Unterschrift der Gutachter

.....

.....

D I S S E R T A T I O N

High-order finite element analysis of the Helmholtz equation and its application in laser physics

Ausgeführt zum Zwecke der Erlangung des akademischen Grades
einer Doktorin der technischen Wissenschaften unter der Leitung von

Univ. Prof. Jens Markus Melenk, PhD
E101 - Institut für Analysis und Scientific Computing

und

Univ. Prof. Dipl.-Ing. Dr.techn. Stefan Rotter
E136 - Institut für Theoretische Physik

eingereicht an der Technischen Universität Wien
Fakultät für Mathematik und Geoinformation

von

Mag. rer. nat. Sofi Esterhazy
Matrikelnummer: 0106175
Krottenbachstraße 29/9, 1190 Wien

Wien, am 13. September 2013

.....
Sofi Esterhazy

Kurzfassung

Diese Dissertation verbindet Methoden der Grundlagenforschung in der Mathematik mit der anwendungsorientierten Forschung auf dem Gebiet der Laserphysik. Dementsprechend ist diese Arbeit in zwei Teile gegliedert.

Im Mittelpunkt steht die numerische Analyse der Helmholtz-Gleichung mit ausstrahlenden Randbedingungen, die für die Modellierung von Phänomenen wie akustische oder elektromagnetische stehende Wellen im freien Raum verwendet wird. Als geeignetes Verfahren zur numerischen Berechnung und Simulation dieses Problems wird hier die Finite-Elemente-Methode (FEM) eingesetzt. Dieses Verfahren leidet jedoch bei hohen Wellenzahlen unter numerischen Dispersionsfehlern. Im Rahmen der mathematischen Grundlagenforschung liegt daher ein besonderer Augenmerk auf der expliziten Abhängigkeit der Konvergenz hinsichtlich der Wellenzahl sowie der Diskretisierungsparameter. Grundlegende Untersuchungen über Regularitätseigenschaften, geometrische Aspekte sowie eine Konvergenz-Analyse in Bezug auf unterschiedliche Normen werden daher im Detail im ersten Teil beschrieben.

Darüber hinaus wird die numerische Untersuchung erweitert/angewendet auf das Gebiet der Laserphysik. Während Laserlicht eine Form von räumlich und zeitlich kohärenten, elektromagnetischen Wellen darstellt, müssen jedoch zusätzliche Effekte, die von der Interaktion mit dem Verstärkungsmaterial kommen, berücksichtigt werden. Die wesentlichen Merkmale eines Lasers werden somit durch ein gekoppeltes System nichtlinearer Helmholtz-Gleichungen beschrieben. Hier lag der Fokus darauf eine effiziente und flexible Lösungsmethode zu finden, um eine breite Palette von experimentellen Anwendungen simulieren zu können. Zu diesem Zweck wurde die Finite-Elemente-Methode erneut für die Diskretisierung verwendet. Darüber hinaus wurde eine stabile und effiziente Iterationsstrategie erarbeitet, die sich im Wesentlichen auf die Lösung eines Systems von nichtlinearen elliptischen PDEs und die Berechnung eines nichtlinearen Eigenwertproblems reduziert. Das nicht-lineare System wird durch eine stabilisierte Form mit der Newton-Methode gelöst. Für die Berechnung des nichtlinearen Eigenwertproblem erwies sich die Konturintegralmethode als zweckdienlich und den Anforderungen aus der Physik am Besten genügend.

Abstract

This thesis forms a bridge between basic research in mathematics and applied research in the field of laser physics, structured accordingly in two parts.

The central focus is the numerical analysis of the Helmholtz equation with radiating boundary condition which is used for modeling of phenomena such as acoustic or electromagnetic standing waves in free space. As a convenient method for the numerical calculations and simulation of this problem, the finite element method (FEM) is used. This method suffers, however, from numerical dispersion errors when increasing the wave number. The explicit dependence on the wave number and the discretization parameters is therefore under a particular focus in the context of basic research. Basic investigations concerning regularity properties, geometric aspects, as well as a convergence analysis with respect to different norms are therefore discussed in detail in the first part.

In addition to that, the numerical study is extended/applied to the field of laser physics. While laser light constitutes a form of spatially and temporally coherent electromagnetic waves, major effects coming from the interaction with the gain material, however, have to be taken into account. The significant characteristics of a laser are thus described by a coupled system of nonlinear Helmholtz-type equations. Here, the focus was to find an efficient and flexible solution method in order to be able to simulate a wide range of experimental applications. To this end, the finite element method has again been used for the discretization. Furthermore, a stable and efficient iteration strategy has been built up which essentially reduces to the solution of a set of nonlinear elliptic PDEs and the computation of a nonlinear eigenvalue problem. The non-linear system is solved by a stabilized form with the Newton method. For the calculation of the nonlinear eigenvalue problem the contour integral method was proved to be useful, satisfying the requirements from physics.

Acknowledgements

In first place I want to thank my supervisor Jens Markus Melenk for giving me the opportunity to accomplish this work and become specialized in the field of finite element methods. With his patience and kindness he introduced me to the field of high order finite element methods through many inspiring and motivating discussions. Here, I also want to thank his former students Markus Aurada for an enriching and professional exchange.

I also particularly like to express my gratitude to Stefan Rotter as my second supervisor who gave me the possibility to work in an interdisciplinary environment. He kindly integrated me into his work group and introduced me to the very interesting field of laser physics. He also gave me the chance to meet many experts in physics as well as in mathematics. I also want to express my joy to the members of his research group who welcomed and integrated me in a very friendly way in their group activities. Special thanks goes to Matthias Liertzer for his help in technical and programming questions as well as his patience when introducing me into the the physical connections in the laser theory. Also I want to thank Dmitry Krimer and Stephan Burkhardt for their input from a physical point of view.

I also want to express my gratitude to Prof. Schöberl, parts of my numerical computations are based on his FEM code NETGEN/NGSOLVE, as well as his group members. Their doors were always open to me when technical issues occurred (specially in the beginning). Here I would also like to especially thank Lothar Nannen for many enlightening and professional discussions.

I also want to thank the members of the research groups of Dirk Praetorius, Ansgar Jüngel and Anton Arnold. Academic life would not have been so much fun as it was without you.

At last I want to give many many thanks to my big big family for their endless support through these last years and their unbreakable faith in my abilities.

Financial support under the graduate school "PDE-Tech" of the Vienna University of Technology and by the Vienna Science and Technology Fund (WWTF) through Project No. MA09-030 "Light coupling to light: nonlinear interactions in semiconductor micro-

lasers" is gratefully acknowledged.

Contents

1. Introduction

1.1. Motivation and overview

1.1.1. Physical motivation

One of the most notable and challenging phenomena in science is the propagation of waves. Many important examples from acoustics, electrodynamics, radar/sonar detection or even medical imaging involve time-harmonic waves also known as stationary or standing waves. The underlying equation to describe these is the Helmholtz equation named after Hermann von Helmholtz, a German polymath of the 19th century [82]. The range of applications of this equation is also reflected on the mathematical level. In fact, this equation can be derived as a specialized case from some major equations in physics. The classical straightforward motivation arises when the boundary value problem governed by the wave equation

$$\frac{\partial^2 w}{\partial t^2} - \Delta w = g$$

has a time periodic inhomogeneity $g(x,t) = f(x)e^{ikt}$ and the solution gets the form $w(x,t) = u(x)e^{ikt}$ so that the (complex) amplitude u solves the Helmholtz equation

$$-\Delta u - k^2 u = f.$$

Likewise the Helmholtz equation can be derived from the stationary Schrödinger equation with free potential or from the time-harmonic Maxwell equations in a charge-free space. The latter describes a wide range of harmonic electromagnetic waves covering the spectrum from gamma rays to X-rays, visible light, microwaves, radio waves and long waves.

Of particular interest in this work is the application to laser light. In that case the propagation of the electromagnetic field is still given by the classic Maxwell equation, but in addition quantum mechanical effects have to be taken into account to describe the interaction of the field with the gain medium. To be more precise we assume that the gain medium can be modeled by an ensemble of identical two-level atoms such that the dynamical properties can be described by the Bloch equations. Together they lead to the Maxwell-Bloch (MB) equations which form the center of the semi-classical laser theory [54]. The MB equations are a set of time-dependent coupled nonlinear equations and are typically difficult to solve analytically. For the case of steady-state lasing a much more efficient theory for solving multi-periodic solutions of the MB equations was found in 2006 in the form of the steady-state ab-initio lasing theory (SALT) [45, 96, 98]. In

one dimension (as well as for transverse magnetic modes in two dimensions) the resulting equations are a set of Helmholtz equations with an additional nonlinear coupling term. The study of such a nonlinear problem mostly relies on properties of a linearized problem. However an extension of the results of the linear Helmholtz equation to the SALT equations has not been accomplished yet in a rigorous way. For the transition from linear to nonlinear problems we refer to [18].

1.1.2. Challenges in numerics and numerical analysis

For both, linear and nonlinear problems, numerical challenges occur on two levels. The first one is to find an appropriate discretization, here by using the finite element method, and the second one is to use a reliable method for the solution of the resulting system of linear or nonlinear algebraic equations. Hereby the study of linear problems is also helpful in the nonlinear case as it often gives a first insight for the nonlinear ones. Moreover, many iterative solution methods for the nonlinear problems are generally based upon the corresponding linearized problems. In fact, the linearization of the SALT equations take the form of a set of Helmholtz equations which are well understood from a mathematical/analytical point of view.

The linear Helmholtz equation is a basic equation for treating wave propagation problems in a time-harmonic setting. In many high frequency situations with a large wave number k the solution u is highly oscillatory. Then the conditions of the discretization for a numerical computation are stringent due to the requirement to resolve the oscillatory nature of the solution. The general "rule of thumb" for the classic discretization which is mostly used in engineering claims the condition $hk \sim 1$ to be satisfied in order to obtain a feasible resolution. This can be motivated as follows: The number of resolution points per wave length is obviously related to the discrete mesh by $n_{res} = \lambda/h$ with λ being the wave length and h the mesh size. Furthermore the wave length is related to the wave number by $\lambda = 2\pi/k$. Hence we have $hk = 2\pi/n_{res}$. The minimal resolution of a wave requires at least $n_{res} = 2$ point per wave length (as stated by Nyquist [79]), thus $hk = 2\pi/2 \sim 3$. A trustworthy resolution would be obtained by $n_{res} = 10$ which means $hk \sim 0.6$. For higher dimensions the minimal resolution would require fine meshes with at least $N = k^d$ degrees of freedom, where d is the spatial dimension. At least this is true for the standard interpolation. But more striking/obtrusive is that the classical discretization of the Helmholtz equation suffers strongly from dispersion errors, i.e. the exact wave number differs from the discrete wave number.

On the level of numerical analysis, it is clearly of interest to provide an error analysis which is explicit in the discretization parameters as well as in the wave number k. We focus here on the finite element method [19, 22, 28, 33], more specifically on the the *hp*-version of finite element methods (*hp*-FEM) as it turns out that high order methods provide a significantly better convergence behavior for large wave numbers and better numerical dispersion properties.

In principle one can follow two different strategies to understand the convergence behavior of (FEM-)discretizations.

One common way is to decompose the FEM-error into the interpolation error and the so called pollution error (the error between the interpolant of the exact solution and the FEM-solution)

$$\|u - u_{FEM}\| \leq \|u - u_I\| + \|u_I - u_{FEM}\|.$$

In fact, for the classical discretizations of the Helmholtz equation it is the pollution error that dominates the error in a large regime of mesh sizes h. This pollution effect refers to the fact that this part of the error is of non-local nature (see [58, Remark 4.5]). Furthermore this behavior appears numerically when increasing the wave number k with a constant number of degrees of freedom. Then the numerical solution becomes less and less accurate. While the interpolation error obeys the rule of thumb mentioned above, the pollution error is small, if additionally by hk^2 is small. The main reason for the pollution error is that the finite element solution suffers from *dispersion* errors. The analysis of these dispersion errors, being of a global nature, is not easy. A classical approach is to study (infinite) translation invariant meshes so that tools from Fourier analysis can be used [3, 34, 58]. This analysis reveals for the Galerkin method that its solutions feature a phase lead. A similar observation concerning the dispersion error has been made for spectral methods [4], producing a phase lag instead. The recent proposal of [5] shows in particular that a suitable combination of the Galerkin FEM and the spectral element method (SEM) can lead to new methods with significantly reduced dispersion errors.

Another strategy with the aim to be applicable for general, not necessarily translation invariant meshes, originates from the question under which condition the approximation error is quasi-optimal, i.e. it is asymptotically as good as the best approximation up to a constant factor

$$\|u - u_{FEM}\| \leq C \inf_v \|u - v\|.$$

It is now of interest how C depends on discretization parameters and the wave number k. This technique reduces the error analysis essentially to two separate questions, namely, the best approximation and the stability, which in the present case, can be obtained in terms of the adjoint solution operator of the Helmholtz problem [84, 87]; this is due to the fact that the Helmholtz problems can be treated as a coercive problem with an additional compact perturbation. Throughout this approach the *regularity theory* for the Helmholtz problem (and its adjoint problem) becomes a crucial ingredient to yield optimal error estimates, explicit in the wave number and the discretization parameters. A k-explicit convergence analysis based on this approach has been discussed by J. M. Melenk and S. Sauter in free space [69] and on analytic, bounded domains as well as for convex polygons [72], giving rise to the resolution conditions

$$\frac{hk}{p} \leq c_1 \quad \text{and} \quad p \geq c_2 \log(k),$$

where c_1 sufficiently small and c_2 sufficiently large and additional mesh refinements for non-smooth geometries (e.g. polygons). For these problem classes the two mentioned articles [69, 72] gave a rigorous proof of the observation that high order methods are

superior to low order methods.

Another numerical challenge comes from the fact that many physical examples take place in free space. However, numerical computations can only be provided for bounded domains. Thus it is a major task to find appropriate methods for the approximation on an unbounded domain, including a proper translation of the radiation conditions onto the artificial boundary. This topic is not part of the main focus of this work, but for completeness some options are given in condensed form.

