Solving an Eigenvalue Problem in Laser Simulation

Dissertation zur Erlangung des
naturwissenschaftlichen Doktorgrades
der Bayerischen Julius-Maximilians-Universität Würzburg

Vorgelegt von

David Seider

aus

Dshetyssai/Kasachstan

Würzburg 2004

Eingereicht am: 8. Juni 2004

bei der Fakultät für Mathematik und Informatik

1. Gutachter: Prof. Dr. Christoph Pflaum

2. Gutachter: Prof. Dr. Manfred Dobrowolski

Tag der mündlichen Prüfung: 6. August 2004

Jesaja 43,2-3:
Ich will vor dir hergehen und das Bergland eben machen, ich will die ehernen Türen zerschlagen und die eisernen Riegel zerbrechen
und will dir heimliche Schätze geben und verborgene Kleinode, damit du erkennst, dass ich der HERR bin, der dich beim Namen ruft, der Gott Israels.

Isaiah 43,2-3:
I will go before you and will level the mountains; I will break down gates of bronze and cut through bars of iron.
I will give you the treasures of darkness, riches stored in secret places, so that you may know that I am the LORD, the God of Israel, who summons you by name.

Acknowledgments

First of all, I want to thank my wife Eleonore for her patience and her never ending encouragement. She is the most precious gift that God could have given to me.
I'm very grateful to Prof. Dr. Christoph Pflaum (University of Erlangen, Germany) who first started me on the problem of laser simulation. In this thesis he will recognize many of the uncountable ideas that we have discussed during his time in Würzburg.
Furthermore, I want to thank Prof. Dr. Manfred Dobrowolski (University of Würzburg, Germany) for many valuable hints and discussions. It was a privilege to be allowed the freedom I enjoyed under his supervision.
I also want to mention Dr. Konrad Altmann (Las-Cad GmbH, Munich, Germany), to whom I am indebted for patiently answering my questions on laser physics and for making the simulation software LASCAD^{TM} freely available.
Special thanks go to Alexander Linke (now at the Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany), for implementing the version of EXPDE that I used for the numerical tests, and to Michael L. Flegel (University of Würzburg, Germany), for doing a very good job in administrating the computer system and proof reading the language of parts of this thesis.
But most of all, I want to thank my Lord and Saviour Jesus Christ for loving me, independently of whether I achieve a great deal or not, and for giving me the strength for everyday life.

Würzburg, June 2004 David Seider

Contents

Introduction

"$\grave{\alpha}\nu\alpha\lambda\tilde{\upsilon}\sigma\iota\varsigma$", the Greek word for "analysis" means dissection. A dissection is typically done for all types of real-world problems. The problem is divided into sub-problems which can be handled easier and one hopes that some recombination of the solutions will lead to a solution of the original problem.
It is the same with numerical laser simulation. There are people who model the laser behavior, people who deal with mathematical aspects of the model, as e.g. the approximation by finite elements, and finally, people who analyze and implement solution methods, such as a solver for the resulting system of linear equations.
A standing wave of a laser resonator, for instance, can be modeled by the homogeneous Helmholtz equation

$$-\Delta E - k^2 E = 0.$$

Much effort has been spent on analyzing the properties of a finite element approximation of the Helmholtz equation. And also much research has been done on algorithms for solving the discrete equation as efficiently as possible. However, combining all results leads to a relatively poor yield with respect to computing standing waves of a laser resonator.
Similarly, the results of methods using the so-called paraxial approximation (a Schrödinger-type equation) of the Helmholtz equation leave a lot to be desired in many real-world examples.
We are very pleased to present a new approach for computing laser cavity eigenmodes in this thesis covering modeling, as well as numerical and computational aspects. Therefore, some parts of this thesis are written in a rigorous mathematical manner and other parts in an informal style (from the viewpoint of a mathematician).
The thesis is organized as follows: Chapter 1 contains a short overview on solving the Helmholtz equation with the help of finite elements. The main part of Chapter 2 is dedicated to the analysis of a one-dimensional model problem containing the main idea of a new model for laser cavity eigenmodes which is derived in detail in Chapter 3. Chapter 4 comprises a convergence theory for the approximate solution of quadratic eigenvalue problems. In Chapter 5, a stabilized finite element discretization of the new model is described and its convergence is proved by applying the theory of Chapter 4. Chapter 6 contains computational aspects of solving the resulting system of equations and, finally, Chapter 7 presents numerical results for various configurations, demonstrating the practical relevance of our new approach.

1 The Helmholtz Equation and Finite Element Methods - A Survey

1.1 The Helmholtz Equation in Two Examples

Many real-world phenomena can be described by the famous wave equation

$$\Delta W(x,y,z;t) = \frac{1}{c^2}\frac{\partial^2 W(x,y,z;t)}{\partial t^2} \tag{1}$$

where the wave $W(x,y,z;t)$ is a real-valued function depending on space coordinates $(x,y,z) \in \mathbb{R}^3$ and time coordinate $t \in \mathbb{R}_+ := \{t \in \mathbb{R} | t \geq 0\}$. Here, Δ denotes the Laplace operator with respect to the spatial coordinates (x,y,z) and c is the velocity of the wave in the medium. At first, we do not specify the domain and the exact function space for W. Later on, when a rigorous mathematical formulation has to be given, this will be done.
If a time-harmonic behavior of the wave is assumed, i.e. if W is of the form

$$W(x,y,z;t) = \tilde{W}(x,y,z) \cdot e^{i\omega t} \tag{2}$$

with time-frequency ω, equation (1) is equivalent to the so-called Helmholtz equation or scalar wave equation for $\tilde{W}$

$$-\Delta \tilde{W}(x,y,z) - \frac{\omega^2}{c^2}\tilde{W}(x,y,z) = 0. \tag{3}$$

Actually, the wave W is real-valued, such that, more precisely, one has

$$W(x,y,z;t) = \mathrm{Re}\left[\tilde{W}(x,y,z) \cdot e^{i\omega t}\right],$$

where $\mathrm{Re}[\cdot]$ stands for taking the real part of a complex number. But as it is usually done, we will assume the representation (2) with complex-valued $\tilde{W}(x,y,z)$. Instead of representation (2), one also could use

$$W(x,y,z;t) = \hat{W}(x,y,z) \cdot e^{-i\omega t},$$

which would lead to the equation (3) for $\hat{W}$ as well. This observation shows, that the Helmholtz equation does not incorporate information on whether the wave W moves forward or backward in time. So, when we speak of $\tilde{W}$ traveling e.g. in positive z-direction, we mean this with respect to representation (2) where the exponent is $+i\omega t$.
Let us now consider two examples for the occurrence of the Helmholtz equation.

1.1.1 Example: Acoustic Scattering Problem

The following presentation is based on [41]; so, we refer to this book for the notions and the proofs of the statements in this example. As it is explained in [41], acoustic waves can under certain conditions be described by a small perturbation $P(x, y, z; t)$ of a reference pressure P_0. This perturbation has to fulfill the equation

$$\Delta P - \frac{1}{c^2}\frac{\partial^2 P}{\partial t^2} = 0. \tag{4}$$

If P is time-harmonic, i.e. if it can be written as

$$P(x, y, z; t) = p(x, y, z) \cdot e^{i\omega t},$$

equation (4) is equivalent to the equation

$$-\Delta p - k^2 p = 0, \tag{5}$$

with wave number $k := \omega/c$.
For scattering wave problems, additional conditions for the wave p are imposed. One type are boundary conditions on the surface $\Gamma_s \subset \partial\Omega_s$ of the (solid) scatterer $\Omega_s \subset \mathbb{R}^3$, which we abbreviate by

$$\mathcal{B}_s p = 0, \tag{6}$$

where usually the operator $\mathcal{B}_s$, which maps p to a function on the boundary Γ_s, is chosen as $\mathcal{B}_s p := p|_{\Gamma_s}$ or $\mathcal{B}_s p := \partial_n p|_{\Gamma_s}$. Here, $\partial_n p$ denotes the derivative of p in the direction of the normal of Ω_s on the boundary Γ_s. Furthermore, in order to avoid spurious reflections "from infinity", the so-called Sommerfeld condition has to be satisfied

$$p = O(R^{-1}),\ ikp - \frac{\partial p}{\partial R} = o(R^{-1}) \qquad \text{for } R \to \infty. \tag{7}$$

Here, p is viewed in spherical coordinates, $\partial/\partial R$ denotes the derivative in radial direction, and $O(\cdot)$ and $o(\cdot)$ are the well-known Landau symbols. It can be shown, that a function that fulfills both the Helmholtz equation and the (second) radiation condition in (7) also satisfies the (first) decay condition in (7).
Naturally, the acoustic scattering problem (5), (6), (7) is formulated on an infinite domain $\Omega_\infty = \mathbb{R}^3 \setminus \Omega_s$, see Figure 1.
To reduce this problem to a problem on a finite domain, an artificial, e.g. spherical, boundary Γ_r that envelops the scatterer Ω_s is introduced. If the coefficient k is constant outside the sphere confined by Γ_r, by the use of a so-called Dirichlet-to-Neumann map $\mathcal{B}_r$ (see [43]) on this boundary, the problem (5), (6), (7) can equivalently be formulated on the finite domain Ω as

$$\begin{aligned} -\Delta p - k^2 p &= 0 && \text{in } \Omega \\ \mathcal{B}_s p &= 0 && \text{on } \Gamma_s \\ \mathcal{B}_r p &= \partial_n p && \text{on } \Gamma_r, \end{aligned} \tag{8}$$

where $\partial_n p$ stands for the derivative in the direction of the outer normal of Ω on the boundary Γ_r.

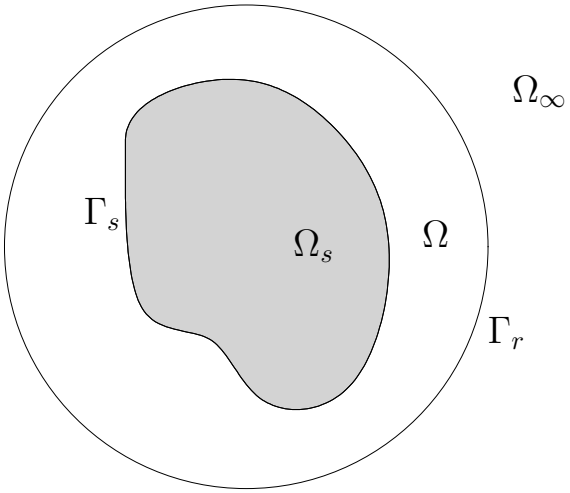

Figure 1: Domain for acoustic scattering problem.

It is important to remark, that the Dirichlet-to-Neumann map $\mathcal{B}_r$ leads to exact non-reflecting, but also, in general, non-local boundary conditions. So, for an efficient numerical computation, the non-reflecting boundary conditions are approximated by so-called absorbing boundary conditions, that depend on the boundary data locally. An usual approximation is the Robin boundary condition with

$$\mathcal{B}_r p := i\alpha p = \partial_n p \quad \text{on } \Gamma_r, \tag{9}$$

where $\alpha \in \mathbb{C}$ is chosen as $\alpha = k$.
Thus, equation (8) with boundary condition (9) describes a strongly formulated approximation of an acoustic scattering problem on a finite domain Ω.

1.1.2 Example: Laser Cavity Eigenmodes

A second example concerns the eigenmodes of a laser cavity that is filled with a real Gaussian duct, see e.g. [64, Chap. 20.3], and that possesses two planar end mirrors. In Chapter 3.3, we derive an Helmholtz eigenvalue problem for a standing electrical wave $\tilde{E}(x, y, z)$ in a laser resonator of length L. Formulated on the three-dimensional domain $\Omega := \Psi \times]0; 2L[$ with simply connected and open $\Psi \subset \mathbb{R}^2$ (see Figure 2), the wave has to satisfy the PDE eigenvalue problem

$$-\Delta \tilde{E}(x, y, z) - k^2(x, y, z)\, \tilde{E}(x, y, z) = \xi\, \tilde{E}(x, y, z) \quad \text{in } \Omega \tag{10}$$

and the periodical and absorbing boundary conditions

$$\tilde{E}(x, y, 0) \;=\; \tilde{E}(x, y, 2L) \quad \text{for } (x, y) \in \Psi,$$

$$\begin{aligned} \frac{\partial}{\partial z}\tilde{E}(x,y,0) &= \frac{\partial}{\partial z}\tilde{E}(x,y,2L) \quad \text{for } (x,y) \in \Psi, \\ \text{and} \quad \mathcal{B}_r\tilde{E}(x,y,z) &= \partial_n\tilde{E}(x,y,z) \quad \text{on } (\partial\Psi)\times]0;2L[\end{aligned}$$

with a Robin boundary condition

$$\mathcal{B}_r\tilde{E}(x,y,z) := ik_0\tilde{E}(x,y,z) = \partial_n\tilde{E}(x,y,z) \text{ on } (\partial\Psi)\times]0;2L[, \tag{11}$$

i.e. on the open part of the boundary. Furthermore, the coefficient function is of the form

$$k^2(x,y,z) = \left(k_0^2 - k_0\frac{2\pi}{\lambda}\frac{n_2(z)}{n_0}(x^2+y^2)\right).$$

where λ is the wave-length in the vacuum, k_0 is a reference value with $k_0 \approx n_0 2\pi/\lambda$, and $n_0 \approx 1$ and $n_2(z) \ll 1$ are parameters for the refractive index distribution of the duct.
Since in this example we deal with electrical waves in or near to the range of visible light, we have

$$k^2 \gg 1.$$

This property makes the numerical solution of this Helmholtz eigenvalue problem challenging or impossible if $L \gg \lambda$, as will be reviewed in Chapter 1.3.

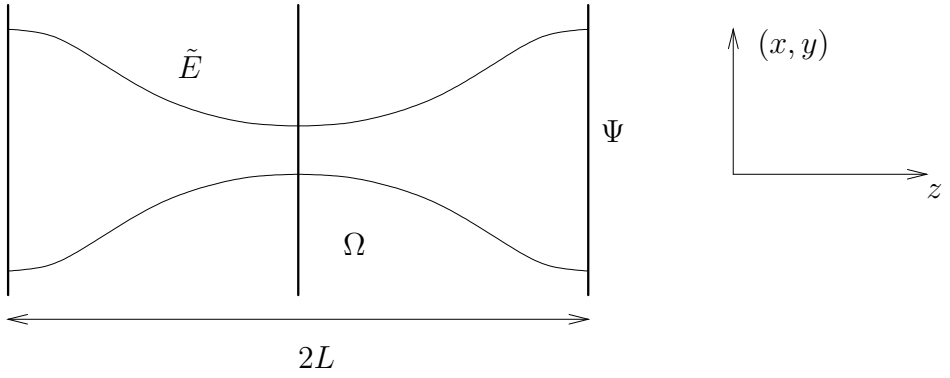

Figure 2: Domain for eigenvalue problem.

1.2 Weak Formulation and Well-Posedness of the Helmholtz Equation on Finite Domains

To solve an Helmholtz problem by finite elements, usually a variational formulation of the problem in the context of Sobolev spaces is utilized. For detailed definitions of the used notions, we refer to the books [3], [35] [38], [41], [73], or [75].
Let Ω be a bounded domain with sufficiently regular boundary, say Lipschitz boundary. By $C^k(\Omega)$, $k \in \mathbb{N}_0$, we denote the space of complex-valued k-times

continuously differentiable functions on Ω. The space $C^\infty(\Omega)$ contains the functions that have derivatives of all orders. $C_0^\infty(\Omega)$ is the subspace of functions in $C^\infty(\Omega)$ with compact support in Ω.
Let $L_2(\Omega)$ be the space of complex-valued square-integrable functions on Ω and let $H^m(\Omega) \subset L_2(\Omega)$ be the subspace of functions that posses weak derivatives up to order $m \in \mathbb{N}_0$. Finally, by $H^m_{\text{loc}}(\Omega)$ we denote the functions u for which

$$u \in H^m(\Omega')$$

holds for all bounded $\Omega' \subset \Omega$ with $\overline{\Omega'} \subset \Omega$.
Endowed with the scalar product $(\cdot,\cdot) : L_2(\Omega) \times L_2(\Omega) \to \mathbb{C}$ defined by

$$(u, v) := \int_\Omega u\overline{v}\; d(x, y, z),$$

$L_2(\Omega)$ is an Hilbert space. We abbreviate the induced norm by

$$\|u\|_L := (u, u)^{1/2}.$$

The natural norms on the spaces $H^m(\Omega)$ are

$$\|u\|_m := \left(\sum_{|\mu|=0}^{m} \|\partial^\mu u\|_L^2 \right)^{1/2},$$

where $\mu = (\mu_x, \mu_y, \mu_z)$ is a multi-index and $\partial^\mu u$ denotes the weak partial derivative

$$\partial^{(\mu_x,\mu_y,\mu_z)} u = \frac{\partial^{\mu_x}}{\partial x^{\mu_x}} \frac{\partial^{\mu_y}}{\partial y^{\mu_y}} \frac{\partial^{\mu_z}}{\partial z^{\mu_z}} u.$$

Furthermore, we define semi-norms on $H^m(\Omega)$ by

$$|u|_m := \left(\sum_{|\mu|=m} \|\partial^\mu u\|_L^2 \right)^{1/2}.$$

For ease of presentation, let us assume that we have a constant coefficient $k \in \mathbb{C}$ and that the (Lipschitz) boundary $\partial\Omega$ is composed of the two $\mathcal{C}^2$-boundaries Γ_s and $\Gamma_r = \partial\Omega \setminus \Gamma_s$ such that Γ_r and the interior of Γ_s with respect to $\partial\Omega$ are open sets in $\partial\Omega$. On Γ_s we prescribe the homogeneous Dirichlet condition

$$\mathcal{B}_s u := u = 0 \text{ on } \Gamma_s, \tag{12}$$

and on Γ_r the Robin boundary condition

$$\partial_n u - \mathcal{B}_r u := \partial_n u - i\alpha u = 0 \text{ on } \Gamma_r \tag{13}$$

with fixed $\alpha \in \mathbb{C}$.
Let us consider the following strong formulation of the Helmholtz equation with boundary conditions (12) and (13): Find $u \in H^2_{\mathrm{loc}}(\Omega)$ with

$$\begin{aligned} -\Delta u - k^2 u &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \Gamma_s \\ \partial_n u - i\alpha u &= 0 \quad \text{on } \Gamma_r \end{aligned} \tag{14}$$

for given right-hand side $f \in L_2(\Omega)$.
For a weak formulation of (14), we define the sesquilinear form $a(\cdot,\cdot) : H^1(\Omega) \times H^1(\Omega) \to \mathbb{C}$ by

$$a(u,v) := \int_\Omega (\nabla u \nabla \overline{v} - k^2 u\overline{v})\; d(x,y,z) - \int_{\Gamma_r} i\alpha u\overline{v}\; d\sigma(x,y,z).$$

Furthermore, let

$$H := H^1_{0,\Gamma_s} := \left\{ u \in H^1(\Omega) \,\middle|\, u|_{\Gamma_s} = 0 \right\}$$

be the subspace of functions in $H^1(\Omega)$, whose trace on Γ_s vanishes.
Restricting the form $a(\cdot,\cdot)$ to H, we obtain a variational formulation of (14) in the Sobolev space H: Find $u \in H$ such that

$$a(u,v) = (f,v) \qquad \text{for all } v \in H \tag{15}$$

for given $f \in L_2(\Omega)$.
Obviously, the sesquilinear form $a(\cdot,\cdot)$ is continuous (or bounded) and H-coercive, i.e. there exist positive real numbers c and C, such that the Gårding inequality

$$\mathrm{Re}\,[a(u,u)] \geq c\|u\|_1^2 - C\|u\|_L^2 \tag{16}$$

holds for all $u \in H$.
Since Ω is assumed to be bounded and sufficiently regular (we have assumed a Lipschitz boundary), the space H can be embedded compactly into $L_2(\Omega)$. Therefore, the Gårding inequality (16) implies that $a(\cdot,\cdot)$ satisfies the Fredholm alternative. This means that either for every $f \in L_2(\Omega)$ the problem (15) possesses a unique solution in H or there exists a finite-dimensional subspace of non-trivial solutions of the homogeneous problem, i.e. of equation (15) with $f \equiv 0$.
In the following, two sufficient conditions for the well-posedness of problem (15) are given. They base on considering the real and the imaginary part, Re[·] and Im[·], of the coefficients k^2 and α.

- Analogously to Theorem 3.2 in [41], i.e. using the analytical continuation principle, it follows that for

  $$\mathrm{Re}[\alpha] \neq 0 \quad \text{and} \quad \mathrm{Im}[k^2] = 0$$

  the weak homogeneous problem (15), meaning with $f = 0$, only possesses the trivial solution $u = 0$ or, equivalently, that the problem (15) is uniquely solvable.

- In the case where
$$\mathrm{Re}[\alpha] = 0 \quad \text{and} \quad \mathrm{Im}[k^2] \neq 0,$$
a unique solution also exists, which can easily be proved by choosing $v = u$ in (15), taking the imaginary part of the resulting equation, and using the positive definiteness of the norm $\|\cdot\|_L$ on $H \subset L_2(\Omega)$.

Although these two conditions are not a complete statement on the well-posedness of problem (15), we can say that for some situations the weak Helmholtz equation with absorbing boundary conditions is well posed. Particularly, that means that only the trivial solution satisfies the homogeneous problem. This fact will be dealt with in Chapter 3.2 where the current methods of analyzing laser cavities are discussed.

1.3 On the Numerical Solution of the Helmholtz Equation by Finite Elements

A standard way of solving the variational equation (15) numerically, is to apply a Galerkin method, i.e. to choose an appropriate finite element subspace $S_h \subset H$ and to solve for $u_h \in S_h$ with

$$a(u_h, v_h) = (f, v_h)_0 \qquad \forall v_h \in S_h. \tag{17}$$

Such finite element approximations (17) of the Helmholtz equation (15) have been studied extensively. A monograph that reviews a lot of important results is, for instance, the book of Ihlenburg [41].
Essentially, two problems arise in solving (15) by a finite element discretization. The first one is an *approximation problem* and the second one a *computational problem*.
The relative *approximation error* $\|u - u_h\|_1/\|u\|_1$ in the H^1-norm is composed of an interpolation error and a so-called pollution error, which comes from a phase lead or phase lag of the finite element solution u_h of (17) with respect to u. It is well known that for real k and a uniform discretization with mesh size h, the interpolation error is of order $O(kh)$. For some model problems a pollution error of order $O(k^3h^2)$ has been proved rigorously under the condition $hk < 1$, see e.g. [41] and the references cited therein.
In one-dimensional problems, it is possible to modify the Galerkin method such that the pollution vanishes. Then, for a satisfactory approximation the value of kh has to be small. In higher dimensions, this generally is not possible, c.f. [11]. However, in this case the pollution error can at least be reduced.
For this purpose, many sophisticated methods have been proposed. We mention a few of them, without claiming to present an exhaustive list: the generalized or quasi-stabilized finite element method [11], the Galerkin least-squares stabilization [39], and the residual-free bubbles finite element method [31]. In the article [20], these methods are compared.

Another way to overcome the pollution error, is to use a least-squares formulation of the Helmholtz equation, see [48]. Therein an $O(kh)$-convergence in the least-squares norm is stated and numerical tests are presented that indicate the same order of convergence in the H^1-norm.
This short overview shows that, in order to obtain a satisfactory finite element approximation, a mesh size at least of order of magnitude of

$$h \approx \frac{1}{k} \tag{18}$$

has to be used.
Having discretized equation (15) by (17) obeying condition (18) one faces a *computational problem*: The resulting system of linear equations is very large, indefinite, non-hermitian, and ill-conditioned in general. (See, for instance, [61] for the meaning of these properties.)
A lot of elaborate methods for solving the discrete Helmholtz equation have been developed, ranging from adapted multigrid methods ([16], [24], [44], [48]) over domain decomposition methods ([28], [33]) to sophisticated preconditioning techniques ([25] with [26], [34], [42], [50], [51], [54], [58]).
In [21], a completely different approach to solve an Helmholtz eigenvalue problem is described. Therein, a non-linear multigrid eigenproblem solver is developed, based on a Schur decomposition, and is applied to a two-dimensional eigenvalue problem.
However, all above mentioned solution methods fail when applied to a truly three-dimensional, discrete Helmholtz problem with large $\mathrm{Re}[k^2]$ (as in laser simulation, where $\mathrm{Re}[k^2] \approx 10^7$), because they would require unrealistic storage and time resources.

2 A Transformation of the Helmholtz Boundary Value Problem

The main topic of this thesis is to compute eigenmodes of a laser resonator which are modeled by an Helmholtz eigenvalue problem, see Chapter 3. The interesting eigenmodes u, which are very oscillatory in z, can be represented as

$$u(x, y, z) = \tilde{u}(x, y, z) \cdot \exp[-i\,\kappa(x, y, z)], \tag{19}$$

where $\tilde{u}$ is smooth with respect to z and κ is also a sufficiently smooth, but real-valued function.
In this chapter, we justify – by the analysis of a one-dimensional model problem – that such eigenmodes can be computed satisfactorily by finite elements. More specifically, we transform a model Helmholtz boundary value problem by the use of representation (19) and prove that the obtained finite element approximation of u does not have a so-called pollution error, if following smoothness assumption is fulfilled:

$$k\|\tilde{u}\|_L > |\tilde{u}|_1 + |\tilde{u}|_2,$$

i.e. the k-fold of the L_2-norm of the reduced function $\tilde{u}$ dominates the H^1- and H^2-semi-norms.

2.1 The Idea: Separating Oscillations from the Solution

Let $\Omega \subset \mathbb{R}^3$ be a bounded, three-dimensional domain with $\mathcal{C}^2$-boundary.
To explain the idea, we consider the Helmholtz equation in strong form, at first disregarding boundary conditions: Find $u \in H^2(\Omega)$, such that

$$-\Delta u - k^2 u = f \tag{20}$$

with $f \in L_2(\Omega)$. (For some boundary conditions the solution u can be less regular; then, the appropriate space for u is $H^2_{\text{loc}}(\Omega) \supset H^2(\Omega)$.)
In this chapter we assume the wave number k to be a real quantity with

$$k \gg 1.$$

As explained in Chapter 1.3, for a discretization of (20) with large k by finite elements a fine mesh size and, consequently, a large number of grid points is needed in general.
Let us write the solution u as

$$u(x, y, z) = \tilde{u}(x, y, z) \cdot \exp[-i\kappa(x, y, z)] \tag{21}$$

with real-valued $\kappa \in C^\infty(\Omega) \cap C^1(\overline{\Omega})$. If u is oscillatory and the order of magnitude of the oscillation is known in advance, κ can be chosen such that the expression $\exp[-i\kappa(x, y, z)]$ contains the main part of the oscillation.

A simple computation shows that the reduced function

$$\tilde{u} = u(x,y,z) \cdot \exp[i\kappa(x,y,z)] \in H^2(\Omega) \qquad (\text{or } H^2_{\text{loc}}(\Omega))$$

satisfies the equation

$$\begin{aligned} -\Delta\tilde{u} + 2i\langle\nabla\kappa(x,y,z), \nabla\tilde{u}\rangle + (\langle\nabla\kappa(x,y,z), \nabla\kappa(x,y,z)\rangle - k^2 + i\Delta\kappa(x,y,z))\tilde{u} \\ = f \cdot \exp[i\kappa(x,y,z)] =: \tilde{f}, \end{aligned} \tag{22}$$

where $\langle(a_1,a_2,a_3),(b_1,b_2,b_3)\rangle := a_1b_1 + a_2b_2 + a_3b_3$ and $\tilde{f} \in L_2(\Omega)$.
Formally, equation (22) is a singularly perturbed problem for $\tilde{u}$ with right-hand side $\tilde{f}$, if for the norm of $\nabla\kappa$ holds: $|\nabla\kappa(x,y,z)| \gg 1$. In the case of real-valued functions and real coefficients an equation like (22) is often called *Diffusion-Convection-Reaction equation*. If $|\nabla\kappa(x,y,z)| \gg 1$, it is said to be convection dominated. In [45], [53], and [59] many important results concerning theory and numerical solution of such equations are presented.
Here, we deal with complex-valued functions and equations with complex coefficients. Therefore, we encounter different properties in general. Since, for instance, the coefficient $2i\nabla\kappa$ of the gradient of $\tilde{u}$

$$2i\langle\nabla\kappa(x,y,z), \nabla\tilde{u}\rangle$$

is purely imaginary, in the case of vanishing $\Delta\kappa$ the operator on the left-hand side of (22) is hermitian, e.g., for homogeneous Dirichlet conditions, in contrast to the non-symmetric operator of the Diffusion-Convection-Reaction equation.
Furthermore, our numerical experiments and the analysis of the one-dimensional model problem in the following Section 2.2, seem to indicate that, different from the real equation, a standard finite element discretization of equation (22) does not lead to approximate solutions with wrong amplitudes or with unwanted oscillations of order of magnitude of the exact solution. It rather seems that the main contribution to the approximation error comes from wrong frequencies of the oscillating portions of the solution. However, it remains to analyze the properties of a standard finite element solution of equation (22) more detailed, meaning beyond the basic statements given in Chapter 2.2.
Nevertheless, the real analogy will help us to stabilize the transformed eigenvalue equation in order to obtain an efficient numerical method for the computation of laser cavity eigenmodes, see Chapter 5.2 and Chapter 6.
Let us now consider a boundary value problem (BVP) for equation (20) on Ω with the boundary $\partial\Omega = \Gamma_s \cup \Gamma_r$, consisting of two disjoint parts Γ_s and Γ_r. We impose the boundary conditions

$$\begin{aligned} u - g &= 0 \quad \text{on } \Gamma_s \\ \partial_n u - i\alpha u &= 0 \quad \text{on } \Gamma_r, \end{aligned} \tag{23}$$

with $g \in H^{3/2}(\Gamma_s)$ and $\alpha \in \mathbb{C}$. The expression ∂_n denotes the derivative in the direction of the outward normal of Ω on Γ_r.
By the application of relation (21), the equations (23) are transformed to following boundary conditions for $\tilde{u}$:

$$\begin{aligned} \tilde{u} - g \cdot \exp[i\kappa(x,y,z)] &= 0 \quad \text{on } \Gamma_s \\ \partial_n \tilde{u} - i(\alpha - \partial_n \kappa(x,y,z))\tilde{u} &= 0 \quad \text{on } \Gamma_r. \end{aligned} \tag{24}$$

It is obvious, that the BVP (20), (23) is equivalent to the BVP (22), (24). In general, the transformed BVP does not have any advantages over the original one. If, however, the term $\exp[-i\kappa(x,y,z)]$ contains – roughly spoken – the main part of the oscillations of the solution u and, therefore, $\tilde{u}$ can be expected to be quite smooth, even a relatively coarse discretization of (22), (24) can lead to a satisfactory numerical approximation. In the following Section 2.2, this heuristic statement will be proved rigorously for a model problem.

2.2 Analysis of a Model Problem

In [41], a one-dimensional model problem is analyzed which reveals some of the main difficulties of solving the Helmholtz equation by finite elements.
This BVP is formulated on the domain

$$\Omega := (0;1) \subset \mathbb{R}$$

and reads

$$\begin{aligned} -u'' - k^2 u &= f \qquad \text{in } \Omega \\ u(0) &= 0 \\ u'(1) &= ik\,u(1) \end{aligned} \tag{25}$$

with $k \gg 1$ and $f \in L_2(\Omega)$. Applying the transformation in the previous Section 2.1 with $\kappa(x) := kx$, one obtains the transformed problem for $\tilde{u} \in H^2(\Omega)$:

$$\begin{aligned} -\tilde{u}'' + 2ik\tilde{u}' &= f \cdot e^{ikx} =: \tilde{f} \qquad \text{in } \Omega \\ \tilde{u}(0) &= 0 \\ \tilde{u}'(1) &= 0 \end{aligned} \tag{26}$$

where $\tilde{f} \in L_2(\Omega)$.
Let us consider a variational formulation of problem (26). For this purpose, we utilize the Sobolev space

$$H := \left\{ u \in H^1(0;1) \,\middle|\, u(x)|_{x=0} = 0 \right\}.$$

On H the sesquilinear forms

$$(u,v)_1 := \int_0^1 u'\bar{v}'\ dx \quad \text{and} \quad (u,v) := \int_0^1 u\bar{v}\ dx$$

are inner products, inducing the norms

$$|u|_1 := \|u\|_H^2 := (u,u)_1 \quad \text{and} \quad \|u\|_L^2 := (u,u),$$

respectively. The standard H^1-norm and the (semi-)norm $|\cdot|_1$ are equivalent on H. (The proof uses the Poincaré inequality on H.)
By the sesquilinear form $a(\cdot,\cdot) : H \times H \to \mathbb{C}$ with

$$a(u,v) := \int_0^1 u'\bar{v}' + 2iku'\bar{v}\ dx,$$

the model problem (26) can be written in a weak form as: Find $\tilde{u} \in H$ such that

$$a(\tilde{u},\tilde{v}) = (\tilde{f},\tilde{v}) \quad \forall \tilde{v} \in H \tag{27}$$

for $\tilde{f} \in L_2(\Omega)$.
We remark that $a(\cdot,\cdot)$, defined on $H \times H$, is not hermitian, because it is

$$a(u,v) \neq \overline{a(v,u)}$$

if $u, v \in H$ are such that $u(1) \neq 0 \neq v(1)$.
Let $S_h \subset H$ be the subspace of linear finite element functions on Ω originating from a uniform discretization with mesh size h. (Since $S_h \subset H$, the functions in S_h vanish at $x = 0$.) We approximate (27) by: Find $\tilde{u}_h \in S_h$ such that

$$a(\tilde{u}_h,\tilde{v}_h) = (\tilde{f},\tilde{v}_h)_0 \quad \forall \tilde{v}_h \in S_h. \tag{28}$$

The aim of the analysis in this chapter is to estimate the overall error

$$e := u - \tilde{u}_h \cdot \exp[-ikx] = (\tilde{u} - \tilde{u}_h)\exp[-ikx] =: \tilde{e}\exp[-ikx]. \tag{29}$$

More precisely, we want to derive upper bounds for the relative $|\cdot|_1$-error

$$\frac{|e|_1}{|u|_1}.$$

Since $u(x) = \tilde{u}(x) \cdot \exp[-ikx] \in H$ implies $\tilde{u} \in H$, we obtain

$$\begin{aligned} |u|_1^2 &= \int_0^1 |(\tilde{u}\exp[-ikx])'|^2\ dx = \int_0^1 |\tilde{u}'\exp[-ikx] - ik\tilde{u}\exp[-ikx]|^2\ dx \\ &= \int_0^1 |\tilde{u}' - ik\tilde{u}|^2\ dx \geq \int_0^1 |k|\tilde{u}| - |\tilde{u}'|\,|^2\ dx = \||k|\tilde{u}| - |\tilde{u}'|\|_L^2 \\ &\geq (k\|\tilde{u}\|_L - |\tilde{u}|_1)^2. \end{aligned} \tag{30}$$

As shown in the following section, the solution $\tilde{u}$ of (27) for $\tilde{f} \in L_2(\Omega)$ is in $H^2(\Omega)$. For the estimation of the relative error, we assume $\tilde{u}$ to be smooth and free of small-scale oscillations, i.e.

$$k\|\tilde{u}\|_L > |\tilde{u}|_1 + |\tilde{u}|_2. \tag{31}$$

However, even for very smooth right-hand sides $\tilde{f}$ in equation (27), a solution $\tilde{u}$ of (27) can fail to satisfy the smoothness condition (31). (This can easily be seen by the use of the integral representation (34) for $\tilde{u}$ with $\tilde{f} = 1$.)
But as mentioned before, we actually are interested in solutions of an eigenvalue problem (see the example in Chapter 1.1.2). For these eigenmodes it is reasonable to assume a representation $u = \tilde{u} \cdot \exp[-i\kappa(x, y, z)]$ with $\tilde{u}$ satisfying condition (31).

