (A x - RQ M x);

           if i == 1
               p = -g_1;
           else
               ε_1 = ((g_1' - g_2') g_1) / (g_2' g_2);
               p = -g_1 + ε_1 p_1;
               p = orthoL^2(col, S, p);
           end;
            (Ax, Mx, xAx, xMx, t) = lineMin(A, M, Ax, Mx, xAx, xMx, p);
                         d          = d + t · p;
                         p_1        = p;
                         g_2        = g_1;
                         RQ         = xAx / xMx;
  end;

end;
```

FIG. 10. *Subspace nonlinear conjugate-gradient minimization (Polak-Ribiere)*

```
function   (Ax, Mx, xAx, xMx, d) = NLGSn(col, S, A, M, Ax, Mx, xAx, xMx)

        d    = 0;
     xAx_0   = xAx;
     xMx_0   = xMx;
      Ax_0   = Ax;
      Mx_0   = Mx;

  for i = 1 : N   // all nodes
                         p                  = 0;
                         p(i)               = 1;
           (Ax, Mx, xAx, xMx, t) = lineMin(A, M, Ax, Mx, xAx, xMx, p);
                         d                  = d + t · p;
  end;

                       //defect correction via reorthogonalization

         d = orthoL^2(col, S, d);

                                   // correction of Ax,Mx,xAx,xMx

         Ax    =  Ax_0 + Ad;
         Mx    =  Mx_0 + Md;
        xAx    =  xAx_0 + 2dAx_0 + dAd;
        xMx    =  xMx_0 + 2dMx_0 + dMd;