The good properties of the high order methods for the Helmholtz problems strongly suggest to apply them to nonlinear problems as those appearing in the SALT. So far, the relevant system of PDEs there has been solved using the corresponding integral formulation with a finite difference discretization [45, 96, 98]. A solution strategy based on PDE methods was still pending until this work. Roughly speaking, the nonlinear problem is computed with the Newton method while the initial data is obtained by solving the SALT equations as a nonlinear eigenvalue problem (NEVP). Thus, besides the discretization error from the FEM approximation, it is also the accuracy of the Newton solver as well as the eigenvalue solver that have an impact on the quality of the solution.

The Newton method has been studied by a large number of authors and is now well understood, converging quadratically if started with an initial guess close to the real solution. Several Newton variants have been derived and proved to be numerically stable. One of the most common schemes is the Newton-Raphson method that we are using here for vector-valued functions.

However, there is more of an issue when it comes to choice of the eigensolver, especially for eigenvalue problems which are nonlinear in the eigenvalue. The interest in an efficient solver for such NEVPs has increased dramatically in the recent years. At the same time, the currently available solution packages are not as numerous and as developed as for the linear case. A major class of NEVPs are those where the eigenvalues appear polynomially. The classical approach for solving such eigenvalue problems is to reduce the problem to a generalized linear eigenvalue problem. In the context of eigenvalue problems this practice is known by the term of *linearization*. Generally, for linear problems one distinguishes between solvers for small, densely populated matrices and large, sparse matrices; FEM matrices belong to the latter. There exists a large number of linearization techniques depending on the properties of the original matrices (real valued, complex valued, symmetric, Hermitian, invertible, etc.) However, this practice becomes easily inconvenient as it increases the problem size in dependence on the polynomial degree and is mostly not able to preserve the sparsity of the original matrices (here the FEM-matrices) to sufficient extent. In addition, often only a particular part of the spectrum is of interest. All these aspects should be reflected and exploited in the choice of the right eigensolver.

1.2. Objectives and contributions

As part of the PhD program "PDEtech" the objective of this thesis was to develop scientific principles of applied mathematics and to investigate applications in an interdisciplinary environment. The latter came to an expression in the field of laser physics as part of the WWTF-project "Light coupling to light" .

The central focus of this dissertation, under the guidance of Prof. Melenk, deals with the numerical and analytic treatment of the Helmholtz equation which describes time-harmonic wave phenomena such as those found in acoustics and electromagnetics. For the numerical simulation of these elliptic partial differential equation the finite element method (FEM) of higher order is used for the discretization. In order to ensure the quality of that method we provide the numerical analysis in terms of an *a priori* convergence analysis. In this context, this work extends in particular on the results made by J. M. Melenk and S. Sauter in [69,72]. There, the analysis was performed in the H^1-like norm $\|\cdot\|_{\mathcal{H}}$ for the free space, analytic domains and convex polygons. We were able to extend these results for arbitrary polygons which have been published by the author and J. M. Melenk in [40]. In addition, for the case of analytic domains, we focused on weaker norms, namely the convergence in the L^2- and the H^{-1}-norm (published in [39]). While the asymptotic convergence rates are, of course, the ones to be expected, the novel aspect of [**?**] over [40, 68, 69, 72] is that we obtain better estimates in the wave number k. Indeed, we obtain an *a priori* bound that is better by a factor k than what a straightforward application of [40,68,69,72] would yield. In addition, the analytic results are corroborated and illustrated with numerical calculations in 1D with MATLAB and in 2D with the *hp*-FEM code NETGEN/NGSOLVE by J. Schöberl, [88,89]. As an additional aspect, we looked at the convergence with respect to the above mentioned dispersion errors. There, we restricted ourselves to the case of one dimension on translation invariant meshes. In that case, the discrete formulation of the Green's function and powerful tools such as Fourier techniques are available to understand and analyze the discretizations and to design new schemes with good dispersion properties. While the recent proposal of [5] concentrates on the presentation of a constructive numerical scheme and the dispersion analysis, we directly focus on the numerical error analysis in dependence of the dispersion error. For a 1D model problem on regular grids, we provide an actual error analysis for the lowest order discretization and show that the greatly reduced dispersion error of the method [5] leads to a gain in accuracy by a factor k as compared with the Galerkin FEM. Again, these results have been highlighted with computational examples made with MATLAB. This topic was also published by the author and J. M. Melenk in [39].

In further cooperation with Prof. Rotter and his group from the Institute for Theoretical Physics it was our intention to realize an implementation of a 1D-FEM code for use in laser physics. In particular, it was our aim to establish an algorithm that covers a wide range of experimental settings, from standard configurations like the 1D slab

cavity to more elaborate configurations from micro-disk lasers, photonic crystal lasers and random lasers including an active control of the pump spatial distribution. The underlying model is described by a set of non-linear Helmholtz-type PDEs with radiating boundary conditions. We were able to implement an efficient and stable Newton solver by exploiting the analytic expression of the corresponding Jacobi matrix. The corresponding nonlinear eigenvalue problem that has to be solved is computed using the contour integral method. All in all, we have created a feasible solution strategy that has a high degree of flexibility and can be expanded to higher dimensions in a reasonable way. The obtained results are collected in [38].

1.3. Outline

The remainder of this thesis is organized in two parts.

The first part focuses on the linear Helmholtz equation and is divided into five chapters.

In Chapter 2, we first introduce the Helmholtz model in free space with radiating boundary conditions as it is the origin of many applications in physics. Since numerical computations can only be established for bounded domains, we review how to truncate the free space and briefly motivate some ways how the boundary conditions at infinity can be transferred to the artificial boundary. Then we focus on the Helmholtz problem on bounded domains with the Robin boundary condition and formulate some properties and results which are mandatory for a unique solvability.

Chapter 3 is dedicated to the regularity theory of the Helmholtz problem. We introduce some geometric assumptions for analytic and polygonal domains as well as the proper function spaces. Furthermore, we discuss the concept of polynomial wellposedness and review some properties for the high and low frequency filtered data. Finally we present a refined regularity theory for arbitrary polygons by decomposition which generalized the results of [72] as mentioned above. The chapter ends with the description of a few more regularity properties by exploiting additional regularity of the source function.

Chapter 4 gives an introduction to the requirements for numerical approximation. First we briefly review the concept of Galerkin discretization. Then we give some details on the high order finite element method, including properties of the triangulation, the finite element space and the quadrature rules. In addition, we briefly describe the spectral element method.

In Chapter 5, we concentrate on the convergence analysis based on the regularity theory given in Chapter 3. Therefore we review the development of quasi-optimality which reduces the error analysis to the task of finding estimates for the best approximation and the adjoint approximability. These are handled in detail for domains with analytic boundaries. Then quasi-optimality is summed up for polygons too and the *hp*-convergence is discussed for polygons, retrieving the exponential rate of convergence on geometrically refined triangulations. For smooth domains with analytic boundaries the convergence theory is depicted in several norms for problems with data of higher

regularity. In addition, we consider the h-convergence for the latter setting, highlighting the poor convergence rate of low order methods for high wave numbers.

In Chapter 6 we focus on the convergence analysis with respect to the dispersion. We recall the dispersive nature of the Galerkin finite element method as well as of the spectral element method and study a suitable combination of both which leads to a significant reduction in the dispersion error. For this blended method, which was introduced in [5], we present an error analysis that takes into account the dispersion rate explicitly. Finally, the theoretical results are confirmed with some numerical examples.

The second part of this thesis is dedicated to the application in laser physics, organized in six chapters.

In Chapter 7, we start with some background information on the physical mechanism of lasers and introduce the describing model of the steady-state ab-initio laser theory in a mathematical setting.

In Chapter 8, we present the direct solution strategy of the coupled system of the nonlinear SALT equations introduced in Chapter 7. In particular, we describe two different ways of solving these equations. The first algorithm serves for the solution of the SALT equations when one is interested in the 'evolution' of the solutions, as the external pump is modulated. The other algorithm is used to solve the SALT equations when one is only interested in the solutions for a given pumping strength and can also be used to accelerate the process.

Chapter 9 concentrates on the properties of the Newton solver for the solution of the nonlinear model. First the standard Newton scheme is described in a formal way and then extended by some additional stability constraints. In addition, the Jacobian matrix is used explicitly and derived in all details.

Chapter 10 focuses on the computation of the nonlinear eigenvalue problem. Under the wide and versatile range of this topic three different solution methods are discussed. The first solution method is a standard technique for polynomial eigenvalue problems known as the first companion linearization. The second approach is a technique that is specialized on the solution of rational eigenvalue problems. The third and most suitable solver for our problem presented here is the contour integral method, [16].

In Chapter 11, many different numerical computations are presented in order to illustrate the reliability of our new direct SALT solver. Therefore we first tested the accuracy of all relevant components of our solver, i.e. the eigensolver, the Newton solver and the convergence rate of the FEM discretization for that nonlinear case. At last we sum up a few cases of applications that represent the versatility of our direct solver.

In Chapter 12, we give a brief outlook on extensions that might be of broader interest and other interesting questions that have not been answered so far.

Part I.

High order FEM for the linear Helmholtz model

2. The model problem

2.1. The Helmholtz equation with radiating boundary condition

The mathematical description of propagating waves is often governed by the linear time-dependent wave equation

$$\frac{\partial^2 w}{\partial t^2} - \frac{1}{c^2}\Delta w = g,$$

where $w = w(x,t), g = g(x,t)$ are real-valued functions of space and time and c describes the wave speed (e.g. speed of light). In the case of time-harmonic electromagnetic waves, the solution as well as the data on the right hand side (*rhs*) are assumed to be of the form $w(t,x) = \Re(u(x)e^{-i\omega t})$ resp. $g = e^{-i\omega t} f$ where ω is the angular frequency and $u(x), f(x)$ are the complex amplitude of the unknown wave function resp. source function. The resulting equation for the unknown variable u is then the Helmholtz equation

$$-\Delta u - k^2 u = f,$$

where k is the wave number, related to the wave length by $k = \omega/c = 2\pi/\lambda$. In order to solve this equation uniquely, we need to specify a boundary condition at infinity. Typically this condition describes some radiating propagation. Here we postulate strictly outgoing behavior. The mathematical expression for this far-field condition is best described by the Sommerfeld condition

$$\partial_{|x|} u - iku = o(\|x\|^{\frac{d-1}{2}}), \ \|x\| \to \infty.$$

In consequence, we formulate the Helmholtz problem in the full space:

Definition 2.1.1. *The Helmholtz problem in the full space $\mathbb{R}^d, d = 1,2,3$, with Sommerfeld condition is given by*

$$\begin{array}{rcll} -\Delta u - k^2 u & = & f & \text{in } \mathbb{R}^d \\ \lim_{r\to\infty} r^{\frac{d-1}{2}} (\partial_r u - iku) & = & 0 & \end{array} \tag{2.1}$$

Here, ∂_r denotes the derivative in radial direction $r := x/\|x\|$. Additionally, we assume f to have local support, i.e. there exists a d-ball $B_R := \{x \in \mathbb{R}^d | \ \|x\| < R\}$ that satisfies $supp(f) \subset B_R$ and denote by

$$S_k^{\mathbb{R}^d} : f \mapsto u \tag{2.2}$$

the solution operator of (2.1).

In terms of operator theory, the solution operator $S_k^{\mathbb{R}^d}$ is also known as the Newton potential and for $f \in L^2(\mathbb{R}^d)$ with compact support the solution can be expressed in terms of the Green's function G_k by

$$u(x) = S_k^{\mathbb{R}^d}(f) = G_k \star f = \int_{\mathbb{R}^d} G_k(x-y) f(y) dy \quad \forall x \in \mathbb{R}^d \tag{2.3}$$

where the Green's function satisfies the fundamental problem

$$\begin{array}{rcl} -\Delta G_k(x-y) - k^2 G_k(x-y) & = & \delta(x-y) \quad \forall x, y \in \mathbb{R}^d \\ \lim_{|x-y|\to\infty} G_k(x-y) & = & 0. \end{array} \tag{2.4}$$

In particular, the Green's function can be given explicitly by $G_k(z) = g_k(\|z\|)$, where

$$g_k(z) := \begin{cases} -\frac{e^{ikz}}{2ik} & d = 1 \\ \frac{i}{4} H_0^{(1)}(kz) & d = 2 \\ \frac{e^{ikz}}{4\pi z} & d = 3. \end{cases} \tag{2.5}$$

and $H_0^{(1)}$ denotes the first kind Hankel function of order zero [1].

In order to study the solutions of the model problem (2.1) we have to analyze some mapping properties of the solution operator $S_k^{\mathbb{R}^d}$. To that end, the following decomposition result is known:

Lemma 2.1.2 ([69, Lemma 3.5]). *For every $f \in L^2(\mathbb{R}^d)$ with $supp(f) \subset\subset B_R$ there holds*

$$k^{-1}\|S_k^{\mathbb{R}^d}(f)\|_{H^2(B_R)} + \|S_k^{\mathbb{R}^d}(f)\|_{H^1(B_R)} + k\|S_k^{\mathbb{R}^d}(f)\|_{L^2(B_R)} \leq C_R \|f\|_{L^2(\mathbb{R}^d)}.$$

Here, the Sobolev spaces H^s are defined in a standard way [2, 95].