2.2.1 Properties of the Transformed Problem

Essentially by the Cauchy-Schwarz inequality and by the Poincaré inequality, it is shown that $a(\cdot, \cdot)$ is continuous (or, equivalently, bounded), i.e. there exists a constant $C_b > 0$ such that

$$|a(u, v)| \leq C_b\|u\|_H\|v\|_H \qquad \forall u, v \in H.$$

Furthermore, the sesquilinear form $a(\cdot, \cdot)$ is H-coercive:

Lemma 1 (Gårding inequality) *The sesquilinear form $a(\cdot, \cdot)$ fulfills a Gårding inequality, i.e. there exist constants $c_g, C_g > 0$ such that*

$$\mathrm{Re}\,[a(\tilde{u}, \tilde{u})] \geq c_g\|\tilde{u}\|_H^2 - C_g\|\tilde{u}\|_L^2.$$

Proof: For the proof, we need the so-called generalized Young inequality for $a, b \geq 0$:

$$ab \leq \frac{\varepsilon}{2}a^2 + \frac{1}{2\varepsilon}b^2 \qquad \forall \varepsilon > 0. \tag{32}$$

By this, it holds

$$\begin{aligned}\mathrm{Re}\,[a(\tilde{u}, \tilde{u})] &\geq \int_0^1 |\tilde{u}'|^2 - 2k|\tilde{u}'||\tilde{u}|\; dx \geq \int_0^1 |\tilde{u}'|^2 - k(\varepsilon|\tilde{u}'|^2 + |\tilde{u}|^2/\varepsilon)\; dx \\ &= (1 - k\varepsilon)\int_0^1 |\tilde{u}'|^2\; dx - k/\varepsilon\int_0^1 |\tilde{u}|^2\; dx = (1 - k\varepsilon)\|\tilde{u}\|_H^2 - k/\varepsilon\|\tilde{u}\|_L^2\end{aligned}$$

for arbitrary $\varepsilon > 0$. Choosing e.g. $\varepsilon = \frac{1}{2k}$, one obtains the stated inequality

$$\mathrm{Re}\,[a(\tilde{u}, \tilde{u})] \geq c_g\|\tilde{u}\|_H^2 - C_g\|\tilde{u}\|_L^2$$

with $c_g = 1/2$ and $C_g = 2k^2$. □

Analogously to the arguments in Chapter 1.2, it can easily be seen, that equation (27) satisfies the Fredholm alternative. By this, well-posedness of problem (27), can be proved:

Lemma 2 (Well-Posedness and Regularity) *For every $\tilde{f} \in L_2(\Omega)$ there exists a unique $\tilde{u} \in H^2(\Omega) \cap H$ such that equation (27) is satisfied.*

PROOF: Let $\tilde{u} \in H$ be a solution of (27). Then, it satisfies the equation

$$\int_0^1 \tilde{u}'v' \, dx = -\int_0^1 (2ik\tilde{u}' - \tilde{f})v \, dx$$

for all $v \in H$, and particularly, for all $v \in C_0^\infty(\Omega)$. This implies that we have $\tilde{u} \in H^2(\Omega) \cap H$, which means that the strong and the weak form of the BVP, (26) and (27), respectively, are equivalent.
Furthermore, by the embedding lemma of Sobolev (see, e.g. [3] or [73]), we have the continuous embedding $H^2(\Omega) \subset C^1(\overline{\Omega})$, which implies that $\tilde{u}$ is in $C^1([0;1])$.
Let us consider the homogeneous problem, i.e. equation (27) with $\tilde{f} = 0$. From this equation, we obtain with $v := \tilde{u}$

$$\begin{aligned} 0 &= \int_0^1 |\tilde{u}'|^2 + 2ik\tilde{u}'\bar{\tilde{u}} \, dx \\ &= \int_0^1 \underbrace{|\tilde{u}'|^2}_{\in\mathbb{R}} \, dx + ik \int_0^1 \underbrace{\frac{d}{dx}(u_R^2) + \frac{d}{dx}(u_I^2)}_{\in\mathbb{R}} \, dx + 2k \int_0^1 \underbrace{-u_I'u_R + u_R'u_I}_{\in\mathbb{R}} \, dx, \end{aligned}$$

where $\tilde{u} = u_R + iu_I$ with real part u_R and imaginary part u_I.
Due to $\tilde{u} \in C^1([0;1])$ the following equation holds

$$0 = \int_0^1 \frac{d}{dx}(u_R^2) + \frac{d}{dx}(u_I^2) \, dx = u_R^2(1) - u_R^2(0) + u_I^2(1) - u_I^2(0) = u_R^2(1) + u_I^2(1),$$

and we obtain $u_R(1) = u_I(1) = 0$.
The regularity theory for the Laplace operator (cf. [35] or [38]) and the lemma of Sobolev imply that $\tilde{u}$ satisfies the (classical) initial value problem

$$\begin{aligned} -\tilde{u}'' + 2ik\tilde{u}' &= 0 \quad \text{in } \Omega \\ \tilde{u}(1) = \tilde{u}'(1) &= 0. \end{aligned} \tag{33}$$

By the theory of ODEs, it follows that $\tilde{u} = 0$ is the unique solution.
Since (27) satisfies the Fredholm alternative, this uniqueness result yields the existence of the solution. □

If $\tilde{f}$ is continuous on $[0;1]$, obviously the function $\tilde{u} \in C^2([0;1])$ defined by

$$\tilde{u}(x) := \int_0^x \left(\int_y^1 \tilde{f}(t) \exp[2ik(1-t)] \, dt \right) \exp[-2ik(1-y)] \, dy \tag{34}$$

is the classical solution of the boundary value problem (26). Furthermore, it is

$$\tilde{u}'(x) = \left(\int_x^1 \tilde{f}(t) \exp[2ik(1-t)] \, dt \right) \exp[-2ik(1-x)] \tag{35}$$

on the interval $[0;1]$.
This also holds in a weak sense for $\tilde{f} \in L_2(0;1)$. By standard arguments it can be shown, that the function

$$\hat{u}(x) := \left(\int_x^1 \tilde{f}(t) \exp[2ik(1-t)] \, dt \right) \exp[-2ik(1-x)] \tag{36}$$

is in $H^1(\Omega)$ and possesses the weak derivative

$$\frac{d}{dx}\hat{u}(x) = -\tilde{f}(x) + 2ik\hat{u}(x). \tag{37}$$

By the lemma of Sobolev, we have $\hat{u} \in C([0;1])$.
So, the function

$$\tilde{u}(x) := \int_0^x \hat{u}(y) \, dy \in C^1([0;1]) \tag{38}$$

is in H and satisfies the problem (26).
Furthermore, by standard arguments (see [4]) and from the continuity of $\hat{u}$, it follows that

$$\hat{u}(x) - \hat{u}(0) = \int_0^x \frac{d}{dy}\hat{u}(y) \, dy. \tag{39}$$

Using these representations, we show:

Lemma 3 (Stability Estimates) *For $\tilde{f} \in L_2(\Omega)$, the solution $\tilde{u}$ of (27) satisfies following stability estimates:*

$$\begin{aligned} \|\tilde{u}\|_L &\leq \frac{C}{k}\|\tilde{f}\|_L, & (40)\\ |\tilde{u}|_1 &\leq \|\tilde{f}\|_L, & (41)\\ |\tilde{u}|_2 &\leq Ck\|\tilde{f}\|_L & (42) \end{aligned}$$

with constants C independent of k.

Proof: By (38), (37), (39), (36), and standard estimates, we obtain

$$\begin{aligned} |\tilde{u}(x)| &= \left| \int_0^x \hat{u}(y) \, dy \right| = \left| \int_0^x \frac{1}{2ik} \left(\frac{d}{dy}\hat{u}(y) + \tilde{f}(y) \right) \, dy \right| \\ &= \frac{1}{2k} \left| \hat{u}(x) - \hat{u}(0) + \int_0^x \tilde{f}(y) \, dy \right| \\ &\leq \frac{1}{2k} \left(|\hat{u}(x)| + |\hat{u}(0)| + \int_0^x |\tilde{f}(y)| \, dy \right) \\ &\leq \frac{1}{2k} \left(\int_x^1 |\tilde{f}(y)| \, dy + \int_0^1 |\tilde{f}(y)| \, dy + \int_0^x |\tilde{f}(y)| \, dy \right) \\ &\leq \frac{C}{k} \int_0^1 |\tilde{f}(y)| \, dy \leq \frac{C}{k}\|\tilde{f}\|_L. \end{aligned}$$

Taking the square and integration yields inequality (40).
Furthermore, we have

$$|\tilde{u}'(x)| = |\hat{u}(x)| \leq \int_x^1 |\tilde{f}(t)|\ dt \leq \int_0^1 |\tilde{f}(t)|\ dt \leq \|\tilde{f}\|_L,$$

which implies (41), again by squaring and integrating.
Finally, from

$$\|\tilde{u}''\|_L = \|2ik\tilde{u}' - \tilde{f}\|_L \leq 2k\|\tilde{u}'\|_L + \|\tilde{f}\|_L,$$

it follows

$$|\tilde{u}|_2 \leq (1+2k)\|\tilde{f}\|_L \leq Ck\|\tilde{f}\|_L$$

with constant C independent of k. □

2.2.2 Asymptotic Analysis

Analogously to [41, Chaps. 4.4 and 4.5], we examine the error of the approximation $\tilde{u}_h \cdot e^{-ikx}$ of u.
In this section, we perform an asymptotic analysis, where asymptotic means that we assume k and h to be such that

$$k^2 h \leq \alpha < 1$$

for appropriate α.
Before examining the error, we mention some approximation properties of the finite element space $S_h \subset H$, which will be needed for the analysis. If $\mathcal{I}_h : H \to S_h$ is the interpolation operator which maps $v \in H$ to its piecewise linear nodal interpolant $\mathcal{I}_h v \in S_h$ the following estimates hold for every $v \in H^2(\Omega) \cap H$:

$$\begin{aligned} \|v - \mathcal{I}_h v\|_L &\leq Ch^2|v|_2, && (43)\\ \|v - \mathcal{I}_h v\|_L &\leq Ch|v - \mathcal{I}_h v|_1, && (44)\\ |v - \mathcal{I}_h v|_1 &\leq Ch|v|_2, && (45) \end{aligned}$$

with constants C independent of the mesh size h, see e.g. [17].
The following estimates are standard, but in deriving them we pay special attention to the question, how h and k determine the constants in the inequalities.
Following lemma states that for fixed k the finite element solution $\tilde{u}_h$ is quasi-optimal in the norm $|\cdot|_1$.

Lemma 4 *Let the finite element error $\tilde{e}$ be defined as $\tilde{e} := \tilde{u} - \tilde{u}_h$, where $\tilde{u}$ is the exact solution of (27) and $\tilde{u}_h$ the finite element solution of (28). Let, furthermore, k and h be such that the constant C_H mentioned below in this lemma possesses a positive denominator.*

Then, we have

$$|\tilde{e}|_1 \leq C_H \inf_{v \in S_h} |\tilde{u} - v|_1, \tag{46}$$

where

$$C_H = \frac{1}{1 - (c_1 k^2 h + c_2 k^3 h^2 + c_3 k^3 h^3 + c_4 k^4 h^4)}$$

with constants c_1, c_2, c_3, c_4 independent of k and h.

PROOF: First, we estimate $\|\tilde{e}\|_L$. For this purpose, we apply a duality argument. Let $z \in H$ be the solution of the (adjoint) problem

$$a(v, z) = (v, \tilde{e})_0 \qquad \forall v \in H. \tag{47}$$

Then, it is $z \in H^2(0;1) \cap H$. For arbitrary $w \in S_h$, the orthogonality relation $a(\tilde{e}, w) = 0$ holds, and we obtain

$$\begin{aligned} \|\tilde{e}\|_L^2 &= (\tilde{e}, \tilde{e})_0 = a(\tilde{e}, z) = a(\tilde{e}, z - w) = \left| \int_0^1 \tilde{e}'(\overline{z - w})' + 2ik\tilde{e}'(\overline{z - w})\, dx \right| \\ &\leq \|(z - w)'\|_L \|\tilde{e}'\|_L + 2k\|z - w\|_L \|\tilde{e}'\|_L. \end{aligned}$$

The choice $w := \mathcal{I}_h z$ and the application of the interpolation estimates (43) and (45), and of the stability estimate (42) imply

$$\|\tilde{e}\|_L^2 \leq (\tilde{C}_1 h + \tilde{C}_2 k h^2)|z|_2 \|\tilde{e}'\|_L \leq (C_1 hk + C_2 k^2 h^2)\|\tilde{e}\|_L \|\tilde{e}'\|_L.$$

with C_1, C_2 independent of k and h. So, we obtain

$$\|\tilde{e}\|_L \leq \underbrace{(C_1 hk + C_2 k^2 h^2)}_{C_3 :=} |\tilde{e}|_1. \tag{48}$$

Now, we derive an estimate for $|\tilde{e}|_1$. From the trivial equation

$$a(\tilde{e}, \tilde{e}) = a(\tilde{e}, \tilde{u} - \tilde{u}_h) = a(\tilde{e}, \tilde{u} - v)$$

for arbitrary $v \in S_h$, it follows

$$\begin{aligned} |\tilde{e}|_1^2 &\leq \|\tilde{e}'\|_L \|(\tilde{u} - v)'\|_L + 2k\|\tilde{e}'\|_L \|\tilde{e}\|_L + 2k\|\tilde{e}'\|_L \|\tilde{u} - v\|_L \\ &\leq \|\tilde{e}'\|_L \|(\tilde{u} - v)'\|_L + 2k\|\tilde{e}'\|_L C_3 \|\tilde{e}'\|_L + 2k\|\tilde{e}'\|_L \|\tilde{u} - v\|_L. \end{aligned}$$

Choosing $v := \mathcal{I}_h \tilde{u}$ in $\|\tilde{u} - v\|_L$ and applying the interpolation and stability estimates, one obtains

$$\begin{aligned} |\tilde{e}|_1^2 &\leq |\tilde{e}|_1 |\tilde{u} - v|_1 + 2k|\tilde{e}|_1 C_3 |\tilde{e}|_1 + 2k|\tilde{e}|_1 \|\tilde{u} - v\|_L \\ &\leq |\tilde{e}|_1 |\tilde{u} - v|_1 + 2k|\tilde{e}|_1 C_3 |\tilde{e}|_1 + C 2kh^2 |\tilde{e}|_1 |\tilde{u}|_2 \\ &\leq |\tilde{e}|_1 |\tilde{u} - v|_1 + 2kC_3(1 + \tilde{C}h^2 k)|\tilde{e}|_1^2 \\ &= |\tilde{e}|_1 |\tilde{u} - v|_1 + (c_1 k^2 h + c_2 k^3 h^2 + c_3 k^3 h^3 + c_4 k^4 h^4)|\tilde{e}|_1^2 \end{aligned}$$

with constants c_1, c_2, c_3, c_4 independent of h and k. If k and h are such that

$$c_1k^2h + c_2k^3h^2 + c_3k^3h^3 + c_4k^4h^4 < 1,$$

this leads to the estimate

$$|\tilde{e}|_1 \leq \underbrace{\frac{1}{1 - (c_1k^2h + c_2k^3h^2 + c_3k^3h^3 + c_4k^4h^4)}}_{C_H :=} |\tilde{u} - v|_1, \tag{49}$$

which is the inequality

$$|\tilde{e}|_1 \leq C_H \inf_{v \in S_h} |\tilde{u} - v|_1, \tag{50}$$

as stated in the lemma. □

The condition on h and k in Lemma 4 essentially depends on the expression k^2h. Thus, it is sufficient for quasi-optimality of the finite element solution, that k^2h is bounded by a constant which is small enough. In Section 2.2.4, numerical experiments indicate that for a uniform quasi-optimality of the error $\tilde{e}$ the boundedness of k^2h is also necessary.
With the help of Lemma 4, we, derive an asymptotic bound for the overall error $|e|_1$.

Lemma 5 *Under the conditions of Lemma 4, we have the estimate*

$$|e|_1 \leq cC_Hh(1 + kC_3)|\tilde{u}|_2. \tag{51}$$

The expressions C_3 and C_H are of the form

$$C_3 = C_1hk + C_2k^2h^2$$

and

$$C_H = \frac{1}{1 - (c_1k^2h + c_2k^3h^2 + c_3k^3h^3 + c_4k^4h^4)}$$

with constants $c, C_1, C_2, c_1, c_2, c_3, c_4$ independent of h and k.

Proof: By standard estimates, it follows

$$\begin{aligned}
|e|_1^2 &= \int_0^1 |(\tilde{e} \cdot \exp[-ikx])'|^2 \, dx = \int_0^1 |\tilde{e}' \cdot \exp[-ikx] - ik\tilde{e} \cdot \exp[-ikx]|^2 \, dx \\
&\leq \int_0^1 (|\tilde{e}' \cdot \exp[-ikx]| + k|\tilde{e} \cdot \exp[-ikx]|)^2 \, dx \\
&\leq 2\int_0^1 |\tilde{e}'|^2 + k^2|\tilde{e}|^2 \, dx = 2(|\tilde{e}|_1^2 + k^2\|\tilde{e}\|_L^2) \\
&\leq 2(|\tilde{e}|_1 + k\|\tilde{e}\|_L)^2.
\end{aligned} \tag{52}$$

Substituting (48) and (46) into (52) and using the approximation properties of S_h in H, we obtain

$$|e|_1 \leq \sqrt{2} C_H (1 + kC_3) \inf_{v \in S_h} |\tilde{u} - v|_1 \leq c C_H h (1 + kC_3) |\tilde{u}|_2 \tag{53}$$

with stated C_H and C_3, and c independent of h and k. □

Similar to [41, Chap. 4.4], we want to derive an estimate for the relative $|\cdot|_1$-error, i.e. for $|u - \tilde{u}_h \exp[-ikx]|_1 / |u|_1$.
Let us assume that

$$k^2 h \leq \alpha < 1. \tag{54}$$

From $k \gg 1$, it follows that $h < 1$ and that with sufficiently small α (e.g. such that $c_1\alpha + (c_2 + c_3 + c_4)\alpha^2 < 1/2$), we have that C_H is bounded independently of h and k.
Then, from the inequalities (51) and (30), and from the assumptions (31) and (54) for large k, we obtain the estimate

$$\begin{aligned} \frac{|e|_1}{|u|_1} &\leq \tilde{C} \frac{(\tilde{c}h + \tilde{c}k^2h^2 + \tilde{c}k^3h^3)|\tilde{u}|_2}{|k\|\tilde{u}\|_L - |\tilde{u}|_1|} < \tilde{C} \frac{(\tilde{c}h + \tilde{c}k^2h^2 + \tilde{c}k^3h^3)}{1} \\ &\leq \tilde{C}\alpha \left(\frac{\tilde{c}}{k^2} + \frac{\tilde{c}\alpha}{k^2} + \frac{\tilde{c}\alpha^2}{k^3}\right) \leq \hat{C}\alpha \frac{1}{k^2} \end{aligned} \tag{55}$$

with generic positive constants $\tilde{c}$ and $\tilde{C}$ independent of h and k. Obviously, $\hat{C}$ is also independent of h and k.
Let us summarize these results:

Proposition 1 (Asymptotic Estimate) *Let $\tilde{u}$ satisfy the smoothness assumption*

$$k\|\tilde{u}\|_L > |\tilde{u}|_2 + |\tilde{u}|_1 \tag{56}$$

and let k and h be such that

$$k^2 h \leq \alpha < 1$$

for sufficiently small α.
Then, for large k the relative error satisfies the estimate

$$\frac{|e|_1}{|u|_1} \leq C \frac{1}{k^2} \tag{57}$$

with constant C independent of h and k.

The estimate (57) is of order $1/k^2$ whereas the corresponding one in [41, p. 122] is of order $1/k$. The smoothness assumption (56) implies in some sense that the term $\exp[-ikx]$ contains the main oscillation of u; this is more specific than the assumptions $|u|_2/|u|_1 \leq Ck$ and $|u|_2/\|u\|_L \leq Ck^2$, that underly the

estimate in [41]. It remains to more carefully work out the relation between the approximation obtained by a finite element discretization of the Helmholtz equation (as considered in [41]) and the approximation $\tilde{u}_h \cdot e^{-ikx}$. This, however, will not be done in this thesis.
In the following, we will perform an in-depth analysis of the error and show, that under assumption (56) no so-called pollution error arises.

2.2.3 Pre-Asymptotic Analysis

In this section, we analyze the pre-asymptotic case, i.e. we examine the finite element approximation more detailed and give an upper bound for the relative error without a condition on h or k (except for the trivial assumption $h < 1$).
For this analysis, we specify the finite element discretization. Let the interval $[0; 1]$ be discretized by a uniform mesh with $N+1$ nodes $x_j := j \cdot h, j = 0, \ldots, N$, where $h := 1/N$ is the mesh size. By $S_h \subset H$ we denote the space of continuous, piecewise linear functions which vanish at $x_0 = 0$. Obviously, a function $\tilde{u}_h \in S_h$ is completely described by its values at the nodes $u_j := \tilde{u}_h(x_j), j = 1, \ldots, N$, see Figure 3. Furthermore, for the representation of S_h we choose the nodal basis functions $\psi_j, (j = 1, \ldots, N)$, which are continuous on [0;1] and linear on each interval $[x_j; x_{j+1}]$, and which satisfy $\psi_j(x_l) = \delta_{jl}$ for $0 \leq j, l \leq N$.

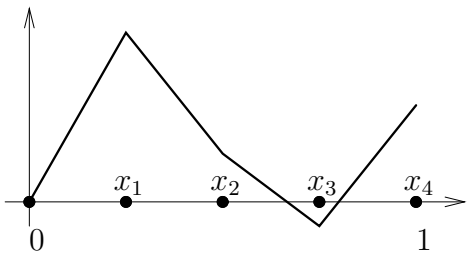

Figure 3: A function in $S_{1/4}$.

As already mentioned at the beginning of Chapter 2.2, the discretized variational problem reads: Find $\tilde{u}_h \in S_h$ such that

$$a(\tilde{u}_h, \tilde{v}_h) = (\tilde{f}, \tilde{v}_h) \quad \forall \tilde{v}_h \in S_h, \tag{58}$$

with sesquilinear forms

$$a(u, v) = \int_0^1 u'\bar{v}' + 2iku'\bar{v}\ dx$$

and

$$(u, v) = \int_0^1 u\bar{v}\ dx.$$

Using the nodal basis, equation (58) can equivalently be written in a matrix-vector formulation for the unknown vector $(u_j) := (u_1, u_2, \ldots, u_N) \in \mathbb{C}^N$

$$A_h \cdot (u_j) = (f_j), \tag{59}$$

where the stiffness matrix $A_h \in \mathbb{C}^{N\times N}$ takes the tri-diagonal form

$$A_h = \begin{pmatrix} 2 & -1+ikh & & & \\ -1-ikh & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & 2 & -1+ikh \\ & & & -1-ikh & 1+ikh \end{pmatrix}, \tag{60}$$

and the right-hand side $(f_j) := (f_1, f_2, \ldots, f_N) \in \mathbb{C}^N$ is given by

$$f_j := h \cdot (\tilde{f}, \psi_i) \qquad i = 1, \ldots, N, \tag{61}$$

where the ψ_i are the nodal basis functions.

Lemma 6 (Discrete Stability) *The finite element solution $\tilde{u}_h$ of (58) satisfies following stability estimates:*

$$\|\tilde{u}_h\|_L \leq \frac{C_L}{k}\|\tilde{f}\|_L \tag{62}$$

$$|\tilde{u}_h|_1 \leq C_1\|\tilde{f}\|_L. \tag{63}$$

Proof: The proof consists of three steps. First, we introduce some notations. Then, an explicit representation of the finite element solution $\tilde{u}_h$, or, equivalently, of $(u_j) \in \mathbb{C}^N$, is computed by applying the $\mathcal{Z}$-transformation (see e.g. [68]). Finally, this representation is used to prove the estimates.

I. Preparations.
We define the discrete norms

$$|||(u_j)||| := \left(h \cdot \sum_{l=1}^{N} |u_l|^2\right)^{1/2} \quad \text{and} \quad |||(u_j)|||_\infty := \max_{l=1,\ldots,N} |u_l|.$$

Obviously, it holds $|||(u_j)||| \leq |||(u_j)|||_\infty$ and $\|\tilde{u}_h\|_L \leq C|||(u_j)|||$ with C independent of h and k.
We expand $(f_j) \in \mathbb{C}^N$ to the infinite sequence $\vec{f} := (f_0, f_1, f_2, \ldots)$, where we define $f_0 := 0$ and $f_l := 0$ for $l > N$.
Furthermore, we use the infinite sequence $\vec{u} := (u_0, u_1, u_2, \ldots)$ with $u_0 := 0$.

II. Representation of the discrete solution.
For ease of presentation, we set $u_1 := \gamma$, with $\gamma \in \mathbb{C}$ to be determined later by enforcing the solution (u_j) to fulfill the last equation in the system (59).
Using the right-shift operator T defined by

$$(T\vec{v})_j := \begin{cases} 0 & \text{if } j = 0 \\ v_{j-1} & \text{if } j > 0, \end{cases}$$

from the rows of (59) we obtain the relation

$$(-1 - ikh)\vec{u} + 2T\vec{u} + (-1 + ikh)T^2\vec{u} = T\vec{f}. \tag{64}$$

Due to $h < 1$, the members of the sequence $\vec{f}$ are bounded independently of h and k

$$|f_j| \le h|(\tilde{f}, \psi_j)| \le \frac{2}{3}h^{(3/2)}\|\tilde{f}\|_L \le \|\tilde{f}\|_L.$$

Therefore, the $\mathcal{Z}$-transform $F(z) := \mathcal{Z}\vec{f}$ of $\vec{f}$ exists, see [68, Theorem 6.4].
Furthermore, if a solution of (59) exists, then it can easily be seen that the members $u_j, j > N$, of the expanded infinite sequence $\vec{u}$ satisfying relation (64) can be bounded as

$$|u_j| \le 3^j \cdot \max_{l=1,\dots,N} |u_l| \quad \text{for } j > N.$$

So, in this case $\vec{u}$ also possesses a $\mathcal{Z}$-transform $U(z) := \mathcal{Z}\vec{u}$ (again, see [68, Theorem 6.4]).
The application of the $\mathcal{Z}$-transformation to equation (64) and decomposition into partial fractions yield

$$U(z) = \frac{1}{2ikh}\left(\gamma(-1 + ikh)\left(\frac{1}{z-1} - \frac{\beta}{z-\beta}\right) + \frac{F(z)}{z-1} - \beta\frac{F(z)}{z-\beta}\right),$$

where $\beta = (-1 - ikh)/(-1 + ikh)$.
Applying the inverse transformation, we obtain a representation of the members u_j, namely

$$u_j = \frac{1}{2ikh}\left(\gamma(-1 + ikh)(1 - \beta^j) + \sum_{l=1}^{j-1}(1 - \beta^l)f_{j-l}\right) \quad \text{for } j > 0.$$

The last equation in (59) imposes the condition

$$f_N = (1 + ikh)(u_N - u_{N-1})$$

by which γ is determined as

$$\gamma = \frac{1}{1 + ikh}\sum_{l=0}^{N-1}\beta^{l-N+1}f_{N-l}.$$

Finally, we obtain the explicit representation

$$u_j = \frac{1}{2ikh}\Big(-(1-\beta^j)\sum_{l=0}^{N-1}\beta^{l-N} f_{N-l} + \sum_{l=1}^{j-1}(1-\beta^l) f_{j-l}\Big) \tag{65}$$

with $\beta = (-1-ikh)/(-1+ikh)$.

III. Proofs of the estimates.
Substituting the right-hand side (61) into representation (65), we obtain

$$\begin{aligned} |u_j| &= \frac{1}{2k}\left| -(1-\beta^j)\sum_{l=0}^{N-1}\beta^{l-N}(\tilde f,\psi_{N-l}) + \sum_{l=1}^{j-1}(1-\beta^l)(\tilde f,\psi_{j-l})\right| \\ &= \frac{1}{2k}\left|\sum_{l=1}^{N}(\tilde f,\alpha_l^{(j)}\psi_l)\right| \end{aligned} \tag{66}$$

with

$$\alpha_l^{(j)} = \begin{cases} (1-\beta^{l-j}) - (1-\beta^{-j})\beta^l & \text{for } 1\le l\le j-1 \\ -(1-\beta^{-j})\beta^l & \text{for } j\le l\le N. \end{cases}$$

Since $|\beta| = 1$, it follows that

$$|\alpha_l^{(j)}| \le 4.$$

Then, representation (66) implies

$$\begin{aligned} |u_j| &= \frac{1}{2k}\left|\left(\tilde f, \sum_{l=1}^{N}\alpha_l^{(j)}\psi_l\right)\right| = \frac{1}{2k}\left|\int_0^1 \tilde f\cdot\sum_{l=1}^{N}\overline{\alpha_l^{(j)}\psi_l}\,dx\right| \\ &\le \frac{1}{2k}\int_0^1 |\tilde f|\cdot 4\,dx \le C\frac{1}{k}\|\tilde f\|_L. \end{aligned} \tag{67}$$

Thus, we have

$$\|\tilde u_h\|_L \le C\|\,|(u_j)|\,\|_\infty \le \frac{C_L}{k}\|\tilde f\|_L$$

with constant $C_L > 0$ independent of h and k, which is inequality (62).
Furthermore, it is

$$\begin{aligned} |\tilde u_h|_1 &= \left(\int_0^1 |\tilde u_h'|^2\,dx\right)^{1/2} = \left(\sum_{j=1}^{N}\int_{x_j-h}^{x_j}\left|\frac{u_j-u_{j-1}}{h}\right|^2\,dx\right)^{1/2} \\ &= \left(h\cdot\sum_{j=1}^{N}\left|\frac{u_j-u_{j-1}}{h}\right|^2\right)^{1/2} \le \frac{1}{h}\max_{j=1,\dots,N}|u_j-u_{j-1}|. \end{aligned} \tag{68}$$

Analogously to (66), we obtain

$$\begin{aligned} |u_j - u_{j-1}| &= \frac{1}{2kh}|1-\beta| \left| -\sum_{l=0}^{N-1} \beta^{j-1+l-N} f_{N-l} + \sum_{l=1}^{j-1} \beta^{l-1} f_{j-l} \right| \\ &= \frac{1}{2k}|1-\beta| \left| \left(\tilde{f}, \sum_{l=1}^{N} \alpha_l^{(j)} \psi_l \right) \right|, \end{aligned} \tag{69}$$

where

$$\alpha_l^{(j)} = \begin{cases} \beta^{l-j+1} - \beta^{l-j+1} & \text{for } 1 \le l \le j-1 \\ -\beta^{l-j+1} & \text{for } j \le l \le N. \end{cases}$$

By

$$|\alpha_l^{(j)}| \le 2 \quad \text{and} \quad |1-\beta| = \left| \frac{2ikh}{-1+ikh} \right| \le 2kh,$$

we obtain from (69)

$$|u_j - u_{j-1}| \le Ch\|\tilde{f}\|_L. \tag{70}$$

Now, inequality (63) follows by combining (68) and (70)

$$|\tilde{u}_h|_1 \le \frac{1}{h} \max_{j=1,\dots,N} |u_j - u_{j-1}| \le \frac{1}{h} Ch\|\tilde{f}\|_L =: C_1\|\tilde{f}\|_L$$

with constant C_1 independent of h and k. □

The function $z := \tilde{u}_h - \mathcal{I}_h\tilde{u} \in S_h$ is the solution of the equation

$$a(z,v) = (2ik(\tilde{u} - \mathcal{I}_h\tilde{u})', v) \quad \forall v \in S_h, \tag{71}$$

since for all $v \in S_h$ it holds

$$\begin{aligned} a(z,v) &= a(\tilde{u}_h - \mathcal{I}_h\tilde{u}, v) \\ &= a(\tilde{u}_h - \tilde{u}, v) + a(\tilde{u} - \mathcal{I}_h\tilde{u}, v) \\ &= 0 + a(\tilde{u} - \mathcal{I}_h\tilde{u}, v) \\ &= \int_0^1 (\tilde{u} - \mathcal{I}_h\tilde{u})'\bar{v}' + 2ik(\tilde{u} - \mathcal{I}_h\tilde{u})'\bar{v}\ dx \\ &= 0 + \int_0^1 2ik(\tilde{u} - \mathcal{I}_h\tilde{u})'\bar{v}\ dx = (2ik(\tilde{u} - \mathcal{I}_h\tilde{u})', v), \end{aligned}$$

where we have used that

$$\int_0^1 (\tilde{u} - \mathcal{I}_h\tilde{u})'\bar{v}'\ dx = 0 \quad \text{for all } v \in S_h. \tag{72}$$

(Equation (72) can be proved by element-wise integration by parts and observing that $\tilde{u} - \mathcal{I}_h\tilde{u}$ vanishes at the nodes.)

Applying the triangle inequality, relation (71), and the discrete stability estimates (62) and (63) of Lemma 6, one obtains

$$\|\tilde{u} - \tilde{u}_h\|_L \leq (Ch + C)|\tilde{u} - \mathcal{I}_h\tilde{u}|_1 \tag{73}$$

and

$$|\tilde{u} - \tilde{u}_h|_1 \leq (1 + Ck)|\tilde{u} - \mathcal{I}_h\tilde{u}|_1, \tag{74}$$

where C shall denote a generic positive constants independent of h and k.
Using inequality (52), and the estimates (73) and (74), it follows for $h < 1$

$$\begin{aligned} |e|_1 &\leq (C + Ck + Ckh + Ck)|\tilde{u} - \mathcal{I}_h\tilde{u}|_1 \leq (Ch + Ckh + Ckh^2 + Ckh)|\tilde{u}|_2 \\ &\leq Chk|\tilde{u}|_2, \end{aligned} \tag{75}$$

and we obtain by the estimate (30):

Proposition 2 (Asymptotic Estimate) *Let $\tilde{u}$ satisfy the smoothness assumption*

$$k\|\tilde{u}\|_L > |\tilde{u}|_2 + |\tilde{u}|_1. \tag{76}$$

Then, the relative error satisfies the estimate

$$\frac{|e|_1}{|u|_1} \leq ckh \tag{77}$$

with constant c independent of h and k.

Inequality (77) gives us a pre-asymptotic estimate where no pollution term arises, i.e. the relative finite element error is of the order of the interpolation error. (The corresponding estimate (4.5.15) in [41] possesses an additional term of order $O(k^3h^2)$.)
Thus, for this one-dimensional model problem the heuristic statements of Section 2.1 hold.

2.2.4 Numerical Experiments

Now, we present some numerical experiments for solving the model problem (26) by uniform finite elements, as specified in Section 2.2.3.
First, we examine the quasi-optimality constant C_H in Lemma 4. For right-hand side $\tilde{f} = 1$, the ratio of the finite element error and of the best approximation error (which is the error of the nodal interpolant in the H^1-semi-norm) are computed for constant kh and for constant k^2h.
In Figure 4, the quotient $|\tilde{u}-\tilde{u}_h|_1/|\tilde{u}-\mathcal{I}_h\tilde{u}|_1$ – which is a lower bound for the quasi-optimality constant C_H – is plotted for $hk \in \{0.1, 0.05, 0.025\}$, whereas Figure 5 depicts the relative error $|\tilde{u} - \tilde{u}_h|_1/|\tilde{u}|_1$ for the same values of hk. (We suppose

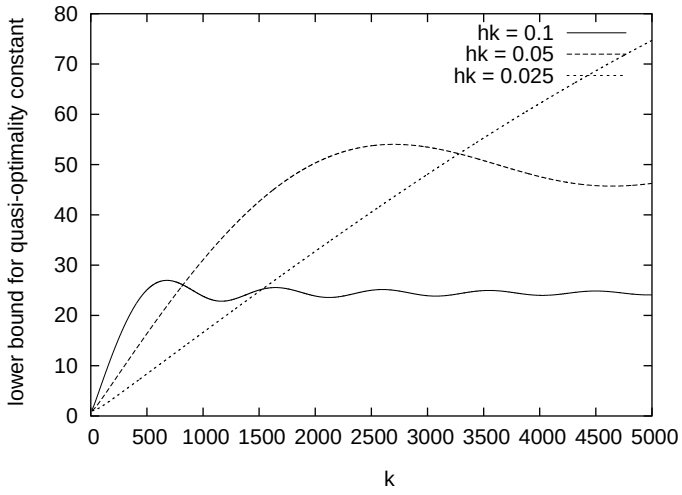

Figure 4: Lower bound for quasi-optimality constant C_H for $hk \equiv \text{const}$.

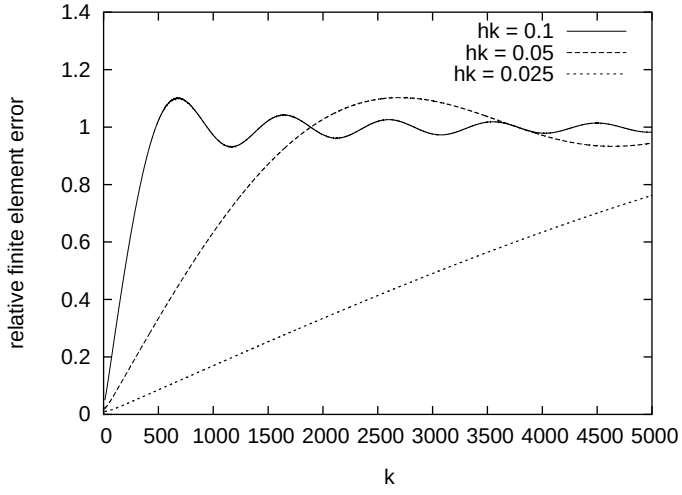

Figure 5: Relative error of finite element solution for $hk \equiv \text{const}$.

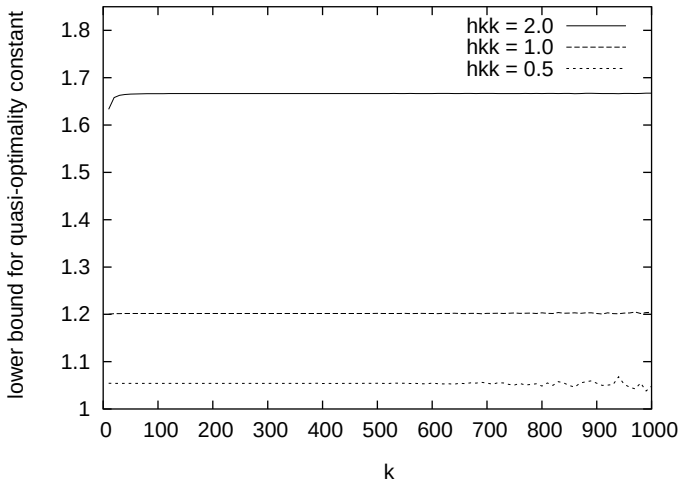

Figure 6: Lower bound for quasi-optimality constant C_H for $hk^2 \equiv \text{const}$.