end;
```

FIG. 11. *Subspace nonlinear Gauß-Seidel minimization*

5.3 Simultaneous Nonlinear Preconditioned CG

The preconditioned nonlinear cg demonstrated here uses the nonlinear Gauß-Seidel smoother as preconditioner. The algorithm differs only slightly from the basic routines. The standard preconditioner NLGS has to be replaced by its subspace variant NLGSn. In comparison with the unpreconditioned subspace cg, the procedure does not need an explicit orthogonalisation of the search direction p, because the preconditioned gradient pg_1 is already orthogonalized.

6 Discussion and Application of the Algorithms in 2D

The main field of applications of the algorithms described so far will be the analysis of the lowest eigenmodes of 2D cross-sections of integrated optics components. As example we investigate the four lowest modes of a strip-loaded coupling structure . This problem has been turned out to be a difficult one [8].
The geometry and the initial triangulation are given in Fig. 15. The normalized dielectric coefficients are $\epsilon_D = 16.43222$ in air, $\epsilon_D = 165.1258$ in cladding and $\epsilon_D = 187.728$ in guide and substrate. These numbers results from a vacuum-wavelength of $1.55\mu m$. The intensity plot of the fundamental mode, given in Fig. 16, shows that the essential part of the solution is concentrated in the waveguide slab, but a slight part is distributed around the ribs. Fig. 13 shows the related number of iterations for all four modes simulated simultaneously versus the number of nodes. The preconditioned cg-version needs the smallest number of iterations, followed by the standard multigrid cg-minimization and the point Gauß-Seidel minimization. However, if we consider the accumulated CPU-time Fig. 14, it turns out, that the minimization using the pure cg-method is the most effective one.

Conclusions

The following results have been presented:

1. An effective implementation of the multilevel based Rayleigh quotient minimization was developed.
2. The efficiency of the nonlinear cg-method as basic minimization procedure has been confirmed.
3. The multilevel minimization algorithm has been extended to a subspace-multilevel minimization procedure.
4. In analogy to linear preconditioned cg-algorithms a hybrid method consisting of the nonlinear cg-method with an embedded nodal basis smoother has been proposed.
5. Numerical experiments based on realistic integrated optics components have shown the applicability of the proposed algorithms.

function $(Ax, Mx, xAx, xMx, d) = \mathbf{NLPCG}(col, S, A, M, Ax, Mx, xAx, xMx)$

$d = 0;$