In order to treat the unbounded problem (2.1) numerically, we have to truncate the full space to a bounded domain $\Omega \subset \mathbb{R}^d$. Studying such truncated boundary value problems, we will have to make use of the restriction $u|_{\partial\Omega}$ as an element of a Sobolev space on $\partial\Omega$. Therefore we require the following result [51, 66]:

Lemma 2.1.3 (trace operator). *Let Ω be a bounded Lipschitz domain and $1/2 < s < 3/2$. Then there exists a unique bounded linear trace mapping $\gamma_0 : H^s(\Omega) \to H^{s-1/2}(\partial\Omega)$ defined by $\gamma_0(u) := u|_{\partial\Omega}$ such that*

$$\|\gamma_0 u\|_{H^{s-1/2}(\partial\Omega)} \leq C_{tr} \|u\|_{H^s(\Omega)}. \tag{2.6}$$

The operator γ_0 is then called the trace operator *and it holds in particular for $s = 1$*

$$\|\gamma_0 u\|_{H^{1/2}(\partial\Omega)} \leq C_{tr} \|u\|_{H^1(\Omega)} \tag{2.7}$$

$$\|\gamma_0 u\|_{L^2(\partial\Omega)} \leq C_{tr} \|u\|_{L^2(\Omega)}^{1/2} \|u\|_{H^1(\Omega)}^{1/2} \tag{2.8}$$

In addition, we will require the co-normal derivative of the restriction $u|_\Omega$:

Lemma 2.1.4 (co-normal trace operator)**.** *Let $\Omega \subset \mathbb{R}^d$ be a Lipschitz domain and let n be the unit exterior normal vector. For $u \in C^\infty(\bar{\Omega})$ define $\gamma_1(u) := n \cdot \nabla u$. This map $u \mapsto \gamma_1(u)$ can be extended to a linear, continuous mapping $\gamma_1 : H^1_\Delta(\Omega) \to H^{-1/2}(\partial\Omega)$ on $H^1_\Delta(\Omega) := \{u \in H^1(\Omega) \mid \Delta u \in L^2(\Omega)\}$ and*

$$\|\gamma_1 u\|_{H^{-1/2}(\partial\Omega)} \leq C\|u\|_{H^1(\Omega)}. \tag{2.9}$$

With this notation the normal derivative can be written as $\partial_n u := \gamma_1(u)$.

2.2. Truncation methods

By reducing the free space problem (2.1) to a bounded domain Ω we have to supply in addition some artificial boundary conditions at $\Gamma := \partial\Omega$ which account for the far field behaviour. Several truncation methods are discussed inter alia in [58] and references therein. In the following we give a short overview of some of theses truncation methods. The presentation here is strongly influenced by the review of this topic in [76].

Domain decomposition and DtN operator

The bounded domain Ω might, in practice, be chosen in dependence of the local source support. Here, we assume Ω to be a d-ball of radius a defined by $B_a := \{x \in \mathbb{R}^d : \|x\| \leq a\}$ such that $\mathrm{supp} f \subset \Omega$ and denote the complement of Ω by $\Omega^+ := \mathbb{R}^d \backslash \overline{\Omega}$, see Figure 2.1. Then (2.1) can be formulated in an equivalent way as the following transmission problem

$$\begin{aligned} -\left(\Delta + k^2\right) u_i &= f && \text{in } \Omega \\ -\left(\Delta + k^2\right) u_e &= 0 && \text{in } \Omega^+ \\ u_i = u_e,\ \partial_n u_i = \partial_n u_e & && \text{on } \Gamma \\ \lim_{r\to\infty} r^{\frac{d-1}{2}} \left(\partial_r u_e - iku_e\right) &= 0 \end{aligned} \tag{2.10}$$

Here, $u_i := u\big|_\Omega$ is a function on the bounded domain Ω, whereas $u_e := u\big|_{\Omega^+}$ is defined on the exterior domain Ω^+ such that $u = u_i + u_e$ and n denotes the normal vector pointing into Ω^+.

According to this domain decomposition we can split the problem (2.10) formally into the following two problems:

$$\begin{aligned} -\left(\Delta - k^2\right) u_e &= 0 && \text{in } \Omega^+ \\ u_e &= u_0 && \text{on } \Gamma \\ \lim_{r\to\infty} r^{\frac{d-1}{2}} \left(\partial_r u_e - iku_e\right) &= 0 \end{aligned} \tag{2.11}$$

and

$$\begin{aligned} -\left(\Delta - k^2\right) u_i &= f && \text{in } \Omega \\ \partial_n u_i &= u_0' && \text{on } \Gamma \end{aligned} \tag{2.12}$$

Both problems are coupled by the interface conditions $u_i = u_e,\ \partial_n u_i = \partial_n u_e$ which is reflected in the relationship between the boundary data u_0, u_0'. This relation is described

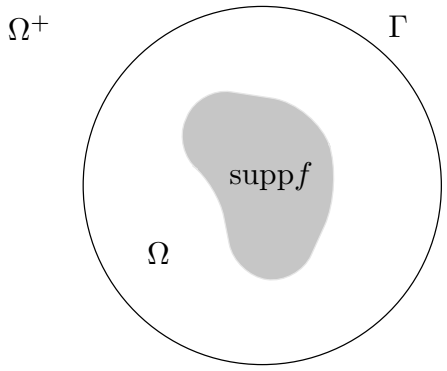


Figure 2.1.: Domain decomposition

by the *Dirichtlet-to-Neumann* (DtN) operator $T_k : H^{1/2}(\Gamma) \to H^{-1/2}(\Gamma), u_0 \mapsto u_0'$ which will be discussed in the following.

In the one dimensional case the solution of (2.11) has the form $u_e(r) = u_0 e^{ik(r-a)}$. Additionally, we see that the outgoing radiation is already attained at finite $a \leq R < \infty$ and the DtN operator in 1D is thus given by

$$T_k f := ikf. \tag{2.13}$$

In this case the boundary condition of the interior problem (2.12) is equivalent to the Robin boundary condition

$$\partial_n u(x) - iku(x) = 0 \quad \forall x \in \Gamma. \tag{2.14}$$

In higher dimensions ($d = 2, 3$) the exterior domain can be expressed in terms of polar coordinates by $\Omega^+ = [a, \infty) \times S^{d-1}$, where $S^{d-1} := \{x \in \mathbb{R}^d : \|x\| = 1\}$ is the $(d-1)$-sphere with radius 1. Then we can rewrite the Helmholtz equation in spherical coordinates as

$$\partial_r^2 u + \frac{d-1}{r}\partial_r u + \frac{1}{r^2}\Delta_{S^{d-1}} u + k^2 u = 0 \tag{2.15}$$

where the spherical Laplace operator $\Delta_{S^{d-1}}$ is defined recursively by

$$\Delta_{S^{d-1}} := \sin^{2-d}(\theta_{d-1})\frac{\partial}{\partial\theta_{d-1}}\left(\sin^{d-2}(\theta_{d-1})\frac{\partial}{\partial\theta_{d-1}}\right) + \sin^{-2}(\theta_{d-1})\Delta_{S^{d-2}}$$

for $\hat{\theta} := (\theta_1, \ldots, \theta_{d-1}) \in [0, 2\pi]^{d-1}$. Using the method of separation of variables we assume the solutions to be of the form $u(r, \hat{\theta}) = v(r)w(\hat{\theta})$. Then equation (2.15) is equivalent to the pair of ordinary differential equations:

$$\begin{aligned} r^2\partial_r^2 v + (d-1)r\partial_r v + (k^2r^2 + \alpha)v &= 0 \quad \text{on } [a, \infty) \\ \Delta_{S^{d-1}} w + \alpha w &= 0 \quad \text{on } S^{d-1} \end{aligned} \tag{2.16}$$

The second equation in (2.16) provides the eigensolutions of the spherical Laplace equation for $d = 2, 3$. These lead to the "exponential" Harmonics Φ_n ($d = 2$) respectively the

spherical Harmonics Y_n^m ($d = 3$) as eigenfunctions. A detailed derivation is available in literature [30,58,78]. In consequence, the second equation has the form of the (spherical) Bessel equation. A solution to this equation is inter alia based on the (spherical) Hankel functions $H_n(r) := r^{1-d/2} H_{n-1+d/2}(r)$ [76].

Finally, the following statement is known from Colton & Kress:

Proposition 2.2.1. *[30, Thm. 2.14] Let u be a solution of (2.11). Then u has an expansion with respect to the spherical wave functions of the form*

$$u(r,\hat{\varphi}) = \begin{cases} \displaystyle\sum_{n=-\infty}^{\infty} \frac{1}{H_{|n|}(ak)} (u_0, \Phi_n)_{L^2(\Gamma)} H_{|n|}(kr) \Phi_n(\hat{\theta}), & d = 2 \\ \displaystyle\sum_{n=0}^{\infty} \frac{1}{h_n(ak)} \sum_{m=-n}^{n} (u_0, Y_n^m)_{L^2(\Gamma)} h_n(kr) Y_n^m(\hat{\theta}), & d = 3 \end{cases} \tag{2.17}$$

As a consequence of the previous result in Proposition 2.2.1 we can write the DtN operator in an explicit way:

Definition 2.2.2 (DtN operator)**.** *The Γ be a d-sphere with radius a and $f \in H^{1/2}(\Gamma)$ a function on Γ. Then the Dirichlet-to-Neumann operator $T_k : H^{1/2}(\Gamma) \to H^{-1/2}(\Gamma)$ is given by*

$$T_k f := \begin{cases} \displaystyle\sum_{n=-\infty}^{\infty} \frac{k H'_{|n|}(ak)}{H_{|n|}(ak)} (f, \Phi_n)_{L^2(\Gamma)} \Phi_n(\hat{\varphi}), & d = 2 \\ \displaystyle\sum_{n=0}^{\infty} \frac{k h'_n(ak)}{h_n(ak)} \sum_{m=-n}^{n} (f, Y_n^m)_{L^2(\Gamma)} Y_n^m(\hat{\varphi}), & d = 3. \end{cases} \tag{2.18}$$

This operator links the interior to the exterior problem. However in both cases ($d = 2, 3$) the operator is non-local and consists of infinite sums, and thus is not numerically feasible.

Also truncating the sum in the exact DtN-operator in order to receive a good approximation on the artificial boundary leads to a non-local operator and therefore the existing solution theory is not applicable.

Localized boundary conditions

One possibility to approximate the DtN operator by some local boundary conditions is done by Feng [41]. Let us here concentrate on the two dimensional case, which is also generalizable to three dimensions, but becomes much more technical. Here, the Hankel functions are first replaced by its asymptotic expansion for large arguments [1]

$$H_n(kr) \longrightarrow \left(\frac{2}{k\pi r} \right)^{1/2} e^{i(kr - \frac{\pi}{2}(n+\frac{1}{2}))}. \tag{2.19}$$

Then the m-th truncation leads to the following absorbing boundary condition (ABC) of m-th order:

$$T_k^m f := -ik \sum_{p=0}^{m} \left(\frac{i}{2ka} \right)^p a_p \left(-\frac{\partial^2}{\partial \varphi^2} \right) \tag{2.20}$$

with the coefficients $a_p(n^2)$ defined recursively in [41]. For both cases ($d = 2, 3$), the absorbing boundary condition of $0th$ order leads again to the same approximation of the DtN-operator already obtained in the one dimensional case

$$T_k^0 f := ikf. \tag{2.21}$$

Again this approximation leads to some Robin-type boundary conditions of the form (2.14).

These boundary conditions involve only partial differential operators. This leads to a compact support of the approximate DtN operator and thus is to be considered as a local ABC in contrast to, for instance, the exact DtN operator which is of integral form and thus considered to be of non-local.

Other approximation methods have been discussed by various authors [13,37]. But as far as they lead to a localized structure, this has to be realized by additional degrees of freedom.

Further approximation methods

Another possibility is to treat the exterior problem via *boundary integral equations methods*, which reduce the problem in Ω^+ to equations on the bounded surface Γ. This method is based on the representation formula

$$u(x) = -\int_\Gamma G_k(x-y)\gamma_1 u(y) ds_y + \int_\Gamma \gamma_1 G_k(x-y)\gamma_0 u(y) ds_y$$

with G_k being the Green's function as introduced in Section 2.1. Applying the trace mappings γ_0, γ_1 (as introduced in Section 2.1) yields the so-called Calderón system

$$\begin{pmatrix} \gamma_0 u \\ \gamma_1 u \end{pmatrix} = \begin{pmatrix} \frac{1}{2} - K_k & V_k \\ W_k & \frac{1}{2} + K_k' \end{pmatrix} \begin{pmatrix} \gamma_0 u \\ \gamma_1 u \end{pmatrix}. \tag{2.22}$$

This provides further representations of the DtN mapping by means of boundary integral operators K_k, K_k', W_k, W_k'. For a detailed definition see [26, 85, 93]. Based on the Calderon system (2.22) various stable realizations of the DtN operator have been studied in the literature, e.g. [36, 56].

In contrast, the *infinite element method* [8,23,47] is built on the representation ansatz $u(x) = \sum c_n b_n(x)$ with basis functions set up via tensor product elements $b_n(x) = b_n^1(r) \otimes b_n^2(\hat{\theta})$. Here, the radial components are described by functions of infinite support based on the Atkinson-Wilcox expansion

$$u(x) \sim \frac{e^{ikr}}{r} \sum_{n=0}^{\infty} \frac{u_n(\hat{\theta})}{r^n}$$

which forces outgoing radiation in the far field [58].

An alternative approach to express the outgoing radiation behaviour is based on the corresponding definition in the frequency space, namely the radial component of the Laplace transformation (in radial direction) has to lie in the Hardy space $H^-(\mathbb{R})$. This property is called *pole condition* as it refers to the singularities of the Laplace transformation. Further transmission onto the complex unit disk and a proper separation technique lead again to a localized boundary structure. A detailed introduction to this topic is available in the dissertation of Lothar Nannen [76] and the works [57, 77].

A final concept mentioned here is the *perfectly matched layer* (PML) method. There, an artificial layer is set up around the inner domain (domain of interest). Inside this layer the partial differential operator is modified by a complex scaling in radial direction. This causes an exponential damping of the outgoing wave such that the solution in the interior is not disturbed by any reflections of this artificial boundary. This approach goes back to the work of J. Berenger [14], which has been discussed for the Helmholtz problem among others in the works [15, 21, 55, 99].

2.3. The Helmholtz equation with Robin boundary condition

Corresponding to the discussions in Section 2.2 we introduce the model problem which represents the mathematical foundation for the rest of this work.

Definition 2.3.1. *Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain. Then the Helmholtz problem with Robin boundary condition is given by*

$$\begin{aligned} -\Delta u - k^2 u &= f \text{ in } \Omega \\ \partial_\nu u - iku &= g \text{ on } \partial\Omega \end{aligned} \tag{2.23}$$

where $k \geq k_0 > 0$, ν being the unit exterior normal vector on the boundary $\partial\Omega$ pointing out of Ω. Further we denote by

$$S_k : (f, g) \mapsto u \tag{2.24}$$

the solution operator of (2.23).