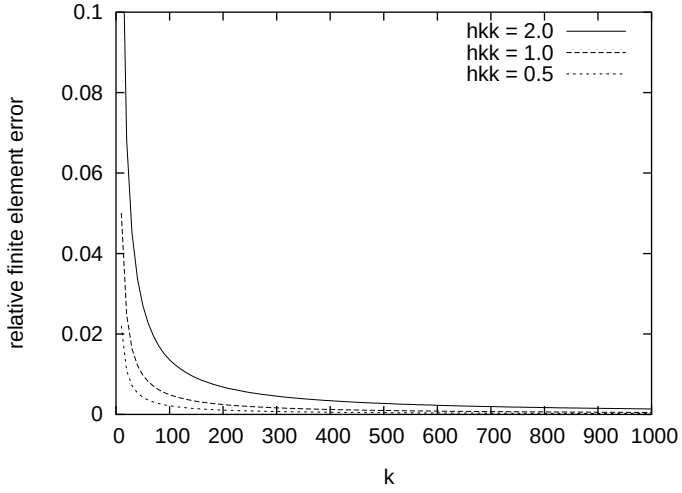

Figure 7: Relative error of finite element solution for $hk^2 \equiv \text{const}$.

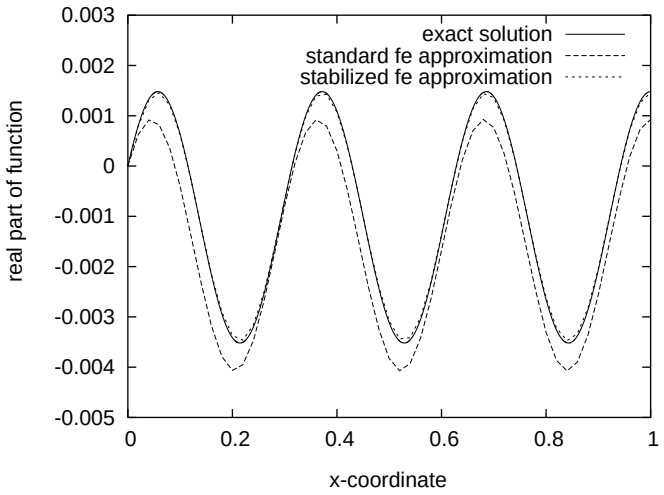

Figure 8: Comparison of exact solution, standard finite element approximation, and stabilized finite element approximation for $k = 10$ and $h = 0.02$.

that in Figures 4 and 5 the behavior of the curves $hk = 0.1$ and $hk = 0.05$ for large k comes from the restricted resolution of the oscillations of the solutions.)
For $k^2h \in \{2.0, 1.0, 0.5\}$, Figures 6 and 7 depict the ratios $|\tilde{u} - \tilde{u}_h|_1/|\tilde{u} - \mathcal{I}_h\tilde{u}|_1$ and $|\tilde{u} - \tilde{u}_h|_1/|\tilde{u}|_1$, respectively.
These numerical results indicate that for quasi-optimality of the standard finite element solution, as stated in Lemma 4, the value of k^2h has to be bounded. Furthermore, it is not sufficient to bound kh when the relative error shall be confined by a value less that unity uniformly in h and k, as Figure 5 shows.
The question arises, if – as for the real diffusion-convection equation – a condition on the magnitude of kh suffices to guarantee quasi-optimality of the finite element solution of equation

$$\int_0^1 \tilde{u}_h'\bar{v}_h' + 2ik\tilde{u}_h'\bar{v}_h \, dx = \int_0^1 \tilde{f}\bar{v}_h \, dx \quad \forall v_h \in S_h \tag{78}$$

after some stabilization.
Let us consider the standard finite element solution of equation (78) for $\tilde{f} = 1$ and $k = 10$. Figure 8 shows the exact solution and the finite element solution obtained by the standard discretization of (78) with mesh size $h = 0.02$ as described by equation (59). It can be observed that the finite element solution possesses a slightly larger wave length and that its derivative at the left boundary is estimated wrongly. The amplitude, however, seems to be accurate.
A streamline diffusion approach, as explained in Chapter 5.2 and used for the computation of laser cavity eigenmodes is not appropriate for the present problem since by this stabilization the amplitude of the oscillating component of the solution is affected, which is irrelevant here.

Our experiments led us to a modification of equation (78) which we describe shortly in the following.
A solution of the model problem satisfies for $v \in H \subset H^1(\Omega)$ the weak equation

$$\int_0^1 \tilde{u}''\bar{v}' + 2ik\tilde{u}'\bar{v}' \, dx = \int_0^1 \tilde{f}\bar{v}' \, dx. \tag{79}$$

Approximating this equation in S_h, we obtain

$$\int_0^1 2ik\tilde{u}_h'\bar{v}_h' \, dx = \int_0^1 \tilde{f}\bar{v}_h' \, dx \tag{80}$$

due to $\tilde{u}_h'' = 0$.
Adding equation (78) multiplied by $(1 + \sigma)$ and equation (80) multiplied by $\sigma \cdot i/2k$, we obtain the stabilized equation: Find $\tilde{u}_h \in S_h$ such that

$$\int_0^1 \tilde{u}_h'\bar{v}_h' + (1+\sigma)2ik\tilde{u}_h'\bar{v}_h \, dx = (1+\sigma)\int_0^1 \tilde{f}\bar{v}_h \, dx + \sigma\frac{i}{2k}\int_0^1 \tilde{f}\bar{v}_h' \, dx \quad \forall v_h \in S_h, \tag{81}$$

where we choose $\sigma = \tau(hk)^2$ with appropriate real parameter τ.
Figure 8 also depicts the finite element solution $\tilde{u}_h$ obtained from solving the stabilized equation (81) with $\tilde{f} = 1$, $k = 10$ and $h = 0.02$ using $\tau \approx 0.33$. Obviously, this approximation almost coincides with the interpolate $\mathcal{I}_h\tilde{u}$.
We have tested the quality of the solutions of (81) for $hk \in \{0.2, 0.1, .0.05, 0.025\}$. An appropriate choice of the parameter τ leads to a constant relative finite element error for $hk \equiv$ const. Table 1 shows that the optimal parameter τ slightly depends on hk and that the relative error behaves like hk, which is the order of the interpolation error. Similar results are obtained for more oscillating right-hand sides.
Thus, the stabilization in equation (81) seems to be a promising approach for solving equation (78) by finite elements and deserves further investigation.

hk	τ	$\lvert\tilde{e}\rvert_1/\lvert\tilde{u}\rvert_1$
0.2	0.338755	0.0814
0.1	0.33467	0.0407
0.05	0.33367	0.0204
0.025	0.33342	0.0102

Table 1: Relative error of stabilized equation with parameter τ for $hk \equiv$ const.

3 Modeling the Eigenmodes of a Laser Resonator

In this chapter we shortly describe the physics of a laser, outline the current methods for analyzing laser cavities, and explain how we model the eigenmodes of a laser resonator.

3.1 An Overview of the Working Principles of a LASER and the Governing Equation

The abbreviation 'LASER' stands for 'Light Amplification by Stimulated Emission of Radiation'. This describes the main principles of a laser which we explain very shortly below. For details, we refer, for instance, to the standard monograph [64].
Appropriately supplying an active medium with energy (pumping) raises the atoms, molecules, ions, or semiconducting crystals, depending on the type of the laser, into a higher metastable energy level. This energy is radiated again either in a stochastic manner (spontaneous emission), i.e. after different delays and in different directions, or by the mechanism of stimulated emission, where a short wave or a photon causes an excited atom (molecule, etc.) to radiate a copy of itself which is coherent in space and time. If a feedback mechanism is installed, for instance by two mirrors forming a main axis, the waves that travel in the direction of this axis are kept inside the resonator cavity and a beam traveling along the main axis can be amplified, see Figure 9, if one has a so-called population inversion in the active medium.

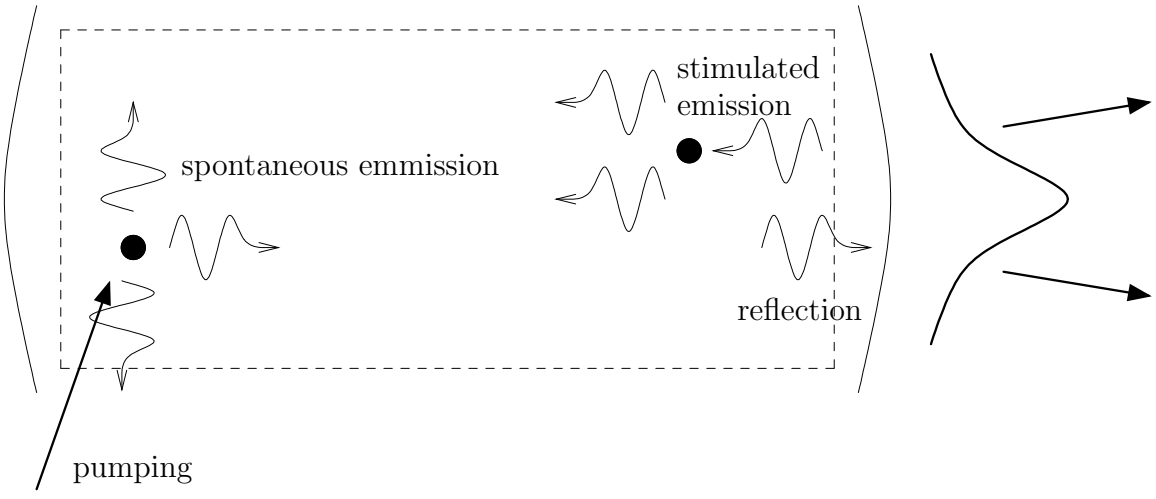

Figure 9: Absorption, spontaneous emission, feedback and stimulated emission.

For the design of lasers, it is very important to know and to influence the phase distribution of a laser beam and its transverse intensity distribution, where transverse is meant with respect to the propagation axis, because these mainly affect the efficiency of the laser and the propagation properties of the beam over longer distances.
In many cases, these profiles can approximately be computed from the eigenmodes of the laser cavity, i.e. from the electrical field distributions $E \not\equiv 0$ (more exactly, the phasor amplitudes E of time-harmonic electrical fields) which fulfill the homogeneous Helmholtz equation

$$-\Delta E - k^2 E = 0 \tag{82}$$

and which satisfy certain conditions imposed at the end mirrors of the cavity.
Roughly spoken, the existing numerical methods for analyzing laser cavities compute approximations of non-trivial solutions E of equation (82).

3.2 Drawbacks of Current Methods for Analyzing Laser Cavities

In this section, we outline the ideas and drawbacks of the present numerical methods for analyzing laser cavities, as also done in [7] or [8].
A *direct discretization* of the Helmholtz equation or of the Helmholtz eigenvalue problem as derived in Section 3.3 suffers from the difficulties mentioned in Chapter 1.3. Particularly, for simulating lasers a small mesh size has to be chosen which leads to a huge number of unknowns, making a numerical solution impossible.
Equation (82) can also be viewed as an eigenvalue problem for k and E

$$-\Delta E = k^2 E. \tag{83}$$

So, sometimes a discretization by finite elements of (83) or of an eigenproblem formulation of Maxwell's equations is applied, see e.g. [69].
This approach, however, only works for small geometries (compared with the wave length) or when a reduction to a 2D problem can be applied by certain symmetry assumptions, since otherwise the number of discretization points would be too large. So, this approach is applicable in a very limited number of cases.
Most of the numerical methods for the analysis of lasers base on an integral formulation (including a so-called round-trip condition) or, equivalently spoken, on the so-called *paraxial approximation* of the Helmholtz equation. Since the important eigenmodes $E(x, y, z)$ are waves propagating mainly in one direction, here we choose the positive z-direction, they can be represented as

$$E(x, y, z) = u(x, y, z)\, e^{-i\beta z}, \tag{84}$$

where the propagation constant $\beta \in \mathbb{R}$ is chosen such that u varies only very slowly in z. Substitution of (84) into the homogeneous Helmholtz equation (82)

and division by $e^{-i\beta z}$ yield

$$-\Delta u + 2i\beta\frac{\partial}{\partial z}u + (\beta^2 - k^2)u = 0. \tag{85}$$

Under the assumption that

$$\left|\frac{\partial^2 u}{\partial z^2}\right| \ll \left|2\beta\frac{\partial u}{\partial z}\right| \text{ or } \left|\frac{\partial^2 u}{\partial x^2}\right| \text{ or } \left|\frac{\partial^2 u}{\partial y^2}\right|,$$

equation (85) can be approximated by the paraxial wave equation

$$-\frac{\partial^2}{\partial x^2}u - \frac{\partial^2}{\partial y^2}u + 2i\beta\frac{\partial}{\partial z}u + (\beta^2 - k^2)u = 0. \tag{86}$$

As explained e.g. in [64, Chap. 7, Chap. 16], equation (86) is a good approximation of equation (85) if u describes a paraxial wave, i.e. a wave that travels at a small angle to the z-axis.
A solution of (86) can, in some sense, also be described by the so-called Fresnel approximation of the Huygens' integral. The principle of Huygens says that a scattered wave is composed of spherical waves with source points on the surface of the scattering object. Writing this out, one obtains an integral expression for the scattered wave. Thus, the application of the Fresnel approximation leads to following propagation rule: if $u(x_0, y_0, z_0)$ describes the wave at the plane $z = z_0$, the wave at the plane $z = z_0 + L$ can be computed by

$$u(x, y, z) = \frac{i}{L\lambda}\iint u(x_0, y_0, z_0)\exp\left[-i\beta\frac{(x - x_0)^2 + (y - y_0)^2}{2L}\right]dx_0 dy_0. \tag{87}$$

The paraxial approximation is, for instance, used in the *gaussian mode analysis*. In the Chapters 16, 17, and 20 of the monograph [64], this method is explained very detailed. In some cases, equation (86) possesses analytic solutions which are called gaussian modes, according to their shape. A fundamental gaussian mode has the form

$$u(x, y, z) = \frac{1}{\tilde{q}(z)}\exp\left[-ik\frac{x^2 + y^2}{2\tilde{q}(z)}\right], \tag{88}$$

where the z-dependent quantity

$$\frac{1}{\tilde{q}(z)} = \frac{1}{R(z)} - i\frac{\lambda}{\pi w^2(z)}$$

is composed of the radius of curvature $R(z)$ and the gaussian spot size $w(z)$. In this representation, as throughout the remainder of this chapter, it is assumed that the propagation direction coincides with the z-axis.

Alternatively, the fundamental gaussian mode can be written in a normalized form as

$$u(x,y,z) = \left(\frac{2}{\pi}\right)^{1/2} \frac{\exp[-ikz + i\psi(z)]}{w(z)} \exp\left[-\frac{x^2+y^2}{w^2(z)} - ik\frac{x^2+y^2}{2R(z)}\right] \qquad (89)$$

where $\psi(z)$ is the so-called Guoy phase shift.
In practice, the elements in a laser configuration are approximated by parabolic (or gaussian) elements, i.e. by elements that keep a beam gaussian meaning in the form (88). By complex ABCD-matrices (see [64]) the effects of (parabolic) apertures, ducts, lenses and other elements on $\tilde{q}(z)$ are described. This is e.g. implemented in the code LASCAD$^{\text{TM}}$ [2].
However, in many practical situations this is no satisfactory approximation, as for instance the numerical example in Chapter 7.6 shows.
The last type of methods for analyzing laser cavities is based on the *Fox-Li-Approach*, see [30]. The idea is to choose a normalized, more or less arbitrary start distribution $u_0(x,y)$ on a reference plane at $z = z_0$, to compute the distribution $u_1(x,y)$ at the same reference plane, but after one round-trip of this front, by a *Beam-Propagation-Method* (BPM), and to normalize the obtained distribution again. For the propagation, either equation (86) is discretized and the front is propagated in positive or negative z-direction by finite difference or finite element methods, until after one round-trip the reference plane has been reached, or a Fourier transformation is applied to the integral representation (87) of a paraxial wave (see e.g. [64, Chap. 14]). Iterating this procedure one hopes to obtain a steady state, i.e. a distribution that does not change its pattern in a round-trip, except for a reduction in amplitude and a phase shift, see again e.g. [64].
Using the integral formulation, this can be expressed as searching for distributions u_{nm} that fulfill the eigenvalue equation

$$\iint K(x,y,x_0,y_0)u_{nm}(x_0,y_0)\; dx_0\; dy_0 = \gamma_{nm}u_{nm}(x,y)$$

where γ_{nm} is a complex eigenvalue and K is the propagation kernel describing the effect of one round-trip on the distribution at the reference plane.
The different transverse modes, essentially, are identified with the help of Fourier analysis applied to a sample of many successive patterns. There exist two different methods for extracting the higher-order modes: the *Prony method* suggested by Siegman and Miller [65] and the approach of Feit and Fleck [29]. As explained in [7], the method of Feit and Fleck fails for long resonators. To our knowledge, no investigations of the accuracy of the eigenvalues and eigenmodes obtained by these methods have been published until now.
Actually, many practitioners are unsatisfied with these BPMs, because the iteration sometimes converges very slowly or does not converge at all. And until now, this behavior is not understood very well. Furthermore, practical experience with

BPM shows that small changes in the start pattern can strongly affect the mode patterns.
The results on the solvability of the Helmholtz boundary value problem that are reviewed in Section 1.2, state that in general the homogeneous BVP for the Helmholtz equation

$$\begin{aligned} -\Delta u - k^2 u &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \Gamma_s \\ \partial_n u - i\alpha u &= 0 \quad \text{on } \Gamma_r \end{aligned} \tag{90}$$

on a finite domain Ω with (approximate) absorbing boundary conditions has no non-trivial solution. This essentially depends on the radiating boundary conditions at Γ_r. Very roughly spoken, using approximate radiating boundary conditions for an electrical wave can lead to a homogeneous boundary value problem which only possesses the trivial solution. Or in other words: the eigenmodes are eigensolutions of the Helmholtz operator with eigenvalue zero. Imposing approximate boundary conditions at a finite radiating boundary, leads to a rejection of the eigenmodes and only admits the trivial solution $u = 0$.
The approach to propagate a wave front in a window of finite width and then to normalize it, is a way to obtain an approximation of a solution of the paraxial wave equation (86) at one plane. Rigorously viewed, this restriction to a finite domain imposes, if explicitly stated or not, approximate boundary conditions on the solution. The statements in the previous paragraph indicate that in many cases the Fox-Li approach is not appropriate for computing eigenmodes of laser cavities.
However, it remains to analyze the beam propagation methods more detailed; particularly, investigating well-posedness and the influence of the boundary conditions will help to understand the behavior of these methods.

3.3 Derivation of a Two-Wave Eigenvalue Problem for the Laser Resonator

Section 3.2 shows, that the existing numerical methods do not yield reliable results or cannot be applied in many cases. Particularly, from the comments on the beam propagation methods, it follows that a model for the eigenmodes of a laser cavity has to be developed which is not sensitive to small changes in the boundary conditions. For this purpose, we approximate the homogeneous Helmholtz equation

$$-\Delta E(x,y,z) - k^2(x,y,z)\ E(x,y,z) = 0 \tag{91}$$

by an eigenvalue problem

$$-\Delta \tilde{E}(x,y,z) - k^2(x,y,z)\tilde{E}(x,y,z) = \xi \tilde{E}(x,y,z) \tag{92}$$

as explained below.

A nontrivial solution $E(x,y,z) \neq 0$ of equation (91) can be represented as

$$E(x,y,z) = u(x,y,z)\, e^{-i(k_f-\epsilon)z}, \tag{93}$$

where, at first, k_f and ϵ are arbitrary real numbers with $\epsilon \ll k_f$. Inserting representation (93) into the homogeneous Helmholtz equation (91) and canceling out the term $e^{-ik_f z}$, we obtain

$$\begin{aligned} -\Delta u(x,y,z) + 2i(k_f-\epsilon)\frac{\partial}{\partial z}u(x,y,z) + (k_f^2 - k^2(x,y,z))u(x,y,z) \\ = \epsilon(2k_f-\epsilon)u(x,y,z) \end{aligned}$$

or, equivalently,

$$\begin{aligned} -\Delta u(x,y,z) + 2ik_f\frac{\partial}{\partial z}u(x,y,z) + (k_f^2 - k^2(x,y,z))u(x,y,z) \\ = 2k_f\epsilon u(x,y,z) + 2i\epsilon\frac{\partial}{\partial z}u(x,y,z) - \epsilon^2 u(x,y,z). \end{aligned} \tag{94}$$

In [7], the representation (93) is justified physically and u, k_f, and ϵ are given a concrete interpretation. Assuming that

$$\left| 2i\epsilon\frac{\partial}{\partial z}u(x,y,z) - \epsilon^2 u(x,y,z) \right| \ll |2k_f\epsilon u(x,y,z)|\,, \tag{95}$$

as it is done in [7], we neglect the two terms of the left-hand side of (95) in the equation (94) and obtain the approximation

$$-\Delta\tilde{u}(x,y,z) + 2ik_f\frac{\partial}{\partial z}\tilde{u}(x,y,z) + (k_f^2 - k^2(x,y,z))\tilde{u}(x,y,z) = \xi\tilde{u}(x,y,z), \tag{96}$$

where $\tilde{u}(x,y,z) \approx u(x,y,z)$ and $\xi \approx 2k_f\epsilon$. The equation (96) can be seen as an eigenvalue problem for $\tilde{u}$ and ξ. If u, k_f, and ϵ fulfill condition (95), equation (96) will be a good approximation of equation (94).
Multiplication of the eigenvalue equation (96) by $e^{-ik_f z}$ and simple manipulation gives an Helmholtz eigenvalue problem for $\tilde{E}(x,y,z) := \tilde{u}(x,y,z)\, e^{-ik_f z}$

$$-\Delta\tilde{E}(x,y,z) - k^2\tilde{E}(x,y,z) = \xi\tilde{E}(x,y,z). \tag{97}$$

We see, that $\tilde{E}$ satisfies a perturbed homogeneous Helmholtz equation

$$\left[-\Delta - (k^2+\xi)\right]\tilde{E}(x,y,z) = 0.$$

The derivation shows, that we, furthermore, have the approximate relation

$$E \approx \tilde{E}e^{i\epsilon z}. \tag{98}$$

Thus, a solution E of the homogeneous Helmholtz equation (91) can be approximated well by solving the eigenvalue problem (97) for $\tilde{E}$ and applying relation (98), if condition (95) holds.
Now, we explain the model, by which we compute the modes of a laser cavity.
Here, for ease of description and computation, we describe the cavity by a cuboid, see Figure 10. The mirrors shall be located at the left and right faces. We denote the cuboid of size $W \times W \times L$ by Ω, and the boundaries by Γ_0, Γ_1, and Γ_r, which stand for the left mirror, the right mirror, and the remaining open part of the boundary, respectively.

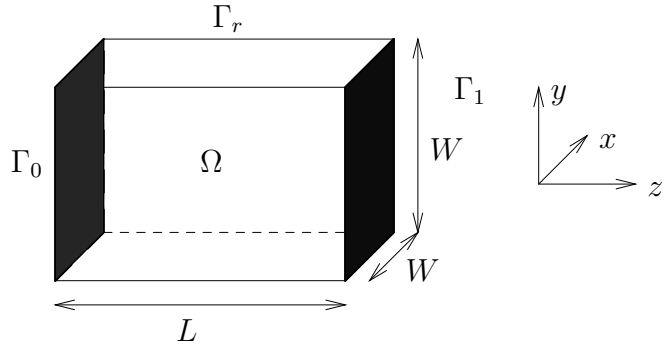

Figure 10: The computational domain for a laser cavity.

To represent a mode in a laser resonator cavity, we use the two-wave ansatz

$$\tilde{E}(x,y,z) = \underbrace{\exp\left[-i(k_f-\epsilon)z\right]\tilde{u}_r(x,y,z)}_{\tilde{E}_r(x,y,z)\cdot\exp[i\epsilon z]:=} + \underbrace{\exp\left[-i(k_f-\epsilon)(L-z)\right]\tilde{u}_l(x,y,z)}_{\tilde{E}_l(x,y,z)\cdot\exp[-i\epsilon z]:=}. \tag{99}$$

with waves $\tilde{E}_r$ and $\tilde{E}_l$ that are eigensolutions of (97) with the same eigenvalue ξ and that are appropriately coupled at the end mirrors, i.e. at the boundaries Γ_0 and Γ_1.
The ansatz (99) reflects the well-known idea (see e.g. [64]) that a mode in a resonator can be seen as a periodic wave with periodicity $2L$ in z. ($\tilde{E}_l$ is the reflected wave for $L < z < 2L$.)
However, solving for the eigensolutions $\tilde{E}_r$ and $\tilde{E}_l$ directly, demands for a numerical solution of two coupled Helmholtz eigenvalue problems (97), which leads to the difficulties mentioned in Section 1.3.
The ansatz (99) and the derivation of the Helmholtz eigenvalue problem reveal that $\tilde{u}_r$ and $\tilde{u}_l$ are the interesting quantities. These fulfill the eigenvalue equation (96), mutatis mutandis for $\tilde{u}_l$. So, we compute $\tilde{E}$ by computing $\tilde{u}_r$ and $\tilde{u}_l$.
We remark, that computing for $\tilde{u}_r$ and $\tilde{u}_l$ instead of for $\tilde{E}_r$ and $\tilde{E}_l$ corresponds to the idea of separating oscillations which is presented in Chapter 2.

As $\tilde{E}_r$ and $\tilde{E}_l$, the functions $\tilde{u}_r$ and $\tilde{u}_l$ must meet coupling boundary conditions at the end mirrors Γ_0 and Γ_1:

$$\begin{aligned} \tilde{u}_r(x,y,0) &= \phi_0(x,y)\cdot\tilde{u}_l(x,y,0), \\ \tilde{u}_l(x,y,L) &= \phi_1(x,y)\cdot\tilde{u}_r(x,y,L), \end{aligned} \tag{100}$$

and

$$\begin{aligned} \frac{\partial\tilde{u}_r}{\partial z}(x,y,0) &= -\phi_0(x,y)\cdot\frac{\partial\tilde{u}_l}{\partial z}(x,y,0), \\ \frac{\partial\tilde{u}_l}{\partial z}(x,y,L) &= -\phi_1(x,y)\cdot\frac{\partial\tilde{u}_r}{\partial z}(x,y,L), \end{aligned} \tag{101}$$

where the functions $\phi_0(x,y)$ and $\phi_1(x,y)$ model the phase shifts of the waves due to reflection and due to the curvature of the mirrors. We approximate the exact shifts by choosing

$$\phi_0(x,y) := \exp\left[ik_f\left(\frac{x^2+y^2}{R_0}\right) - i\pi\right] \tag{102}$$

and

$$\phi_1(x,y) := \exp\left[ik_f\left(\frac{x^2+y^2}{R_1}\right) - i\pi\right], \tag{103}$$

where R_0 is the (parabolic) radius of curvature of the left mirror and R_1 of the right mirror, respectively. In the following Chapter 3.4 the modeling of the boundary conditions is explained more detailed.
At the open part of the boundary Γ_r we impose Robin boundary conditions on $\tilde{E}_r$ and $\tilde{E}_l$. As explained for the example in Chapter 1.1.1, this condition is an approximation for the exact non-reflecting boundary condition. Since the outward normal of the domain Ω on the boundary Γ_r always is perpendicular to the z-direction, the Robin boundary conditions can equivalently be applied to $\tilde{u}_r$ and $\tilde{u}_l$.
Let us now combine all considerations above. An eigenmode of a laser resonator is modeled as an eigensolution $(\tilde{u}_r, \tilde{u}_l)$ with complex eigenvalue ξ of following PDE eigenvalue problem:

$$\begin{aligned} -\Delta\tilde{u}_r + 2ik_f\frac{\partial\tilde{u}_r}{\partial z} + (k_f^2 - k^2)\tilde{u}_r &= \xi\tilde{u}_r \quad \text{and} \\ -\Delta\tilde{u}_l - 2ik_f\frac{\partial\tilde{u}_l}{\partial z} + (k_f^2 - k^2)\tilde{u}_l &= \xi\tilde{u}_l \quad \text{in } \Omega \end{aligned} \tag{104}$$

with boundary conditions

$$\begin{aligned} \tilde{u}_r - \phi_0\tilde{u}_l &= 0 \quad \text{on } \Gamma_0, \\ \tilde{u}_r - \bar{\phi}_1\tilde{u}_l &= 0 \quad \text{on } \Gamma_1, \end{aligned}$$

$$\begin{aligned}
\frac{\partial \tilde{u}_r}{\partial z} + \phi_0 \frac{\partial \tilde{u}_l}{\partial z} &= 0 \quad \text{on } \Gamma_0, \\
\frac{\partial \tilde{u}_r}{\partial z} + \bar{\phi}_1 \frac{\partial \tilde{u}_l}{\partial z} &= 0 \quad \text{on } \Gamma_1, \\
\frac{\partial \tilde{u}_r}{\partial \vec{n}} - iC_b \tilde{u}_r &= 0 \quad \text{on } \Gamma_r, \\
\frac{\partial \tilde{u}_l}{\partial \vec{n}} - iC_b \tilde{u}_l &= 0 \quad \text{on } \Gamma_r,
\end{aligned} \tag{105}$$

where C_b can be chosen as $C_b = k_f$.
This two-wave eigenvalue problem [(104) and (105)] is our model for the determination of the eigenmodes of a laser cavity.
It is noteworthy, that ξ is not a simple eigenvalue in general. To demonstrate this, we assume the parameters to be

$$\phi_0 \equiv \phi_1 \equiv -1 \qquad \text{and} \qquad e^{i4k_f z} = 1.$$

Then, if $(\tilde{u}_r, \tilde{u}_l)$ is an eigensolution of the eigenvalue problem (104), (105) with eigenvalue ξ, simple computation shows that the pair

$$(\hat{u}_r, \hat{u}_l) := (e^{-2ik_f z} \cdot \tilde{u}_l, e^{2ik_f z} \cdot \tilde{u}_r)$$

satisfies the boundary conditions (105) and the equations

$$\begin{aligned}
-\Delta \hat{u}_r + 2ik_f \frac{\partial \hat{u}_r}{\partial z} + (k_f^2 - k^2)\hat{u}_r - 4ik_f e^{-2ik_f z} \frac{\partial \tilde{u}_l}{\partial z} &= \xi \hat{u}_r \quad \text{and} \\
-\Delta \hat{u}_l - 2ik_f \frac{\partial \hat{u}_l}{\partial z} + (k_f^2 - k^2)\hat{u}_l + 4ik_f e^{2ik_f z} \frac{\partial \tilde{u}_r}{\partial z} &= \xi \hat{u}_l \quad \text{in } \Omega.
\end{aligned} \tag{106}$$

If we, furthermore, assume that $\partial \tilde{u}_r / \partial z = \partial \tilde{u}_l / \partial z = 0$, obviously $(\hat{u}_r, \hat{u}_l)$ is also an eigensolution with eigenvalue ξ. In contrast to the constant solution $(\tilde{u}_r, \tilde{u}_l)$, the eigensolution $(\hat{u}_r, \hat{u}_l)$ is very oscillatory.

3.4 Description of Boundary Conditions and Interior Boundary Conditions

In this section, we derive boundary and interior boundary conditions by which we describe the effects of lenses, dielectric interfaces and mirrors in a laser configuration. We model these elements by applying appropriate phase shifts to the waves $\tilde{u}_r$ and $\tilde{u}_l$. (The formulas are the results of a personal communication [6].) However, we do not specify conditions for the first order derivatives of $\tilde{u}_r$ and $\tilde{u}_l$, which could be necessary for obtaining a well-defined strongly formulated PDE. The variational formulation, as used for the finite element analysis, yields so-called natural boundary conditions. For the end mirrors, these natural boundary conditions are equivalent with the first order coupling conditions in (105), see Chapter 5.1.

3.4.1 Conditions for Dielectric Interfaces and Lenses

In the following, we derive the condition that has to be imposed on a wave which is passing a *dielectric interface*.

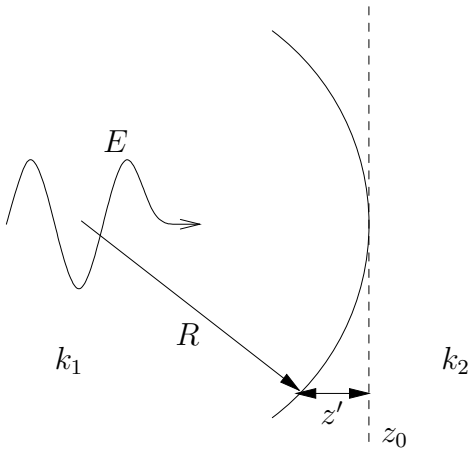

Figure 11: Interior boundary condition for a dielectric interface.

Let E be a wave moving to the right with propagation constants k_1 for the domain left from the parabolic dielectric interface with apex at $z = z_0$ and k_2 for right-hand part of the domain, as shown in Figure 11. We can represent E as

$$E(x,y,z) = \begin{cases} u_1(x,y,z)\exp(-ik_1(z-z_0)) & \text{if } z \le z_0 - z'(x,y) \\ u_2(x,y,z)\exp(-ik_2(z-z_0)) & \text{if } z > z_0 - z'(x,y), \end{cases} \tag{107}$$

where u_1, u_2 vary slowly in z and $z'(x,y) = (x^2+y^2)/(2R)$ describes the interface with radius of curvature R. For a convex interface as in Figure 11 the radius of curvature shall, by convention, be positive $R > 0$.
Since the wave E is continuous, particularly, at $z = z_0 - z'$, following relation must hold

$$u_2(x,y,z_0+z') = u_1(x,y,z_0+z')\exp(i(k_1-k_2)z'). \tag{108}$$

We want to approximate E by $\tilde{E}$ defined by

$$\tilde{E}(x,y,z) = \begin{cases} \tilde{u}_1(x,y,z)\exp(-ik_1(z-z_0)) & \text{if } z \le z_0 \\ \tilde{u}_2(x,y,z)\exp(-ik_2(z-z_0)) & \text{if } z > z_0. \end{cases} \tag{109}$$

If we assume that the interface has not a large curvature, the relation (108) can be applied at the plane $z = z_0$ yielding

$$\tilde{u}_2(x,y,z_0) := \tilde{u}_1(x,y,z_0)\exp(i\varphi_{\mathrm{DI}}(x,y)). \tag{110}$$

with

$$\varphi_{\mathrm{DI}}(x,y) = (k_1-k_2)\left(\frac{x^2+y^2}{2R}\right). \tag{111}$$

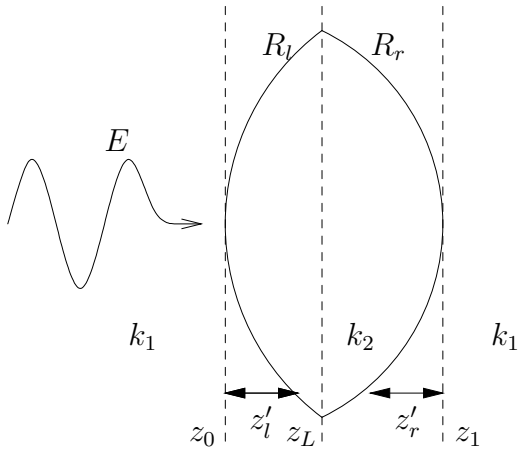

Figure 12: Interior boundary condition for a lens.