for $i = 1 : cgSteps$

$$RQ = \frac{xAx}{xMx};$$

$$g_1 = \frac{2}{xMx}(A\,x - RQ\,Mx);$$

$$\begin{array}{rclrcl} Ax_c & = & Ax; & Mx_c & = & Mx; \\ xAx_x & = & xAx; & xMx_x & = & xMx; \end{array} \quad \texttt{// dummies}$$

$$(Ax_c, Mx_c, xAx_c, xMx_c, pg_1) \;=\; \mathbf{NLGSn}(col, S, A, M, Ax_c, Mx_c, xAx_c, xMx_c);$$

if $i == 1$

$p = -pg_1;$

else

$$\epsilon_1 = \frac{(g_1' - g_2')pg_1}{g_2' pg_2};$$

$p = -pg_1 + \epsilon_1 p_1;$

end;

$$(Ax, Mx, xAx, xMx, t) = \mathbf{lineMin}(A, M, Ax, Mx, xAx, xMx, p);$$

$$\begin{array}{rcl} d & = & d + t \cdot p; \\ p_1 & = & p; \\ g_2 & = & g_1; \\ pg_2 & = & pg_1; \end{array}$$

end;

end;

FIG. 12. *Preconditioned nonlinear conjugate-gradient minimization*

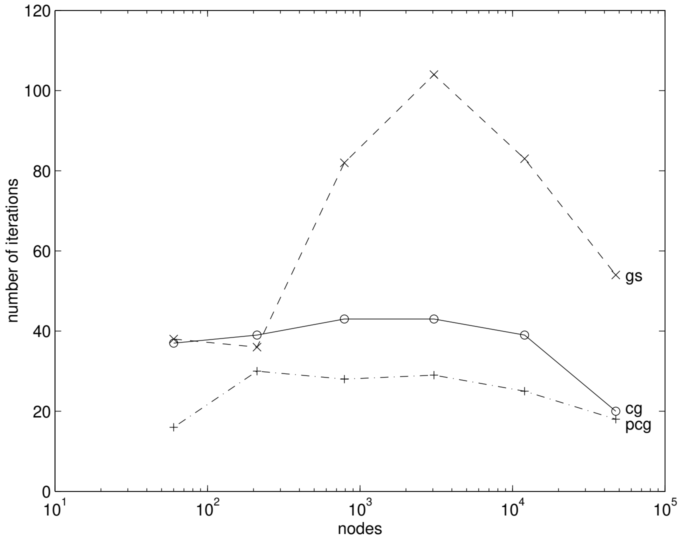

FIG. 13. *Number of iterations for the 4 lowest modes vs. number of nodes, different minimization strategies are compared*

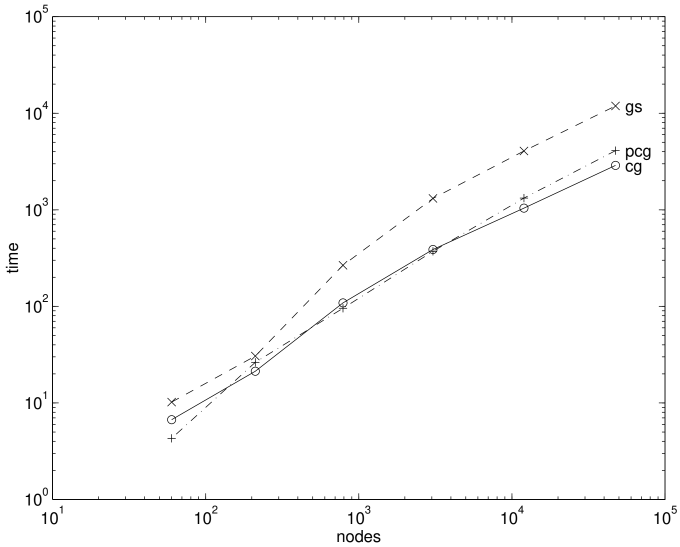

FIG. 14. *Accumulated CPU-time vs. number of nodes.*

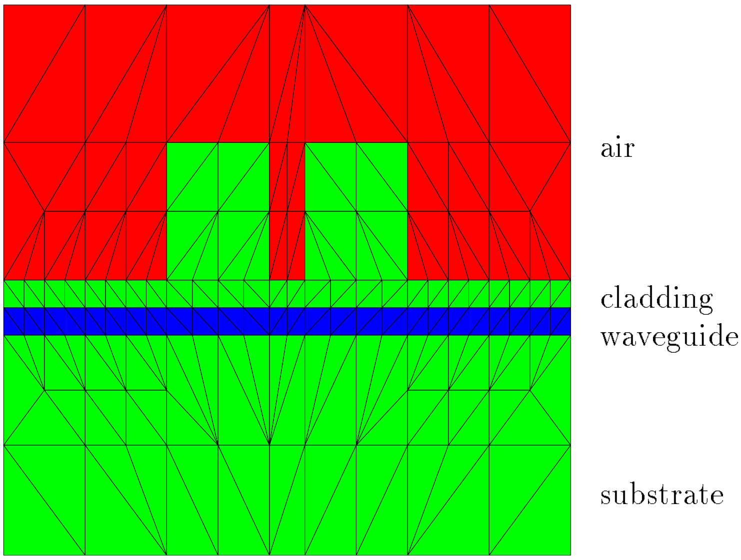

Fig. 15. *Geometry and initial triangulation for the strip loaded coupling structure*

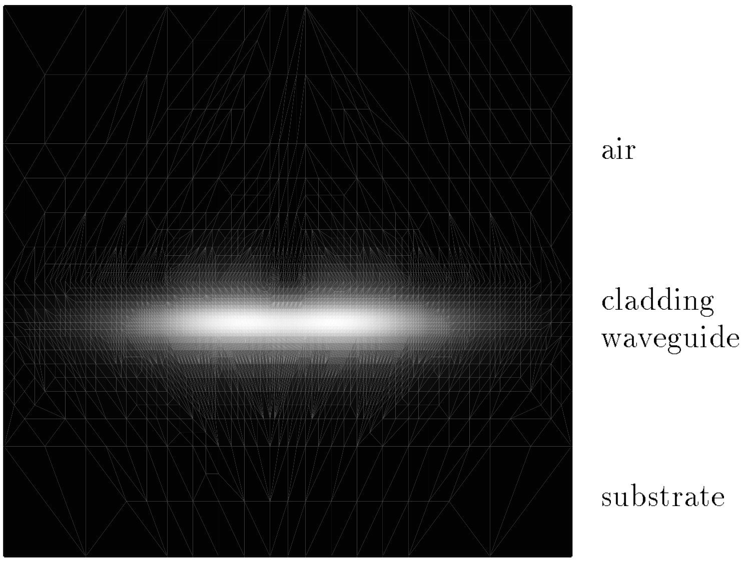

Fig. 16. *Intensity plot of the fundamental mode of the strip loaded coupling structure*

Acknowledgment

This project is funded by the *Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie*, Germany, Grant Number DE7ZIB, in the program *Anwendungsorientierte Verbundprojecte auf dem Gebiet der Mathematk.* I gratefully acknowledge the advice given by Prof. P. Deuflhard and Dr. R. Kornhuber (WIAS).

REFERENCES

[1] D. K. Faddejew, W. N. Faddejewa, *Computational methods of linear algebra*, San Francisco, 1963

[2] S. McCormick, *Multilevel Projection Methods for Partial Differential Equations*, CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM, Philadelphia, 1992

[3] E. Polak, *Computational Methods in Optimization*, Academic Press, New York, 1971

[4] W. W. Bradbury, R. Fletcher, *New iterative methods for the solution of the eigenproblem*, Numer. Math. 9, 259-267, 1966

[5] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, London, 1965

[6] R. Courant, D. Hilbert,*Methoden der mathematischen Physik*, Springer Verlag, Berlin, 1937

[7] P. Deuflhard, A. Hohmann *Numerische Mathematik*, Walter de Gruyter, Berlin, New York, 1991

[8] R. Accornero et al *Finite different methods for the analysis of integrated optical wavegides*, Electronics Letters, Vol. 26 No. 3, 1959-1960, 1990

[9] F. Sartoretto, G. Pini, and G. Gambolati *Accelerated Simultaneous Iterations for Large Finite Element Eigenproblems*, J. Comp. Physics 81, 53-69 1989