Following the discussion of linear elliptic partial differential equations (PDEs) in [18] a "classical solution" $u_0 \in C^2(\Omega)$, where C^2 denotes the space of 2-times continuously differentiable functions, does not always exist, if we only think, for example, of a piecewise continuous right-hand side $f \in C_{pw}(\Omega)$. Even for a continuous right-hand side $f \in C_{comp}(\Omega)$ with compact support this is not always the case [53]. For compact supported, integrable right-hand sides $f \in L^2_{comp}(\Omega)$ the straightforward way to solve (2.23) by determinating $u_0 \in H^2_{loc}(\Omega)$ is still too restrictive. From the calculus of variations it is known that the condition $u_0' \in L^2$ is often sufficient for the problem (2.23) to be well-posed. This motivates to search for solutions of (2.23) in the weak sense:

Definition 2.3.2. *Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain. Then a weak solution $u \in H^1(\Omega)$ of the Helmholtz problem (2.23) has to satisfy*

$$B(u, v) = l(v) \qquad \forall v \in H^1(\Omega), \tag{2.25}$$

where, for $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega)$, B and l are given by

$$B(u,v) := \int_\Omega (\nabla u \cdot \nabla \overline{v} - k^2 u\overline{v}) - ik \int_{\partial\Omega} u\overline{v}, \quad l(v) := \int_\Omega f\overline{v} + \int_{\partial\Omega} g\overline{v}. \tag{2.26}$$

In this context we employ the space $H^1(\Omega)$ with the norm

$$\|u\|_{\mathcal{H}}^2 := k^2\|u\|_{L^2(\Omega)}^2 + \|\nabla u\|_{L^2(\Omega)}^2. \tag{2.27}$$

Based on this k-weighted norm continuity of the bilinear form B can be shown as follows:

Lemma 2.3.3. *The bilinear form $B(\cdot,\cdot)$ as defined in* (2.26) *is continuous resp. bounded. That is, there exists $C_B > 0$ independent of k such that*

$$|B(u,v)| \le C_B \|u\|_{\mathcal{H}} \|v\|_{\mathcal{H}} \qquad \forall u, v \in H^1(\Omega). \tag{2.28}$$

Proof. By using the multiplicative trace inequality (2.8) we can estimate

$$\begin{aligned} |B(u,v)| &\le |u|_{H^1(\Omega)}|v|_{H^1(\Omega)} + k^2\|u\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} + k\|u\|_{L^2(\partial\Omega)}\|v\|_{L^2(\partial\Omega)} \\ &\le \left(|u|_{H^1(\Omega)} + k\|u\|_{L^2(\Omega)}\right)\left(|v|_{H^1(\Omega)} + k\|v\|_{L^2(\Omega)}\right) \\ &\quad + C_{tr}^2 k^{1/2}\|u\|_{L^2(\Omega)}^{1/2}\|u\|_{\mathcal{H}}^{1/2} k^{1/2}\|v\|_{L^2(\Omega)}^{1/2}\|v\|_{\mathcal{H}}^{1/2} \\ &\le C_B\|u\|_{\mathcal{H}}\|v\|_{\mathcal{H}} \end{aligned}$$

as we do estimate $k\|u\|_{L^2(\Omega)} \le \|u\|_{\mathcal{H}}$ and $\|u\|_{H^1(\Omega)} \le \|u\|_{\mathcal{H}}$. □

The next essential property of bilinear forms is the coercivity resp. ellipticity:

Definition 2.3.4 (Coercivity and ellipticity)**.** *(i) Let $B(\cdot,\cdot)$ be a bilinear form on $H^1(\Omega)$. It is called* coercive*, if it is bounded and there exists a constant $C_c > 0$ such that*

$$B(u,u) > C_c\|u\|_{\mathcal{H}}^2 \quad \forall u \in H^1(\Omega).$$

(ii) A bounded bilinear form $B(\cdot,\cdot)$ on $H^1(\Omega)$ is called elliptic*, if there exist constants $C_g \in \mathbb{R}$ and $\alpha > 0$ such that*

$$B(u,u) \ge \alpha\|u\|_{\mathcal{H}}^2 - C_g\|u\|_{L^2(\Omega)}^2 \quad \forall u \in H^1(\Omega).$$

This inequality is also known as the Gårding inequality*.*

The bilinear form in (2.26) is not coercive, but satisfies a Gårding inequality as it holds

$$\operatorname{Re} B(u,u) + 2k^2(u,u)_{L^2(\Omega)} = \|u\|_{\mathcal{H}}^2. \tag{2.29}$$

Additionally, uniqueness of the solution of (2.25) can be shown via the unique continuation principle for elliptic problems (see, e.g., [63, Chap. 4.3]). This can be best illustrated by the following example in one dimension:

Example 2.3.5. *We consider the one-dimensional Helmholtz problem*

$$u''(x) + k^2 u(x) = f \text{ on } \Omega = [a,b],\ u'(x) - iku(x) = 0 \text{ for } x \in \{a,b\}.$$

As the weak formulation satisfies the Gårding inequality (with $\gamma > k^2$ and $\alpha = min\{1, \gamma - k^2\}$), we know that the induced operator of the corresponding bilinear form is of the form "coercive + compact perturbation". Based on the fact that $H^1(\Omega) \subset\subset L^2(\Omega)$ we could then get from the Riesz theory [62] the existence, if the homogeneous equation has only the trivial solution $u \equiv 0$. Thus, testing the homogeneous weak formulation

$$-\int_\Omega u'v' + k^2 \int_\Omega uv + ik(u(b)v(b) - u(a)v(a)) = 0 \quad \forall v \in H^1(\Omega)$$

with $v = u$ and considering the imaginary part, leads to $u(a) = u(b) = 0$ for any solution $u \in H^1(\Omega)$ of (2.23). Consequently, the trivial extension $\tilde{u}$ to $\mathbb{R}$ satisfies $\tilde{u} \in H^1(\tilde{\Omega})$ for any $\tilde{\Omega} \subset \mathbb{R}$ with $\Omega \subset \tilde{\Omega}$ and $(k^2, \tilde{u}) \in \mathbb{R} \times H^1(\tilde{\Omega})$ satisfies the eigenvalue problem

$$-\int_{\tilde{\Omega}} \tilde{u}'v' = k^2 \int_{\tilde{\Omega}} \tilde{u}v \quad \forall v \in H^1(\tilde{\Omega}).$$

By choosing $\tilde{\Omega} := [-cR, cR]$ with $c > 1$ and applying integration by substitution for $u_c(x) := \tilde{u}(x/c)$, this would then lead to uncountably many eigenpairs $((k/c)^2, u_c) \in \mathbb{R} \times H^1([-R,R])$ for the Laplace operator on $[-R,R]$. However, there exist only countably many eigenvalues of the Laplace operator on a bounded domain. As $k > 0$, this implies that $u_c \equiv 0$, which in turn yields $u \equiv 0$.

The previous result is directly generalizable to higher dimensions. Consequently, we can obtain unique solvability using the classical Riesz-Fredholm theory. To this end the following result is known from [70]:

Proposition 2.3.6. *[70, Prop. 8.1.3] Let Ω be a bounded Lipschitz domain. Then there is a constant $C(\Omega, k) > 0$ such that for every $f \in H^1(\Omega)'$, $g \in H^{-1/2}(\partial\Omega)$ a unique solution u of the problem exists and depends continuously on the data such that*

$$\|u\|_{\mathcal{H}} \leq C(\Omega, k)\big(\|f\|_{H^1(\Omega)'} + \|g\|_{H^{-1/2}(\partial\Omega)}\big).$$

The regularity result based on the Fredholm alternative as stated in Prop. 2.3.6 does not give any indication of how the solution operator depends on the wavenumber k. Yet, it is of interest to know how k influences the regularity behavior of the solution operator. It turns out that both the geometry and the type of boundary conditions have a strong impact. The following chapter will treat this aspects in more detail and provide a refined regularity analysis which is explicit in the wave number k.

3. Regularity analysis

3.1. Geometric assumptions

For the Helmholtz problem in one dimension the domain Ω is in fact an interval and the integral acting on the boundary is reduced to the evaluation at the two end points of Ω.

For the model in higher dimensions we need to say a few more words concerning the geometry of the domain. Most commonly the boundary can be seen as the graph of a function $\phi : \mathbb{R}^{d-1} \to \mathbb{R}$ or as a sub-manifold embedded in $\mathbb{R}^d$.

According to the discussions in [51] we say that the boundary is *analytic*, if for every $x \in \Gamma$ in a neighborhood $\mathcal{U} \subset \mathbb{R}^d$ there exists an analytic, bijective mapping φ from $\mathcal{U}$ onto $\mathcal{U}(\varphi) \subset \mathbb{R}^d$ such that $\Gamma := \{y \in \Omega | \varphi_d(y) = 0\}$ where φ_d is the d-th component of φ. Note that this excludes e.g. domains with a cut. For bounded domains with analytic boundary, it is known that the classic shift theorem holds [42].

However in the context of numerical applications the domain is rather a triangulation of a (curvilinear) polygon. For our purposes, we consider a *(curvilinear) polygon* as sub-manifold with corners A_j, $j = 1, \ldots, J$ and assume that it is defined piece-wise. We denote by Γ_j the (curvilinear) edge which has the endpoint A_j, and ω_j are the interior angles at A_j. In that case the standard shift theorem fails and singularities appear at corners. To that end, we define the $H^{1/2}$-Sobolev space on the boundary $\partial\Omega$ of a polygon edge-wise by

$$H^{1/2}_{pw}(\partial\Omega) := \{g \in L^2(\partial\Omega) : g|_{\Gamma_i} \in H^{1/2}(\Gamma_i)\} \tag{3.1}$$

and illustrate this issue in more detail by recalling the following result harking back to the work by Kondratiev and Grisvard [51]:

Lemma 3.1.1. *Let $\Omega \subset \mathbb{R}^d$ be a polygon with vertices A_j, $j = 1, \ldots, J$, and interior angles ω_j, $j = 1, \ldots, J$. Define for each vertex A_j the singularity function S_j by*

$$S_j(r_j, \varphi_j) = r_j^{\pi/\omega_j} \cos\left(\frac{\pi}{\omega_j}\varphi_j\right), \tag{3.2}$$

where (r_j, φ_j) are polar coordinates centered at the vertex A_j such that the edges of Ω meeting at A_j correspond to $\varphi_j = 0$ and $\varphi_j = \omega_j$. Then every solution u of

$$-\Delta u = f \quad \text{in } \Omega, \qquad \partial_n u = g \quad \text{on } \partial\Omega,$$

can be written as $u = u_0 + \sum_{j=1}^J a_j^\Delta(f, g) S_j$ with the a priori bounds

$$\|u_0\|_{H^2(\Omega)} + \sum_{j=1}^J |a_j^\Delta(f, g)| \leq C \left[\|f\|_{L^2(\Omega)} + \|g\|_{H^{1/2}_{pw}(\partial\Omega)} + \|u\|_{H^1(\Omega)}\right]. \tag{3.3}$$

The a_j^Δ are linear functionals, and $a_j^\Delta = 0$ for convex corners A_j (i.e., if $\omega_j < \pi$).

Proof. This classical result is comprehensively treated in [51] or [31]. A sketch of the proof for the homogeneous Dirichlet problem is given in [83]. □

The lack of smoothness of these boundaries influences the regularity properties of solutions of elliptic equations near these points of non-smoothness. One way to describe the regularity of this solutions is to replace the standard Sobolev spaces with weighted Sobolev spaces. In order to embed the solution of our model problem into the right function space we require the countably normed spaces introduced in [67]. These function spaces are defined with the aid of weight functions $\Phi_{p,\overrightarrow{\beta},k}$ that we now define. For $\beta \in [0,1)$, $n \in \mathbb{N}_0$, and $k > 0$ we set

$$\Phi_{n,\beta,k}(x) = \min\left\{1, \frac{|x|}{\min\left\{1, \frac{|n|+1}{k+1}\right\}}\right\}^{n+\beta}.$$

For given $\overrightarrow{\beta} \in [0,1)^J$, we define

$$\Phi_{n,\overrightarrow{\beta},k}(x) \quad = \quad \prod_{j=1}^{J} \Phi_{n,\beta_j,k}(x - A_j). \tag{3.4}$$

In order to cover the case of polygons as well as domains with analytic boundary by the same notation, we define $\Phi_{n,\overrightarrow{\beta},k} \equiv 1$, if Ω has an analytic boundary.

Furthermore it can be shown that the singularity functions (3.2) weighted by theses functions (3.4) are analytic:

Lemma 3.1.2. *Let $\beta_i \in [0,1)$ satisfy $\beta_i > 1 - \frac{\pi}{\omega_i}$. Then the singularity functions S_i of (3.2) satisfy $\|S_i\|_{H^1(\Omega)} \leq C$ and*

$$\|\Phi_{n,\overrightarrow{\beta},k} \nabla^{n+2} S_i\|_{L^2(\Omega)} \quad \leq \quad C k^{-(2-\beta_i)} \gamma^n \max\{n,k\}^{n+2} \quad \forall \in \mathbb{N}_0$$

for some C, $\gamma > 0$ independent of k.

Proof. Follows from a direct calculation. For details see [40]. □

3.2. Polynomial wellposedness

An important ingredient of the regularity and stability theory will be the concept of *polynomial well-posedness* by which we mean polynomial-in-k-bounds for the norm of the solution operator. To be more precise, this means that we assume that there exists a constant $C_{sol}(k)$ which is polynomially bounded, i.e.

$$C_{sol}(k) \leq \widetilde{C}_{sol} k^\theta \tag{3.5}$$

such that for all $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega)$ the solution u of (2.23) satisfies

$$\|u\|_{\mathcal{H}} \leq C_{sol}(k) \left[\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\partial\Omega)}\right]. \tag{3.6}$$

For *star-shaped* domains Ω, [70] (for $d = 2$) and [32] (for $d = 3$) established the k-explicit stability bound

$$k^{-1}\|u\|_{H^2(\Omega)} + \|u\|_{\mathcal{H}} \quad \leq \quad C\left[\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\partial\Omega)}\right] \tag{3.7}$$

for a $C > 0$ that is independent of k. This shows polynomial wellposedness with $\theta = 0$.