As for dielectric interfaces, we want to describe the effect of a *lens* (Figure 12) by an appropriately chosen phase shift $\varphi_{\mathrm{L}}(x, y)$. The wave E is approximated by

$$\tilde{E}(x,y,z) = \begin{cases} \tilde{u}_1(x,y,z)\exp\left(-ik_1(z-z_L)\right) & \text{if } z \le z_L \\ \tilde{u}_2(x,y,z)\exp\left(-ik_1(z-z_L)\right) & \text{if } z_L < z. \end{cases} \tag{112}$$

Since the propagation of the wave through a lens can be seen as passing two dielectric interfaces with radii R_l and R_r, the phase shift φ_{L} that a wave which passes a lens is subjected to can be approximated by the sum of the shift φ^l_{DI} from entering the lens, the shift φ^r_{DI} from leaving the lens, and shifts due to different propagation constants:

$$\begin{aligned} \varphi_{\mathrm{L}}(x,y) &:= \varphi^l_{\mathrm{DI}}(x,y) - (k_2-k_1)(z_L-z_0) + \varphi^r_{\mathrm{DI}}(x,y) - (k_2-k_1)(z_1-z_L) \\ &= (k_2-k_1)\left(\frac{x^2+y^2}{-2R_l} + \frac{x^2+y^2}{2R_r} + (z_0-z_1)\right). \end{aligned} \tag{113}$$

Thus, we have the interior boundary condition

$$\tilde{u}_2(x,y,z_L) = \tilde{u}_1(x,y,z_L)\exp\left(i\varphi_{\mathrm{L}}(x,y)\right) \tag{114}$$

with φ_{L} as in (113).
From the phase shift (113) a condition for a *thin lens* can be derived. For a thin lens we assume that $z_1 - z_0 \approx 0$ and $R_l \approx \infty$. If, furthermore, the refractive index n_1 outside the lens is $n_1 = 1$, then instead of using R_r, k_1, and k_2, we describe the effect of a thin lens by its focal distance f, and obtain the phase shift

$$\varphi_{\mathrm{TL}}(x,y) := k_1\frac{x^2+y^2}{2f}. \tag{115}$$

3.4.2 Conditions for End Mirrors

To derive the phase shift for the *reflection at end mirrors*, we consider the configuration in Figure 13. Let E_l be the incoming wave traveling to the left and E_r the reflected one propagating to the right.
We describe the incoming wave E_l by

$$E_l(x, y, z) = u_l(x, y, z) \exp\,(\,+ ik(z - z_0)) \tag{116}$$

and the reflected wave E_r by

$$E_r(x, y, z) = u_r(x, y, z) \exp\,(\,- ik(z - z_0)). \tag{117}$$

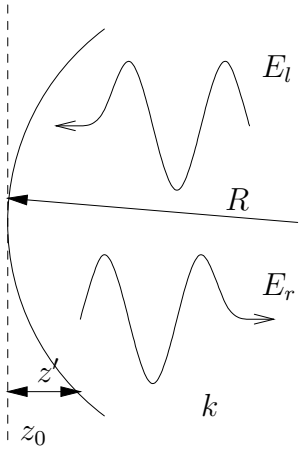

Figure 13: Boundary condition for a reflecting mirror.

The reflection at a mirror can be described by a phase shift of $-\pi$ between incoming and reflected wave as known from textbooks of optics. That means for the parabolic end mirror in Figure 13, that

$$E_r(x, y, z_0 + z') = \exp(-i\pi) E_l(x, y, z_0 + z'). \tag{118}$$

Using (116) and (117) we obtain the relation

$$\tilde{u}_r(x, y, z_0 + z') \exp\,(\,- ikz') = \exp(-i\pi) \tilde{u}_l(x, y, z_0 + z') \exp\,(\,+ ikz') \tag{119}$$

or, equivalently,

$$\tilde{u}_r(x, y, z_0 + z') = \tilde{u}_l(x, y, z_0 + z') \exp\,(i\varphi_0(x, y)) \tag{120}$$

with

$$\varphi_0(x, y) = 2kz' - \pi = k\frac{x^2 + y^2}{R} - \pi. \tag{121}$$

As for the dielectric interface, if the radius of curvature R is large, we obtain a good approximation when we apply this locally varying phase shift at the plane $z = z_0$ instead of the mirror $z = z_0 + z'$.
In our numerical tests in Chapter 7, we restricted ourselves to configurations which can be described without interior boundary conditions. As can be seen in Chapter 5.1, the curvature of the end mirrors was modeled by incorporating the phase shifts into the test and trial space of the variational formulation.

3.5 Extracting the Spot Size and the Guoy Phase Shift from the Fundamental Eigenmode

Let $(\tilde{u}_r, \tilde{u}_l)$ be the solution of the eigenvalue problem (104), (105) with eigenvalue $\xi(= 2k_f\epsilon)$, which describes a nearly gaussian, fundamental (or lowest-order) eigenmode. Then, on every plane which is perpendicular to the propagation axis, i.e. to the z-axis, $\tilde{u}_r$ and $\tilde{u}_l$ assume their maximal absolute values on this axis.
The *spot size* $w(z_0)$ (see also equation (89)) of a fundamental wave $\tilde{u}$ at the plane with $z = z_0$ is defined as the radius $r = \sqrt{x^2 + y^2}$ where the squared modulus of $\tilde{u}$ has decreased to $1/e^2$ of its maximum value (which is assumed on the z-axis) with respect to this plane, i.e. where

$$|\tilde{u}(x, y, z_0)|^2 = \frac{1}{e^2} \cdot |\tilde{u}(0, 0, z_0)|^2.$$

Thus, the z-dependent spot size of the finite element solution can easily be computed by interpolation, as it has been done for the numerical tests in Chapter 7.
Furthermore, the *Guoy phase shift* $\psi(z)$ can be computed from $\tilde{\epsilon}$ and $\tilde{u}_r$ or $\tilde{u}_l$, respectively. As equation (89) shows, on the propagation axis the relation

$$\tilde{u}(0, 0, z) = \exp\left[-i(\tilde{\epsilon}z - \psi(z))\right] |\tilde{u}(0, 0, z)| \tag{122}$$

holds. Since $\psi(z)$ is only determined except for a constant, an appropriate normalization has to be applied. For the numerical example in Chapter 7.2, the Guoy phase shift computed from the finite element solution is compared with analytical one showing very good correspondence.

4 Abstract Convergence Proof for an Approximate Solution of a Quadratic Eigenvalue Problem

Eigenvalue problems for partial differential equations, as for instance the two-wave eigenvalue problem in Chapter 3.3, are a very important class of problems in mathematics and in many fields of science. A standard way of numerical approximation of these problems is to discretize them using the finite element method (FEM). Doing so, immediately the question of existence and convergence of the finite element solution arises.
For the linear eigenvalue problem, the answers are well-known. For instance in [10] or [38] a complete convergence theory is presented. Furthermore, we mention just a few articles that contain some additional methods and ideas concerning linear eigenvalue problems: [13], [15], [19], [40], [46], [71], and [74].
For quadratic eigenvalue problems, the situation is completely different: On the convergence of finite element solutions of quadratic PDE eigenvalue problems almost nothing has been proved. As far as we know, solely in [14] convergence has been shown for a special problem and a special discretization. However, the theory of finite-dimensional quadratic eigenvalue problems has developed to a mature status, see e.g. [70] for an overview.
In this chapter, we present an abstract – and hence, quite general – proof of convergence for a discretized quadratic eigenvalue problem.
In Chapter 5.3 we will apply this theory to show convergence of the (linear) two-wave eigenvalue problem formulated in Chapter 3.3.

4.1 The Idea: An Abstract Quadratic Eigenvalue Problem in Hilbert Spaces and its Linearization

Let us, first, consider the abstract quadratic eigenvalue problem

$$\mathcal{A}u = \lambda\mathcal{B}u + \lambda^2\mathcal{C}u, \tag{123}$$

where $\mathcal{A}, \mathcal{B}$ and $\mathcal{C}$ are bounded linear operators on a complex Hilbert Space $\mathcal{H}$, and $\mathcal{B}, \mathcal{C}$ additionally are assumed to be compact. For the notions and statements of functional analysis in this chapter, we refer e.g. to [4], [38], and [73].
Let $(u, \lambda) \in \mathcal{H} \times \mathbb{C}$ be the eigenpair (i.e. the eigensolution and the corresponding eigenvalue) of interest and let $w \in \mathcal{H}$ be a fixed approximation of u. To norm u, we additionally impose the condition

$$\langle u, w \rangle = 1$$

with $\langle \cdot, \cdot \rangle$ being a scalar product on the space $\mathcal{H}$.

Then, (u, λ) fulfills

$$F(u,\lambda) := \begin{pmatrix} \mathcal{A}u - \lambda\mathcal{B}u - \lambda^2\mathcal{C}u \\ \langle u, w\rangle - 1 \end{pmatrix} = 0. \tag{124}$$

Using the Frechét derivative F_L of F in (u, λ), we can write

$$F(v,\mu) = \underbrace{F(u,\lambda)}_{=0} + F_L(v-u, \mu-\lambda) + G(v,\mu), \tag{125}$$

where

$$F_L(x,\xi) = \begin{pmatrix} \mathcal{A}x - \xi\mathcal{B}u - \lambda Bx - 2\lambda\xi\mathcal{C}u - \lambda^2 Cx \\ \langle x, w\rangle \end{pmatrix}$$

and

$$G(v,\mu) = \begin{pmatrix} -(\mu-\lambda)\mathcal{B}(v-u) - (\mu-\lambda)(\mu+\lambda)\mathcal{C}(v-u) - (\mu-\lambda)^2\mathcal{C}u \\ 0 \end{pmatrix}.$$

Obviously, $G(v,\mu)$ is of order $o(\|v-u\| + |\mu-\lambda|)$.
Roughly spoken, the operator F_L describes the behavior of F in a small neighborhood of (u, λ). Since F_L is a linear operator, it is called the *linearization* of F in (u, λ).
To estimate the error

$$e := (v-u, \mu-\lambda),$$

one aims for an appropriate representation of it. Under the assumption that the inverse operator $(F_L)^{-1}$ exists, from (125) the equation

$$e = (F_L)^{-1}(F(v,\mu) - G(v,\mu)) \tag{126}$$

follows.
The idea behind representation (126) will help us to estimate the error and to prove convergence in a variational context.
If the operator F_L, for instance, satisfies a Fredholm alternative, the regularity of this operator (or the existence of the inverse operator $(F_L)^{-1}$) is equivalent to the statement that $F_L(x,\xi) = 0$ implies $(x,\xi) = 0$, or more detailed, that the equations

$$\mathcal{A}x - \lambda\mathcal{B}x - \lambda^2\mathcal{C}x = \xi\mathcal{B}u + 2\lambda\xi\mathcal{C}u \tag{127}$$

$$\langle x, w\rangle = 0 \tag{128}$$

imply $(x,\xi) = 0$. This is true, if the eigenvalue λ is of simple geometric and algebraic multiplicity, i.e. if

$$\dim\big(\ker(\mathcal{A} - \lambda\mathcal{B} - \lambda^2\mathcal{C})\big) = 1 \tag{129}$$

holds and no generalized eigensolution for (u, λ) exists:

$$\nexists x \in \mathcal{H} : (\mathcal{A} - \lambda\mathcal{B} - \lambda^2\mathcal{C})x = -\frac{d}{d\tau}(\mathcal{A} - \tau\mathcal{B} - \tau^2\mathcal{C})\Big|_{\tau=\lambda} u = (\mathcal{B} + 2\lambda\mathcal{C})u, \quad (130)$$

see [47] or [70] for the finite dimensional analogon.
Conditions (129) and (130) are a generalization of the well-known conditions for linear eigenvalue problems.

4.2 A Variational Formulation of the Eigenvalue Problem in Hilbert Spaces and its Linearization

In this section, we deal with a variational quadratic eigenvalue problem and apply the ideas of the previous section in this context.
For this purpose, let us consider the Gelfand triple

$$\mathcal{H} \subset \mathcal{L} \subset \mathcal{H}',$$

where the embedding $\mathcal{H} \subset \mathcal{L}$ is continuous, dense, and compact and $\mathcal{H}'$ is the dual space of $\mathcal{H}$. Let $\|\cdot\|_L$, $(\cdot,\cdot)$ and $\|\cdot\|_H, \langle\cdot,\cdot\rangle$ be the norms and inner products of $\mathcal{L}$ and $\mathcal{H}$, respectively.
We consider the following eigenvalue problem, which can be seen as a variational formulation of the eigenvalue problem (123):
Find $v \in \mathcal{H}$ and $\mu \in \mathbb{C}$, such that

$$a(v, \varphi) = \mu\, b(v, \varphi) + \mu^2 c(v, \varphi) \qquad \forall \varphi \in \mathcal{H} \quad (131)$$

under the restriction

$$\langle v, w\rangle = 1. \quad (132)$$

Here, $a(\cdot,\cdot), b(\cdot,\cdot)$, and $c(\cdot,\cdot)$ are sesquilinear forms on $\mathcal{H}$ corresponding to the operators $\mathcal{A}, \mathcal{B}$, and $\mathcal{C}$, respectively, in Chapter 4.1. Let us suppose, that there exist constants $c_a, c_b, c_c > 0$ such that

$$|a(v, \varphi)| \leq c_a \|v\|_H \|\varphi\|_H, \quad (133)$$

$$|b(v, \varphi)| \leq c_b \|v\|_H \|\varphi\|_L, \quad (134)$$

$$\text{and} \quad |c(v, \varphi)| \leq c_c \|v\|_H \|\varphi\|_L \quad (135)$$

for all $v, \varphi \in \mathcal{H}$. Condition (133) expresses that $a(\cdot,\cdot)$ is a continuous (or bounded) sesquilinear form on $\mathcal{H}$, whereas (134) and (135) are stronger conditions which by the continuous embedding $\mathcal{H} \subset \mathcal{L}$, however, imply continuity of $b(\cdot,\cdot)$ and $c(\cdot,\cdot)$.
According to the conditions (129) and (130), we suppose that (u, λ) is an eigenpair of (131) which fulfills following condition of simplicity

(A_S): The space $\mathcal{H}$ can be decomposed into the direct sum

$$\mathcal{H} = \mathcal{H}^\perp \oplus \operatorname{span}_{\mathbb{C}} \{u\}$$

such that

(i) the equation

$$a(v,\varphi) - \lambda\, b(v,\varphi) - \lambda^2 c(v,\varphi) = 0 \quad \forall \varphi \in \mathcal{H}$$

implies

$$v \in \operatorname{span}_{\mathbb{C}} \{u\},$$

(ii) for $z \in \mathcal{H} \setminus \{0\}$ holds: From

$$\exists v \in \mathcal{H} : b(z,\varphi) + 2\lambda\, c(z,\varphi) = a(v,\varphi) - \lambda\, b(v,\varphi) - \lambda^2 c(v,\varphi) \quad \forall \varphi \in \mathcal{H}$$

follows

$$z \in \mathcal{H}^\perp \setminus \{0\}.$$

In the sequel, we use the product spaces $\mathcal{H} \times \mathbb{C}$ and $\mathcal{L} \times \mathbb{C}$ endowed with the norms

$$\|[x,\xi]\|_{\mathcal{H}\times\mathbb{C}} := \|x\|_H + |\xi|$$

and

$$\|[x,\xi]\|_{\mathcal{L}\times\mathbb{C}} := \|x\|_L + |\xi|.$$

Variationally formulating the linearization indicated in Section 4.1, we obtain the sesquilinear form

$$B : (\mathcal{H} \times \mathbb{C}) \times (\mathcal{H} \times \mathbb{C}) \to \mathbb{C}$$

defined by

$$\begin{aligned} &B([x,\xi],[\varphi,\zeta]) := \\ &\quad a(x,\varphi) - \lambda\, b(x,\varphi) - \xi\, b(u,\varphi) - 2\lambda\xi\, c(u,\varphi) - \lambda^2 c(x,\varphi) + \bar{\zeta}\langle x, w\rangle. \end{aligned} \tag{136}$$

It is standard, to prove that B also is a continuous sesquilinear form. Furthermore, a simple computation shows, that $B([v-u,\mu-\lambda],[\varphi,\zeta])$ can be written as

$$\begin{aligned} &B([v-u,\mu-\lambda],[\varphi,\zeta]) = \\ &\quad a(v,\varphi) - \mu\, b(v,\varphi) - \mu^2 c(v,\varphi) \\ &\quad +(\mu-\lambda)\, b(v-u,\varphi) + (\mu-\lambda)(\mu+\lambda)\, c(v-u,\varphi) + (\mu-\lambda)^2 c(u,\varphi) \\ &\quad +\bar{\zeta}\langle v-u, w\rangle. \end{aligned} \tag{137}$$

In the following lemma we prove a condition for the validity of the Fredholm alternative.

Lemma 7 (Gårding inequality) *Let $a(\cdot,\cdot)$ satisfy a Gårding inequality on $\mathcal{H}$, i.e. let there exist constants $c_g, C_g > 0$ such that*

$$\operatorname{Re}\left[a(v,v)\right] \geq c_g \|v\|_H^2 - C_g \|v\|_L^2 \tag{138}$$

holds for all $v \in \mathcal{H}$. Let, furthermore, inequalities (134) and (135) be satisfied. Then, the sesquilinear form B defined by (136) also satisfies a Gårding inequality on $\mathcal{H} \times \mathbb{C}$

$$\operatorname{Re}\left[B([x,\xi],[x,\xi])\right] \geq \tilde{c}_g \left(\|x\|_H + |\xi|\right)^2 - \tilde{C}_g \left(\|x\|_L + |\xi|\right)^2 \tag{139}$$

with constants $\tilde{c}_g, \tilde{C}_g > 0$.

Proof: By simple estimates, inequalities (134) and (135), and the Cauchy-Schwarz inequality for the inner product $\langle\cdot,\cdot\rangle$ on $\mathcal{H}$, one obtains

$$\begin{aligned}
&\operatorname{Re}\left[B([x,\xi],[x,\xi])\right] \\
&\quad\geq \operatorname{Re}\left[a(u,u)\right] - \left|-\lambda\, b(x,x) - \xi\, b(u,x) - 2\lambda\xi\, c(u,x) - \lambda^2 c(x,x) + \bar{\xi}\langle x,w\rangle\right| \\
&\quad\geq \operatorname{Re}\left[a(u,u)\right] - \left(|\lambda|\,|b(x,x)| + |\xi|\,|b(u,x)| + 2|\lambda|\,|\xi|\,|c(u,x)| + |\lambda|^2|c(x,x)|\right. \\
&\qquad \left.+|\xi|\,|\langle x,w\rangle|\right) \\
&\quad\geq \operatorname{Re}\left[a(u,u)\right] - \left(|\lambda|\, c_b\, \|x\|_H\|x\|_L + |\xi|\, c_b\, \|u\|_H\|x\|_L + 2|\lambda|\,|\xi|\, c_c\, \|u\|_H\|x\|_L\right. \\
&\qquad \left.+|\lambda|^2 c_c\, \|x\|_H\|x\|_L + |\xi|\,\|x\|_H\|w\|_H\right).
\end{aligned}$$

Using the Gårding inequality (138) and appropriately applying the generalized Young inequality (32) to the products in the set of parentheses, gives

$$\begin{aligned}
&\operatorname{Re}\left[B([x,\xi],[x,\xi])\right] \\
&\quad\geq c_g\|x\|_H^2 - C_g\|x\|_L^2 - |\lambda|c_b\left(\frac{\varepsilon}{2}\|x\|_H^2 + \frac{1}{2\varepsilon}\|x\|_L^2\right) - c_b\|u\|_H\left(\frac{1}{2}|\xi|^2 + \frac{1}{2}\|x\|_L^2\right) \\
&\qquad -2|\lambda|c_c\|u\|_H\left(\frac{1}{2}|\xi|^2 + \frac{1}{2}\|x\|_L^2\right) - |\lambda|^2 c_c\left(\frac{\varepsilon}{2}\|x\|_H^2 + \frac{1}{2\varepsilon}\|x\|_L^2\right) \\
&\qquad -\|w\|_H\left(\frac{\varepsilon}{2}\|x\|_H^2 + \frac{1}{2\varepsilon}|\xi|^2\right) \\
&\quad= c_g - \underbrace{\left(\frac{\varepsilon}{2}(|\lambda|c_b + |\lambda|^2 c_c + \|w\|_H)\right)}_{c_1:=}\|x\|_H^2 \\
&\qquad \underbrace{-\frac{1}{2}\left(c_b\|u\|_H + 2|\lambda|c_c\|u\|_H + \frac{\|w\|_H}{\varepsilon}\right)}_{c_2:=}|\xi|^2 \\
&\qquad \underbrace{-\left(C_g + \frac{1}{2}\left(\frac{|\lambda|c_b}{\varepsilon} + c_b\|u\|_H + 2|\lambda|c_c\|u\|_H + \frac{|\lambda|^2 c_c}{\varepsilon}\right)\right)}_{c_3:=}\|x\|_L^2
\end{aligned}$$

for all $\varepsilon > 0$. Choosing ε such that $c_1 = c_g/2 > 0$, we obtain

$$\mathrm{Re}\,[B([x,\xi],[x,\xi])] \geq \frac{c_g}{2}\left(\|x\|_H^2 + |\xi|^2\right) - \left(c_2 + \frac{c_g}{2}\right)|\xi|^2 - c_3\|x\|_L^2.$$

With $c_4 := \max\{c_2 + c_g/2, c_3\}$, this implies

$$\mathrm{Re}\,[B([x,\xi],[x,\xi])] \geq \frac{c_g}{2}\left(\|x\|_H^2 + |\xi|^2\right) - c_4\left(\|x\|_L^2 + |\xi|^2\right).$$

Obviously, the norms $\left(\|x\|_H^2 + |\xi|^2\right)^{1/2}$ and $\left(\|x\|_H + |\xi|\right)$ on $\mathcal{H}\times\mathbb{C}$, and the norms $\left(\|x\|_L^2 + |\xi|^2\right)^{1/2}$ and $\left(\|x\|_L + |\xi|\right)$ on $\mathcal{L}\times\mathbb{C}$ are equivalent. Therefore, there exist constants $\tilde{c}_g, \tilde{C}_g > 0$ such that

$$\mathrm{Re}\,[B([x,\xi],[x,\xi])] \geq \tilde{c}_g\left(\|x\|_H + |\xi|\right)^2 - \tilde{C}_g\left(\|x\|_L + |\xi|\right)^2,$$

which proves the statement. □

Lemma 7 states that the sesquilinear form $B(\cdot,\cdot)$ is $\mathcal{H}\times\mathbb{C}$-coercive. In the following theorem we show that the form is also regular, if λ is a simple eigenvalue, i.e. if assumption ($\mathrm{A_S}$) holds.

Theorem 1 (Regularity) *Let the sesquilinear form B be $\mathcal{H}\times\mathbb{C}$-coercive and let assumption (A_S) be fulfilled.*
Then, B also is regular, i.e. the problem

$$B([x,\xi],[\varphi,\zeta]) = f(\varphi) + \alpha\overline{\zeta} \qquad \forall[\varphi,\zeta]\in\mathcal{H}\times\mathbb{C} \tag{140}$$

with arbitrary $f\in\mathcal{H}', \alpha\in\mathbb{C}$, is uniquely solvable. Furthermore, there exists a constant $m_c > 0$ such that

$$m_c\left(\|x\|_H + |\xi|\right) \leq \sup_{\|\varphi\|_H+|\zeta|=1} |B([x,\xi],[\varphi,\zeta])| \tag{141}$$

holds for all $[x,\xi]\in\mathcal{H}\times\mathbb{C}$.

Proof: Since B is $\mathcal{H}\times\mathbb{C}$-coercive, the Fredholm alternative applies to it, see e.g. [38] or [73]. So, the regularity of B can be shown by proving that the homogeneous problem only possesses the trivial solution.
Let $[x,\xi]$ be a solution of (140) with $f = 0$ and $\alpha = 0$. It follows, that x and ξ fulfill

$$\begin{aligned} a(x,\varphi) - \lambda\, b(x,\varphi) - \lambda^2 c(x,\varphi) - \xi\, b(u,\varphi) - 2\xi\lambda\, c(u,\varphi) &= 0 \quad \forall\varphi\in\mathcal{H} \\ \overline{\zeta}\langle x, w\rangle &= 0 \quad \forall\zeta\in\mathbb{C} \end{aligned}$$

and, consequently,

$$a(x,\varphi) - \lambda\, b(x,\varphi) - \lambda^2 c(x,\varphi) = b(\xi u,\varphi) + 2\lambda\, c(\xi u,\varphi) \quad \forall\varphi\in\mathcal{H} \tag{142}$$

$$\langle x, w\rangle = 0. \tag{143}$$

If ξ was not zero, i.e. $\xi u \in \operatorname{span}_{\mathbb{C}}\{u\} \setminus \{0\}$, equation (142) would contradict assumption $(\mathrm{A_S})$(ii).
For $\xi = 0$, assumption $(\mathrm{A_S})$(i) implies $x = \nu u$, with $\nu \in \mathbb{C}$. By (143), this yields $0 = \langle \nu u, w\rangle = \nu\langle u, w\rangle = \nu$, which means $x = 0$.
Thus, we have shown, that the homogeneous problem only possesses the trivial solution $[x, \xi] = [0, 0]$, and, hence, that the sesquilinear form is regular.
The stated inequality is a basic property of regular sesquilinear forms on Hilbert spaces, cf. [38] or [73]. □

4.3 Regularity of the Discretized Linearization

Let $\mathcal{H}_h \subset \mathcal{H}$ be finite-dimensional approximation spaces which depend on a parameter $h > 0$. We assume

$(\mathrm{A_A})$: For $h \to 0$ the spaces $\mathcal{H}_h$ approximate $\mathcal{H}$ arbitrarily well or, more precisely,

$$\lim_{h\to 0} \inf_{v_h \in \mathcal{H}_h} \|v - v_h\|_H = 0 \qquad \forall v \in \mathcal{H}. \tag{144}$$

For a standard finite element discretization the parameter h can be thought of as the mesh size h.
Assuming the validity of $(\mathrm{A_A})$, we obtain:

Theorem 2 (Regularity of Discretized Problem) *Under the conditions of Theorem 1 and assumption (A_A), there exists a constant $h_0 > 0$ such that for all $h < h_0$ the sesquilinear form B is $\mathcal{H}_h \times \mathbb{C}$-regular, i.e. the problem*

$$B([x_h, \xi_h], [\varphi_h, \zeta_h]) = f(\varphi_h) + \alpha\overline{\zeta_h} \qquad \forall [\varphi_h, \zeta_H] \in \mathcal{H}_h \times \mathbb{C} \tag{145}$$

with arbitrary $f \in \mathcal{H}', \alpha \in \mathbb{C}$, is uniquely solvable. Then, all $[x_h, \xi_h] \in \mathcal{H}_h \times \mathbb{C}$ satisfy

$$m_d\left(\|x_h\|_H + |\xi_h|\right) \le \sup_{\|\varphi_h\|_H + |\zeta_h| = 1} |B([x_h, \xi_h], [\varphi_h, \zeta_h])| \tag{146}$$

with a constant $m_d > 0$ independent of $[x_h, \xi_h]$ and h.

Proof: The statement of the theorem will be proved by contradiction in three steps. In a preparation step, an appropriate sequence $([v_j, \mu_j])_{j\ge 0}$ with limit $[v, \mu]$ is constructed. Then, it is proved that the limit is $[v, \mu] = [0, 0]$. Finally, a contradiction is derived.
For the statements from functional analysis in this proof, we refer to [4], [38], or [73].

Step I. Preparation.
Assume the contrary of the theorem. Then, there exists a sequence $(h_j)_{j\geq 0}$ with $h_j \to 0$ for $j \to \infty$, and corresponding sequences $(v_j)_{j\geq 0}$, $v_j \in \mathcal{H}_{h_j}$, and $(\mu_j)_{j\geq 0}$, $\mu_j \in \mathbb{C}$, with $\|v_j\|_H + |\mu_j| = 1$, such that

$$\sup_{\|\varphi_{h_j}\|_H+|\xi|=1} B([v_j,\mu_j],[\varphi_{h_j},\xi]) \underset{j\to\infty}{\longrightarrow} 0. \tag{147}$$

Particularly, the estimates $\|v_j\|_H \leq 1$ and $|\mu_j| \leq 1$ hold.
Due to the boundedness of the μ_j in $\mathbb{C}$, it follows, that there exists a $\mu \in \mathbb{C}$ and a subsequence of $(\mu_j)_{j\geq 0}$, which will be denoted identically, such that $\mu_j \to \mu$ for $j \to \infty$.
Furthermore, a basic lemma on bounded sequences in Hilbert spaces states that there exists a $v \in \mathcal{H}$ and a further subsequence $(v_j)_{j\geq 0}$ such that v_j converges weakly to v in $\mathcal{H}$

$$\langle v_j,\psi\rangle \underset{j\to\infty}{\longrightarrow} \langle v,\psi\rangle \quad \forall \psi \in \mathcal{H}.$$

Since the embedding $\mathcal{H} \subset \mathcal{L}$ is assumed to be compact and the sequence $(v_j)_{j\geq 0}$ is bounded, there exists a further subsequence, again denoted as $(v_j)_{j\geq 0}$, which converges strongly in $\mathcal{L}$, i.e. particularly

$$\|v_j\|_L \underset{j\to\infty}{\longrightarrow} \|v\|_L.$$

Let us consider the last sequence $([v_j,\mu_j])_{j\geq 0}$. The statements

$$\mu_j \underset{j\to\infty}{\longrightarrow} \mu \quad \text{and} \quad \langle v_j,\psi\rangle \underset{j\to\infty}{\longrightarrow} \langle v,\psi\rangle \quad \forall \psi \in \mathcal{H}.$$

imply the weak convergence

$$\langle v_j,\psi\rangle + \mu_j\overline{\eta} \underset{j\to\infty}{\longrightarrow} \langle v,\psi\rangle + \mu\overline{\eta} \quad \forall [\psi,\eta] \in \mathcal{H}\times\mathbb{C}$$

in the Hilbert space $\mathcal{H} \times \mathbb{C}$.

Step II.
In this step we show that the limit $[v,\mu]$ of the sequence $([v_j,\mu_j])_{j\geq 0}$ is equal to zero. We prove that

$$B([v,\mu],[\varphi,\zeta]) = 0 \quad \forall [\varphi,\zeta] \in \mathcal{H}\times\mathbb{C}. \tag{148}$$

Then, by Theorem 1, from equation (148), it follows that $[v,\mu] = [0,0]$.
So, let $[\varphi,\zeta] \in \mathcal{H}\times\mathbb{C}$ be such that $\|\varphi\|_H + |\zeta| = 1$. Let $\varepsilon > 0$ be given. For, at first, arbitrary $j \in \mathbb{N}$ and $[\varphi_{h_j},\zeta_j] \in \mathcal{H}_{h_j}\times\mathbb{C}$ with $\|\varphi_{h_j}\|_H + |\zeta_j| = 1$, we have

$$\begin{aligned} &|B([v,\mu],[\varphi,\zeta])| \leq \\ &\quad\leq \underbrace{|B([v-v_j,\mu-\mu_j],[\varphi,\zeta])|}_{(\mathrm{B}_1):=} + \underbrace{|B([v_j,\mu_j],[\varphi-\varphi_{h_j},\zeta-\zeta_j])|}_{(\mathrm{B}_2):=} + \underbrace{|B([v_j,\mu_j],[\varphi_{h_j},\zeta_j])|}_{(\mathrm{B}_3):=} \end{aligned}$$

Step IIa.
The limit (147) yields that there exists a j_3 with

$$(\mathrm{B}_3) = |B([v_j, \mu_j], [\varphi_{h_j}, \zeta_j])| < \varepsilon/3 \quad \forall j \geq j_3$$

for arbitrary $[\varphi_{h_j}, \zeta_j] \in \mathcal{H}_{h_j} \times \mathbb{C}$ with $\|\varphi_{h_j}\|_H + |\zeta_j| = 1$.

Step IIb.
For given $0 < \varepsilon' < 1/2$, assumption $(\mathrm{A_A})$ implies the existence of a j' such that there exists a $\tilde{\varphi}_{h_j} \in \mathcal{H}_{h_j}$ with

$$\|\tilde{\varphi}_{h_j} - \varphi\|_H \leq \varepsilon'$$

for all $j \geq j'$. With $\nu_j := \|\tilde{\varphi}_{h_j}\|_H + |\zeta| \geq 1 - \varepsilon'$ we define $[\varphi_{h_j}, \zeta_j] := \frac{1}{\nu_j}[\tilde{\varphi}_{h_j}, \zeta] \in \mathcal{H}_{h_j} \times \mathbb{C}$. Then, it is $\|\varphi_{h_j}\|_H + |\zeta_j| = 1$. Trivial estimates show that

$$\|\varphi_{h_j} - \varphi\|_H + |\zeta_j - \zeta\| \leq \frac{2\varepsilon'}{1-\varepsilon'} + \frac{\varepsilon'}{1-\varepsilon'} < 6\varepsilon'$$

for all $j \geq j'$. Thus, by the continuity of B and with ε', for instance, chosen as $\varepsilon' := \min\{\frac{1}{6C_B \cdot 3}\varepsilon, \frac{1}{3}\}$, the existence of a j_2 follows such that for $j \geq j_2$ there exists a $[\varphi_{h_j}, \zeta_j] \in \mathcal{H}_{h_j} \times \mathbb{C}$ with $\|\varphi_{h_j}\|_H + |\zeta_j| = 1$ and

$$\begin{aligned}(\mathrm{B}_2) &= |B([v_j, \mu_j], [\varphi - \varphi_{h_j}, \zeta - \zeta_j])| \leq C_B(\|v_j\|_H + |\zeta_j|)(\|\varphi_{h_j} - \varphi\|_H + |\zeta_j - \zeta|) \\ &\leq C_B 6\varepsilon' \leq \varepsilon/3.\end{aligned}$$

Step IIc.
Since

$$T : \mathcal{H} \times \mathbb{C} \to \mathbb{C},\ T[x, \xi] := B([x, \xi], [\varphi, \zeta])$$

is a bounded linear functional on the Hilbert space $\mathcal{H} \times \mathbb{C}$, by the representation theorem of Riesz, the existence of a $[\psi_T, \eta_T] \in \mathcal{H} \times \mathbb{C}$ follows which satisfies

$$T[x, \xi] = \langle x, \psi_T \rangle + \xi \overline{\eta}_T$$

for all $[x, \xi] \in \mathcal{H} \times \mathbb{C}$. The weak convergence of $([v_j, \mu_j])_{j \geq 0}$ implies

$$B([v_j, \mu_j], [\varphi, \zeta]) = \langle v_j, \psi_T \rangle + \mu_j \overline{\eta}_T \underset{j \to \infty}{\longrightarrow} \langle v, \psi_T \rangle + \mu \overline{\eta}_T = B([v, \mu], [\varphi, \zeta]).$$

In other words, there exists a j_1 such that for all $j \geq j_1$ we have

$$(\mathrm{B}_1) = |B([v - v_j, \mu - \mu_j], [\varphi, \zeta])| \leq \varepsilon/3.$$

Step IId.
Combining this three estimates, we obtain that for every $\varepsilon > 0$ there exists a $j_0 := \max\{j_1, j_2, j_3\}$ such that for all $j \geq j_0$ there exists a $[\varphi_{h_j}, \zeta_j]$ with $\|\varphi_{h_j}\|_H + |\zeta_j| = 1$ and

$$|B([v, \mu], [\varphi, \zeta])| \leq (\mathrm{B}_1) + (\mathrm{B}_2) + (\mathrm{B}_3) \leq \varepsilon/3 + \varepsilon/3 + \varepsilon/3 = \varepsilon.$$

This particularly implies, that

$$B([v,\mu],[\varphi,\zeta]) = 0 \tag{149}$$

for $[\varphi,\zeta] \in \mathcal{H}\times\mathbb{C}$ with $\|\varphi\|_H + |\zeta| = 1$. From the anti-linearity of the sesquilinear form B in the second argument $[\varphi,\zeta]$ the statement (148) follows.