For general Lipschitz domains polynomial wellposedness can be shown with a rate of $\theta = 5/2$ as follows:

Theorem 3.2.1. *Let $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$ be a bounded Lipschitz domain. Then there exists $C > 0$ (independent of k) such that for $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega)$ the solution $u \in H^1(\Omega)$ of (2.23) satisfies*

$$\|u\|_{\mathcal{H}} \leq C\left[k^2\|g\|_{L^2(\partial\Omega)} + k^{5/2}\|f\|_{L^2(\Omega)}\right]. \tag{3.8}$$

Proof. We first consider the full space problem by extending the homogeneous right-hand side f in the standard way. A particular solution of the equation (2.1) is given by the Newton potential $u_0 := G_k \star f$ with the Green's function G_k as introduced in Section 2.1. Then $u_0 \in H^2_{loc}(\mathbb{R}^d)$ and by the analysis of the Newton potential given in [69, Lemma 3.5] we have

$$k^{-1}\|u_0\|_{H^2(\Omega)} + \|u_0\|_{H^1(\Omega)} + k\|u_0\|_{L^2(\Omega)} \leq C\|f\|_{L^2(\Omega)}. \tag{3.9}$$

The difference $\widetilde{u} := u - u_0$ then satisfies

$$-\Delta\widetilde{u} - k^2\widetilde{u} \quad = \quad 0 \quad \text{in } \Omega, \tag{3.10a}$$

$$\partial_n\widetilde{u} - ik\widetilde{u} \quad = \quad g - (\partial_n u_0 - iku_0) =: \widetilde{g}. \tag{3.10b}$$

Then we have

$$\|\widetilde{u}\|_{L^2(\partial\Omega)} \leq k^{-1}\|g\|_{L^2(\partial\Omega)} \tag{3.11}$$

which can be seen by selecting $v = \widetilde{u}$ in the weak formulation (2.25) and then applying the Cauchy-Schwarz inequality to the imaginary part

$$k\|\widetilde{u}\|^2_{L^2(\partial\Omega)} = \operatorname{Im}\int_{\partial\Omega} \widetilde{g}\overline{\widetilde{u}} \leq \|\widetilde{g}\|_{L^2(\partial\Omega)}\|\widetilde{u}\|_{L^2(\partial\Omega)}.$$

Next we have with the multiplicative trace inequality (2.8)

$$\begin{aligned}\|\widetilde{g}\|_{L^2(\partial\Omega)} \quad &\leq \quad C\left[\|g\|_{L^2(\partial\Omega)} + \|u_0\|^{1/2}_{H^2(\Omega)}\|u_0\|^{1/2}_{H^1(\Omega)} + k\|u_0\|^{1/2}_{H^1(\Omega)}\|u_0\|^{1/2}_{L^2(\Omega)}\right] \\ &\leq \quad C\left[\|g\|_{L^2(\partial\Omega)} + k^{1/2}\|f\|_{L^2(\Omega)}\right]. \end{aligned} \tag{3.12}$$

To get bounds on $\widetilde{u}$, we employ (3.11) and (3.12) to conclude

$$\begin{aligned} \|\widetilde{u}\|_{L^2(\partial\Omega)} &\leq Ck^{-1}\|\widetilde{g}\|_{L^2(\partial\Omega)} \leq C\left[k^{-1}\|g\|_{L^2(\partial\Omega)} + k^{-1/2}\|f\|_{L^2(\Omega)}\right], && (3.13)\\ \|\partial_n\widetilde{u}\|_{L^2(\partial\Omega)} &\leq C\left[\|\widetilde{g}\|_{L^2(\partial\Omega)} + k\|\widetilde{u}\|_{L^2(\partial\Omega)}\right] \leq C\left[\|g\|_{L^2(\partial\Omega)} + k^{1/2}\|f\|_{L^2(\Omega)}\right]. && (3.14)\end{aligned}$$

By the use of the results on layer potentials for the Helmholtz equation from [71] and the generous estimate $\|\partial_n\widetilde{u}\|_{H^{-1}(\partial\Omega)} \leq C\|\partial_n\widetilde{u}\|_{L^2(\partial\Omega)}$ we get

$$\|\widetilde{u}\|_{H^1(\Omega)} + k\|\widetilde{u}\|_{L^2(\Omega)} \leq C\left[k^2\|g\|_{L^2(\partial\Omega)} + k^{5/2}\|f\|_{L^2(\Omega)}\right]. \tag{3.15}$$

Combining (3.9), (3.15) provides the statement for $u = \widetilde{u} + u_0$. □

Remark 3.2.2. *The arguments in the proof show us that it is not the geometry but the type of boundary conditions in our model problem (2.23), namely, the Robin boundary conditions that makes it polynomially well-posed.*

While (3.11) resp. (3.7) do not make minimal assumptions on the regularity of f and g they can be used to show that the sesquilinear form B of (2.25) satisfies an inf-sup condition with an inf-sup constant $\gamma = O(k^{-(\theta+1)})$. This again makes it possible to obtain regularity estimates for $f \in H^1(\Omega)'$ and $g \in H^{-1/2}(\partial\Omega)$:

Theorem 3.2.3. *Let $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$ be a bounded Lipschitz domain. Then there exists $C > 0$ (independent of k) such that the sesquilinear form B of (2.26) satisfies*

$$\inf_{0\neq u\in H^1(\Omega)} \sup_{0\neq v\in H^1(\Omega)} \frac{\operatorname{Re} B(u,v)}{\|u\|_{\mathcal{H}}\|v\|_{\mathcal{H}}} \geq Ck^{-7/2}. \tag{3.16}$$

Furthermore, for every $f \in (H^1(\Omega))'$ and $g \in H^{-1/2}(\partial\Omega)$ the problem (2.25) is uniquely solvable, and its solution $u \in H^1(\Omega)$ satisfies the a priori bound

$$\|u\|_{\mathcal{H}} \leq Ck^{7/2}\left[\|f\|_{(H^1(\Omega))'} + \|g\|_{H^{-1/2}(\partial\Omega)}\right]. \tag{3.17}$$

If Ω is convex or if Ω is star-shaped and has a smooth boundary, then the following, sharper estimate holds:

$$\inf_{0\neq u\in H^1(\Omega)} \sup_{0\neq v\in H^1(\Omega)} \frac{\operatorname{Re} B(u,v)}{\|u\|_{\mathcal{H}}\|v\|_{\mathcal{H}}} \geq Ck^{-1}. \tag{3.18}$$

Proof. The proof relies on standard arguments for sesquilinear forms satisfying a Gårding inequality. Given $u \in H^1(\Omega)$ we define $z \in H^1(\Omega)$ as the solution of

$$2k^2(\cdot,u)_{L^2(\Omega)} = B(\cdot,z).$$

Theorem 3.2.1 implies $\|z\|_{\mathcal{H}} \leq Ck^{9/2}\|u\|_{L^2(\Omega)}$, and $v = u + z$ satisfies

$$\operatorname{Re} B(u,v) = \operatorname{Re} B(u,u) + \operatorname{Re} B(u,z) = \|u\|_{\mathcal{H}}^2 - 2k^2\|u\|_{L^2(\Omega)}^2 + \operatorname{Re} B(u,z) = \|u\|_{\mathcal{H}}^2.$$

Thus

$$\begin{aligned} \operatorname{Re} B(u,v) &= \|u\|_{\mathcal{H}}^2, \\ \|v\|_{\mathcal{H}} &= \|u+z\|_{\mathcal{H}} \le \|u\|_{\mathcal{H}} + \|z\|_{\mathcal{H}} \le \|u\|_{\mathcal{H}} + Ck^{9/2}\|u\|_{L^2(\Omega)} \le Ck^{7/2}\|u\|_{\mathcal{H}}. \end{aligned}$$

Therefore,

$$\operatorname{Re} B(u,v) = \|u\|_{\mathcal{H}}^2 \ge \|u\|_{\mathcal{H}} \frac{1}{Ck^{7/2}} \|v\|_{\mathcal{H}},$$

which concludes the proof of (3.16). Example 2.3.5 provides unique solvability for (2.23) so that (3.16) gives the *a priori* estimate (3.17). Finally, (3.18) is shown by the same arguments using (3.7). □

3.3. Frequency splitting

As the behavior of the Helmholtz operator changes essentially for high wave numbers it seems reasonable to treat the low and highly oscillatory contribution separately. To begin with we employ this frequency spitting to the inhomogeneous data f and g.

In order to perform a splitting in the Fourier space for functions $f \in L^2(\Omega)$ we introduce the following lifting operator by [92]:

Proposition 3.3.1 (Extension operator of Stein)**.** *Let $\Omega \subset \mathbb{R}^d$ be a Lipschitz domain, $m \in \mathbb{N}_0$ and $p \in [1,\infty]$. Then there exists a bounded linear extension operator $E : W^{m,p}(\Omega) \to W^{m,p}(\mathbb{R}^d)$, i.e. $Eu|_\Omega = u$ for all $u \in W^{m,p}(\Omega)$ and there exists a constant $C \ge 0$ such that for all $u \in W^{m,p}(\Omega)$*

$$\|Eu\|_{W^{m,p}(\Omega)} \le C\|u\|_{W^{m,p}(\mathbb{R}^d)}.$$

With the help of this extension operator we can apply the Fourier transformation and define (similar to [69, 72]) the low and high frequency filter $L_{\Omega,\eta} f : L^2(\Omega) \to L^2(\Omega)$ and $H_{\Omega,\eta} f : L^2(\Omega) \to L^2(\Omega)$ by

$$L_{\Omega,\eta} f := \mathcal{F}^{-1}(\chi_{B^d_{\eta k}(0)} \mathcal{F}(E_\Omega f))\big|_\Omega, \tag{3.19}$$

$$H_{\Omega,\eta} f := \mathcal{F}^{-1}(\chi_{\mathbb{R}^d \setminus B^d_{\eta k}(0)} \mathcal{F}(E_\Omega f))\big|_\Omega, \tag{3.20}$$

where χ_ω is the characteristic function of a set ω and $B^d_{\eta k}(0)$ the d-ball with radius ηk centered at the origin 0.

For the case of polygons in 2D with boundary functions $g \in L^2_{pw}(\partial\Omega)$ we define $L_{\partial\Omega,\eta} g : L^2(\partial\Omega) \to L^2(\partial\Omega)$ and $H_{\partial\Omega,\eta} : L^2(\partial\Omega) \to L^2(\partial\Omega)$ in an *edgewise* fashion:

$$L_{e,\eta} g := \mathcal{F}^{-1}(\chi_{B^{d-1}_{\eta k}(0)} \mathcal{F}(E_e g))\big|_e, \tag{3.21}$$

$$H_{e,\eta} g := \mathcal{F}^{-1}(\chi_{\mathbb{R}^{d-1} \setminus B^{d-1}_{\eta k}(0)} \mathcal{F}(E_e g))\big|_e, \tag{3.22}$$

where $E_e : L^2(e) \to L^2(\mathbb{R})$ is the extension operator of Stein for an edge $e \subset \partial\Omega$. Of course it holds $L_{\Omega,\eta} + H_{\Omega,\eta} = \mathrm{Id}$ on $L^2(\Omega)$ and we have

$$\|L_{\Omega,\eta} f\|_{L^2(\Omega)} + \|H_{\Omega,\eta} f\|_{L^2(\Omega)} \leq C\|f\|_{L^2(\Omega)} \qquad \forall f \in L^2(\Omega),$$

where $C > 0$ depends solely on Ω [72].

Next, we review some properties for the low and high frequency part of the splitting from [69,72]. Here, the implied constants in the estimates are independent of f, g and n.

Lemma 3.3.2. *Let $f \in L^2(\Omega), g \in L^2(\partial\Omega)$. Then the following estimates hold for all $n \in \mathbb{N}_0$:*

$$\begin{aligned}
\|\nabla^n L_{\Omega,\eta} f\|_{L^2(\Omega)} &\lesssim (\eta k)^n \|f\|_{L^2(\Omega)} \\
\|\nabla^n L_{\partial\Omega,\eta} g\|_{L^2(\partial\Omega)} &\lesssim \begin{cases} (\eta k)^{n-3/2} \|g\|_{L^2(\partial\Omega)}, \ n \geq 3/2, & \text{if } \partial\Omega \text{ is smooth} \\ (\eta k)^{n-2} \|g\|_{H^{1/2}_{pw}(\partial\Omega)}, & \text{if } \partial\Omega \text{ is a polygon} \end{cases}
\end{aligned}$$

Proof. [72, Lemma 4.3] □

Remark 3.3.3. *Since it holds $\mathrm{supp}(\widehat{L_{\Omega,\eta} f}) \subset B_{\eta k}(0)$, this implies that $L_{\Omega,\eta} f$ resp. $L_{\Omega,\eta} g$ are analytic.*

Lemma 3.3.4. *Let $f \in H^s(\Omega)$ for $s \in \{0, 1\}$ and $g \in H^{1/2}_{pw}(\partial\Omega)$. Then the following estimates hold:*

$$\begin{aligned}
\|H_{\Omega,\eta} f\|_{L^2(\Omega)} &\lesssim (\eta k)^{-s} \|f\|_{H^s(\Omega)} \\
\|H_{\partial\Omega,\eta} g\|_{L^2(\partial\Omega)} &\lesssim \begin{cases} (\eta k)^{-s} \|g\|_{H^s(\partial\Omega)}, & \text{if } \partial\Omega \text{ is smooth} \\ (\eta k)^{-1/2} \|g\|_{H^{1/2}_{pw}(\partial\Omega)}, & \text{if } \partial\Omega \text{ is a polygon} \end{cases}
\end{aligned}$$

Proof. [72, Lemma 4.2] □

The operators $H_{\Omega,\eta}$ and $H_{\partial\Omega,\eta}$ have furthermore approximation properties if the function they are applied to has some Sobolev regularity. We illustrate this for the operator $H_{\partial\Omega,\eta}$:

Lemma 3.3.5. *Let $\Omega \subset \mathbb{R}^2$ be a polygon. Then there exists $C > 0$ independent of k and $\eta > 1$ such that for all $g \in H^{1/2}_{pw}(\partial\Omega)$*

$$k^{1/2}(1 + \eta^{1/2}) \|H_{\partial\Omega,\eta} g\|_{L^2(\partial\Omega)} + \|H_{\partial\Omega,\eta} g\|_{H^{1/2}_{pw}(\partial\Omega)} \leq C\|g\|_{H^{1/2}_{pw}(\partial\Omega)}.$$

Proof. We only show the estimate for $\|H_{\partial\Omega,\eta}g\|_{L^2(\partial\Omega)}$. We consider first the case of an interval $I \subset \mathbb{R}$. We define $H_{I,\eta}g$ by $H_{I,\eta}g = \mathcal{F}^{-1}\chi_{\mathbb{R}\setminus B_{\eta k}(0)}\mathcal{F}E_I g$, where $\chi_{\mathbb{R}\setminus B_{\eta k}(0)}$ is the characteristic function for $\mathbb{R}\setminus(-\eta k, \eta k)$ and E_I is the Stein extension operator for the interval I. Since, according to Parseval, $\mathcal{F}$ is an isometry on $L^2(\mathbb{R})$ we have

$$\begin{aligned}\|H_{I,\eta}g\|^2_{L^2(I)} \le \|H_{I,\eta}g\|^2_{L^2(\mathbb{R})} &= \int_{\mathbb{R}\setminus B_{\eta k}(0)} |\mathcal{F}E_I g|^2\, d\xi \\ = \int_{\mathbb{R}\setminus B_{\eta k}(0)} \frac{(1+|\xi|^2)^{1/2}}{(1+|\xi|^2)^{1/2}}|\mathcal{F}E_I g|^2\, d\xi &\le \frac{1}{(1+(\eta k)^2)^{1/2}}\int_{\mathbb{R}}(1+|\xi|^2)^{1/2}|\mathcal{F}E_I g|^2\, d\xi.\end{aligned}$$

The last integral can be bounded by $C\|E_I g\|^2_{H^{1/2}(\mathbb{R})}$. The stability properties of the extension operator E_I then imply furthermore $\|E_I g\|_{H^{1/2}(\mathbb{R})} \le C\|g\|_{H^{1/2}(I)}$. In total, we arrive at

$$\|H_{I,\eta}g\|_{L^2(I)} \le C\frac{1}{(1+(\eta k)^2)^{1/4}}\|g\|_{H^{1/2}(I)} \le Ck^{-1/2}(1+\eta)^{-1/2}\|g\|_{H^{1/2}(I)},$$

where, in the last estimate, the constant C depends additionally on k_0. From this estimate, we obtain the desired bound for $\|H_{\partial\Omega,\eta}g\|_{L^2(\partial\Omega)}$ by identifying each edge of Ω with an interval. □

To simplify notation we drop the indices of the frequency splitting operators throughout the remaining of this chapter and just write L and H for the low and high frequency filters. If these operators are acting on a function f defined on the domain Ω, then they have to be understood as $Lf = L_{\Omega,\eta}f$ and $Hf = H_{\Omega,\eta}f$. Accordingly, in the case of a function g defined on the boundary $\partial\Omega$ we use $Lf = L_{\partial\Omega,\eta}f$ and $Hf = H_{\partial\Omega,\eta}f$.

3.4. Regularity by decomposition (for polygons)

In order to make use of the frequency splitting we decompose the solution of the model problem (2.23) to obtain a refined regularity result formulated in Theorem 3.4.6. An illustration of that decomposition scheme can be seen in Figure 3.1. According to this scheme a sketch of the proof is as follows:

First we split off the Helmholtz problem according to the low frequency filtered data (Lf, Lg). As the low frequency filtered data is (piecewise) analytic, the corresponding solution of the Helmholtz operator $u_{\mathcal{A}} := S_k(Lf, Lg)$ where S_k is the operator defined in (2.24), is again analytic on Ω (Lemma 3.4.1). Concerning the high frequency filtered data (Hf, Hg) the Helmholtz operator $-\Delta - k^2$ acts very similarly to the modified Helmholtz operator $-\Delta + k^2$. However, the latter is positive definite and thus much easier to analyze. In detail, Hf and Hg are treated separately. The high frequency inhomogeneity Hf is handled by the full space model of the modified operator (3.26). For $Hf \in L^2$ the concerning solution $S^+_{\mathbb{R}^d}(Hf)$ is then H^2-regular as the classic shift theorem applies (Lemma 3.4.3). However, we are actually interested in the solution restricted to Ω. The resulting boundary data has to be included in a subsequent decomposition step. Thus, we

consider next the modified homogeneous model on Ω where the boundary data contains Hg and corresponding boundary data of the modified full space model. The concerning solution can again be split into an H^2-regular part and an analytic part emerging from potential corners in the geometry. This is shown in Lemma 3.4.4, motivated by the discussions in Section 3.1. The remaining problem has the same structure as the original model problem recouping the alterations of the modified problems by a correction term on its right-hand-side. To this end it can be shown that the resulting inhomogeneity $\widetilde{f}$ results from a contraction of the initial boundary data (f, g), see Lemma 3.4.5. This is a crucial component for the refined regularity theory as it allows to employ a geometric series argument in order to get to the main result stated in Theorem 3.4.6. This statement and the preliminary work have been published in [40], while here they have been organized a little differently.

$$
\boxed{\begin{aligned} -\Delta u - k^2 u &= Lf + Hf \quad \text{in} \quad \Omega \\ \partial_n u - iku &= Lg + Hg \quad \text{on} \quad \partial\Omega \end{aligned}}
$$

$$
=
$$

$$
\boxed{\begin{aligned} -\Delta u_{\mathcal{A}} - k^2 u_{\mathcal{A}} &= Lf \quad \text{in} \quad \Omega \\ \partial_n u_{\mathcal{A}} - iku_{\mathcal{A}} &= Lg \quad \text{on} \quad \partial\Omega \end{aligned}}
$$

$$
+
$$

$$
\boxed{-\Delta u_1 + k^2 u_1 = Hf \quad \text{in} \quad \mathbb{R}^2}
$$

$$
+
$$

$$
\boxed{\begin{aligned} -\Delta u_2 + k^2 u_2 &= 0 \\ \partial_n u_2 - iku_2 = Hg - \partial_n u_1 + iku_1 &=: h \end{aligned}}
$$

$$
+
$$

$$
\boxed{\begin{aligned} -\Delta \tilde{u} - k^2 \tilde{u} &= 2k^2(u_1 + u_2) =: \tilde{f} \\ \partial_n \tilde{u} - ik\tilde{u} &= 0 \end{aligned}}
$$

Figure 3.1.: Decomposition scheme, $u = u_1|_\Omega + u_{\mathcal{A}} + u_2 + \tilde{u}$

3.4.1. The Helmholtz problem with analytic data

First we consider the Helmholtz problem for (piecewise) analytic data. This case has been studied extensively by J. M. Melenk in [67]. We repeat here the corresponding statement without proof:

Lemma 3.4.1 ([72, Lemma 4.12], analytic regularity of $S(f, g)$)**.** *Let Ω be a polygon. Let f be analytic on Ω and $g \in L^2(\partial\Omega)$ be piecewise analytic and satisfy for some constants*

$\widetilde{C}_f$, $\widetilde{C}_g$, γ_f, $\gamma_g > 0$

$$\|\nabla^n f\|_{L^2(\Omega)} \leq \widetilde{C}_f \gamma_f^n \max\{n,k\}^n \qquad \forall n \in \mathbb{N}_0 \tag{3.23a}$$

$$\|\nabla_T^n g\|_{L^2(e)} \leq \widetilde{C}_g \gamma_g^n \max\{n,k\}^n \qquad \forall n \in \mathbb{N}_0 \quad \forall e \in \mathcal{E}, \tag{3.23b}$$

where $\mathcal{E}$ denotes the set of edges of Ω and ∇_T tangential differentiation. Then there exist $\overrightarrow{\beta} \in [0,1)^J$ (depending only on Ω) and constants C, $\gamma > 0$ (depending only on Ω and γ_f, γ_g) such that the following is true for the constant $C_{sol}(k)$ of (3.6):

$$\|u\|_{\mathcal{H}} \leq C_{sol}(k)(\widetilde{C}_f + \widetilde{C}_g) \tag{3.24}$$

$$\|\phi_{n,\overrightarrow{\beta},k} \nabla^{n+2} u\|_{L^2(\Omega)} \leq C C_{sol}(k) k^{-1} (\widetilde{C}_f + \widetilde{C}_g) \gamma^n \max\{n,k\}^{n+2} \quad \forall n \in \mathbb{N}_0. \tag{3.25}$$

Remark 3.4.2. *(i) We remark that the L^2-estimate is not sharp with respect to the wave number k. The constant $C_{sol}(k)$ from our stability assumption (3.6) is motivated by the estimates available for the star-shaped case, but could clearly be replaced with other assumptions.*
(ii) The low frequency filtered data (Lf, Lg) satisfy the requirements of Lemma 3.4.1 given their properties in Lemma 3.3.2.

3.4.2. The modified full space Helmholtz problem

In order to circumvent the fact that the solution operator $S_k(f,g)$ is indefinite we consider the modified problem

$$-\Delta u + k^2 u = f \quad \text{in} \quad \mathbb{R}^2 \tag{3.26}$$

and denote the corresponding solution operator by $S^+_{\mathbb{R}^2} : f \mapsto u$. Accordingly, we denote with $\|\cdot\|_{\mathcal{H}}$, here in this subsection, the weighted H^1-norm (2.27) on the entire space $\mathbb{R}^2$.

Lemma 3.4.3 (properties of $S^+_{\mathbb{R}^2}$)**.** *There exists $C > 0$ such that for every $\eta > 1$ and every $f \in L^2(\mathbb{R}^2)$ whose Fourier transform $\mathcal{F}f(\xi) := \frac{1}{\sqrt{2\pi}} \int e^{-i\xi\cdot x} f(x) dx$ satisfies $\operatorname{supp} \mathcal{F}f \subset \mathbb{R}^2 \setminus B_{\eta k}(0)$, the solution $u = S^+_{\mathbb{R}^2} f$ of (3.26) satisfies*

$$\|u\|_{\mathcal{H}} \leq k^{-1} \frac{1}{\sqrt{1+\eta^2}} \|f\|_{L^2(\mathbb{R}^2)} \qquad \text{and} \qquad \|u\|_{H^2(\mathbb{R}^2)} \leq C \|f\|_{L^2(\mathbb{R}^2)}.$$

Proof. By using Parseval's theorem, Cauchy-Schwarz inequality and the notation $\widehat{f} = \mathcal{F}f$ and $\widehat{u} = \mathcal{F}u$ for the Fourier transforms we get

$$\begin{aligned}
\|u\|^2_{\mathcal{H}} &= (f,u)_{L^2(\mathbb{R}^2)} = (\widehat{f}, \widehat{u})_{L^2(\mathbb{R}^2)} \\
&= \left((|\xi|^2 + k^2)^{-1} \widehat{f}, (|\xi|^2 + k^2) \widehat{u} \right)_{L^2} \\
&\leq \|(|\xi|^2 + k^2)^{-1} \widehat{f}\|_{L^2} \|(|\xi|^2 + k^2)\widehat{u}\|_{L^2} \\
&= \sqrt{\int_{\mathbb{R}^2 \setminus B_{\eta k}(0)} (|\xi|^2 + k^2)^{-1} |\widehat{f}|^2 \, d\xi} \|u\|_{\mathcal{H}}
\end{aligned}$$

where, in the last step, we used the support properties of $\widehat{f}$. Applying again Parseval's theorem, we get

$$\|u\|_{\mathcal{H}} \leq ((\eta k)^2 + k^2)^{-1/2} \|f\|_{L^2(\mathbb{R}^2)} \leq k^{-1}(1+\eta)^{-1/2} \|f\|_{L^2(\mathbb{R}^2)}.$$

The estimate for $\|u\|_{H^2(\mathbb{R}^2)}$ now follows from $f \in L^2(\mathbb{R}^2)$ and the standard interior regularity for the Laplacian:

$$\|u\|_{H^2} \leq \|f\|_{L^2} + k^2 \|u\|_{L^2} \leq Ck \|u\|_{\mathcal{H}} \leq C(1+\eta)^{-1/2} \|f\|_{L^2}.$$

□

3.4.3. The modified Helmholtz problem on a bounded domain

Next we discuss the modified Helmholtz problem on a bounded domain defined by

$$\begin{aligned} -\Delta u + k^2 u &= f \quad \text{in } \Omega && (3.27) \\ \partial_n u - iku &= g \quad \text{on } \partial\Omega && (3.28) \end{aligned}$$

and denote the corresponding solution operator by $S_\Omega^+ : (f, g) \mapsto u$.