Step III. Contradiction.
However, the Gårding inequality (139)

$$0 < \tilde{c_g} = \tilde{c_g}\left(\|v_j\|_H + |\mu_j|\right)^2 \le |B([v_j,\mu_j],[v_j,\mu_j])| + \tilde{C}_K\left(\|v_j\|_L + |\mu_j|\right)^2$$

and the limit $\|v_j\|_L + |\mu_j| \to \|v\|_L + |\mu| = 0$ for $j\to\infty$ yield, that for large j

$$|B([v_j,\mu_j],[v_j,\mu_j])| \ge \tilde{c}_g/2 > 0.$$

This is a contradiction to the assumption

$$\sup_{\|\varphi_h\|_H+|\xi|=1} B([v_j,\mu_j],[\varphi_h,\xi]) \underset{j\to\infty}{\longrightarrow} 0.$$

So, the statement of the theorem is proved. □

4.4 A Parameter Dependent Perturbation of the Discretized Eigenvalue Problem

Let us, now, consider a discretized and perturbed version of the eigenvalue problem (131), (132). We search for an eigenpair $[v_h,\mu_h] \in \mathcal{H}_h \times \mathbb{C}$ that satisfies

$$\begin{aligned} &a(v_h,\varphi_h) + a'_\rho(v_h,\varphi_h) = \\ &\quad \mu_h\Big(b(v_h,\varphi_h) + b'_\rho(v_h,\varphi_h)\Big) + \mu_h^2\Big(c(v_h,\varphi_h) + c'_\rho(v_h,\varphi_h)\Big) \qquad \forall\varphi_h \in \mathcal{H}_h \end{aligned} \tag{150}$$

under the restriction

$$\langle v_h, w\rangle = 1, \tag{151}$$

where the sesquilinear forms $a'_\rho(\cdot,\cdot), b'_\rho(\cdot,\cdot)$, and $c'_\rho(\cdot,\cdot)$, describing the perturbations, depend on a parameter ρ. For these perturbations we assume that there exist constants $C_a, C_b, C_c > 0$ such that

$$\begin{aligned} a'_\rho(v_1,v_2) &\le \rho\, C_a \|v_1\|_H \|v_2\|_H, \\ b'_\rho(v_1,v_2) &\le \rho\, C_b \|v_1\|_H \|v_2\|_H, \\ \text{and}\quad c'_\rho(v_1,v_2) &\le \rho\, C_c \|v_1\|_H \|v_2\|_H. \end{aligned} \tag{152}$$

Corollary 1 (Fixed Point Characterization) *Let the parameter h satisfy $h < h_0$, where h_0 is the constant of Theorem 2.*
Then, for every $[\tilde{v}_h, \tilde{\mu}_h] \in \mathcal{H}_h \times \mathbb{C}$ there exists a unique $[\hat{v}_h, \hat{\mu}_h] \in \mathcal{H}_h \times \mathbb{C}$ such that

$$\begin{aligned} &B([\hat{v}_h - u, \hat{\mu}_h - \lambda], [\varphi_h, \zeta_h]) = \\ &\quad (\tilde{\mu}_h - \lambda)\, b(\tilde{v}_h - u, \varphi_h) + (\tilde{\mu}_h - \lambda)(\tilde{\mu}_h + \lambda)\, c(\tilde{v}_h - u, \varphi_h) + (\tilde{\mu}_h - \lambda)^2 c(u, \varphi_h) \\ &\quad -a'_\rho(\tilde{v}_h, \varphi_h) + \tilde{\mu}_h\, b'_\rho(\tilde{v_h}, \varphi_h) + \tilde{\mu}_h^2\, c'_\rho(\tilde{v}_h, \varphi_h) =: \tilde{f}_{\tilde{v}_h, \tilde{\mu}_h}(\varphi_h) \end{aligned} \tag{153}$$

for all $[\varphi_h, \zeta_h] \in \mathcal{H}_h \times \mathbb{C}$.
This defines a continuous (non-linear) operator $\tilde{\mathcal{R}}_h : \mathcal{H}_h \times \mathbb{C} \to \mathcal{H}_h \times \mathbb{C}$ via

$$\tilde{\mathcal{R}}_h[\tilde{v}_h, \tilde{\mu}_h] := [\hat{v}_h, \hat{\mu}_h]. \tag{154}$$

Furthermore, $[v_h, \mu_h]$ is an eigenpair of the discretized and perturbed problem (150), (151), if and only if $[v_h, \mu_h]$ is a fixed point of the operator $\tilde{\mathcal{R}}_h$, i.e. if

$$\tilde{\mathcal{R}}_h[v_h, \mu_h] = [v_h, \mu_h].$$

Proof: Obviously, it is $\tilde{f}_{\tilde{v}_h, \tilde{\mu}_h} \in \mathcal{H}'$ for fixed $[\tilde{v}_h, \tilde{\mu}_h] \in \mathcal{H}_h \times \mathbb{C}$. Then, in

$$B([\hat{v}_h, \hat{\mu}_h], [\varphi_h, \zeta_h]) = \tilde{f}_{\tilde{v}_h, \tilde{\mu}_h}(\varphi_h) + B([u, \lambda], [\varphi_h, \zeta_h]) \tag{155}$$

the right-hand side is an anti-linear functional on $\mathcal{H}_h \times \mathbb{C}$. So, Theorem 2 yields that the operator $\tilde{\mathcal{R}}_h$ is well-defined.
The fixed point relation is proved by a simple computation using (137) with $[v, \mu] = [v_h, \mu_h]$ and equations (150), (151). □

4.5 Existence and Convergence of a Discrete Eigensolution

The last abstract assumption is, that the eigensolution u and the approximation spaces $\mathcal{H}_h \subset \mathcal{H}$ are such that the following statement holds:

(A$_\text{I}$): There exists a constant $C_I > 0$ such that

$$\|u - \mathcal{I}_h u\|_H \leq C_I h,$$

where $\mathcal{I}_h : \mathcal{H} \to \mathcal{H}_h$ denotes an appropriate interpolation (or approximation) operator.

Now, we can state

Theorem 3 (Convergence of an Approximate Eigenpair) *Let the conditions of Theorem 2, inequalities (152) and assumption (A_I) be satisfied. Let, furthermore, the perturbation parameter be of the form $\rho = ch$.*
Then, there exists a constant $h_1 > 0$ such that for all $h < h_1$ the problem (150), (151) possesses a solution $[u_h, \lambda_h] \in \mathcal{H}_h \times \mathbb{C}$ for which

$$\|u - u_h\|_H + |\lambda - \lambda_h| \leq c_e h \tag{156}$$

holds with a constant $c_e > 0$ independent of h.

Proof: *Step I. An inequality for the error.*
By the triangle inequality, we obtain

$$\|\hat{v}_h - u\|_H + |\hat{\mu}_h - \lambda| \leq \|\hat{v}_h - \mathcal{I}_h u\|_H + \|\mathcal{I}_h u - u\|_H + |\hat{\mu}_h - \lambda|.$$

Let $\|\varphi_h\| + |\zeta_h| = 1$. Then, by the continuity of B, relation (153), and by conditions (152), we get

$$\begin{aligned}
&|B([\hat{v}_h - \mathcal{I}_h u, \hat{\mu}_h - \lambda], [\varphi_h, \zeta_h])| \\
&\quad\leq |B([u - \mathcal{I}_h u, 0], [\varphi_h, \zeta_h])| + |B([\hat{v}_h - u, \hat{\mu}_h - \lambda], [\varphi_h, \zeta_h])| \\
&\quad\leq C\|u - \mathcal{I}_h u\|_H + |\tilde{f}(\varphi_h)| \\
&\quad\leq C\|u - \mathcal{I}_h u\|_H + |\tilde{\mu}_h - \lambda|\, C\|\tilde{v}_h - u\|_H\|\varphi_h\|_H \\
&\qquad + |\tilde{\mu}_h - \lambda|\, |\tilde{\mu}_h + \lambda|\, C\|\tilde{v}_h - u\|_H\|\varphi_h\|_H + |\tilde{\mu}_h - \lambda|^2 C\|u\|_H\|\varphi_h\|_H \\
&\qquad + \rho\, C_a\|\tilde{v}_h - u\|_H\|\varphi_h\|_H + \rho\, C_b\|\tilde{v}_h - u\|_H\|\varphi_h\|_H + \rho\, C_c\|\tilde{v}_h - u\|_H\|\varphi_h\|_H \\
&\qquad + \rho\, C_a\|u\|_H\|\varphi_h\|_H + \rho\, C_b\|u\|_H\|\varphi_h\|_H + \rho\, C_c\|u\|_H\|\varphi_h\|_H \\
&\quad\leq C\|u - \mathcal{I}_h u\|_H + |\tilde{\mu}_h - \lambda|\, C\|\tilde{v}_h - u\|_H + |\tilde{\mu}_h - \lambda|\, |\tilde{\mu}_h + \lambda|\, C\|\tilde{v}_h - u\|_H \\
&\qquad + |\tilde{\mu}_h - \lambda|^2 C\|u\|_H + \rho(C_a + C_b + C_c)\|\tilde{v}_h - u\|_H + \rho(C_a + C_b + C_c)\|u\|_H
\end{aligned}$$

with C being a positive generic constant.
Assuming that $h < h_0$, from the $\mathcal{H}_h \times \mathbb{C}$-regularity of B, particularly from inequality (146) in Theorem 2, we obtain

$$\begin{aligned}
&\|\hat{v}_h - \mathcal{I}_h u\|_H + |\hat{\mu}_h - \lambda| \qquad (157)\\
&\quad\leq C\|u - \mathcal{I}_h u\|_H + |\tilde{\mu}_h - \lambda|\, C\|\tilde{v}_h - u\|_H + |\tilde{\mu}_h - \lambda|\, |\tilde{\mu}_h + \lambda|\, C\|\tilde{v}_h - u\|_H \\
&\qquad + |\tilde{\mu}_h - \lambda|^2 C\|u\|_H + \rho\, C\|\tilde{v}_h - u\|_H + \rho\, C\|u\|_H.
\end{aligned}$$

Step II. Substituting the discretization parameter h.
Choosing $\rho = ch$, using condition (A_I), and the inequalities of *Step I*, it follows for $\|\tilde{v}_h - u\|_H + |\tilde{\mu}_h - \lambda| \leq 2c_1 h$ that

$$\begin{aligned}
&\|\hat{v}_h - u\|_H + |\hat{\mu}_h - \lambda| \qquad (158)\\
&\quad\leq C_1 C_I h + 4C_2 c_1^2 h^2 + 4C_3\, 2(c_1 + |\lambda|)c_1^2 h^2 + 4C_4 c_1^2 h^2 + C_5 2c_1 h^2 + C_6 h.
\end{aligned}$$

To derive the third term of the right-hand side of the above inequality, amongst other things, following estimate has been used

$$|\tilde{\mu}_h + \lambda| \leq |\tilde{\mu}_h - \lambda| + 2|\lambda| \leq 2(c_1 + |\lambda|).$$

Let $c_1 := C_1 C_I + C_6$ and $\tilde{h}_1 := c_1 \cdot (4C_2 c_1^2 + 8C_3(c_1 + |\lambda|)c_1^2 + 4C_4 c_1^2 + C_5 2c_1)^{-1}$. Then, for $h < h_1 := \min\{h_0, \tilde{h}_1\}$, it is

$$4C_2 c_1^2 h^2 + 8C_3(c_1 + |\lambda|)c_1^2 h^2 + 4C_4 c_1^2 h^2 + C_5 2c_1 h^2 \leq c_1 h,$$

and, consequently,

$$\|\hat{v}_h - u\|_H + |\hat{\mu}_h - \lambda| \leq 2c_1 h.$$

Thus, we have proved, that for $h < h_1$ the continuous operator $\tilde{\mathcal{R}}_h$ maps the closed ball

$$\mathcal{B} := \{[v_h, \mu_h] \mid \|v_h - u\|_H + |\mu_h - \lambda| \leq 2c_1 h\} \subset \mathcal{H}_h \times \mathbb{C}$$

into itself.
The application of the fixed point theorem of Brouwer (see e.g. [32]) yields the existence of a fixed point $[u_h, \lambda_h] \in \mathcal{B}$. This fixed point is a solution of the approximated eigenvalue problem, as stated in Corollary 1, and, obviously, satisfies the error inequality (156) with constant $c_e := 2c_1$. □

4.6 Uniqueness of the Discrete Eigensolution

Theorem 4 (Uniqueness of Approximate Eigenpair) *Under the conditions of Theorem 3, there exists a constant $h_2 > 0$ such that for $h < h_2$ the problem (150), (151) possesses a unique solution $[u_h, \lambda_h]$ which satisfies*

$$\|u - u_h\|_H + |\lambda - \lambda_h| \leq c_e h$$

with a constant $c_e > 0$ independent of h.

Proof: Let $h < h_1 \leq h_0$ (see proof of Theorem 3) and let $[u_h, \mu_h]$ and $[u'_h, \mu'_h]$ be two solutions of problem (150), (151) that satisfy inequality (156). Then, $[u_h, \mu_h]$ and $[u'_h, \mu'_h]$ are fixed points of the operator $\tilde{\mathcal{R}}_h$, and from relation (155), it follows

$$\begin{aligned}
B([u_h - u'_h, \mu_h - \mu'_h], [\varphi_h, \zeta_h]) &= \tilde{f}_{u_h,\mu_h}(\varphi_h) - \tilde{f}_{u'_h,\mu'_h}(\varphi_h) \\
&= (\mu_h - \lambda)\, b(u_h - u, \varphi_h) + (\mu_h - \lambda)\,(\mu_h + \lambda)\, c(u_h - u, \varphi_h) \\
&\quad + (\mu_h - \lambda)^2 c(u, \varphi_h) \\
&\quad - a'_\rho(u_h, \varphi_h) + \mu_h\, b'_\rho(u_h, \varphi_h) + \mu_h^2\, c'_\rho(u_h, \varphi_h) \\
&\quad - (\mu'_h - \lambda)\, b(u'_h - u, \varphi_h) - (\mu'_h - \lambda)\,(\mu'_h + \lambda)\, c(u'_h - u, \varphi_h) \\
&\quad - (\mu'_h - \lambda)^2 c(u, \varphi_h)
\end{aligned}$$

$$
\begin{aligned}
& +a'_\rho(u'_h,\varphi_h) - \mu'_h\, b'_\rho(u'_h,\varphi_h) - (\mu'_h)^2 c'_\rho(u'_h,\varphi_h) \\
= \;& (\mu_h-\lambda)\, b(u_h-u'_h,\varphi_h) + (\mu_h-\mu'_h)\, b(u'_h-u,\varphi_h) \\
& +(\mu_h-\lambda)\,(\mu_h+\lambda)\, c(u_h-u'_h,\varphi_h) + (\mu_h+\mu'_h)\,(\mu_h-\mu'_h)\, c(u'_h-u,\varphi_h) \\
& +(\mu_h-\mu'_h)\,((\mu_h-\lambda)+(\mu'_h-\lambda))\, c(u,\varphi_h) \\
& +a'_\rho(u'_h-u_h,\varphi_h) \\
& +(\mu_h-\mu'_h)\, b'_\rho(u_h,\varphi_h) + \mu'_h\, b'_\rho(u_h-u'_h,\varphi_h) \\
& +(\mu_h-\mu'_h)\,(\mu_h+\mu'_h)\, c'_\rho(u_h,\varphi_h) + (\mu'_h)^2 c'_\rho(u_h-u'_h,\varphi_h).
\end{aligned}
$$

Using inequality (156), we obtain for $\|\varphi_h\| + |\zeta_h| = 1$

$$
\begin{aligned}
& |B([u_h-u'_h,\mu_h-\mu'_h],[\varphi_h,\zeta_h])| \\
\leq\; & Ch\|u_h-u'_h\|_H + Ch|\mu_h-\mu'_h| + 2(c_e+|\lambda|)Ch\|u_h-u'_h\|_H + 2(c_e+|\lambda|)Ch|\mu_h-\mu'_h| \\
& +Ch|\mu_h-\mu'_h| + Ch\|u_h-u'_h\|_H + Ch|\mu_h-\mu'_h| + (c_e+|\lambda|)Ch\|u_h-u'_h\|_H \\
& +2(c_e+|\lambda|)Ch|\mu_h-\mu'_h| + 2(c_e+|\lambda|)^2 Ch\|u_h-u'_h\|_H \\
\leq\; & Ch(\|u_h-u'_h\|_H + |\mu_h-\mu'_h|)
\end{aligned}
$$

with a generic constant $C > 0$ independent of h. By Theorem 2, this gives the estimate

$$
\begin{aligned}
m_d\,(\|u_h-u'_h\|_H + |\mu_h-\mu'_h|) &\leq \sup_{\|\varphi_h\|_H+|\zeta_h|=1} |B([u_h-u'_h,\mu_h-\mu'_h],[\varphi_h,\zeta_h])| \\
&\leq Ch(\|u_h-u'_h\|_H + |\mu_h-\mu'_h|)
\end{aligned}
$$

with $m_d, C > 0$ independent of h. Then, for h so small that

$$
Ch \leq \frac{m_d}{2},
$$

we obtain

$$
m_d\,(\|u_h-u'_h\|_H + |\mu_h-\mu'_h|) \leq \frac{m_d}{2}\,(\|u_h-u'_h\|_H + |\mu_h-\mu'_h|)\,,
$$

which implies

$$
\|u_h-u'_h\|_H + |\mu_h-\mu'_h| = 0.
$$

Thus, we have proved that the eigenpair is unique for $h < h_2 := \min\{h_1, m_d/(2C)\}$. □

In this chapter, we have developed a convergence theory for the discretization of quadratic eigenvalue problems.
Existence and uniqueness of the eigensolution $[u_h, \lambda_h]$ of the discretized and perturbed eigenvalue problem have been proved for a sufficiently small discretization parameter h under following conditions

- the eigenvalue of interest λ (with eigensolution u) is algebraically and geometrically simple, see assumption ($\mathrm{A_S}$),
- the approximation assumptions ($\mathrm{A_A}$) and ($\mathrm{A_I}$) hold,
- and the perturbation is of order h.

The convergence follows from the error bound

$$\|u - u_h\|_H + |\lambda - \lambda_h| \leq ch.$$

5 Discretization of the Two-Wave Eigenvalue Problem by Finite Elements

In this chapter, we present a variational formulation of the two-wave eigenvalue problem derived in Chapter 3.3, describe its discretization by finite elements, and prove convergence of the finite element eigensolution using the abstract convergence theory of Chapter 4.

5.1 A Variational Formulation of the Two-Wave Eigenvalue Problem

Let us recall the strongly formulated two-wave eigenvalue problem. One searches for complex eigenvalues ξ and eigensolutions (u_r, u_l) in proper function spaces such that

$$\begin{aligned} -\Delta u_r + 2ik_f \frac{\partial u_r}{\partial z} + (k_f^2 - k^2)u_r &= \xi u_r \quad \text{and} \\ -\Delta u_l - 2ik_f \frac{\partial u_l}{\partial z} + (k_f^2 - k^2)u_l &= \xi u_l \quad \text{in } \Omega \end{aligned} \tag{159}$$

with boundary conditions

$$\begin{aligned} u_r - \phi_0 u_l &= 0 \quad \text{on } \Gamma_0, \\ u_r - \bar{\phi}_1 u_l &= 0 \quad \text{on } \Gamma_1, \\ \frac{\partial u_r}{\partial z} + \phi_0 \frac{\partial u_l}{\partial z} &= 0 \quad \text{on } \Gamma_0, \\ \frac{\partial u_r}{\partial z} + \bar{\phi}_1 \frac{\partial u_l}{\partial z} &= 0 \quad \text{on } \Gamma_1, \\ \frac{\partial u_r}{\partial \vec{n}} - iC_b u_r &= 0 \quad \text{on } \Gamma_r, \\ \frac{\partial u_l}{\partial \vec{n}} - iC_b u_l &= 0 \quad \text{on } \Gamma_r, \end{aligned} \tag{160}$$

where the domain

$$\Omega :=]-W/2; W/2[\times]-W/2; W/2[\times]0; L[$$

and the boundaries

$$\begin{aligned} \Gamma_0 &:= \{(x, y, 0) \in \mathbb{R}^3 \mid \max\{|x|, |y|\} \leq W/2\} \subset \partial\Omega, \\ \Gamma_1 &:= \{(x, y, L) \in \mathbb{R}^3 \mid \max\{|x|, |y|\} \leq W/2\} \subset \partial\Omega, \\ \Gamma_r &:= \partial\Omega \setminus (\Gamma_0 \cup \Gamma_1), \end{aligned}$$

are defined as in Chapter 3.3 (see Figure 10).

For a weak formulation of (159), (160), appropriate spaces have to be chosen. For this purpose, we utilize the standard Lebesgue and Sobolev spaces $L_2(\Omega)$, $L_\infty(\Omega)$, $H^1(\Omega)$, and $H^2(\Omega)$, respectively, equipped with the standard norms. For details of these spaces, see e.g. [3], [38], or [73].
Furthermore, we assume that the complex-valued coefficient functions satisfy

$$k^2(x,y,z) \in L_\infty(\Omega) \quad \text{and} \quad \phi_0(x,y,z), \phi_1(x,y,z) \in C(\partial\Omega).$$

We define the product space

$$\mathcal{H} := \mathcal{H}(\Omega) := \tag{161}$$
$$\left\{ (v_r, v_l) \in (H^1(\Omega) \times H^1(\Omega)) \;\middle|\; v_r\Big|_{\Gamma_0} - v_l\Big|_{\Gamma_0} \cdot \phi_0 = 0 \text{ and } v_r\Big|_{\Gamma_1} - v_l\Big|_{\Gamma_1} \cdot \bar{\phi_1} = 0 \right\}$$

with inner product and norm

$$\langle (v_r, v_l), (\varphi_r, \varphi_l) \rangle := \int_\Omega \nabla v_r \nabla \overline{\varphi}_r + \nabla v_l \nabla \overline{\varphi}_l + v_r \overline{\varphi}_r + v_l \overline{\varphi}_l \; d(x,y,z)$$

and

$$\|(v_r, v_l)\|_H^2 := \langle (v_r, v_l), (v_r, v_l) \rangle,$$

respectively. Furthermore, we use the space

$$\mathcal{L} := L_2(\Omega) \times L_2(\Omega)$$

with inner product

$$((v_r, v_l), (\varphi_r, \varphi_l)) := \int_\Omega v_r \overline{\varphi}_r + v_l \overline{\varphi}_l \; d(x,y,z)$$

and norm

$$\|(v_r, v_l)\|_L^2 := ((v_r, v_l), (v_r, v_l)).$$

At last, we define the sesquilinear forms

$$a((v_r, v_l), (\varphi_r, \varphi_l)) :=$$
$$\int_\Omega \nabla v_r \nabla \overline{\varphi}_r + 2ik_f \frac{\partial}{\partial z} v_r \overline{\varphi}_r + (k_f^2 - k^2) v_r \overline{\varphi}_r \; d(x,y,z) - iC_b \int_{\Gamma_r} v_r \overline{\varphi}_r \; d\sigma(x,y,z)$$
$$+ \int_\Omega \nabla v_l \nabla \overline{\varphi}_l - 2ik_f \frac{\partial}{\partial z} v_l \overline{\varphi}_l + (k_f^2 - k^2) v_l \overline{\varphi}_l \; d(x,y,z) - iC_b \int_{\Gamma_r} v_l \overline{\varphi}_l \; d\sigma(x,y,z)$$

and

$$b((v_r, v_l), (\varphi_r, \varphi_l)) := ((v_r, v_l), (\varphi_r, \varphi_l)) = \int_\Omega v_r \overline{\varphi}_r \; d(x,y,z) + \int_\Omega v_l \overline{\varphi}_l \; d(x,y,z).$$

Then, a variational formulation of (159), (160) reads: Find $(u_r, u_l) \in \mathcal{H}$ and $\xi \in \mathbb{C}$ such that

$$a((u_r, u_l), (\varphi_r, \varphi_l)) = \xi\, b((u_r, u_l), (\varphi_r, \varphi_l)) \quad \forall\, (\varphi_r, \varphi_l) \in \mathcal{H}. \tag{162}$$

In the formulation (162), the first order coupling conditions in (160) on the boundaries Γ_0 and Γ_1 have been incorporated in a weak sense. They coincide with the so-called natural boundary conditions (see e.g. [38]) for problem (162).
Since we want to apply the convergence theory of Chapter 4 to the linear two-wave eigenvalue problem, we choose the sesquilinear forms $c(v, \varphi)$ and $c'_\rho(v, \varphi)$ to vanish for all $v, \varphi \in \mathcal{H}$.

5.2 Stabilized Finite Element Discretization

The aim is, to solve the eigenvalue problem (162) by a finite element method. More precisely, we want to use trilinear finite elements, see e.g. [37] or [63].
So, let $S_{h',h}$ be the space of trilinear finite element functions on a partition of the domain Ω into cuboids with length h' in (x, y)-direction and h in z-direction, see Figure 14. Here, h and h' have to be such that there exist $m, m' \in \mathbb{N}$ with $L = m \cdot h$ and $W = m' \cdot h'$, respectively.

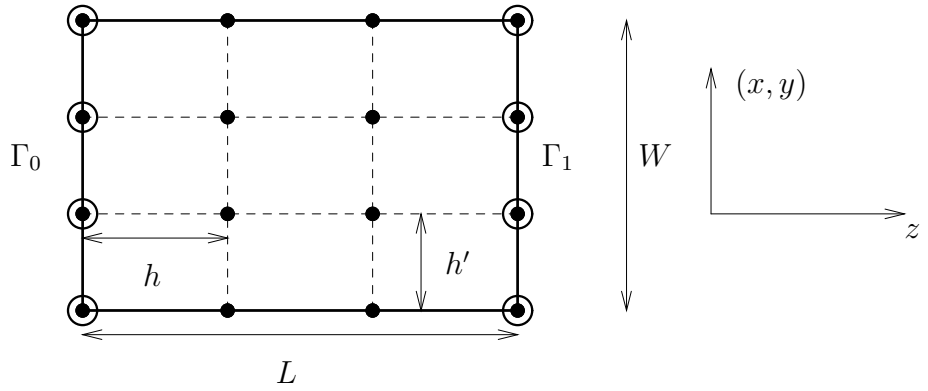

Figure 14: Decomposition.

We approximate $\mathcal{H}$ by

$$\mathcal{H}_{h',h} := \left\{ (V_r, V_l) \in S_{h',h} \times S_{h',h} \;\middle|\; V_r - \phi_0 \cdot V_l\Big|_{\Gamma_{0,h'}} = 0 \text{ and } V_r - \bar{\phi_1} \cdot V_l\Big|_{\Gamma_{1,h'}} = 0 \right\}, \tag{163}$$

where the expressions

$$V_r - \phi_0 \cdot V_l\Big|_{\Gamma_{0,h'}} = 0 \quad \text{and} \quad V_r - \bar{\phi_1} \cdot V_l\Big|_{\Gamma_{1,h'}} = 0$$

are to be understood in the sense that the equations are satisfied on the nodes of the corresponding boundaries.

For general $\phi_0, \phi_1 \in C(\partial\Omega)$, this defines a non-conform approximation, in the sense that $\mathcal{H}_{h',h} \not\subset \mathcal{H}$.
Using the space $\mathcal{H}_{h',h}$, we obtain a discretization of (162): Find $(U_r, U_l) \in \mathcal{H}_{h',h}$ and $\xi \in \mathbb{C}$ such that

$$a((U_r, U_l), (\varphi_r, \varphi_l)) = \xi\, b((U_r, U_l), (\varphi_r, \varphi_l)) \quad \forall\, (\varphi_r, \varphi_l) \in \mathcal{H}_{h',h}. \tag{164}$$

As it has been remarked in Chapter 2.1, in the case of real valued functions and real coefficients, equations of the form (159) are called convection-dominated if k_f is large. It is well known for these equations, that a standard finite element discretization suffers from stability problems. That means that a standard discretization of the large first order terms can lead to solutions which exhibit strange oscillations. To avoid this, a very fine mesh has to be used or a stabilization has to be applied, see e.g. [45], [53], and [59].
For equation (159) it is not known, whether the same problem arises (see remarks in Chapter 2.1). But there are several reason, why a stabilization of (159) – similar to the real case – should be applied nevertheless. We mention them, without giving detailed arguments. As the remark on the multiplicity of eigenvalues of the two-wave problem at the end of Chapter 3.3 shows, there can be a smooth and a very oscillatory eigensolution with nearly the same or the same eigenvalue. By penalizing the oscillatory solution one can extract the smooth eigensolution. Second, the applied stabilization, as derived below, improves the so-called condition of the discrete equation; this is explained in [9]. Furthermore, due to the large first order terms $+2ik_f\frac{\partial}{\partial z}u_r$ and $-2ik_f\frac{\partial}{\partial z}u_l$ a solver which is based on local relaxation for a standard finite element discretization of problem (162) can suffer from stability problems, which we encountered in our numerical computations.
Therefore, we stabilize equation (162) by applying a technique similar to the streamline diffusion approach (which is also called streamline upwind Petrov Galerkin discretization). For more details of the application of this approach to convection-dominated equations, see the references cited above or reference [37]. We explain the idea by considering the example

$$-\Delta u + \beta\frac{\partial}{\partial z}u + \gamma u = f \tag{165}$$

with $\beta, \gamma \in L_\infty$ and $f \in L_2$. Weakly formulated, we search for $u \in H^1$, such that

$$\int_\Omega \nabla u \nabla\overline{\varphi} + \beta\frac{\partial}{\partial z}u\overline{\varphi} + \gamma u\overline{\varphi}\; d(x,y,z) = \int_\Omega f\overline{\varphi}\; d(x,y,z) \quad \forall\varphi \in H^1. \tag{166}$$

Let $\mathcal{P}$ be a partition of Ω into cuboids, see for instance Figure 14. If we assume that a solution u of (166) exists and is even in $H^2(\Omega)$, then on every cuboid C of the partition $\mathcal{P}$ the function u satisfies

$$-\Delta u + \beta\frac{\partial}{\partial z}u + \gamma u\Big|_C = f\Big|_C \tag{167}$$

almost everywhere.
Multiplication by $\tau\overline{\tilde{\beta}\frac{\partial}{\partial z}\varphi}$, where $\tilde{\beta} \in \mathbb{C}$ and $\varphi \in H^1(\Omega)$, and integration on C leads to

$$\tau \int_C -\Delta u \,\overline{\tilde{\beta}\frac{\partial}{\partial z}\varphi} + \beta\frac{\partial}{\partial z}u\,\overline{\tilde{\beta}\frac{\partial}{\partial z}\varphi} + \gamma u\,\overline{\tilde{\beta}\frac{\partial}{\partial z}\varphi}\; d(x,y,z) = \tau \int_C f\,\overline{\tilde{\beta}\frac{\partial}{\partial z}\varphi}\; d(x,y,z). \tag{168}$$

The real quantity $\tau \geq 0$ is called stabilization parameter.
Thus, adding equation (168) with integration domain C replaced by Ω to the weak problem (166), is a consistent transformation.
If we approximate u by a continuous function U which is trilinear on every cuboid $C \in \mathcal{P}$, the first term of the integrand on the left-hand side in equation (168) vanishes. Thus, it has not to be considered in the discretization, and it, therefore, suffices to add the equation

$$\tau\overline{\tilde{\beta}} \int_\Omega \beta\frac{\partial}{\partial z}U\frac{\partial}{\partial z}\overline{\varphi} + \gamma U\frac{\partial}{\partial z}\overline{\varphi}\; d(x,y,z) = \tau\overline{\tilde{\beta}} \int_\Omega f\frac{\partial}{\partial z}\overline{\varphi}\; d(x,y,z) \tag{169}$$

to equation (164).
This modification is applied singly to the two waves u_r and u_l choosing $\tilde{\beta} = 2k_f$ for u_r and $\tilde{\beta} = -2k_f$ for u_l. Thus, the problem (164) is "consistently" transformed into

$$\begin{aligned} &a((U_r,U_l),(\varphi_r,\varphi_l)) + a'_\tau((U_r,U_l),(\varphi_r,\varphi_l)) = \\ &\quad \xi\left(b((U_r,U_l),(\varphi_r,\varphi_l)) + b'_\tau((U_r,U_l),(\varphi_r,\varphi_l))\right) \quad \forall\,(\varphi_r,\varphi_l) \in \mathcal{H}_{h',h}. \end{aligned} \tag{170}$$

with

$$\begin{aligned} &a'_\tau((U_r,U_l),(\varphi_r,\varphi_l)) := \\ &\quad \tau\,2k_f \int_\Omega 2ik_f\frac{\partial}{\partial z}U_r\frac{\partial}{\partial z}\bar{\varphi}_r + (k_f^2-k^2)U_r\frac{\partial}{\partial z}\bar{\varphi}_r\; d(x,y,z) \\ &\quad -\tau\,2k_f \int_\Omega -2ik_f\frac{\partial}{\partial z}U_l\frac{\partial}{\partial z}\bar{\varphi}_l + (k_f^2-k^2)U_l\frac{\partial}{\partial z}\bar{\varphi}_l\; d(x,y,z) \end{aligned}$$

and with

$$b'_\tau((U_r,U_l),(\varphi_r,\varphi_l)) := \tau\,2k_f \int_\Omega U_r\frac{\partial}{\partial z}\bar{\varphi}_r - U_l\frac{\partial}{\partial z}\bar{\varphi}_l\; d(x,y,z).$$

In our computations, the stabilization parameter takes the form

$$\tau = \frac{c_s}{k_f}h, \tag{171}$$

where h is the mesh size in z-direction and $c_s \geq 0$ is an appropriately chosen constant.

5.3 Convergence Proof for the Finite Element Solution

To analyze the eigenvalue problem in the framework of Chapter 4, we use the trilinear finite element space S_h based on the decomposition of Ω depicted in Figure 15, where for ease of description the width W and the length L are assumed to satisfy $W/L \in \mathbb{Q}_+$. For h with $0 < h < \min\{L, W\}$ and $W/h, L/h \in \mathbb{N}$ we define

$$\mathcal{H}_h := \left\{ (V_r, V_l) \in S_h \times S_h \;\middle|\; V_r - \phi_0 V_l\Big|_{\Gamma_{0,h}} = 0 \text{ and } V_r - \bar{\phi_1} V_l\Big|_{\Gamma_{1,h}} = 0 \right\}, \quad (172)$$

where the expressions

$$V_r - \phi_0 V_l\Big|_{\Gamma_{0,h}} = 0 \quad \text{and} \quad V_r - \bar{\phi_1} V_l\Big|_{\Gamma_{1,h}} = 0$$

mean that the equations are satisfied at the nodes lying on the corresponding boundaries; see also definition (163).

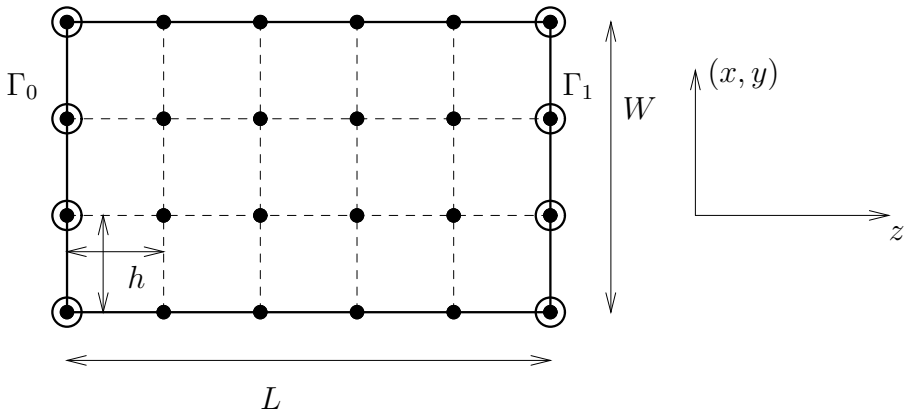

Figure 15: Decomposition for Convergence Proof.

For ease of presentation, we additionally assume that

$$\phi_0 \equiv \phi_1 \equiv -1. \quad (173)$$

The choice (173) guarantees, amongst other things, that the inclusion $\mathcal{H}_h \subset \mathcal{H}$ holds.
First, we prove the Gårding inequality for $a(\cdot, \cdot)$ in $\mathcal{H}$.

Lemma 8 (Gårding inequality) *There exist constants $c_g, C_g > 0$ such that*

$$\mathrm{Re}\,[a((v_r, v_l), (v_r, v_l))] \geq c_g \|(v_r, v_l)\|_H^2 - C_g \|(v_r, v_l)\|_L^2 \quad \textit{for all } (v_r, v_l) \in \mathcal{H}.$$

PROOF: We have

$$\begin{aligned}
&\mathrm{Re}\,[a((v_r,v_l),(v_r,v_l))]\\
&= \int_\Omega |\nabla v_r|^2 + |\nabla v_l|^2 + \mathrm{Re}\left[k_f^2 - k^2\right]\left(|v_r|^2+|v_l|^2\right) + \mathrm{Re}\left[2ik_f\frac{\partial}{\partial z}v_r\overline{v_r}\right]\\
&\quad -\mathrm{Re}\left[2ik_f\frac{\partial}{\partial z}v_l\overline{v_l}\right]\; d(x,y,z)\\
&\geq \int_\Omega |\nabla v_r|^2 + |\nabla v_l|^2 - \left|k_f^2 - k^2\right|\left(|v_r|^2+|v_l|^2\right) - 2k_f\left|\frac{\partial}{\partial z}v_r\right||v_r|\\
&\quad -2k_f\left|\frac{\partial}{\partial z}v_l\right||v_l|\; d(x,y,z).
\end{aligned}$$

Using the generalized Young inequality (32), the expression $2k_f\int_\Omega \left|\frac{\partial}{\partial z}v_r\right||v_r|\,d(x,y,z)$ can be estimated from above as follows:

$$\begin{aligned}
2k_f\int_\Omega\left|\frac{\partial}{\partial z}v_r\right||v_r|\;d(x,y,z) &\leq 2k_f\|\frac{\partial}{\partial z}v_r\|_{L_2}\|v_r\|_{L_2}\\
&\leq \varepsilon k_f\|\frac{\partial}{\partial z}v_r\|_{L_2}^2 + \frac{k_f}{\varepsilon}\|v_r\|_{L_2}^2\\
&\leq \varepsilon k_f\|\nabla v_r\|_{L_2}^2 + \frac{k_f}{\varepsilon}\|v_r\|_{L_2}^2
\end{aligned}$$

for arbitrary $\varepsilon > 0$. An analogous estimate holds for v_l.
Choosing ε properly, i.e. such that $1 - k_f\varepsilon = 1/2 =: c_g$, we obtain the desired inequality

$$\begin{aligned}
&\mathrm{Re}\,[a((v_r,v_l),(v_r,v_l))]\\
&\geq (1-k_f\varepsilon)\int_\Omega |\nabla v_r|^2 + |\nabla v_l|^2\;d(x,y,z) - \|k_f^2-k^2\|_\infty\int_\Omega |v_r|^2+|v_l|^2\;d(x,y,z)\\
&\quad -\frac{k_f}{\varepsilon}(\|v_r\|_{L_2}^2 + \|v_l\|_{L_2}^2)\\
&= c_g\|(v_r,v_l)\|_H^2 - \left(c_g + \|k_f^2-k^2\|_\infty + \frac{k_f}{\varepsilon}\right)\|(v_r,v_l)\|_L^2\\
&= c_g\|(v_r,v_l)\|_H^2 - C_g\|(v_r,v_l)\|_L^2
\end{aligned}$$

with $C_g := c_g + \|k_f^2 - k^2\|_\infty + k_f/\varepsilon > 0$ and $c_g > 0$. □

In order to prove convergence of the finite element solution, we assume H^2-regularity of the eigensolution (u_r, u_l). In Chapter 5.4 an outline for a proof of this property on a slightly different domain Ω' is given.