Lemma 3.4.4 (properties of S_Ω^+)**.** *Let $\Omega \subset \mathbb{R}^2$ be a polygon and $f \in L^2(\Omega)$, $g \in H^{1/2}_{pw}(\partial\Omega)$. Then the solution $u := S_\Omega^+(f, g)$ satisfies*

$$\|u\|_{\mathcal{H}} \leq k^{-1/2} \|g\|_{L^2(\partial\Omega)} + k^{-1} \|f\|_{L^2(\Omega)}. \tag{3.29}$$

Furthermore, there exists $C > 0$ independent of k and the data f, g and there exists a decomposition $u = u_{H^2} + \sum_{i=1}^J a_i^+(f, g) S_i$ for some linear functionals a_i^+ with

$$\|u_{H^2}\|_{H^2(\Omega)} + \sum_{i=1}^J |a_i^+(f, g)| \leq C \left[\|f\|_{L^2(\Omega)} + \|g\|_{H^{1/2}_{pw}(\partial\Omega)} + k^{1/2} \|g\|_{L^2(\partial\Omega)} \right]. \tag{3.30}$$

Proof. The estimate (3.29) for $\|u\|_{\mathcal{H}}$ follows by Lax-Milgram – see [69, Lemma 4.6] for details. Since u satisfies

$$-\Delta u = f - k^2 u \quad \text{in } \Omega, \qquad \partial_n u = g + iku \quad \text{on } \partial\Omega,$$

the standard regularity theory for the Laplacian (see Lemma 3.1.1) permits us to decompose $u = u_{H^2} + \sum_{i=1}^J a_i^\Delta (f - k^2 u, g + iku) S_i$. The continuity of the linear functionals a_i^Δ reads

$$\sum_{i=1}^J |a_i^\Delta (f - k^2 u, g + iku)| \leq C \left[\|f - k^2 u\|_{L^2(\Omega)} + \|g + iku\|_{H^{1/2}_{pw}(\partial\Omega)} \right].$$

Since $(f, g) \mapsto S_\Omega^+(f, g)$ is linear, the map $(f, g) \mapsto a_i^+(f, g) := a_i^\Delta(f - k^2 u, g + iku)$ is linear, and (3.29), (3.3) give the desired estimates for u_{H^2} and $a_i^+(f, g)$. □

3.4.4. The remainder problem

In conclusion to the previous regularity the results we consider the following remaining problem which is of the same structure as original model problem (2.23) with homogeneous boundary condition

$$\begin{aligned} -\Delta u - k^2 u &= \widetilde{f} \text{ in } \Omega \\ \partial_n u - iku &= 0 \text{ on } \partial\Omega \end{aligned} \tag{3.31}$$

and the specific right hand side

$$\widetilde{f} := 2k^2 \left(S^+_{\mathbb{R}^d}(Hf)\big|_\Omega + S^+_\Omega \left(0, Hg - \partial_n S^+_{\mathbb{R}^d}(Hf)\big|_{\partial\Omega} - ikS^+_{\mathbb{R}^d}(Hf)\big|_{\partial\Omega}\right)\right).$$

The following result builds the key argument to the proof of Theorem 3.4.6:

Lemma 3.4.5 (contraction lemma)**.** *Let $\Omega \subset \mathbb{R}^2$ be a polygon. Fix $q \in (0,1)$. Then one can find $\overrightarrow{\beta} \in [0,1)^J$ (depending solely on Ω) and constants C, $\gamma > 0$ independent of k such that for every $f \in L^2(\Omega)$ and every $g \in H^{1/2}_{pw}(\partial\Omega)$, the solution u of (2.23) can be decomposed as $u = u_{H^2} + \sum_{i=1}^J a_i(f,g)S_i + u_{\mathcal{A}} + r$, where $u_{H^2} \in H^2(\Omega)$, the a_i are linear functionals, and $u_{\mathcal{A}} \in C^\infty(\Omega)$. These functions satisfy*

$$\begin{aligned} &k\|u_{H^2}\|_{\mathcal{H}} + \|u_{H^2}\|_{H^2(\Omega)} + \sum_{i=1}^J |a_i(f,g)| \le C\left[\|f\|_{L^2(\Omega)} + \|g\|_{H^{1/2}_{pw}(\partial\Omega)}\right], \\ &\|u_{\mathcal{A}}\|_{\mathcal{H}} \le CC_{sol}(k)\left[\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\partial\Omega)}\right], \\ &\|\Phi_{n,\overrightarrow{\beta},k}\nabla^{n+2}u_{\mathcal{A}}\|_{L^2(\Omega)} \le CC_{sol}(k)k^{-1}\gamma^n \max\{n,k\}^{n+2}\left[\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\partial\Omega)}\right] \end{aligned}$$

for all $n \in \mathbb{N}_0$. Finally, the remainder $\widetilde{u}$ satisfies

$$-\Delta\widetilde{u} - k^2\widetilde{u} = \widetilde{f} \quad \text{in } \Omega, \qquad \partial_n\widetilde{u} - ik\widetilde{u} = \widetilde{g} \quad \text{on } \partial\Omega$$

for some $\widetilde{f} \in L^2(\Omega)$ and $\widetilde{g} \in H^{1/2}_{pw}(\partial\Omega)$ with

$$\|\widetilde{f}\|_{L^2(\Omega)} + \|\widetilde{g}\|_{H^{1/2}_{pw}(\partial\Omega)} \le q\left(\|f\|_{L^2(\Omega)} + \|g\|_{H^{1/2}_{pw}(\partial\Omega)}\right).$$

Proof. We start by decomposing $(f,g) = (L_{\Omega,\eta}f, L_{\partial\Omega,\eta}g) + (H_{\Omega,\eta}f, H_{\partial\Omega,\eta}g)$ with a parameter $\eta > 1$ that will be selected below. We set

$$u_{\mathcal{A}} := S_k(L_{\Omega,\eta}f, L_{\partial\Omega,\eta}g), \qquad u_1 := S^+_{\mathbb{R}^2}(H_{\Omega,\eta}f)|_\Omega,$$

where we tacitly extended $H_{\Omega,\eta}f$ (which is only defined on Ω) by zero outside Ω. First we can retrieve the desired estimates for $u_{\mathcal{A}}$ from Lemma 3.4.1 and the analyticity of the low frequency data in Lemma 3.3.2. For u_1 we get by Lemma 3.4.3 and the estimate for Hf in Lemma 3.3.4 the a priori estimates

$$\begin{aligned} \|u_1\|_{\mathcal{H}} &\le Ck^{-1}(1+\eta^2)^{-1/2}\|H_{\Omega,\eta}f\|_{L^2(\Omega)} \le Ck^{-1}(1+\eta)^{-1}\|f\|_{L^2(\Omega)}, \\ \|u_1\|_{H^2(\Omega)} &\le C\|H_{\Omega,\eta}f\|_{L^2(\Omega)} \le C\|f\|_{L^2(\Omega)}. \end{aligned}$$

The trace inequality (2.7) and the multiplicative trace inequalities (2.8) imply for $g_1 := \partial_n u_1 - iku_1$:

$$k^{1/2}(1+\eta)^{1/2}\|g_1\|_{L^2(\partial\Omega)} + \|g_1\|_{H^{1/2}_{pw}(\partial\Omega)} \quad \leq \quad C\|f\|_{L^2(\Omega)}.$$

For $g_2 := H_{\partial\Omega,\eta}g - g_1$ we then get from Lemma 3.3.5 and the triangle inequality

$$k^{1/2}(1+\eta)^{1/2}\|g_2\|_{L^2(\partial\Omega)} + \|g_2\|_{H^{1/2}_{pw}(\partial\Omega)} \quad \leq \quad C\left[\|g\|_{H^{1/2}_{pw}(\partial\Omega)} + \|f\|_{L^2(\Omega)}\right].$$

Next we apply Lemma 3.4.4 to $u_2 := S^+_\Omega(0, g_2)$ which yields

$$\|u_2\|_{\mathcal{H}} \leq Ck^{-1/2}\|g_2\|_{L^2(\partial\Omega)} \leq Ck^{-1}(1+\eta)^{-1/2}\left[\|f\|_{L^2(\Omega)} + \|g\|_{H^{1/2}_{pw}(\partial\Omega)}\right].$$

Furthermore we can write $u_2 = u_{H^2} + \sum_{i=1}^J a_i^+(0, g_2)S_i$, with

$$\|u_{H^2}\|_{H^2(\Omega)} + \sum_{i=1}^J |a_i^+(0, g_2)| \leq C\left[\|f\|_{L^2(\Omega)} + \|g\|_{H^{1/2}_{pw}(\partial\Omega)}\right].$$

We then define $a_i(f, g) := a_i^+(0, g_2)$ and note that $(f, g) \mapsto a_i(f, g)$ is linear by linearity of the maps a_i^+ and $(f, g) \mapsto g_2$. The above shows that u_{H^2} and the a_i satisfy the required estimates. Finally, the function $\widetilde{u} := u - (u_{\mathcal{A}} + u_1 + u_2)$ satisfies

$$-\Delta\widetilde{u} - k^2\widetilde{u} = 2k^2(u_1 + u_2) =: \widetilde{f}, \qquad \partial_n\widetilde{u} - ik\widetilde{u} = 0 =: \widetilde{g},$$

together with

$$\|\widetilde{f}\|_{L^2(\Omega)} \leq 2k^2\left(\|u_1\|_{L^2(\Omega)} + \|u_2\|_{L^2(\Omega)}\right) \leq C(1+\eta)^{-1/2}\left[\|f\|_{L^2(\Omega)} + \|g\|_{H^{1/2}_{pw}(\partial\Omega)}\right].$$

Hence, selecting $\eta > 1$ sufficiently large so that for the chosen $q \in (0, 1)$ we have $C(1+\eta)^{-1/2} \leq q$ which allows us to conclude the proof. □

Theorem 3.4.6 (main result)**.** *Let $\Omega \subset \mathbb{R}^2$ be a polygon with vertices A_j, $j = 1, \ldots, J$. Then there exist constants C, $\gamma > 0$, $\beta \in [0, 1)^J$ such that for every $f \in L^2(\Omega)$ and $g \in H^{1/2}_{pw}(\partial\Omega)$ the solution u of (2.23) can be decomposed as $u = u_{H^2} + u_{\mathcal{A}}$ with*

$$\begin{aligned}
k\|u_{H^2}\|_{\mathcal{H}} + \|u_{H^2}\|_{H^2(\Omega)} &\leq CC_{f,g} \\
\|u_{\mathcal{A}}\|_{H^1(\Omega)} &\leq (C_{sol}(k) + 1)\, C_{f,g} \\
k\|u_{\mathcal{A}}\|_{L^2(\Omega)} &\leq (C_{sol}(k) + k)\, C_{f,g} \\
\|\Phi_{n,\vec{\beta},k}\nabla^{n+2}u_{\mathcal{A}}\|_{L^2(\Omega)} &\leq C(C_{sol}(k) + 1)k^{-1}\max\{n, k\}^{n+2}\gamma^n C_{f,g} \quad \forall n \in \mathbb{N}_0
\end{aligned}$$

with $C_{f,g} := \|f\|_{L^2(\Omega)} + \|g\|_{H^{1/2}_{pw}(\partial\Omega)}$ and $C_{sol}(k)$ introduced in (3.6).

Proof. We repeat here in extracts the proof given in [40]. According to Lemma 3.4.5 we can rewrite the solution as

$$S_k(f,g) = u_{H^2} + \sum_{i=1}^{J} a_i(f,g) S_i + u_{\mathcal{A}} + S(\widetilde{f},\widetilde{g}).$$

From there we have the desired estimates for u_{H^2}, $a_i(f,g)$, and $u_{\mathcal{A}}$. Previously, it was shown that for our chosen $q \in (0,1)$, we have

$$\|\widetilde{f}\|_{L^2(\Omega)} + \|\widetilde{g}\|_{H^{1/2}_{pw}(\partial\Omega)} \leq q \left[\|f\|_{L^2(\Omega)} + \|g\|_{H^{1/2}_{pw}(\partial\Omega)} \right].$$

By repeated execution of this decomposition scheme to the remainder term $S(\widetilde{f},\widetilde{g})$ a geometric series argument can be employed such that the solution can be rewritten as

$$u = S_k(f,g) = u_{H^2} + \sum_{i=1}^{J} \widetilde{a}_i(f,g) S_i + \widetilde{u}_{\mathcal{A}},$$

where $u_{H^2} \in H^2(\Omega)$, $\widetilde{u}_{\mathcal{A}} \in C^\infty(\Omega)$, and the coefficients $\widetilde{a}_i$ are in fact linear functionals of the data (f,g). In addition, we have with the abbreviation $C_{f,g} := \|f\|_{L^2(\Omega)} + \|g\|_{H^{1/2}_{pw}(\partial\Omega)}$ the estimates

$$\begin{aligned} \|\widetilde{u}_{\mathcal{A}}\|_{\mathcal{H}} &\leq C C_{f,g} \\ \|\Phi_{n,\overrightarrow{\beta},k} \nabla^{n+2} \widetilde{u}_{\mathcal{A}}\|_{L^2(\Omega)} &\leq C C_{sol}(k) k^{-1} C_{f,g} \gamma^n \max\{n,k\}^{n+2} \quad \forall n \in \mathbb{N}_0, \end{aligned}$$

and

$$k\|u_{H^2}\|_{\mathcal{H}} + \|u_{H^2}\|_{H^2(\Omega)} + \sum_{i=1}^{J} |\widetilde{a}_i(f,g)| \leq C C_{f,g}.$$

Finally, Lemma 3.1.2 allows us to absorb the contribution $\sum_{i=1}^{J} \widetilde{a}_i(f,g) S_i$ in the analytic part by setting $u_{\mathcal{A}} := \widetilde{u}_{\mathcal{A}} + \sum_{i=1}^{J} \widetilde{a}_i(f,g) S_i$. In view of $\beta_i < 1$, we have $2 - \beta_i \geq 1$ and arrive at

$$\begin{aligned} \|u_{\mathcal{A}}\|_{H^1(\Omega)} &\leq C(C_{sol}(k)+1) C_{f,g}, \\ k\|u_{\mathcal{A}}\|_{L^2(\Omega)} &\leq C C_{f,g}(C_{sol}(k)+k), \\ \|\Phi_{n,\overrightarrow{\beta},k} \nabla^{n+2} u_{\mathcal{A}}\|_{L^2(\Omega)} &\leq C C_{f,g} \left[C_{sol}(k) k^{-1} + k^{-1} \right] \max\{n,k\}^{n+2} \quad \forall n \in \mathbb{N}_0, \end{aligned}$$

which concludes the argument. □

3.5. Additional regularity properties

The results of this section have been published in [40]. The statements here explore whether additional regularity can be used to improve the k-dependence of the solution. Indeed, it turns out that $f \in H^1$ in conjunction with $g = 0$ provides an improvement.