Theorem 5 (Convergence of the Discrete Two-Wave Eigensolution)
Let $(u_r, u_l) \in \mathcal{H}$ be an eigensolution of (162) with simple eigenvalue λ, i.e. let

the eigenpair $[(u_r, u_l), \lambda]$ *fulfill assumption* (A_S) *in Chapter 4. Furthermore, let* (u_r, u_l) *be in* $H^2(\Omega) \times H^2(\Omega)$.
Then, for sufficiently small mesh sizes h *a unique finite element eigenpair*

$$[(U_r, U_l), \lambda_h] \in \mathcal{H}_h \times \mathbb{C}$$

of the discretized and stabilized equation (170) exists and converges to $[(u_r, u_l), \lambda]$.

PROOF: The statement of the theorem follows directly from Theorem 4 in Chapter 4, if the conditions of Theorem 4 hold. Here, different from the convergence theory in Chapter 4, we have a discrete sequence of approximation spaces $\mathcal{H}_h$. But reviewing the proofs, it can easily be seen that the results also hold if the discrete versions of (A_A) and (A_I) are assumed.
Since for our situation the proofs of the conditions of Theorem 4 are standard, we verify them shortly by pointing to the main ideas or giving references.

I. Continuous, compact and dense embedding $\mathcal{H} \subset \mathcal{L}$.
It is well known, that the embedding $H^1(\Omega) \subset L_2(\Omega)$ is continuous and dense. Since the domain Ω is bounded and possesses the cone property, this embedding also is compact, see for instance [3] or [38].
By basic arguments, it can be shown that the continuity, density, and compactness also holds for the embedding $\mathcal{H} \subset \mathcal{L}$.

II. Approximation property (A_A) *of* $\mathcal{H}_h$ *with respect to* $\mathcal{H}$. *Assumption* (A_I). *Discrete Versions.*
The Scott-Zhang approximation operator $\mathcal{I}_h : H^1(\Omega) \to S_h$ (see e.g. [17] or [22]) can be defined such that the approximation $(\mathcal{I}_h u_r, \mathcal{I}_h u_l) \in S_h \times S_h$ of $(u_r, u_l) \in \mathcal{H}$ satisfies the discrete coupling conditions at the boundaries Γ_0 and Γ_1, i.e. $(u_r, u_l) \in \mathcal{H}$ is approximated by $(\mathcal{I}_h u_r, \mathcal{I}_h u_l) \in \mathcal{H}_h$. Since the eigensolution (u_r, u_l) of (162) is assumed to be H^2-regular, standard estimates of the approximation error imply condition (A_I)

$$\|(u_r, u_l) - \mathcal{I}_h(u_r, u_l)\|_H \leq \tilde{c}h(|u_r|_{H^2} + |u_l|_{H^2}) =: C_I h, \tag{174}$$

where $|\cdot|_{H^2}$ is the H^2-semi norm, see e.g. [17] or [38].
By standard arguments, it can be proved that $(C^\infty(\Omega) \times C^\infty(\Omega)) \cap \mathcal{H}$ is dense in $\mathcal{H}$ (see e.g. the proof in [52]). Combining this with estimate (174), we obtain assumption (A_A)

$$\lim_{h \to 0} \left(\inf_{v_h \in \mathcal{H}_h} \|v - v_h\|_H \right) = 0 \qquad \forall v \in \mathcal{H}. \tag{175}$$

III. Properties of sesquilinear forms $a(\cdot,\cdot), b(\cdot,\cdot)$ *and perturbations* $a'_\tau(\cdot,\cdot), b'_\tau(\cdot,\cdot)$.
The Gårding inequality for $a(\cdot,\cdot)$ has been proved in Lemma 8.
By simple estimates, it follows that the sesquilinear forms $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ satisfy (133) and (134), respectively.

Similarly, it can be shown that the perturbations satisfy (152) with $\rho = \tau = ch$. □

5.4 Outline for a Proof of H^2-Regularity of the Eigensolution

In this section, we present the main ideas for proving that an eigensolution (u_r, u_l) of problem (162) is two times weakly differentiable, if the domain Ω' can be described as a Cartesian product

$$\Omega' = \Psi \times]0; L[, \Psi \subset \mathbb{R}^2,$$

where Ψ is a two-dimensional domain with sufficiently regular boundary (e.g., $C^{2,1}$-boundary; see [38] or [73]), and if

$$\phi_0 \equiv \phi_1 \equiv -1.$$

Some of the used concepts have already been applied in [55] and [56].
First, we transform (162) into a problem on a torus $\mathcal{T} \subset \mathbb{R}^3$, which we, for ease of presentation, describe by

$$\mathcal{T} := \Psi \times [-L; L],$$

where the points (x, y, L) and $(x, y, -L)$ are identified for $(x, y) \in \Psi$.
The Sobolev spaces on $\mathcal{T}$ or, more specifically, on the differentiable three-dimensional manifold $\mathcal{T}$, are denoted as $L_2(\mathcal{T}), H^1(\mathcal{T})$, and $H^2(\mathcal{T})$, see [73]. Let them be equipped with the natural inner products and norms, denoted by the additional index $\mathcal{T}$ as, for instance, in $\| \cdot \|_{L_2,\mathcal{T}}$. The functions in these Sobolev spaces can be seen as $2L$-periodic functions with respect to z on the infinite domain $\Psi \times \mathbb{R}$.
We introduce the mapping $\mathcal{F} : L_2(\Omega') \times L_2(\Omega') \to L_2(\mathcal{T})$ defined by

$$\mathcal{F}(v_r, v_l)(x, y, z) = \begin{cases} v_r(x, y, z) & \text{if } z > 0 \\ -v_l(x, y, -z) & \text{if } z < 0. \end{cases}$$

Obviously, $\mathcal{F}$ is bijective and the equation $\|(v_r, v_l)\|_L = \|\mathcal{F}(v_r, v_l)\|_{L_2,\mathcal{T}}$ holds for $(v_r, v_l) \in \mathcal{L}$. The following lemma implies, that we also have $\|(v_r, v_l)\|_H = \|\mathcal{F}(v_r, v_l)\|_{H^1,\mathcal{T}}$ for $(v_r, v_l) \in \mathcal{H}(\Omega')$, with $\mathcal{H}(\Omega')$ as defined in (161).

Lemma 9 *$\mathcal{F}$ is a bijective mapping from $\mathcal{H}(\Omega')$ to $H^1(\mathcal{T})$.*

Outline of Proof: Let us define

$$\mathcal{T}_+ := \Psi \times]0; L[, \quad \mathcal{T}_- := \Psi \times]-L; 0[, \quad \text{and} \quad \mathcal{T}_{z,\epsilon} := \Psi \times]z - \epsilon; z + \epsilon[$$

for $z \in [-L; L]$, where the points $(x, y, -L)$ and (x, y, L), $(x, y) \in \Psi$, shall be identified.
We consider

$$\int_{\mathcal{T}} \mathcal{F}(v_r, v_l)\, \overline{\partial^\alpha \varphi}\; d(x, y, z)$$

for $(v_r, v_l) \in \mathcal{H}(\Omega'), \varphi \in C_0^\infty(\mathcal{T})$, and multi-index α with $|\alpha| = 1$.
For test functions $\varphi \in C_0^\infty(\mathcal{T}_+) \cup C_0^\infty(\mathcal{T}_-)$, we obtain the relation

$$\begin{aligned}\int_{\mathcal{T}} \mathcal{F}(v_r, v_l)\, \overline{\partial^\alpha \varphi}\; d(x, y, z) &= \int_{\mathcal{T}_+} v_r \overline{\partial^\alpha \varphi}\; d(x, y, z) - \int_{\mathcal{T}_-} v_l \overline{\partial^\alpha \varphi}\; d(x, y, z) \\ &= -\int_{\mathcal{T}_+} \partial^\alpha v_r\, \overline{\varphi}\; d(x, y, z) + \int_{\mathcal{T}_-} \partial^\alpha v_l\, \overline{\varphi}\; d(x, y, z),\end{aligned}$$

where we have used that (v_r, v_l) is in $H^1(\Omega') \times H^1(\Omega')$.
For $\varphi \in C_0^\infty(\mathcal{T} \setminus \overline{\mathcal{T}_{L,\epsilon}})$, where $0 < \epsilon < L$, the slice of $\mathcal{T}$, which corresponds to the boundary Γ_0, has to be taken into account. Since Γ_0 is lying in a plane that is perpendicular to the z-axis, a boundary term in the integration by parts only arises for the z-derivative, i.e. for the case $\alpha = (0, 0, 1)$. Then, we have

$$\begin{aligned}&\int_{\mathcal{T}} \mathcal{F}(v_r, v_l)\, \frac{\partial}{\partial z}\overline{\varphi}\; d(x, y, z) = \int_{\mathcal{T}_+} v_r \frac{\partial}{\partial z}\overline{\varphi}\; d(x, y, z) - \int_{\mathcal{T}_-} v_l \frac{\partial}{\partial z}\overline{\varphi}\; d(x, y, z) \\ &= -\int_{\mathcal{T}_+} \frac{\partial}{\partial z} v_r \overline{\varphi}\; d(x, y, z) + \int_{\Gamma_0} v_r \overline{\varphi}\; d\sigma(x, y, z) \qquad (176) \\ &\quad + \int_{\mathcal{T}_-} \frac{\partial}{\partial z} v_l \overline{\varphi}\; d(x, y, z) + \int_{\Gamma_0} v_l \overline{\varphi}\; d\sigma(x, y, z) \\ &= -\int_{\mathcal{T}_+} \frac{\partial}{\partial z} v_r \overline{\varphi}\; d(x, y, z) + \int_{\mathcal{T}_-} \frac{\partial}{\partial z} v_l \overline{\varphi}\; d(x, y, z) + \int_{\Gamma_0} (v_r + v_l) \overline{\varphi}\; d\sigma(x, y, z).\end{aligned}$$

However, due to the coupling condition on Γ_0 for $(v_r, v_l) \in \mathcal{H}(\Omega')$, the last integral vanishes, and we get

$$\int_{\mathcal{T}} \mathcal{F}(v_r, v_l)\, \frac{\partial}{\partial z}\overline{\varphi}\; d(x, y, z) = -\int_{\mathcal{T}_+} \frac{\partial}{\partial z} v_r \overline{\varphi}\; d(x, y, z) + \int_{\mathcal{T}_-} \frac{\partial}{\partial z} v_l \overline{\varphi}\; d(x, y, z).$$

For the case $\varphi \in C_0^\infty(\mathcal{T} \setminus \overline{\mathcal{T}_{0,\epsilon}})$, one obtains the same relation by using the coupling condition on Γ_1.
Thus, we have shown that for $(v_r, v_l) \in \mathcal{H}(\Omega')$ the image $\mathcal{F}(v_r, v_l) \in L_2(\mathcal{T})$ possesses the weak derivative $\mathcal{F}(\partial^\alpha v_r, \partial^\alpha v_l) \in L_2(\mathcal{T})$ and, therefore, is in $H^1(\mathcal{T})$. So, $(v_r, v_l) \in \mathcal{H}(\Omega')$ is mapped into $H^1(\mathcal{T})$.
Obviously, every $V \in H^1(\mathcal{T})$ has the pre-image $(v_r, v_l) \in H^1(\Omega') \times H^1(\Omega')$, with $v_r(x, y, z) := V(x, y, z)$ and $v_l(x, y, z) := -V(x, y, -z)$ for $(x, y, z) \in \Omega'$. Using equation (176) and standard arguments, one obtains that the pair (v_r, v_l), additionally, satisfies the coupling boundary conditions and, therefore, is in $\mathcal{H}(\Omega')$.

Obviously, the mapping $\mathcal{F}$ is also injective on $\mathcal{H}(\Omega')$. □

Observing the definitions of the sesquilinear forms $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ in Chapter 5.1 and the Gårding inequality for $a(\cdot,\cdot)$ stated in Lemma 8, we define the sesquilinear forms $A(\cdot,\cdot)$ on $H^1(\mathcal{T})$ and $B(\cdot,\cdot)$ on $L_2(\mathcal{T})$ by

$$\begin{aligned} A(V,\Phi) \;:=\; & \int_{\mathcal{T}} \nabla V \nabla \overline{\Phi} + 2ik_f \frac{\partial}{\partial z} V\overline{\Phi} + (k_f^2 - k^2) V\overline{\Phi}\; d(x,y,z) + C_g(V,\Phi)_{L_2,\mathcal{T}} \\ & + iC_b \int_{\partial\mathcal{T}} V\overline{\Phi}\; d\sigma(x,y,z) \end{aligned}$$

and

$$B(V,\Phi) := (V,\Phi)_{L_2,\mathcal{T}} = \int_{\mathcal{T}} V\overline{\Phi}\; d(x,y,z),$$

where $C_g > 0$ is the constant in the Gårding inequality for $a(\cdot,\cdot)$. Obviously, the form $A(\cdot,\cdot)$ is $H^1(\mathcal{T})$-elliptic.
If (u_r, u_l) is a solution of the eigenvalue problem (162), then it satisfies

$$\begin{aligned} & A(\mathcal{F}(u_r,u_l),\mathcal{F}(\varphi_r,\varphi_l)) \\ &= \xi B(\mathcal{F}(u_r,u_l),\mathcal{F}(\varphi_r,\varphi_l)) + C_g(\mathcal{F}(u_r,u_l),\mathcal{F}(\varphi_r,\varphi_l))_{L_2,\mathcal{T}} \\ &= (\xi + C_g)(\mathcal{F}(u_r,u_l),\mathcal{F}(\varphi_r,\varphi_l))_{L_2,\mathcal{T}} \qquad \forall (\varphi_r,\varphi_l) \in \mathcal{H}(\Omega'). \end{aligned}$$

Let us consider the following equation for $V \in H^1(\mathcal{T})$:

$$A(V,\Phi) = (F,\Phi)_{L_2,\mathcal{T}} \qquad \forall \Phi \in H^1(\mathcal{T}) \tag{177}$$

with right-hand side $F \in L_2(\mathcal{T})$.

Lemma 10 (H^2-Regularity on Torus) *If $F \in L_2(\mathcal{T})$ holds, then the unique solution $V \in H^1(\mathcal{T})$ of (177) is in $H^2(\mathcal{T})$ and satisfies*

$$\|V\|_{H^2,\mathcal{T}} \le C\|F\|_{L_2,\mathcal{T}}.$$

Outline of Proof: For the H^2-regularity with respect to the z-direction it has to be used that $\mathcal{T}$ is a torus. For the other derivatives the proof of interior regularity is standard (see e.g. [73]).
The proof of H^2-regularity on the sufficiently regular boundary $(\partial\Psi)\times]-L;L[$, perpendicular to z, is standard and follows, e.g. for a $C^{2,1}$-boundary $\partial\Psi$, from general results in [73]. □

Corollary 2 (H^2-Regularity of Eigensolution) *An eigensolution $(u_r,u_l) \in \mathcal{H}$ of equation (162) satisfies*

$$(u_r,u_l) \in H^2(\Omega') \times H^2(\Omega').$$

OUTLINE OF PROOF: Due to the bijectivity of $\mathcal{F}$ stated in Lemma 9, $\mathcal{F}(u_r, u_l)$ is the solution of (177) with right-hand side $F = (\xi + C_g)\mathcal{F}(u_r, u_l) \in L_2(\mathcal{T})$. By Lemma 10, $U := \mathcal{F}(u_r, u_l)$ is even $H^2(\mathcal{T})$-regular and (u_r, u_l) is in $H^2(\Omega') \times H^2(\Omega')$. □

6 Solving the Discretized Two-Wave Eigenvalue Problem Applying A Shift-and-Invert Technique with Preconditioned GMRES

In this chapter, we describe, how we solve the two-wave eigenvalue problem numerically.
As explained in Chapter 5, this PDE eigenvalue problem is discretized by finite elements which leads to sparse matrices. To solve the discrete problem we, therefore, use an appropriate iterative method, which essentially consists of a Krylov subspace method (see e.g. [61]).
Let us review the discrete problem. The aim is to find a discrete eigensolution $(U_r, U_l) \in \mathcal{H}_{h',h}$ with eigenvalue $\xi_h \in \mathbb{C}$, such that

$$a_{tw}((U_r, U_l), (\varphi_r, \varphi_l)) = \xi_h \; b_{tw}((U_r, U_l), (\varphi_r, \varphi_l)) \quad \forall (\varphi_r, \varphi_l) \in \mathcal{H}_{h',h}. \tag{178}$$

with sesquilinear forms

$$a_{tw}((U_r, U_l), (\varphi_r, \varphi_l)) := a((U_r, U_l), (\varphi_r, \varphi_l)) + a'_\tau((U_r, U_l), (\varphi_r, \varphi_l))$$

and

$$b_{tw}((U_r, U_l), (\varphi_r, \varphi_l)) := b((U_r, U_l), (\varphi_r, \varphi_l)) + b'_\tau((U_r, U_l), (\varphi_r, \varphi_l)),$$

where

$$\begin{aligned} &a((U_r, U_l), (\varphi_r, \varphi_l)) := \\ &\int_\Omega \nabla U_r \nabla \overline{\varphi}_r + 2ik_f \frac{\partial}{\partial z} U_r \overline{\varphi}_r + (k_f^2 - k^2) U_r \overline{\varphi}_r \; d(x,y,z) - iC_b \int_{\Gamma_r} U_r \overline{\varphi}_r \; d\sigma(x,y,z) \\ &+ \int_\Omega \nabla U_l \nabla \overline{\varphi}_l - 2ik_f \frac{\partial}{\partial z} U_l \overline{\varphi}_l + (k_f^2 - k^2) U_l \overline{\varphi}_l \; d(x,y,z) - iC_b \int_{\Gamma_r} U_l \overline{\varphi}_l \; d\sigma(x,y,z), \end{aligned}$$

and

$$b((U_r, U_l), (\varphi_r, \varphi_l)) := \int_\Omega U_r \overline{\varphi}_r \; d(x,y,z) + \int_\Omega U_l \overline{\varphi}_l \; d(x,y,z),$$

describe the weak form of the original equation and

$$\begin{aligned} &a'_\tau((U_r, U_l), (\varphi_r, \varphi_l)) := \\ &\quad \tau 2k_f \int_\Omega 2ik_f \frac{\partial}{\partial z} U_r \frac{\partial}{\partial z} \bar{\varphi}_r + (k_f^2 - k^2) U_r \frac{\partial}{\partial z} \bar{\varphi}_r \; d(x,y,z) \\ &\quad -\tau 2k_f \int_\Omega -2ik_f \frac{\partial}{\partial z} U_l \frac{\partial}{\partial z} \bar{\varphi}_l + (k_f^2 - k^2) U_l \frac{\partial}{\partial z} \bar{\varphi}_l \; d(x,y,z), \end{aligned}$$

and

$$b'_\tau((U_r, U_l), (\varphi_r, \varphi_l)) := \tau 2k_f \int_\Omega U_r \frac{\partial}{\partial z} \bar{\varphi}_r - U_l \frac{\partial}{\partial z} \bar{\varphi}_l \; d(x,y,z)$$

represent the stabilization, see Chapter 5.2.

6.1 A Matrix-Vector Representation of the Two-Wave Eigenvalue Problem

As mentioned in Chapter 5.2, we use the finite element space $S_{h',h}$ of continuous, piecewise trilinear functions on a partition of the cuboid Ω into sub-cuboids of size $h' \times h' \times h$. Let the, say N, vertices be enumerated and let (x_j, y_j, z_j) be the coordinates of the j-th node. Then, the corresponding trilinear nodal basis function $\psi_j : \Omega \to \mathbb{R}$ is given by

$$\psi_j(x,y,z) := \begin{cases} \left(1 - \left|\frac{x-x_j}{h'}\right|\right)\left(1 - \left|\frac{y-y_j}{h'}\right|\right)\left(1 - \left|\frac{z-z_j}{h}\right|\right) & \text{if } |x-x_j|, |y-y_j| \le h' \\ & \quad \text{and } |z-z_j| \le h, \\ 0 & \text{else.} \end{cases} \tag{179}$$

A function $V_r \in S_{h',h}$ can uniquely be written as a linear combination of these nodal basis functions:

$$V_r(x,y,z) = \sum_{j=1}^{N} \nu_j^{(r)} \ \psi_j(x,y,z) \tag{180}$$

with a coefficient vector $(\nu_j^{(r)}) \in \mathbb{C}^N$. Obviously, it is

$$\nu_j^{(r)} = V_r(x_j, y_j, z_j) \qquad \text{for } j = 1, \ldots, N. \tag{181}$$

The nodes can be enumerated as follows: Using numbers $N_{ir}, N_0, N_1 > 0$, with $N_{ir} + N_0 + N_1 = N$, we partition the nodes into the sets

$$\mathcal{N}_{ir} := \{1, \ldots, N_{ir}\}$$

corresponding to the interior of Ω and the radiating boundary Γ_r,

$$\mathcal{N}_0 := \{N_{ir} + 1, \ldots, N_{ir} + N_0\}$$

corresponding to the boundary Γ_0, and

$$\mathcal{N}_1 := \{N_{ir} + N_0 + 1, \ldots, N_{ir} + N_0 + N_1\}$$

corresponding to the boundary Γ_1.
The finite element space $S_{h',h}$ – as described in Chapter 5.2 – is used to construct the test and trial space for the eigenvalue problem (178). We define

$$\mathcal{H}_{h',h} := \left\{ (V_r, V_l) \in S_{h',h} \times S_{h',h} \;\middle|\; V_r - \phi_0 V_l\Big|_{\Gamma_{0,h'}} = 0 \text{ and } V_r - \bar{\phi}_1 V_l\Big|_{\Gamma_{1,h'}} = 0 \right\}.$$

The pair $(V_r, V_l) \in \mathcal{H}_{h',h}$ can be represented by the vector

$$\nu := (\nu_1^{(r)}, \ldots, \nu_N^{(r)}, \nu_1^{(l)}, \ldots, \nu_N^{(l)}) \in \mathcal{H}_d \subset \mathbb{C}^{2N} \tag{182}$$

with $\mathcal{H}_d$ defined as the subspace of $\mathbb{C}^{2N}$, where the relations

$$\nu_j^{(r)} = \phi_0(x_j, y_j, z_j) \cdot \nu_j^{(l)} \quad \text{for } N_{ir} + 1 \leq j \leq N_{ir} + N_0 \tag{183}$$

and

$$\nu_j^{(r)} = \overline{\phi_1}(x_j, y_j, z_j) \cdot \nu_j^{(l)} \quad \text{for } N_{ir} + N_0 + 1 \leq j \leq N \tag{184}$$

hold. Relations (183) and (184) reflect the the coupling conditions in the finite element space $\mathcal{H}_{h',h}$ on the boundaries Γ_0 and Γ_1.
It is obvious, that the subset

$$\mathcal{B} := \mathcal{B}_r \cup \mathcal{B}_l \cup \mathcal{B}_0 \cup \mathcal{B}_1 \tag{185}$$

of $\mathcal{H}_{h',h}$ with

$$\begin{aligned}
\mathcal{B}_r &:= \{(\psi_k, 0) \mid k \in \mathcal{N}_{ir}\} \\
\mathcal{B}_l &:= \{(0, \psi_k) \mid k \in \mathcal{N}_{ir}\} \\
\mathcal{B}_0 &:= \left\{(\psi_k, \overline{\phi_0(x_k, y_k, z_k)} \cdot \psi_k) \mid k \in \mathcal{N}_0\right\} \\
\mathcal{B}_1 &:= \{(\psi_k, \phi_1(x_k, y_k, z_k) \cdot \psi_k) \mid k \in \mathcal{N}_1\}
\end{aligned}$$

forms a basis of $\mathcal{H}_{h',h}$.
In the following, we show how the weak equation (178) can be written as a matrix-vector eigenvalue problem

$$A\nu = \xi B\nu, \tag{186}$$

where $\nu \in \mathbb{C}^{2N}$ is the wanted coefficient vector, $A \in \mathbb{C}^{2N \times 2N}$ the stiffness matrix, and $B \in \mathbb{C}^{2N \times 2N}$ the mass matrix.
For this purpose, we first consider the expressions depending on the wave U_r in the sesquilinear form $a_{tw}((U_r, U_l), (\varphi_r, \varphi_l))$. This gives the sparse matrix $A^{(r)} \in \mathbb{C}^N$ computed by

$$(A^{(r)})_{k,l} := a_{tw}((\psi_l, 0), (\psi_k, 0)) \quad \text{for } 1 \leq k, l \leq N$$

with a structure as depicted in Figure 16.
The blocks of $A^{(r)}$ reflect the partition of the nodes and are obtained by following choices of k and l:

$$\begin{aligned}
&A_1^{(r)} \text{ for } k, l \in \mathcal{N}_{ir} \\
&A_2^{(r)} \text{ for } k \in \mathcal{N}_{ir}, l \in \mathcal{N}_0 \\
&A_3^{(r)} \text{ for } k \in \mathcal{N}_{ir}, l \in \mathcal{N}_1 \\
&A_4^{(r)} \text{ for } k \in \mathcal{N}_0, l \in \mathcal{N}_{ir} \\
&A_5^{(r)} \text{ for } k \in \mathcal{N}_0, l \in \mathcal{N}_0 \\
&A_6^{(r)} \text{ for } k \in \mathcal{N}_1, l \in \mathcal{N}_{ir} \\
&A_7^{(r)} \text{ for } k \in \mathcal{N}_1, l \in \mathcal{N}_1.
\end{aligned}$$

$$A^{(r)} = \begin{pmatrix} A_1^{(r)} & A_2^{(r)} & A_3^{(r)} \\ A_4^{(r)} & A_5^{(r)} & 0 \\ A_6^{(r)} & 0 & A_7^{(r)} \end{pmatrix}$$

Figure 16: Part of the stiffness matrix.

Analogously to $A^{(r)}$, we define

$$(A^{(l)})_{k,l} := a_{tw}((0,\psi_l),(0,\psi_k)) \quad \text{for } 1 \le k,l \le N.$$

Due to the coupling boundary conditions (or, equivalently, due to the restrictions (183) and (184)), the functions in $\mathcal{H}_{h',h}$ only possess $N + N_{ir} = 2N_{ir} + N_0 + N_1$ degrees of freedom. A usage of the basis $\mathcal{B}$ in (185) transforms the variational problem (178) into an equivalent matrix eigenvalue problem of order $(N + N_{ir}) \times (N + N_{ir})$.

If, however, a standard finite element program or library shall be used in the computations, as e.g. the C++-library EXPDE [1], [57], which was our choice, the functions U_r and U_l normally have to be stored separately and the restrictions (183) and (184) have to be enforced additionally.

For this purpose, we blow up the problem (178) to a $(2N \times 2N)$-system of rank $N + N_{ir}$. Figure 17 shows the structure of the resulting stiffness matrix A. The new mass matrix B has the same structure and is computed analogously to the stiffness matrix, the structure of which is explained more detailed in the following.

For an easier description of A, we first define the diagonal matrices

$$\Phi_0 = \mathrm{diag}(\phi_0(x_j,y_j,z_j)), \quad j = N_{ir}+1,\ldots,N_{ir}+N_0, \tag{187}$$

and

$$\Phi_1 = \mathrm{diag}(\phi_1(x_j,y_j,z_j)), \quad j = N_{ir}+N_0+1,\ldots,N, \tag{188}$$

which incorporate the coupling (183) and (184) at the boundaries Γ_0 and Γ_1. The additional sub-matrices of A in Figure 17 are defined by

$$\begin{bmatrix} \tilde{A}_4^{(l)} & \tilde{A}_5^{(l)} \end{bmatrix} := \Phi_0 \begin{bmatrix} A_4^{(l)} & A_5^{(l)} \end{bmatrix}, \tag{189}$$

$$\begin{bmatrix} \tilde{A}_6^{(l)} & \tilde{A}_7^{(l)} \end{bmatrix} := \overline{\Phi_1} \begin{bmatrix} A_6^{(l)} & A_7^{(l)} \end{bmatrix}, \tag{190}$$

$$\begin{bmatrix} \hat{A}_4^{(r)} & \hat{A}_5^{(r)} & \hat{A}_4^{(l)} & \hat{A}_5^{(l)} \end{bmatrix} := \overline{\Phi_0} \begin{bmatrix} A_4^{(r)} & A_5^{(r)} & \tilde{A}_4^{(l)} & \tilde{A}_5^{(l)} \end{bmatrix}, \tag{191}$$

$$\text{and} \quad \begin{bmatrix} \hat{A}_6^{(r)} & \hat{A}_7^{(r)} & \hat{A}_6^{(l)} & \hat{A}_7^{(l)} \end{bmatrix} := \Phi_1 \begin{bmatrix} A_6^{(r)} & A_7^{(r)} & \tilde{A}_6^{(l)} & \tilde{A}_7^{(l)} \end{bmatrix}. \tag{192}$$

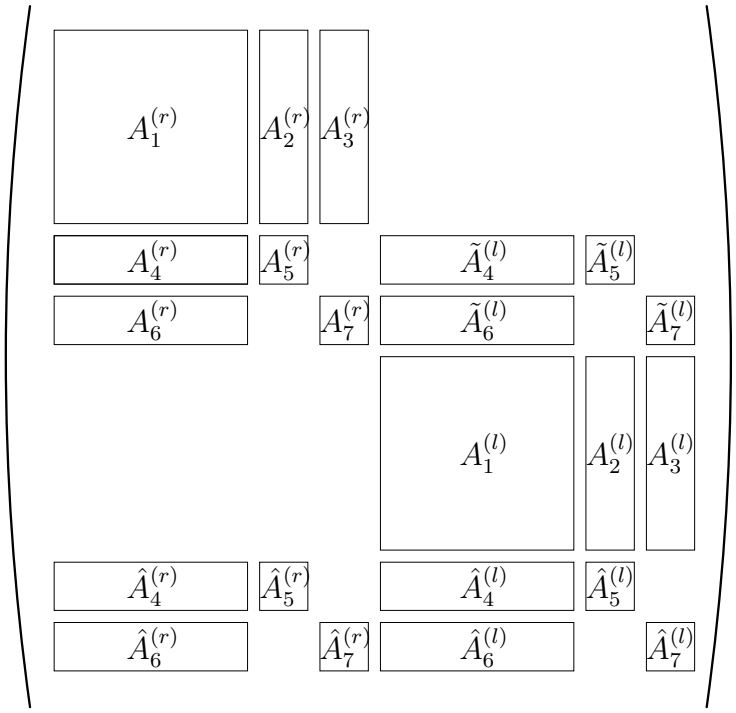

Figure 17: Stiffness matrix A.

Constructing A and B this way, we obtain

Lemma 11 (Properties of Matrix Eigenvalue Problem) *The eigenvalue problem (178) is equivalent to the problem: Find $\nu \in \mathcal{H}_d \subset \mathbb{C}^{2N}$ and $\xi_h \in \mathbb{C}$ with*

$$A\nu = \xi_h \, B\nu. \tag{193}$$

Furthermore, for all $\tilde{\nu} \in \mathbb{C}^{2N}$, it is

$$A\tilde{\nu} \in \mathcal{H}_d \quad \textit{and} \quad B\tilde{\nu} \in \mathcal{H}_d. \tag{194}$$

Proof: The statements (194) follow directly from the definitions (187) and (188) of the diagonal matrices Φ_0 and Φ_1, respectively, and from relations (191) and (192) for A and B.
Due to (194) for $n \in \{N_{ir}+1, \ldots, N\}$ the row n and the row $n+N$ of the system (193) agree with each other except for multiplication by a scalar. For the proof of the equivalence of the eigenvalue problems, it therefore suffices to consider the rows 1 to $2N_{ir} + N_0 + N_1$.
We recall that a function $(U_r, U_l) \in \mathcal{H}_{h',h}$ can – with respect to the basis $\mathcal{B}$ – equivalently be described as a vector $\nu \in \mathcal{H}_d$, see relation (182) in conjunction with (181), and relation (180).
Since the basis $\mathcal{B}$ is a disjoint union of the sets $\mathcal{B}_r, \mathcal{B}_l, \mathcal{B}_0$, and $\mathcal{B}_1$, we distinguish four types of "test functions" (φ_r, φ_l):

$\mathcal{B}_r$: Let $(\varphi_r, \varphi_l) \in \mathcal{B}_r$, i.e. $(\varphi_r, \varphi_l) = (\psi_k, 0)$ with appropriate $k \in \mathcal{N}_{ir}$. Since U_r and φ_l, and U_l and φ_r, respectively, are not coupled in the sesquilinear form $a_{tw}((U_r, U_l), (\varphi_r, \varphi_l))$, we obtain

$$\begin{aligned} a_{tw}((U_r, U_l), (\varphi_r, \varphi_l)) &= a_{tw}((U_r, U_l), (\psi_k, 0)) = a_{tw}((U_r, 0), (\psi_k, 0)) \\ &= a_{tw}((\sum_{j=1}^{N} \nu_j^{(r)} \psi_j, 0), (\psi_k, 0)) = \sum_{j=1}^{N} \nu_j^{(r)} a_{tw}((\psi_j, 0), (\psi_k, 0)). \end{aligned}$$

The last term, however, is exactly the k-th entry of the vector $A\nu \in \mathbb{C}^{2N}$, i.e.

$$a_{tw}((U_r, U_l), (\varphi_r, \varphi_l)) = (A\nu)_k .$$

Analogously, for the right-hand side we obtain

$$b_{tw}((U_r, U_l), (\varphi_r, \varphi_l)) = (B\nu)_k$$

for $k \in \mathcal{N}_{ir}$.

$\mathcal{B}_l$: As for the previous case, for $(\varphi_r, \varphi_l) \in \mathcal{B}_l$, it follows that

$$a_{tw}((U_r, U_l), (\varphi_r, \varphi_l)) = (A\nu)_{N+k} \text{ and } b_{tw}((U_r, U_l), (\varphi_r, \varphi_l)) = (B\nu)_{N+k}$$

holds for the $k \in \mathcal{N}_{ir}$ with $(0, \psi_k) = (\varphi_r, \varphi_l)$.

$\mathcal{B}_0$: With $(\varphi_r, \varphi_l) = (\psi_k, \overline{\phi_0(x_k, y_k, z_k)} \cdot \psi_k)$ for $k \in \mathcal{N}_0$, we have

$$\begin{aligned} a_{tw}((U_r, U_l), (\varphi_r, \varphi_l)) &= a_{tw}((U_r, 0), (\psi_k, 0)) + \phi_0(x_k, y_k, z_k) \cdot a_{tw}((0, U_l), (0, \psi_k)) \\ &= (A\nu)_k \end{aligned}$$

and also

$$b_{tw}((U_r, U_l), (\varphi_r, \varphi_l)) = (B\nu)_k .$$

$\mathcal{B}_1$: Finally, the choice $(\varphi_r, \varphi_l) = (\psi_k, \phi_1(x_k, y_k, z_k) \cdot \psi_k), k \in \mathcal{N}_1$ gives

$$\begin{aligned} a_{tw}((U_r, U_l), (\varphi_r, \varphi_l)) &= a_{tw}((U_r, 0), (\psi_k, 0)) + \overline{\phi_1(x_k, y_k, z_k)} \cdot a_{tw}((0, U_l), (0, \psi_k)) \\ &= (A\nu)_k \end{aligned}$$

and

$$b_{tw}((U_r, U_l), (\varphi_r, \varphi_l)) = (B\nu)_k .$$

Since $\mathcal{B}$ is a basis consisting of $N + N_{ir}$ functions, $(U_r, U_l) \in \mathcal{H}_{h',h}$ solves problem (178) if and only if the corresponding coefficient vector $\nu \in \mathcal{H}_d$ satisfies the first $N + N_{ir}$ rows of the matrix eigenvalue problem (193). Thus, the equivalence of both eigenvalue problems is proved. □

We shortly remark, that the nodal basis functions in $\mathcal{B}_0 \cup \mathcal{B}_1 \subset \mathcal{B}$, corresponding to the boundaries Γ_0 and Γ_1, are a natural choice if we have the coefficient functions $\phi_0 \equiv \phi_1 \equiv -1$. In this case, as e.g. explained in Chapter 3.3 or Chapter 5.4, the two-wave eigenvalue problem can be viewed as a problem on a torus with respect to the z-direction (or as a problem for $2L$-periodic functions in z). From this point of view, the reflecting boundaries Γ_0 and Γ_1 are interior boundaries, and the rows $N_{ir}+1, \ldots, N_0$ and $N_{ir}+N_0, \ldots, N$ exactly describe the integration on a sub-domain of the torus including the reflecting boundaries, as e.g. Figure 18 shows for a node (x_k, y_k, z_k) on the boundary Γ_0.
So, for planar end mirrors ($\phi_0 \equiv \phi_1 \equiv -1$), the described matrix-vector representation of the system (178) is the natural continuation onto the torus.

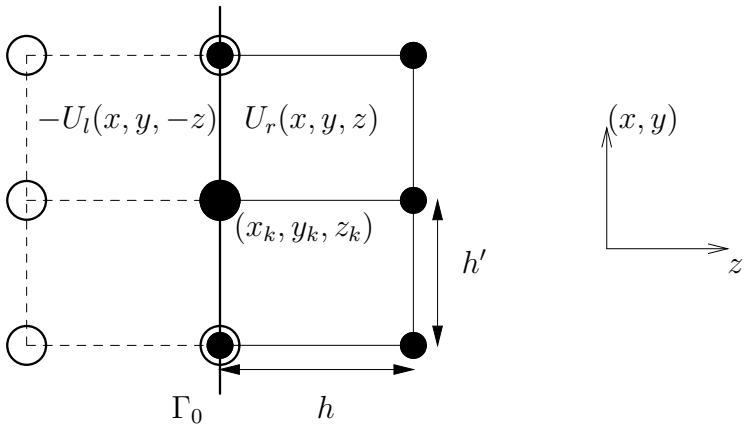

Figure 18: Torus interpretation of integration for a node on the boundary Γ_0.

A further important property of A (and B) is, that for $\tilde{\nu} \in \mathbb{C}^{2N}$ we have $A\tilde{\nu} \in \mathcal{H}_d$ (and $B\tilde{\nu} \in \mathcal{H}_d$). Since in Krylov-type iterative methods the search space is spanned by a repeated application of the system matrix A, this property will keep the intermediate approximate vectors in the space $\mathcal{H}_d$ and guarantee that the final discrete solution satisfies the coupling boundary conditions (183) and (184).

6.2 Shift-and-Invert

In this section, we describe an iterative algorithm for solving the large and sparse matrix eigenvalue problem (193).
For details on matrix eigenvalue problems, we refer to the monographs [18] and [72], and to the books [23], [60], and [67], which rather focus on numerical and practical issues of matrix eigenvalue problems.
Furthermore, we point out three state-of-the-art iterative methods that are able to compute even several eigenvalues at a time or that can be modified to do so,

namely, the implicitly restarted Arnoldi methods (IRA) [49], the adaptive block Lanczos method ABLE [12], and the Jacobi-Davidson methods [66].
In this thesis, however, we focus on the computation of one eigenvalue and use a Wieland iteration (an inverse power method with shift) for generalized matrix eigenvalue problems.
Let us assume, that ξ is a simple eigenvalue of (178) with eigensolution (U_r, U_l) and that for $\sigma \in \mathbb{C}$ with $\sigma \neq \lambda$ in a small neighborhood of σ no other eigenvalue except for ξ is present.
Instead of using $a_{tw}(\cdot,\cdot)$, a shifted eigenvalue problem with sesquilinear form

$$a_{sh}((V_r, V_l), (\varphi_r, \varphi_l)) := a_{tw}((V_r, V_l), (\varphi_r, \varphi_l)) - \sigma b((V_r, V_l), (\varphi_r, \varphi_l)), \qquad (195)$$

and resulting stiffness matrix A_{sh}, can be solved.
Then, (U_r, U_l) is the eigensolution related to the eigenvalue of smallest modulus of the shifted eigenvalue problem (195).
To compute this eigenvalue and the corresponding eigensolution, we apply an inverse power method (see e.g. [67]) to (195). The complete solution method can be outlined as follows:

Shift-and-Invert:

1. Choose σ and generate matrices $A_{sh}, B \in \mathbb{C}^{2N\times 2N}$.
2. Choose start vector $\nu^{(0)}$ with $\|\nu^{(0)}\| = 1$.
3. For $n = 1, \ldots$ until convergence do
4. Solve $A_{sh}\,\tilde{\nu} = B\nu^{(n)}$ for $\tilde{\nu} \in \mathcal{H}_d$ by preconditioned GMRES
5. compute approximate eigenvalue $\mu^{(n)}$ from $\tilde{\nu}$ and $\nu^{(n)}$
6. and assign $\nu^{(n+1)} := \tilde{\nu}/\|\tilde{\nu}\|$.
7. Compute approximate (un-shifted) eigenvalue $\xi_h = \sigma + \mu^{(n)}$.

Obviously, in the case of convergence of this algorithm, we have

$$A\nu = \xi_h\, B\nu. \qquad (196)$$

The solution of the equation

$$A_{sh}\,\tilde{\nu} = B\nu^{(n)} \qquad (197)$$

in line 4 is the part of this method with the largest time and storage requirements. The solver (preconditioned GMRES) is explained in the following Sections 6.3 and 6.4.
The computation of the approximate eigenvalue $\mu^{(n)}$ in line 5 can be performed with the help of an appropriately chosen linear functional $L : \mathbb{C}^{2N} \to \mathbb{C}$ by $\mu^{(n)} := L(\nu^{(n)})/L(\tilde{\nu})$.

6.3 Preconditioned GMRES

The *generalized minimal residual method* (GMRES), was for the first time presented in [62]. It is designed for the solution of a linear system of equations

$$Ax = b \tag{198}$$

with general regular matrix $A \in \mathbb{C}^{m\times m}$, unknown vector $x \in \mathbb{C}^m$ and right-hand side $b \in \mathbb{C}^m$.
When applied as an iterative method for large, sparse matrices, the method computes an approximate

$$y_0 \in x_0 + \mathcal{K}_d(r_0, A),$$

where $x_0 \in \mathbb{C}^m$ is a start approximation, $r_0 := Ax_0 - b$ is the start residual and $\mathcal{K}_d(r_0, A) := \mathrm{span}_{\mathbb{C}}\{r_0, Ar_0, \ldots, A^{d-1}r_0\}$ is the Krylov subspace of order d with respect to r_0 and A. The approximate y_0 is chosen such that it minimizes the l^2-residual in the affine search space $x_0 + \mathcal{K}_d(r_0, A)$:

$$\|Ay_0 - b\|_2 = \min_{y\in x_0+\mathcal{K}_d(r_0,A)} \|Ay - b\|_2$$

with l_2-norm $\|x\|_2 := (\sum_{j=1}^m x_j \cdot \overline{x_j})^{1/2}$. For large problems, the dimension d is chosen much smaller than the dimension of the original system m. For details of the algorithm, see e.g. [61].
Since the basic algorithm does not exploit any structure of the matrix A, the convergence is very poor in general, see [27], [36], or [61], and the references cited therein. For matrices A that are diagonalizable, a bound for the convergence rate can be proved that depends on the distribution of the eigenvalues of the system matrix and on the condition number of the transformation matrix (see e.g. [61]). Particularly, this result shows that GMRES converges fast, if all eigenvalues of the diagonalizable system matrix are, for instance, lying near to $1 \in \mathbb{C}$ in the complex plane and if the normed eigenvectors are pairwise nearly orthogonal.
Due to this observation, one *preconditions* the system meaning that one performs an equivalent transformation of the original equation (198) into a system which (better) satisfies the properties mentioned in the previous sentence, and, therefore, exhibits a better convergence behavior.
For our problem, we use the so called *right preconditioning*, where instead of (198) the system

$$AM^{-1}u = b, \quad x = M^{-1}u \tag{199}$$

is solved. The preconditioner $M \in \mathbb{C}^{m\times m}$ is chosen such that it approximates the system matrix A, but – in contrast to it – can be inverted easily.
The choice of the preconditioner is the crucial step in applying GMRES to a concrete problem. In our case, the preconditioner has been constructed according to the underlying physics of the problem, as we will describe below.

For the preconditioned GMRES algorithm we, again, refer to [61].
In our eigenvalue problem (196), the system matrix $A \in \mathbb{C}^{2N\times 2N}$ is of rank $N + N_{ir} < 2N$, i.e. it is singular. But due to the construction of A and B, the GMRES method solves equation (197) in the $(N + N_{ir})$-dimensional subspace $\mathcal{H}_d \subset \mathbb{C}^{2N}$ (or, in other words, GMRES applied to (197) essentially computes the solution of an $(N + N_{ir}) \times (N + N_{ir})$-system of linear equations).

6.4 The Preconditioner

Before explaining the preconditioner, we prove a lemma.

Lemma 12 *Under the condition*

$$\phi_0 \equiv \phi_1 \equiv -1,$$

the space $\mathcal{H}$ defined in (161) can be represented as a direct sum

$$\mathcal{H}(\Omega) = \mathcal{H}^{(c)} \oplus \mathcal{H}^{(n)} \tag{200}$$

with

$$\mathcal{H}^{(c)} := \left\{ (v_r(x,y,z), v_l(x,y,z)) \in \mathcal{H} \;\middle|\; \frac{\partial}{\partial z} v_r(x,y,z) = \frac{\partial}{\partial z} v_l(x,y,z) = 0 \right\}$$

and

$$\mathcal{H}^{(n)} := \left\{ (v_r(x,y,z), v_l(x,y,z)) \in \mathcal{H} \;\middle|\; \int_0^L v_r(x,y,z) - v_l(x,y,z)\; dz = 0 \right.$$
$$\left. \text{for } (x,y) \in]-W/2; W/2[\times]-W/2; W/2[\right\},$$

where the equations in the definition of the spaces are to be understood in a weak sense.

Proof: Let $(v_r, v_l) \in \mathcal{H}(\Omega)$ be chosen arbitrarily. We define

$$c(x,y,z) := \frac{1}{2} \int_0^L v_r(x,y,t) - v_l(x,y,t)\; dt.$$

Using the boundedness of the domain Ω and standard arguments, one can prove that

$$c \in H^1(\Omega) \qquad \text{with (weak) derivative} \quad \frac{\partial}{\partial z} c = 0.$$

Obviously, it is $(c, -c) \in \mathcal{H}^{(c)} \subset \mathcal{H}$. Since in the equation

$$(v_r, v_l) = (v_r - c, v_l + c) + (c, -c)$$

the pair $(v_r^{(n)}, v_l^{(n)}) := (v_r - c, v_l + c)$ is in $\mathcal{H}^{(n)}$, we have the stated decomposition.

Furthermore, it is clear that the intersection of $\mathcal{H}^{(c)}$ and $\mathcal{H}^{(n)}$ only contains the zero:

$$\mathcal{H}^{(c)} \cap \mathcal{H}^{(n)} = \{(0,0)\} \subset \mathcal{H}.$$

Thus, the decomposition (200) is proved. □

In other words, a pair $(v_r, v_l) \in \mathcal{H}(\Omega)$ is the sum of a pair of functions $(v_r^{(c)}, v_l^{(c)}) \in \mathcal{H}(\Omega)$, where both $v_r^{(c)}$ and $v_l^{(c)}$ are constant in z-direction, and a pair $(u_r^{(n)}, u_l^{(n)}) \in \mathcal{H}(\Omega)$, where $u_r^{(n)}$ and $u_l^{(n)}$ are not constant in z-direction and are normed such that $\int_0^L u_r^{(n)} - u_l^{(n)}\, dz$=0.

Viewing the two-wave eigenvalue problem from a physical perspective yields the idea for the construction of the *preconditioner*. The two-wave ansatz (99) in Chapter 3.3 reflects, that an eigenmode of a laser cavity can be seen as a superposition of two waves traveling in opposite directions. Bearing in mind the remarks on the time-harmonic representation (2) in Chapter 1.1 and the derivation in Chapter 3.3, we can say that in the BVP (159), (160), u_r is the wave moving in positive z-direction and u_l the wave moving in negative z-direction.

This interpretation of an eigenmode suggests to precondition the matrix eigenvalue problem (193) by an appropriate *block Gauss-Seidel iteration* (see for instance [61] or [67]) in the direction of the waves. We remark that by the stabilization described in Chapter 5.2 we obtain a corresponding (nearly upwind or downwind) discretization of the first order terms, see also [9].

However, in the governing equations (159), the first oder derivatives $+2ik_f\,\partial/\partial z$ and $-2ik_f\,\partial/\partial z$ dominate, if k_f is very large. It is clear, that the portion of (u_r, u_l) that is constant in z-direction (see the decomposition in Lemma 12) is smoothed very slowly by relaxation with the Gauss-Seidel method.

So, we use a preconditioner which consists of a relaxation M_c^{-1} of the constant portion of the residual in the subspaces $\mathcal{H}^{(c)}$ and a relaxation M_{GS}^{-1} of the residual in the whole space $\mathcal{H}$ by the Gauss-Seidel method.

The relaxation M_{GS}^{-1} is composed of four steps (as also depicted in Figure 19):

1. relaxation for u_r and u_l on boundary Γ_0,
2. plane-wise relaxation for u_r in the interior and on the radiating boundary from left to right,
3. relaxation for u_r and u_l on boundary Γ_1
4. and plane-wise relaxation for u_l in the interior and on the radiating boundary from right to left.

So, the preconditioner can be described by the matrix product

$$M^{-1} = (M_{GS}^{-1})^{l_2}\,(M_c^{-1})^{l_1},$$

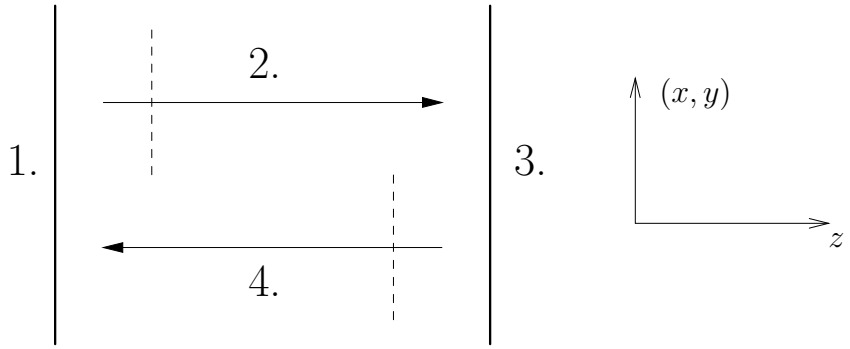

Figure 19: Part of the preconditioner for the two-wave eigenvalue problem.

where $l_1 \in \{0, 1\}$ and $l_2 \in \mathbb{N}$ determine how often the single relaxations are repeated.
With the help of this preconditioner the discrete equation could be solved satisfactorily with acceptable time and memory requirements.

7 Numerical Results for Different Cavity Configurations

In this chapter, we present numerical tests of our new approach, some of which have also been published in [7] or [9]. As it has been shown in Chapter 3, for a wide range of different configurations eigenmodes can be described using the two-wave eigenvalue model (104), (105) at the end of Chapter 3.3. The configurations are specified by the phase shifts at the reflecting end mirrors or at the interior interfaces, as explained in Chapter 3.4, and by the coefficient function $k(x, y, z)$ (see, e.g., (104), (105)). Furthermore, a transformation of the the cuboidal domain Ω can be applied, such that e.g. for curved end mirrors the eigenvalue problem actually is solved on a domain with curved faces.
The accuracy of the two-wave eigensolutions is shown by comparing them with the (analytical) gaussian modes (as mentioned in Chapter 3.2) in configurations where these are appropriate. Particularly, we have chosen configurations that in the majority of cases cannot be computed satisfactorily by the other current numerical methods as discussed in Chapter 3.2 or in the article [7].
Furthermore, in Section 7.6 we present numerical results for a problem which cannot be computed well by the gaussian mode algorithm and which show that the FEM result deviates from the gaussian results as expected.
Altogether, these numerical examples show the correctness and the capability of our new approach.
For all tests, we have chosen a carrier wave length λ with $\lambda = 2.0\ \mu$m; that means that the considered eigenmodes are in the infrared spectrum. The length L of the cavity lies between 0.1 mm and 22.0 mm.

7.1 Cavity with Planar End Mirrors and Parabolic Refractive Index Distribution

The first example concerns a cavity consisting of a real gaussian duct (see e.g. [64], Chap. 20) between planar end mirrors, i.e. a distribution of refractive index showing a parabolic profile perpendicular to the propagation direction z, but being independent of z. In detail, we used the following dimensions: distance between end mirrors $L = 0.1$ mm, width of computational domain $W = 0.2$ mm, and refractive index $n(x, y, z) = n_0 - (n_2 \cdot r^2)/2$, where we choose $n_0 = 1.0$. From a physical point of view, a refractive index $0 < n(x, y, z) < 1$ does not make sense. However, since the mode shapes do not depend on the absolute value of the refractive index n in Ω but on its variation, we used $n_0 = 1.0$ throughout all examples .
Using the two-wave ansatz, we performed the FEM computations on a domain discretized by $40 \times 40 \times 5$ elements ($\approx$ 10 000 grid points). For comparison, Table 2 shows the spot sizes w_{FEM} computed for three different values of n_2 by

the use of our FEM method and w_{G} computed by the gaussian mode algorithm, respectively.
Since the refractive index is independent of z in this configuration, the ansatz (99) shows that the phase shift is contained in the quantity ε for waves $\tilde{u}_r$ and $\tilde{u}_l$ that are constant in z. As for instance explained in [64], Chap. 16.6, for the lowest-order gaussian mode ε_{G} is given by $\varepsilon_{\mathrm{G}} = (n_2)^{1/2}$. Table 2 shows the good correspondence of (the analytical shift) ε_{G} with $\varepsilon_{\mathrm{FEM}}$ deduced from the computed eigenvalue $\xi_h (= 2\,\varepsilon_{\mathrm{FEM}}\,k_f)$.

n_2	w_{FEM} [μm]	w_{G} [μm]	$\varepsilon_{\mathrm{FEM}}$	$\varepsilon_{\mathrm{G}} = \sqrt{n_2}$
0.24	36.220	36.049	0.497	0.490
0.48	30.376	30.313	0.708	0.693
0.72	27.714	27.391	0.871	0.849

Table 2: Comparison between FEM and gaussian results for a gaussian duct with real refractive index.

7.2 Empty Cavity with Concave Mirror: Spot Size and Guoy Phase Shift

Here, we apply our method to an empty cavity of length $L = 1.0$ mm, whose left end mirror is concave with a (parabolic) radius of curvature $R_0 = 5.0$ mm, whereas the right mirror is planar. Figure 20 shows the spot size, as a function of z, of the lowest-order mode as obtained by our FEM code in comparison with the result of the gaussian algorithm. The width of the computational domain was $W = 0.2$ mm; in the computation $41 \times 41 \times 61$ ($\approx 102\,000$) nodes have been involved.
As mentioned in Chapter 3.5, the Guoy phase shift $\psi(z)$ can be computed by the use of the relation

$$\tilde{u}(0,0,z) = \exp\left[-i(\varepsilon z - \psi(z))\right] |\tilde{u}(0,0,z)|,$$

where $\tilde{u}(0,0,z)$ and ε are calculated from the numerically obtained eigenmode $\tilde{u}$ and eigenvalue $\xi (= 2k_f\varepsilon)$, respectively. Since $\psi(z)$ only is determined except for a constant, we set $\psi(L) = 0$ for this example. In Figure 21, the numerically computed phase shift $\psi_{\mathrm{FEM}}(z)$ and the gaussian phase shift $\psi(z) = \arctan(z/z_R)$ with Rayleigh range z_R (see for instance [64], Chap. 19.3) are plotted. The excellent agreement between both results shows that our method is able to predict even fine details exactly.

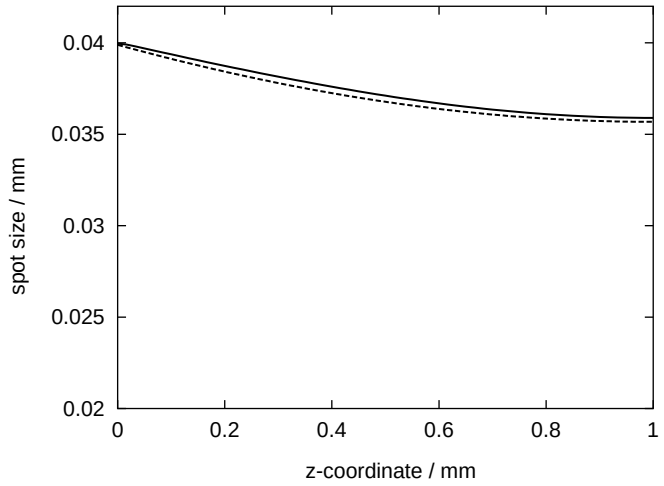

Figure 20: Empty cavity with one concave ($R_0 = 5.0$ mm) and one planar end mirror, FEM (solid curve) and gaussian mode shape (dashed curve).

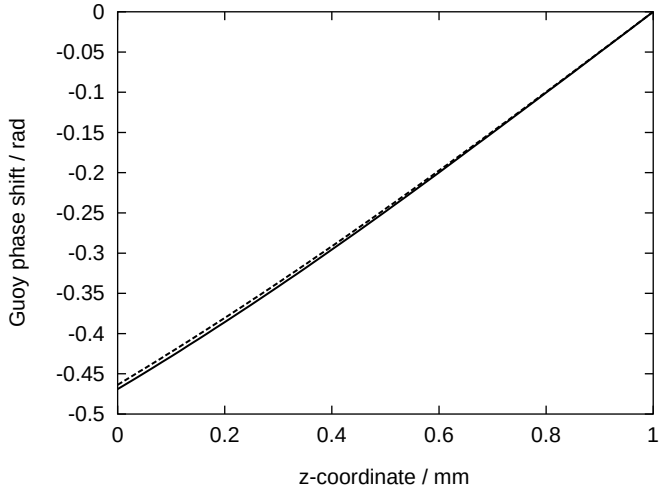

Figure 21: Guoy phase shift for an empty cavity with one concave ($R_0 = 5.0$ mm) and one planar end mirror, FEM (solid curve) and gaussian mode result (dashed curve).

7.3 Long Resonator Cavity with Focusing Element Near One End Mirror

To demonstrate that our method can reliably take into account focusing effects in longer cavities, we analyze in this example a cavity with planar end mirrors having a distance of $L = 10.0$ mm. Only a short section extending from the left mirror over a length of 1.0 mm is filled with a Gaussian duct. That means, more specifically, that the refractive index is defined by $n(x, y, z) = 1 - (n_2(z) \cdot r^2)/2$ with

$$n_2(z) = \begin{cases} 0.06 & \text{if } z \leq 1.0 \\ 0.0 & \text{else.} \end{cases}$$

Figure 22 shows that the z-dependent spot sizes of the lowest-order mode obtained by the use of FEM and by the gaussian code LASCAD$^{\text{TM}}$ [2, 5], respectively, nearly coincide. The FEM computations have been carried through on a domain with width $W = 0.4$ mm. The grid was made up by $20 \times 20 \times 150$ elements, which corresponds to approximately 70 000 points.
Figures 23 and 24 display the normalized intensity of the lowest-order and of one higher-order mode, respectively.

7.4 Gain Guiding Effect and A Geometrically Instable Configuration

By the following examples, we show that our new approach models gain guiding effects correctly. At first, we consider a cavity with planar end mirrors, length $L = 0.1$ mm and width $W = 1.2$ mm with the computational domain being discretized by $96 \times 96 \times 5$ elements. We choose a constant refractive index distribution $n(x, y, z) \equiv 1$ and a parabolic gain/loss distribution with $\alpha = \alpha_2 \cdot r^2$. (The notation corresponds to [64, Chap. 20.3].) In other words, we use the coefficient function

$$k^2 = k_0^2(1 - n_2 \cdot r^2) - i\,\alpha_2 \cdot k_0 r^2 \tag{201}$$

with $n_2 = 0$ and $k_0 := k_f$.
For different values of α_2, Table 3 shows the spot sizes w_{FEM} as obtained by the finite element calculations in comparison to the gaussian mode results w_{G}. As in the example in Section 7.1, the values nearly coincide.
Additionally, we compute two configurations where the left mirror is curved and the length of the domain is $L = 1.0$ mm.
In Figure 25, for a cavity with concave left end mirror ($R_0 = 100$ mm) and width $W = 0.4$ mm the spot size is plotted against the z-coordinate. The result, which was obtained on a grid of approximately 85 000 nodes, also agrees very well with the gaussian mode spot size.
A configuration with planar right and convex left end mirror with radius of curvature $R_0 = -100.0$ mm, is known to be geometrically instable, as for instance

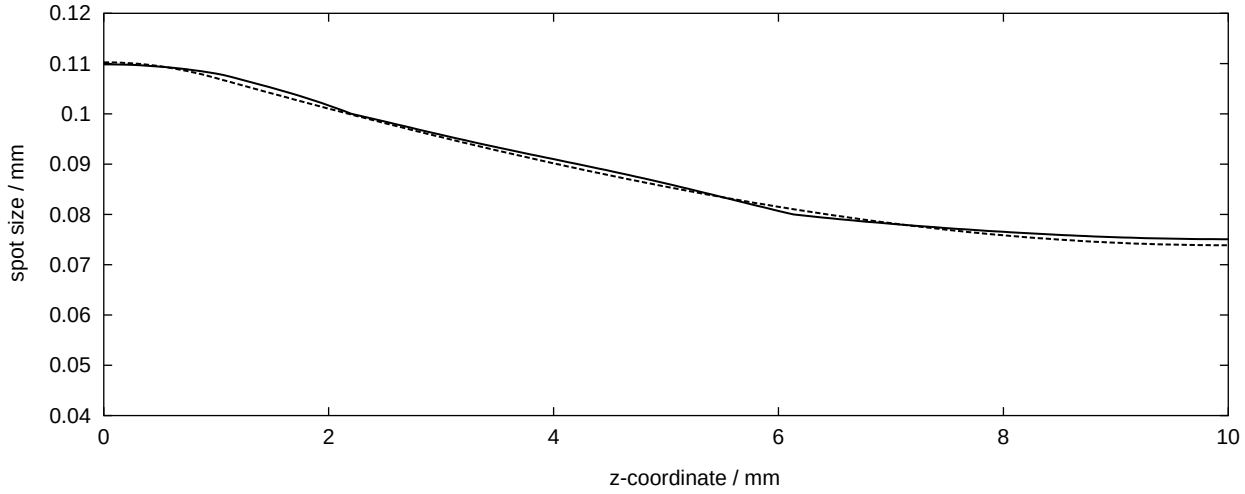

Figure 22: Comparison between FEM (solid curve) and gaussian results (dashed curve) for a long resonator with a short gaussian duct attached to the left mirror.

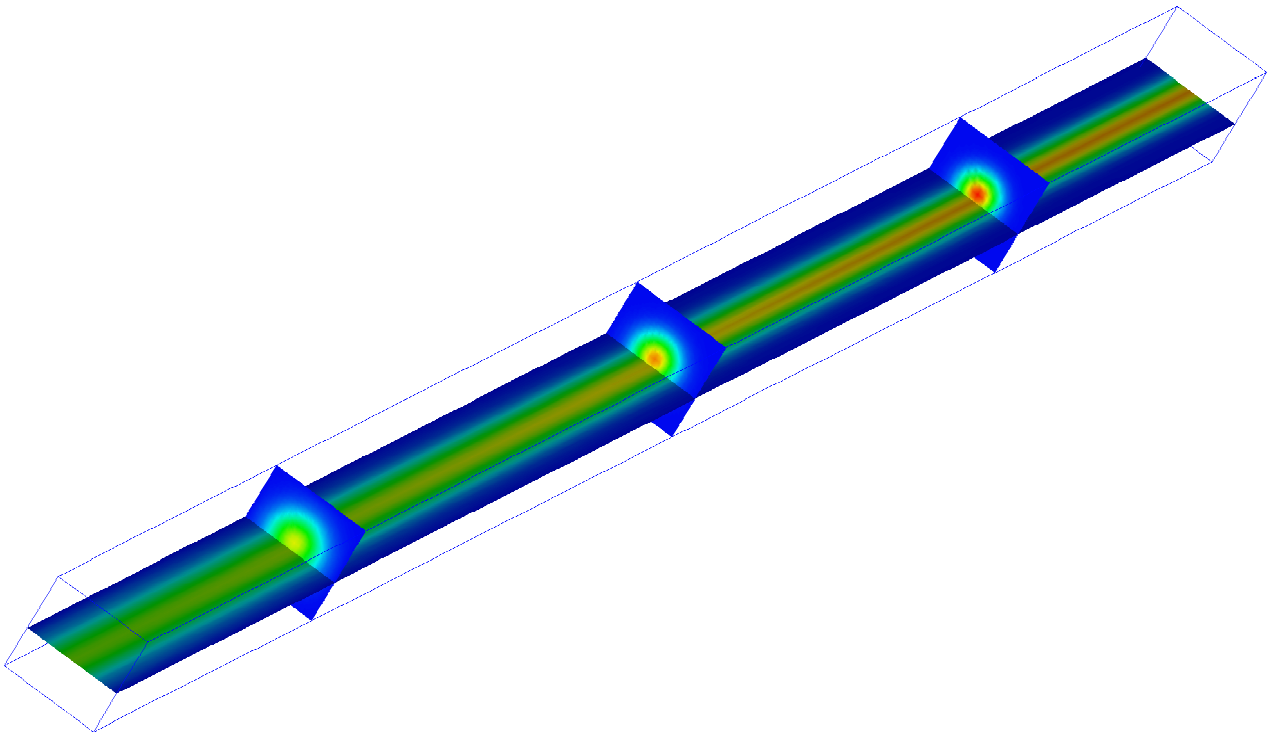

Figure 23: Lowest-order or TEM_{00}-mode in a long resonator.

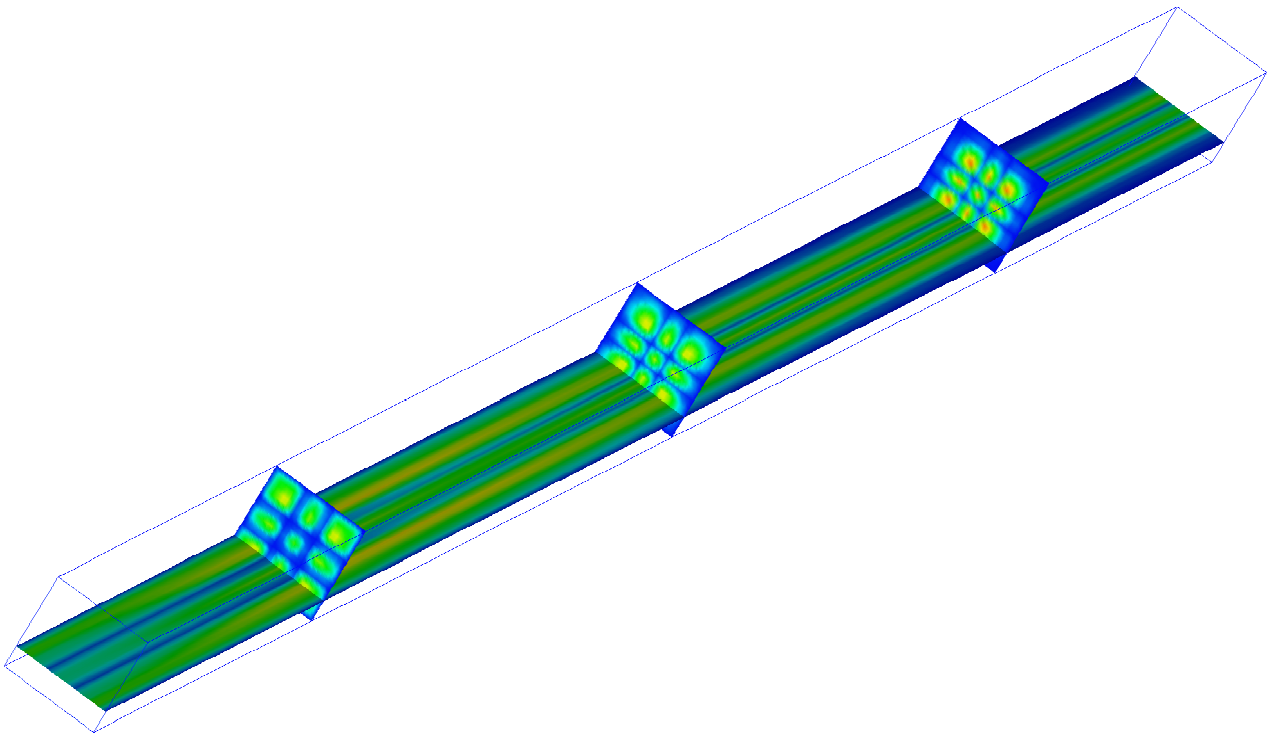

Figure 24: TEM_{22}-mode in a long resonator.

α_2	w_{G} [μm]	w_{FEM} [μm]
1.	224.64	224.72
10.	126.32	126.63
40.	89.32	89.85

Table 3: Comparison between FEM and gaussian results for a gaussian duct with planar end mirrors taking into account gain.

explained in [64], Chapters 19, 21, and 22. In practice, such lasers are quite sensitive because the eigenmodes heavily depend on the gain distribution. Figure 26 depicts the spot size of the fundamental mode for a gain distribution with $\alpha_2 = 10$. Again, the accuracy of our FEM approach is demonstrated by the comparison to the corresponding gaussian mode. Here, the computations were performed on a domain of width $W = 0.8$ mm, which has been discretized by $80 \times 80 \times 50$ elements ($\approx 335\,000$ grid points).

7.5 Splitting of Round-Trip Mode Shape in an Instable Cavity due to Gain

Since in our approach an eigenmode is represented by two waves, configurations can be dealt with where a so-called splitting occurs, i.e. the wave $\tilde{u}_r$ propagating to the right and the wave $\tilde{u}_l$ propagating to the left in the resonator (see the ansatz (99)) have different mode shapes.
To demonstrate this, we consider a cavity of length $L = 10.0$ mm with planar end mirrors and with a short gaussian duct attached to the left mirror. The refractive index and gain distribution is given by the coefficient function (201) with $n_2(z)$ and $\alpha_2(z)$ defined as:

$$n_2(z) := \begin{cases} -0.005 & \text{if } 0 \le z \le 2 \\ 0.0 & \text{if } 2 < z \le 10 \end{cases}$$

and

$$\alpha_2(z) := \begin{cases} 40.0 & \text{if } 0 \le z \le 2 \\ 0.0 & \text{if } 2 < z \le 10. \end{cases}$$

Due to the negative refractive index parameter n_2, this configuration is instable if no gain is taken into account.
We performed the computation on a domain of width $W = 1.0$ mm, discretized by $50 \times 50 \times 100$ elements, which is approximately $260\,000$ grid points. In the Figures 27 and 28 the gaussian mode spot size and the spot size of the finite element solution are compared, namely for $\tilde{u}_r$ and $\tilde{u}_l$, respectively, whereas Figure 29 depicts the spot sizes of the finite element solution in one graph.

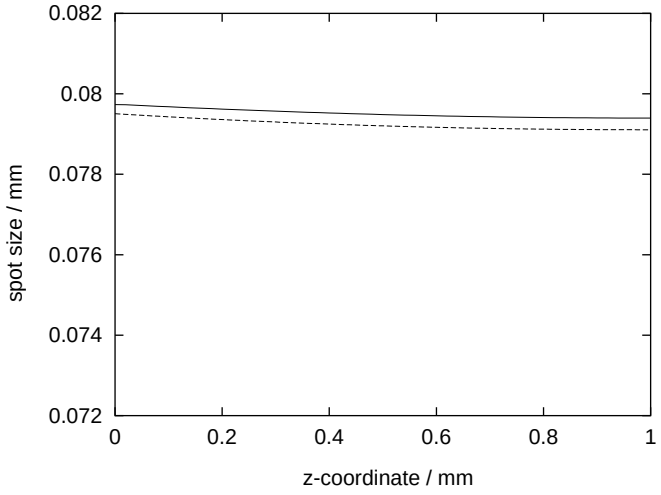

Figure 25: Comparison between FEM (solid curve) and gaussian results (dashed curve) for a cavity with concave left and planar right end mirror with gain.

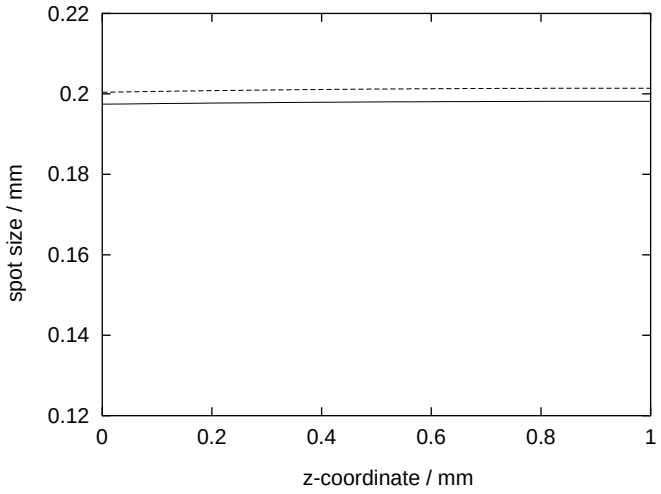

Figure 26: Comparison between FEM (solid curve) and gaussian results (dashed curve) for a geometrically instable resonator with convex left and planar right end mirror with gain.

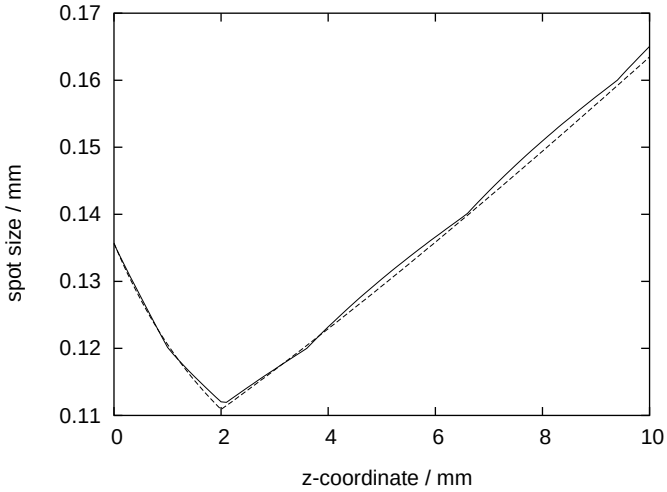

Figure 27: Comparison between FEM (solid curve) and gaussian spot size (dashed curve) of the wave traveling to the right in a configuration with gain where a splitting occurs.

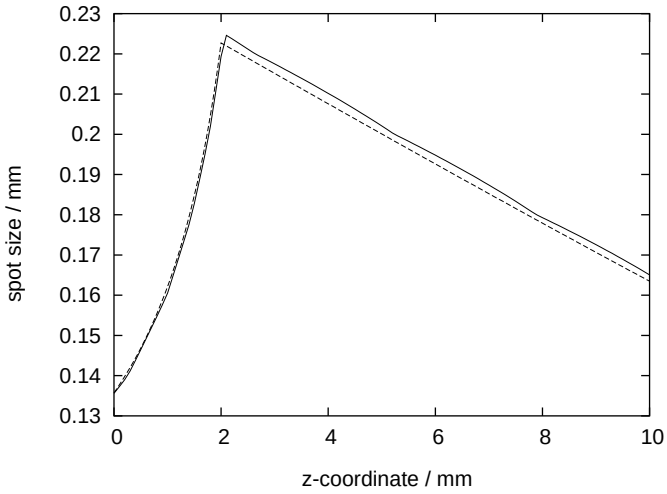

Figure 28: Comparison between FEM (solid curve) and gaussian spot size (dashed curve) of the wave traveling to the left in a configuration with gain where a splitting occurs.

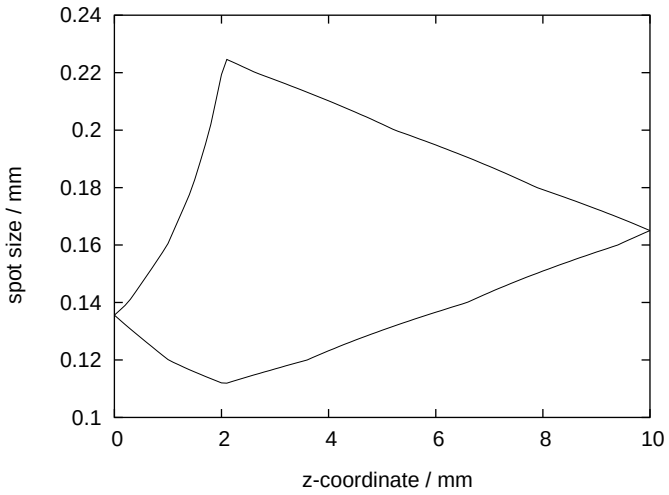

Figure 29: Spot sizes of split eigenmode obtained by the finite element approach.

7.6 Thermal Effects in a Monolithic Laser

In this example we use our method to model a monolithic diode-pumped solid state laser consisting of an end pumped crystal (with material parameters of a Nd:YAG crystal except for the refractive index parameter n_0, see comment on that in Section 7.1), whose deformed end faces represent the end mirrors of the cavity. To take into account thermal lensing due to the temperature dependence of the refractive index and due to thermal distortion a thermal and structural finite element analysis has been carried through by the use of LASCAD$^{\text{TM}}$ [2]. The data obtained in this way for the temperature distribution and deformation of the crystal have been imported into our program. We used a rectangular slab of equal height and width $W = 0.8$ mm and length $L = 8.0$ mm cooled from top and bottom, but not from left and right. The obtained temperature distribution therefore deviates strongly from rotational symmetry. Figure 30 shows the thermally induced refractive index distribution on the whole slab and Figure 31 shows the distribution along x- and y-axis immediately behind the entrance plane of the pump beam (at $z = 0$) as obtained with LASCAD$^{\text{TM}}$ [2]. To compute the mode shape by the use of a gaussian approximation as implemented in LASCAD$^{\text{TM}}$ [2], the refractive index distribution is fitted parabolically for a series of cross sections along the z-axis as also shown in Figure 31 for a cross section close to $z = 0$. The obtained parabolic coefficients are used in a round trip ABCD matrix to compute the mode shape. Figures 32 and 33 show the spot sizes along the z-axis obtained in this way in comparison with the spot sizes obtained by our new approach, that uses the full 3D data of the thermal and structural finite element analysis without parabolic fit. For our computations we used $80 \times 80 \times 32$ elements which equates to approximately 250 000 grid points. As one can see, the results are very close to each other in the y-z-plane, whereas

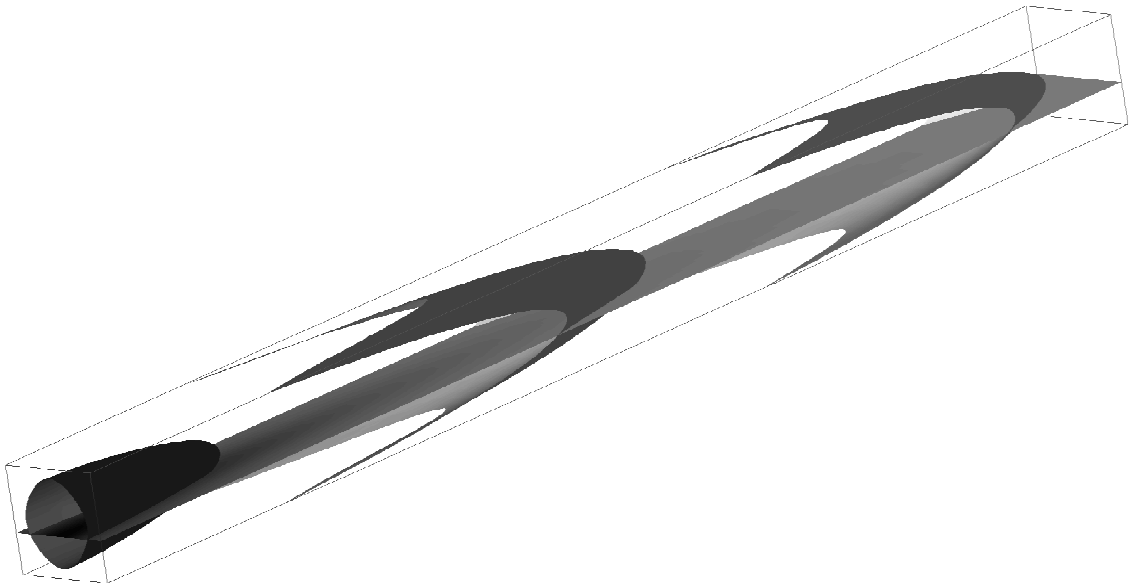

Figure 30: Slice and isosurfaces of a refractive index distribution based on numerical data imported from LASCAD^{TM}.

in x-z-plane the spot size obtained by our new approach is considerably larger. This is expected from the fact that the parabolic fit shown in Fig. 31 is good along the y-axis, but very poor along the x-axis for which the plot shows a bell shaped distribution.
Accordingly, for the transverse mode profile the deviation between the gaussian profile and the result of the 3D approach also is much stronger along the x-axis than along the y-axis as shown in Figures 34 and 35.

7.7 Oscillating Beam in a Gaussian Duct

The last numerical example concerns a resonator of length $L = 22.0$ mm with planar right end mirror and concave left end mirror ($R_0 = 3.0$ mm). The cavity is filled with a long complex gaussian duct of length 20.0 mm attached to the right end mirror. That means we have following parameters for the refractive index and gain distribution:

$$n_2(z) = \begin{cases} 0.0 & \text{if } 0 \leq z \leq 2 \\ 0.0885 & \text{if } 2 < z \leq 22 \end{cases}$$

and

$$\alpha_2(z) = \begin{cases} 0.0 & \text{if } 0 \leq z \leq 2 \\ 40.0 & \text{if } 2 < z \leq 10. \end{cases}$$

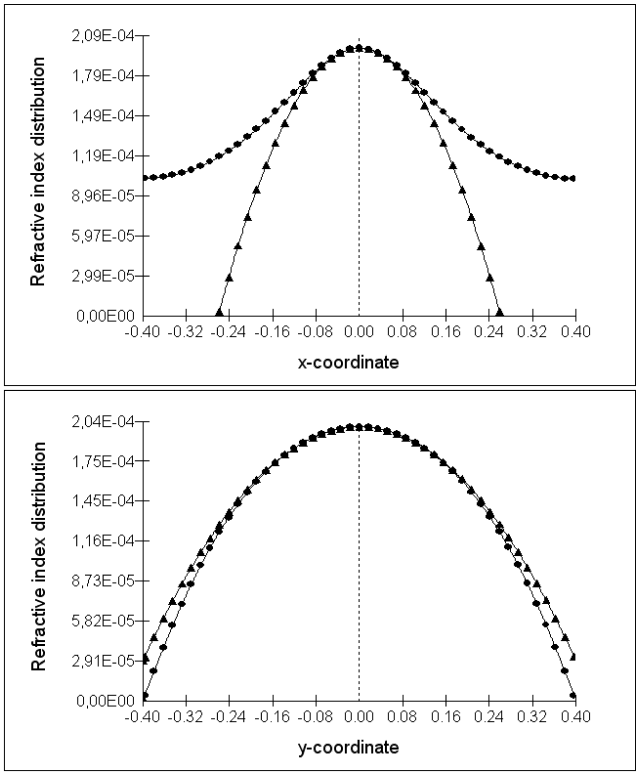

Figure 31: Comparison of numerical (dots) and parabolically fitted (triangles) refractive index. Screen shot of LASCAD$^{\text{TM}}$.

Due to the concavity of the left mirror and due to the gain eigenmodes occur which are oscillating in the long duct and exhibit a splitting as in the example in Section 7.5. To our knowledge, until now these oscillating eigenmodes could not be computed by the existing full numerical methods presented in Chapter 3.2. For the computations on the domain with width $W = 0.4$ mm a grid of $61 \times 61 \times 353$ ($\approx 1\,300\,000$) nodes was used. Figure 36 depicts the spot sizes of the finite element and gaussian mode result, respectively, for the wave moving to the right and Figure 37 for the wave moving to the left. The spot sizes of the two-wave solution are plotted in Figure 38.

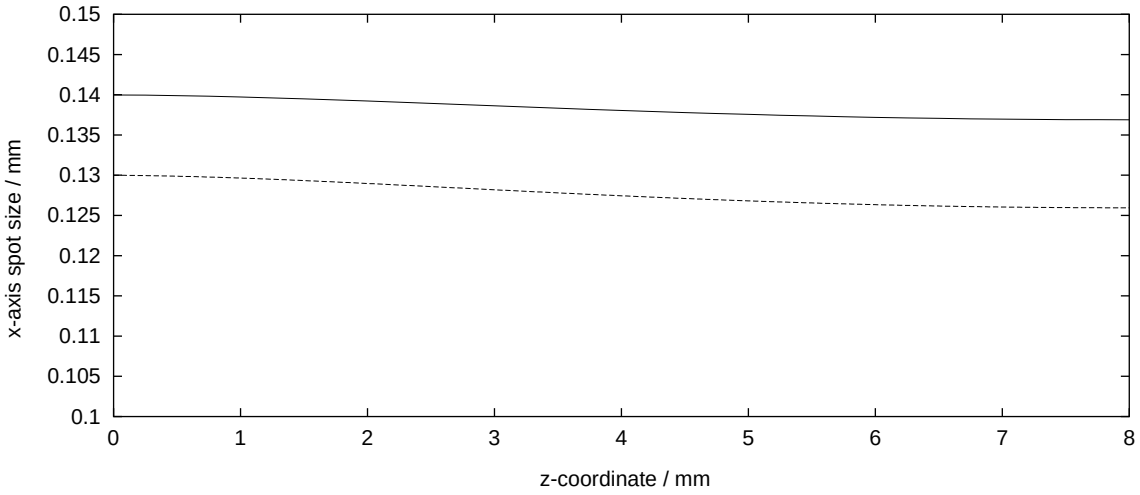

Figure 32: Comparison of FEM (solid curve) and gaussian x-axis spot size (dashed curve) along cavity axis for a monolithic laser with thermal effects taken into account.

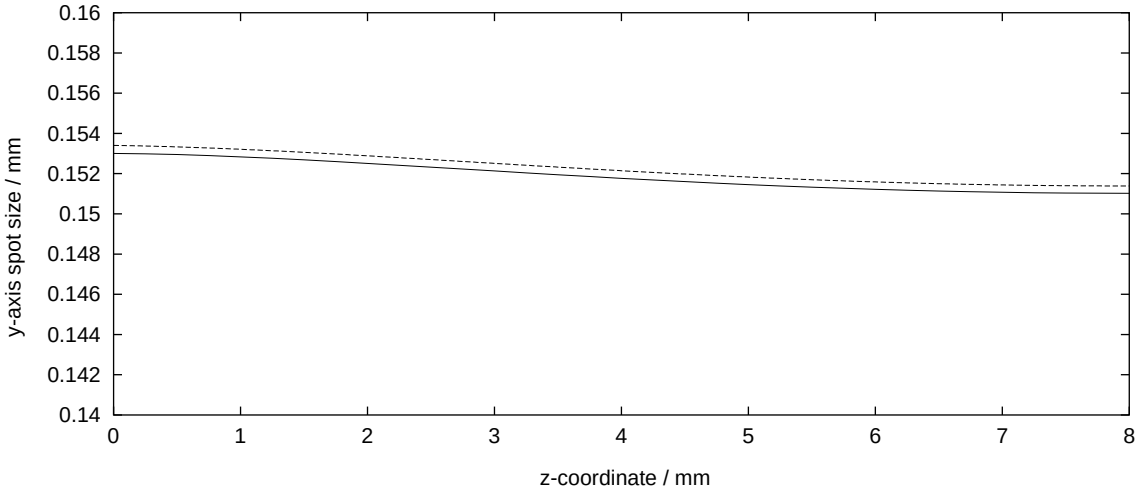

Figure 33: Comparison of FEM (solid curve) and gaussian y-axis spot size (dashed curve) along cavity axis for a monolithic laser with thermal effects taken into account.

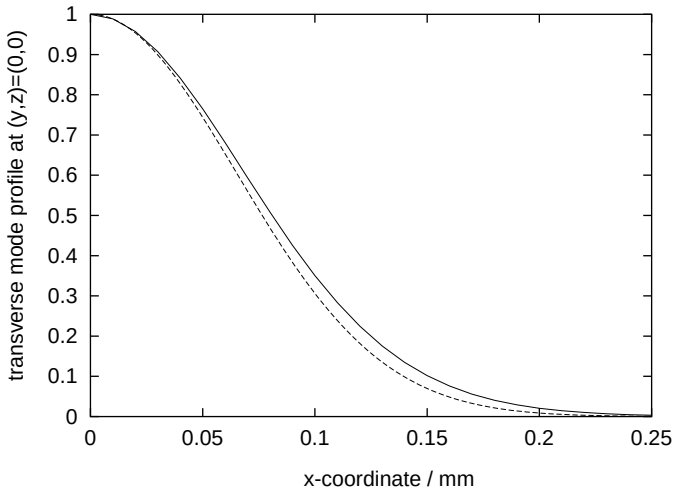

Figure 34: Transverse mode profile along x-axis for a monolithic laser with thermal effects taken into account, FEM (solid curve) and gaussian mode result (dashed curve).

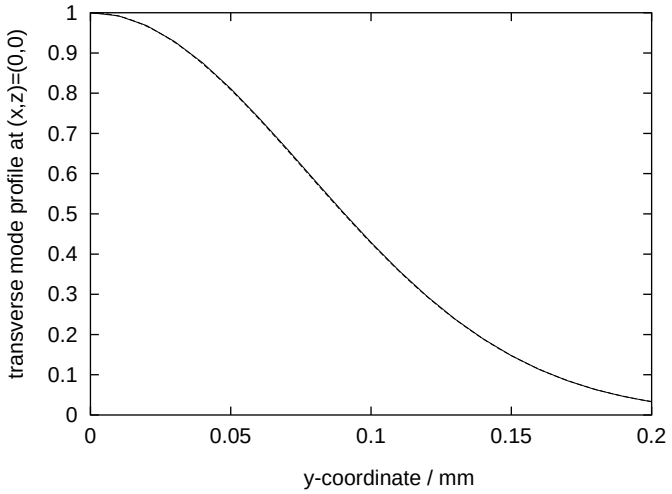

Figure 35: Coinciding transverse mode profiles along y-axis for a monolithic laser with thermal effects taken into account, FEM (solid curve) and gaussian mode result (dashed curve).

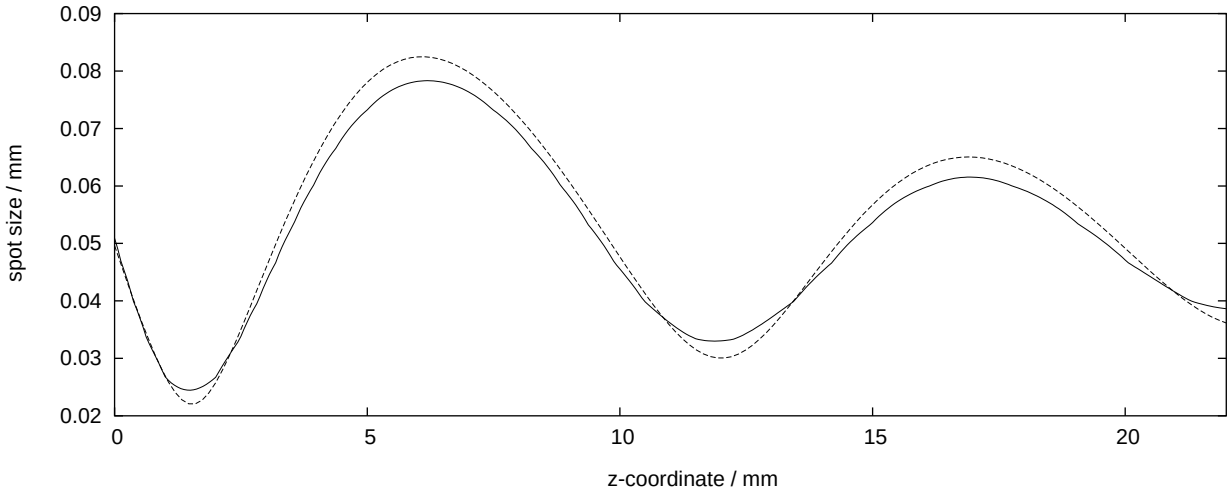

Figure 36: Comparison between FEM (solid curve) and gaussian spot size (dashed curve) of the oscillating wave traveling to the right in a long duct.

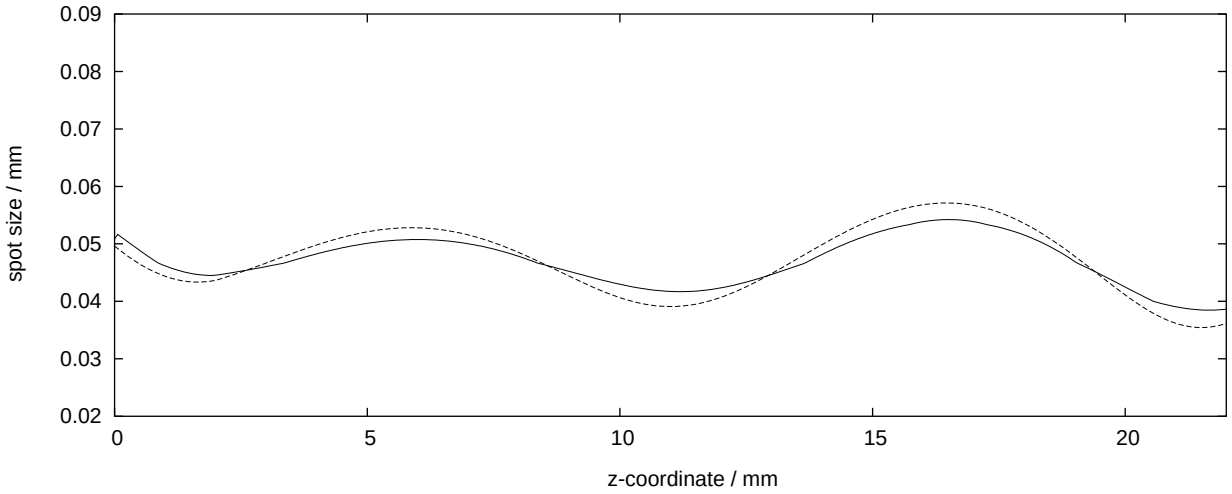

Figure 37: Comparison between FEM (solid curve) and gaussian spot size (dashed curve) of the oscillating wave traveling to the left in a long duct.

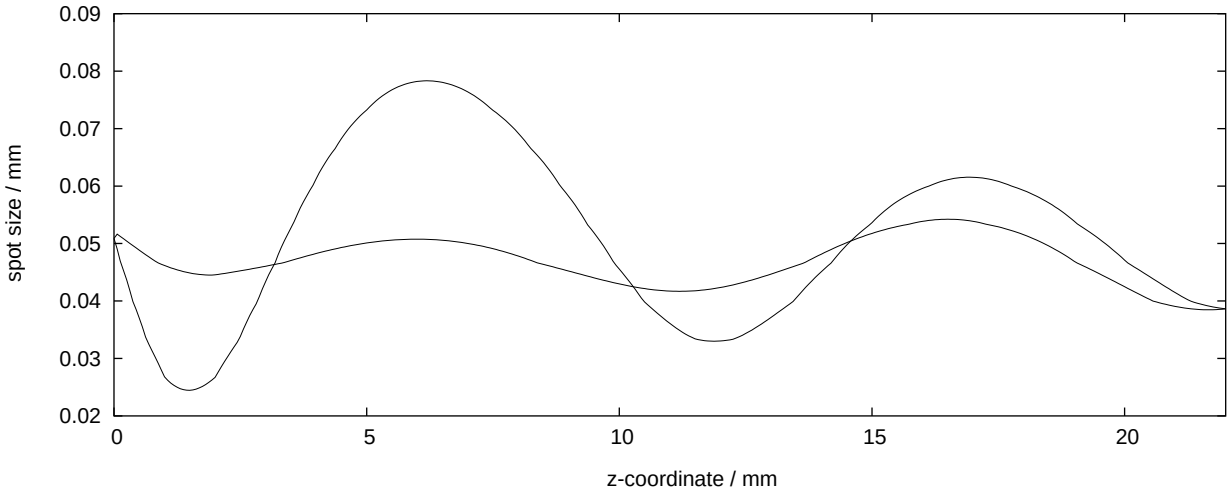

Figure 38: Spot sizes of oscillating eigenmode obtained by the finite element approach.

Conclusions

The numerical results in Chapter 7 demonstrate very clearly the accuracy of the numerical eigenmodes obtained by our two-wave model. Since this model is based on a partial differential equation eigenvalue problem, it can be applied to a wide range of laser cavity configurations as the examples show. By appropriate modifications of the model, it should therefore also be possible to solve more complicated configurations.
From a mathematical point of view the convergence proof in Chapter 4 can conceivably be extended to show convergence for finite element solutions of equations with a more complex nonlinearity. However, that was not the scope of this dissertation.
Since some of the results in this thesis are a first step toward answering problems in mathematics and physics that have not been investigated in detail until now, there remain many open questions concerning mathematical and computational aspects of our new approach and of the existing numerical methods for analyzing laser cavities. We mention just a few of them:

- Which further properties do the solutions and the finite element approximations of the transformed equation in Chapter 2.1 have?
- In what cases can the beam propagation methods be proved to fail? Which influence do the boundary conditions have (Chapter 3.2)?
- What does the spectrum of the two-wave eigenvalue problem look like (Chapters 3.3 and 5.1)?
- In what sense do the stabilizations in Chapter 5.2 and Chapter 2.2.4 change the character of the original equations?

This short overview contains a lot of interesting questions the answers of which can improve our finite element analysis and can lead to a deeper understanding of the existing methods for analyzing laser cavities.
In conclusion, it may be hoped that the success of the two-wave finite element analysis of laser cavities developed in this thesis will be an impetus for researchers and laser designers to rely on numerical simulation more than they have in the past.

References

[1] Expde. `http://ifamus.mathematik.uni-wuerzburg.de/~expde`.

[2] LASCAD^{TM}. `http://www.las-cad.com`.

[3] R. A. Adams. *Sobolev Spaces*. Academic Press, London, 1975.

[4] H. W. Alt. *Lineare Funktionalanalysis*. Springer, Berlin, Heidelberg, 1992.

[5] K. Altmann. Simulation software tackles design of laser resonators. *Laser Focus World*, 36(5):293–294, 2000.

[6] K. Altmann. Personal communication, April – August 2003.

[7] K. Altmann, C. Pflaum, and D. Seider. Third-dimensional finite element computation of laser cavity eigenmodes. *Appl. Opt.*, 43(9):1892–1901, 2004.

[8] K. Altmann, C. Pflaum, and D. Seider. Three-dimensional computation of laser cavity eigenmodes by the use of finite element analysis (FEA). In *SPIE Proc., Photonics West, Symposion LASE 2004, Conference 5333- Laser Resonators and Beam Control VII*, number 5333-16, 2004.

[9] K. Altmann, C. Pflaum, and D. Seider. Modeling and computation of laser cavity eigenmodes. In *SPIE Proc., Photonics Europe 2004, Conference 5460 - Solid State Lasers and Amplifiers*, number 5460-26, 2004, to appear.

[10] I. Babuška and J. Osborn. Eigenvalue Problems. In P.G. Ciarlet and J.L. Lions, editors, *Handbook of numerical analysis. Volume II: Finite element methods (Part 1)*, pages 641–790. North-Holland, Amsterdam, 1991.

[11] I. Babuška and S. Sauter. Is the pollution effect of the FEM avoidable for the Helmholtz equation considering high wave numbers? *SIAM REVIEW*, 42(3):451–484, 2000.

[12] Z. Bai, D. Day, and Q. Ye. ABLE: An adaptive block Lanczos method for non-hermitian eigenvalue problems. *SIAM J. Matr. Anal. Appl.*, 20(4):1060–1082, 1999.

[13] C. Beattie. Galerkin eigenvector approximations. *Math. Comp.*, 69(232):1409–1434, 2000.

[14] A. Bermúdez, R.G. Duán, R. Rodríguez, and J. Solomin. Finite element analysis of a quadratic eigenvalue problem arising in dissipative acoustics. *SIAM J. Numer. Anal.*, 38(1):267–291, 2000.

[15] D. Boffi, F. Brezzi, and L. Gastaldi. On the convergence of eigenvalues for mixed formulations. *Ann. Scuola Norm. Sup. Pisa Cl. Sci.*, XXV:131–154, 1997.

[16] A. Brandt and A. Livshits. Wave-ray multigrid method for standing wave equations. *Electron. Tans. Numer. Anal.*, 6:162–181, 1997.

[17] S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods*. Springer, New York, Berlin, Heidelberg, 1994.

[18] F. Chatelin. *Eigenvalues of matrices*. John Wiley & Sons, Chichester, 1993.

[19] S. Coarsi, P. Fernandes, and M. Raffetto. On the convergence of Galerkin finite element approximations of electromagnetic eigenproblems. *SIAM J. Numer. Anal.*, 38(2):580–607, 2000.

[20] A. Deraemaeker, I. Babuška, and P. Bouillard. Dispersion and pollution of the FEM solution for the Helmholtz equation in one, two and three dimensions. *Int. J. Numer. Meth. Engng.*, 46:471–499, 1999.

[21] P. Deufelhard, T. Friese, and F. Schmidt. A nonlinear multigrid eigenproblem solver for the complex Helmholtz equation. Preprint SC 97-55, Konrad-Zuse-Zentrum für Informationstechnik Berlin, 1997.

[22] M. Dobrowolski. Vorlesungsskript: Finite Elemente, 1999. Available under: `http://ifamus.mathematik.uni-wuerzburg.de/~dobro/pub/finel.ps`.

[23] Z. Bai e.a. *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*. SIAM, Philadelphia, 2000.

[24] H. Elman, O. Ernst, and D. O'Leary. A multigrid method enhanced by Krylov subspace iteration for discrete Helmholtz equations. *SIAM J. Sci. Comp.*, 23(4):1291–1315, 2001.

[25] H. Elman and D. O'Leary. Efficient iterative solution of the three-dimensional Helmholtz equation. *J. Comput. Phys.*, 142(1):163–181, 1998.

[26] H. Elman and D. O'Leary. Eigenanalysis of some preconditioned Helmholtz problems. *Numer. Math.*, 83:231–257, 1999.

[27] M. Embree. How descriptive are GMRES convergence bounds? Numer. Anal. Report 99/08, Computing Laboratory, Oxford University, June 1999. `http://web.comlab.ox.ac.uk/oucl/publications/natr/na-99-08.html`.

[28] C. Farhat, A. Macedo, and M. Lesoinne. A two-level domain decomposition method for the iterative solution of high frequency exterior Helmholtz problems. *Numer. Math.*, 85:283–308, 2000.

[29] M. D. Feit and J.A. Fleck, Jr. Spectral approach to optical resonator theory. *Appl. Opt.*, 20(16):2843–2851, 1981.

[30] A. G. Fox and T. Li. Resonant modes in a maser interferometer. *Bell Sys. Tech. J.*, 40:453–458, 1961.

[31] L.P. Franca, C. Farhat, A.P. Macedo, and M. Lesoinne. Residual-free bubbles for the Helmholtz equation. *Int. J. Numer. Meth. Engng.*, 40:4003–4009, 1997.

[32] J. Franklin. *Methods of Mathematical Economics*. Springer, Berlin, Heidelberg, New York, 1980.

[33] M. Gander, F. Magoulès, and F. Nataf. Optimized Schwarz methods without overlap for the Helmholtz equation. *SIAM J. Sci. Comput.*, 24(1):38–60, 2002.

[34] M. Gander and F. Nataf. AILU for Helmholtz problems: A new preconditioner based on an analytic factorization. *C. R. Acad. Sci., Paris, Sr. I, Math.*, 331(3):261–266, 2000.

[35] D. Gilbarg and N. S. Trudinger. *Elliptic Partial Differential Equations of Second Order*. Springer, Berlin, Heidelberg, New York, 1977.

[36] Anne Greenbaum. *Iterative Methods for Solving Linear Systems*, volume 17 of *Frontiers in Applied Mathematics*. SIAM, Philadelphia, 1997.

[37] Ch. Großmann and H.-G. Roos. *Numerik partieller Differentialgleichungen*. Teubner, Stuttgart, 1992.

[38] W. Hackbusch. *Theorie und Numerik elliptischer Differentialgleichungen*. Teubner, Stuttgart, 2nd edition, 1996.

[39] I. Harari and T.J.R. Hughes. Finite element methods for the Helmholtz equation in an exterior domain: Model problems. *Comput. Methods Appl. Mech. Eng.*, 87:59–96, 1991.

[40] V. Heuveline and R. Rannacher. A posteriori error control for finite element approximations of elliptic eigenvalue problems. *Adv. Comput. Math.*, 15(1-4):107–138, 2001.

[41] F. Ihlenburg. *Finite Element Analysis of Acoustic Scattering*, volume 132 of *Springer Series Applied Mathematical Sciences*. Springer, New York, Paris, London, 1998.

[42] R. Kechroud, A. Soulaimani, and Y. Saad. Preconditionning techniques for the solution of the Helmholtz equation by the finite element method. Technical Report umsi-2003-40, Minnesota Supercomputer Institute, University of Minnesota, 2003.

[43] J.B. Keller and D. Givoli. Exact non-reflecting boundary conditions. *J. Comp. Phys.*, 82:172–192, 1989.

[44] S. Kim and S. Kim. Multigrid simulation for high-frequency solutions of the Helmholtz problem in heterogeneous media. *SIAM J. Sci. Comput*, 24(2):684–701, 2002.

[45] P. Knabner and L. Angermann. *Numerik partieller Differentialgleichungen*. Springer, Berlin, 2000.

[46] W. G. Kolata. Eigenvalue approximation by the finite element method: The method of Lagrange multipliers. *Math. Comput.*, 33:63–76, 1997.

[47] P. Lancester and M. Tismentetsky. *The Theory of Matrices*. Academic Press, San Diego, 1985.

[48] B. Lee, T. Manteuffel, S. McCormick, and J. Ruge. First-order system least-squares for the Helmholtz equation. *SIAM J. Sci. Comput.*, 21(5):1927–1949, 2000.

[49] R.B. Lehoucq, D.C. Sorensen, and C. Yang. *ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*. SIAM, Philadelphia, 1998.

[50] M. M. Monga Made. Incomplete factorization-based preconditionings for solving the Helmholtz equation. *Int. J. Numer. Meth. Engng.*, 50:1077–1101, 2001.

[51] M. M. Monga Made, R. Beauwens, and G. Warzée. Preconditioning of discrete Helmholtz operators perturbed by a diagonal complex matrix. *Commun. Numer. Meth. Engng.*, 16:801–817, 2000.

[52] N. Meyers and J. Serrin. $H = W$. *Proc. Nat. Acad. Sci. USA*, 51:1055–1056, 1964.

[53] K. W. Morton. *Numerical Solution of Convection-Diffusion Problems*. Number 12 in Applied Mathematics and Mathematical Computation. Chapman & Hall, London, New York, 1996.

[54] K. Otto and E. Larsson. Iterative solution of the Helmholtz equation by a second-order method. *SIAM J. Matrix Anal. Appl.*, 21(1):209–229, 1999.

[55] C. Pflaum. *Diskretisierung elliptischer Differentialgleichungen mit dünnen Gittern*. Dissertation, Technische Universität München, 1995.

[56] C. Pflaum. On the regularity of elliptic differential equations using symmetry techniques and suitable discrete spaces. *Electron. J. Differ. Equ.*, 1996(11):1–9, 1996. `http://ejde.math.unt.edu`.

[57] C. Pflaum. Expression templates for partial differential equations. *Computing and Visualization in Science*, 4(1):1–8, 2001.

[58] R.E. Plessix and W.A. Mulder. Separation-of-variables as a preconditioner for an iterative Helmholtz solver. *Appl. Numer. Math.*, 44(3):385–400, 2003.

[59] H.-G. Roos, M. Stynes, and L. Tobiska. *Numerical Methods for Singularly Perturbed Differential Equations*. Springer Series in Computational Mathematics. Springer, Berlin, Heidelberg, 1996.

[60] Y. Saad. *Numerical methods for large eigenvalue problems*. Algorithms and architectures for advanced scientific computing. Univ. Press e.a., Manchester, 1992.

[61] Y. Saad. *Iterative methods for sparse linear systems*. SIAM, Philadelphia, 2nd edition, 2003.

[62] Y. Saad and M. H. Schultz. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Stat. Comput.*, 7:856–869, 1986.

[63] H.R. Schwarz. *Methode der finiten Elemente*. Teubner, Stuttgart, 3rd edition, 1991.

[64] A. E. Siegman. *Lasers*. University Science Books, Mill Valley, 1986.

[65] A. E. Siegman and H. Y. Miller. Unstable optical resonator loss calculation using the Prony method. *Appl. Opt.*, 9(10):2729–2736, 1970.

[66] G. Sleijpen and H. Van der Vorst. A Jacobi-Davidson iteration method for linear eigenvalue problems. *SIAM Review*, 42(2):267–293, 2000.

[67] J. Stoer and R. Bulirsch. *Introduction to Numerical Analysis*. Springer, New York, Heidelberg, Berlin, 2nd edition, 1993.

[68] F. Stopp. *Operatorenrechnung*. Verlag Harri Deutsch, Thun, Frankfurt/Main, 2nd edition, 1978.

[69] M. Streiff, A. Witzig, and W. Fichtner. Computing optical modes for VCSEL device simulation. *IEE Proc. Optoelectron.*, 149:166–173, 2002.

[70] F. Tisseur and K. Meerbergen. The quadratic eigenvalue problem. *SIAM Review*, 43(2):235–286, 2001.

[71] M. Vanmaele and R. Van Keer. Convergence and error estimates for a finite element method with numerical quadrature for a second order elliptic eigenvalue problem. In J. Albrecht, L. Collatz, P. Hagedorn, and W. Velte, editors, *Numerical treatment of eigenvalue problems. Vol. 5. Workshop in Oberwolfach, Germany, February 25-March 3, 1990.*, volume 96 of *International Series of Numerical Mathematics*, pages 225–236. Birkhäuser Verlag, Basel, 1991.

[72] J.H. Wilkinson. *The Algebraic Eigenvalue Problem*. Clarendon Press, Oxford, 1967.

[73] J. Wloka. *Partielle Differentialgleichungen*. Teubner, Stuttgart, 1982.

[74] J. Xu and A. Zhou. A two-grid discretization scheme for eigenvalue problems. *Math. Comp.*, 70(233):17–25, 2001.

[75] J. Yosida. *Functional Analysis*. Classics in Mathematics. Springer, New York, 1998.