Prelude: the 1D situation

Several of the regularity issues for (2.23) can be already be seen in 1D. Therefor we consider the following situation studied already in [58–60]:

$$-u'' - k^2 u = f \quad \text{in } I = (0,1), \qquad u(0) = 0, \qquad u'(1) - iku(1) = g \in \mathbb{C}. \tag{3.32}$$

The Green's function is known explicitly, namely,

$$G(x,y) = \frac{1}{k} \begin{cases} \sin(kx)e^{iky} & 0 \le x \le y \le 1 \\ \sin(ky)e^{ikx} & 0 \le y \le x \le 1 \end{cases} \tag{3.33}$$

so that the solution can be written as

$$u(x) = \int_0^1 G(x,y) f(y)\, dy + g \frac{\sin kx}{k(\cos k - i \sin k)} \tag{3.34}$$

One has the stability estimate (see, e.g., [58, Thm. 4.4])

$$\|u\|_{\mathcal{H}} \le C \left[\|f\|_{L^2(I)} + |g| \right]. \tag{3.35}$$

For smooth f, the solution formula (3.34) is an oscillatory integral (for large k) so that integration by parts is expected to give an additional power of k^{-1}. The following lemma asserts the validity of this expectation. Instead of working with the solution formula, we prove it using arguments that will also be used in the multi-d case:

Lemma 3.5.1. *The solution u of (3.32) satisfies, for a constant C independent of k, f, and g,*

$$\|u\|_{\mathcal{H}} \le C \left[k^{-1} \|f\|_{H^1(I)} + |g| \right]. \tag{3.36}$$

Proof. We may restrict our attention to the case $g = 0$. Define the function $u_0(x) := -k^{-2} f(x) + k^{-2} f(0) \cos kx$. Then $\|u_0\|_{\mathcal{H}} \le C k^{-1} \|f\|_{H^1(I)}$. The difference $\delta := u - u_0$ satisfies

$$\begin{aligned} -\delta'' - k^2 \delta &= -k^{-2} f'', \\ \delta(0) &= 0 \\ \delta'(1) - ik\delta(1) &= -(-k^{-2} f'(1) - k^{-1} f(0) \sin k) + ik(-k^{-2} f(1) + k^{-2} f(0) \cos k). \end{aligned}$$

Applying now the stability estimate (3.35) and the Sobolev embedding theorem gives

$$\begin{aligned} \|\delta\|_{\mathcal{H}} &\le C \left[k^{-2} \|f''\|_{L^2(I)} + k^{-2} |f'(1)| + k^{-1} |f(1)| + k^{-1} |f(0)| \right] \\ &\le C \left[k^{-2} \|f''\|_{L^2(I)} + k^{-1} \|f\|_{H^1(I)} \right]. \end{aligned}$$

Hence, we have obtained

$$\|u\|_{\mathcal{H}} \le \|u_0\|_{\mathcal{H}} + \|\delta\|_{\mathcal{H}} \le C \left[k^{-2} \|f\|_{H^2(I)} + k^{-1} \|f\|_{H^1(I)} \right].$$

The term $k^{-2} \|f\|_{H^2(I)}$ can be reduced to a term of the form $k^{-1} \|f\|_{H^1(I)}$ by interpolation arguments as worked out in the proof of Lemma 3.5.4 below. □

Remark 3.5.2. We note that the term involving $|g|$ in (3.36) is not improved by a factor k^{-1} as compared with (3.35). Inspection of the solution formula (3.34) shows that its k-dependence is sharp. Thus, better estimates (with respect to k) can only be expected for the case of homogeneous boundary conditions.

Concerning the regularity of the solution u of (3.32) we have:

Proposition 3.5.3. *Let $s \in \mathbb{N}_0$. Then there exist constants C, $\lambda > 0$ such that the following is true. For every $f \in H^s(I)$ and $g \in \mathbb{C}$ the solution u of (3.32) can be written as $u = u_{H^{s+2}} + u_{\mathcal{A}}$ where $u_{H^{s+2}} \in H^{s+2}(I)$ and $u_{\mathcal{A}}$ is analytic. Additionally,*

$$\begin{aligned} k^{s+2}\|u_{H^{s+2}}\|_{L^2(I)} + \|u_{H^{s+2}}\|_{H^{s+2}(I)} &\leq C\|f\|_{H^s(I)}, \\ \|u_{\mathcal{A}}\|_{\mathcal{H}} &\leq C\left[\|f\|_{L^2(I)} + |g|\right] \\ \|u_{\mathcal{A}}^{(n+2)}\|_{L^2(I)} &\leq C\lambda^n k^{-1}\max\{k,n\}^{n+2}\left[\|f\|_{L^2(I)} + |g|\right] \qquad \forall n \in \mathbb{N}_0. \end{aligned}$$

Proof. Follows by arguing as in the proof of [68, Thm. 4.5] and the appropriate modifications for the Dirichlet boundary conditions at $x = 0$ ([68, Thm. 4.5]) considers (2.23) with Robin boundary conditions). □

Regularity in higher dimensions

As in the 1D situation, it is possible to obtain a better k-dependence by exploiting additional regularity of the data f. The following result shows this for the multi-dimensional case:

Lemma 3.5.4. *Let $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$ be a bounded Lipschitz domain. Let $C_{sol}(k)$ be given by (3.6). Let $g = 0$. Then there exists $C > 0$ independent of f and k such that*

$$\|u\|_{\mathcal{H}} \leq Ck^{-1}(1 + C_{sol}(k))\|f\|_{H^1(\Omega)}.$$

Proof. Assume first $f \in H^2(\Omega)$. Define the function $u_0 := -k^{-2}f$. Then $\|u_0\|_{\mathcal{H}} \leq Ck^{-1}\|f\|_{L^2(\Omega)} + k^{-2}\|f\|_{H^1(\Omega)}$. Consequently, the function $\delta := u - u_0$ satisfies

$$\begin{aligned} -\Delta\delta - k^2\delta &= f - (-\Delta u_0 - k^2 u_0) = +k^{-2}\Delta f \qquad \text{in } \Omega, \\ \partial_n\delta - ik\delta &= 0 - (\partial_n u_0 - iku_0). \end{aligned}$$

By stability and generous trace estimates we have

$$\begin{aligned} \|\delta\|_{\mathcal{H}} &\leq C_{sol}(k)\left[k^{-2}\|\Delta f\|_{L^2(\Omega)} + k^{-2}\|\partial_n f\|_{L^2(\partial\Omega)} + k^{-1}\|f\|_{L^2(\partial\Omega)}\right] \\ &\leq CC_{sol}(k)\left[k^{-2}\|f\|_{H^2(\Omega)} + k^{-1}\|f\|_{H^1(\Omega)}\right] \end{aligned}$$

and conclude from the triangle inequality $\|u\|_{\mathcal{H}} \leq \|u_0\|_{\mathcal{H}} + \|\delta\|_{\mathcal{H}}$

$$\|u\|_{\mathcal{H}} \leq C\left[k^{-2}\|f\|_{\mathcal{H}} + C_{sol}(k)(k^{-2}\|f\|_{H^2(\Omega)} + k^{-1}\|f\|_{H^1(\Omega)})\right]. \tag{3.37}$$

In order to lower the regularity requirement for f from H^2 to H^1, we employ an interpolation argument [22, 95]. Recognizing that $H^1(\Omega)$ is the interpolation space $H^1(\Omega) = (L^2(\Omega), H^2(\Omega))_{1/2,2}$, we can write, for every $t > 0$, the function $f \in H^1(\Omega)$ as

$$f = (f - f_{H^2}) + f_{H^2},$$

where $f_{H^2} \in H^2(\Omega)$ and the following estimates are true (see [20] for details):

$$\begin{aligned} \|f - f_{H^2}\|_{L^2(\Omega)} + t\|f_{H^2}\|_{H^2(\Omega)} &\leq Ct^{1/2}\|f\|_{H^1(\Omega)}, \\ \|f_{H^2}\|_{H^1(\Omega)} &\leq C\|f\|_{H^1(\Omega)}. \end{aligned}$$

Selecting $t = k^{-2}$, we arrive at

$$\|f - f_{H^2}\|_{L^2(\Omega)} \leq k^{-1}\|f\|_{H^1(\Omega)}, \qquad \|f_{H^2}\|_{H^1(\Omega)} + k^{-1}\|f_{H^2}\|_{H^2(\Omega)} \leq C\|f\|_{H^1(\Omega)},$$

We write $u = u_1 + u_2$, where u_1 and u_2 solve

$$\left\{ \begin{array}{rcll} -\Delta u_1 - k^2 u_1 &=& f - f_{H^2} & \text{in } \Omega \\ \partial_n u_1 - iku_1 &=& 0 & \text{on } \partial\Omega \end{array} \right. \qquad \left\{ \begin{array}{rcll} -\Delta u_2 - k^2 u_2 &=& f_{H^2} & \text{in } \Omega \\ \partial_n u_2 - iku_2 &=& 0 & \text{on } \partial\Omega \end{array} \right.$$

We conclude from (3.6) for u_1 and from (3.37) for u_2 that

$$\begin{aligned} \|u\|_{\mathcal{H}} &\leq \|u_1\|_{\mathcal{H}} + \|u_2\|_{\mathcal{H}} \\ &\leq C_{sol}(k)\|f - f_{H^2}\|_{L^2(\Omega)} + C\left[k^{-2}\|f_{H^2}\|_{\mathcal{H}} + C_{sol}(k)(k^{-2}\|f_{H^2}\|_{H^2(\Omega)} + k^{-1}\|f_{H^2}\|_{H^1(\Omega)})\right] \\ &\leq C(1 + C_{sol}(k))k^{-1}\|f\|_{H^1(\Omega)}. \end{aligned}$$

□

With similar arguments as is Section 3.4 the following refined regularity result has been presented in [68]:

Proposition 3.5.5 ([68, Thm. 4.5])**.** *Let $\Omega \in \mathbb{R}^d$, $d \in \{2, 3\}$ be a bounded Lipschitz domain. Assume additionally that Ω has an analytic boundary. Let $C_{sol}(k)$ be given by (3.6). Fix $s \in \mathbb{N}_0$. Then there exist constants C, $\lambda > 0$ independent of $k \geq k_0 > 0$ such that for every $f \in H^s(\Omega)$ and $g \in H^{s+1/2}(\partial\Omega)$ the solution $u = S(f, g)$ of the Helmholtz problem (2.23) can be written as $u = u_{H^{s+2}} + u_{\mathcal{A}}$, where, for all $n \in \mathbb{N}_0$,*

$$\|u_{\mathcal{A}}\|_{\mathcal{H},\Omega} \leq CC_{sol}(k)\left(\|f\|_{L^2(\Omega)} + \|g\|_{H^{1/2}(\partial\Omega)}\right) \tag{3.38}$$

$$\|\nabla^{n+2} u_{\mathcal{A}}\|_{L^2(\Omega)} \leq C\lambda^n k^{-1} C_{sol}(k) \max\{n, k\}^{n+2}\left(\|f\|_{L^2(\Omega)} + \|g\|_{H^{1/2}(\partial\Omega)}\right) \tag{3.39}$$

$$\|u_{H^{s+2}}\|_{H^{s+2}(\Omega)} + k^{s+2}\|u_{H^{s+2}}\|_{L^2(\Omega)} \leq C\left(\|f\|_{H^s(\Omega)} + \|g\|_{H^{s+1/2}(\partial\Omega)}\right). \tag{3.40}$$

As we have seen in the 1D case, it is possible to improve the estimates by one power of k for the special case of homogeneous boundary conditions and some additional regularity of the right-hand side f. This extends to the multi-dimensional case:

Theorem 3.5.6. *Let $\Omega \in \mathbb{R}^d$, $d \in \{2,3\}$ be a bounded Lipschitz domain. Assume additionally that Ω has an analytic boundary. Fix $s \in \mathbb{N}$. Then there exist constants C, $\lambda > 0$ independent of $k \geq k_0 > 0$ such that for every $f \in H^s(\Omega)$ the solution u of (2.23) with $g = 0$ can be written as $u = u_{H^{s+2}} + u_{\mathcal{A}}$, where, for all $n \in \mathbb{N}_0$,*

$$\|u_{\mathcal{A}}\|_{\mathcal{H},\Omega} \leq Ck^{-1}(1 + C_{sol}(k))\|f\|_{H^1(\Omega)} \tag{3.41}$$
$$\|\nabla^{n+2} u_{\mathcal{A}}\|_{L^2(\Omega)} \leq C\lambda^n k^{-2}(1 + C_{sol}(k)) \max\{n,k\}^{n+2}\|f\|_{H^1(\Omega)} \tag{3.42}$$
$$\|u_{H^{s+2}}\|_{H^{s+2}(\Omega)} + k^{s+2}\|u_{H^{s+2}}\|_{L^2(\Omega)} \leq C\|f\|_{H^s(\Omega)}. \tag{3.43}$$

Proof. We decompose the data f: Using the operators L_Ω and H_Ω of [72, (4.1b)], we can write

$$f = L_\Omega f + H_\Omega f =: f_L + f_H,$$

where, by [72, Lemma 4.2, Lemma 4.3], we have for some C, $\eta > 0$ independent of k the bounds

$$\begin{aligned}
\|f_H\|_{H^{s_1}(\Omega)} &\leq Ck^{s_1 - s_2}\|f\|_{H^{s_2}(\Omega)}, \qquad 0 \leq s_1 \leq s_2 \leq s,\\
\|\nabla^p f_L\|_{L^2(\Omega)} &\leq C(\eta k)^p \|f\|_{L^2(\Omega)} \qquad \forall p \in \mathbb{N}_0,\\
\|\nabla^p f_L\|_{L^2(\Omega)} &\leq C(\eta k)^{p-s}\|f\|_{H^s(\Omega)} \qquad \forall p \geq s.
\end{aligned}$$

We denote by u_L and u_H the solutions to (2.23) with right-hand sides f_L and f_H, respectively. For u_H, we have $f_H \in H^s(\Omega)$ together with $\|f_H\|_{L^2(\Omega)} \leq Ck^{-1}\|f\|_{H^1(\Omega)}$. By Proposition 3.5.5, we may write $u_H = u_{H^{s+2}} + \widetilde{u}_{\mathcal{A}}$ with

$$\begin{aligned}
k^{s+2}\|u_{H^{s+2}}\|_{L^2(\Omega)} + \|u_{H^{s+2}}\|_{H^{s+2}(\Omega)} &\leq C\|f_H\|_{H^s(\Omega)} \leq C\|f\|_{H^s(\Omega)}\\
\|\widetilde{u}_{\mathcal{A}}\|_{\mathcal{H}} &\leq C_{sol}(k)\|f_H\|_{L^2(\Omega)}\\
\|\nabla^n \widetilde{u}_{\mathcal{A}}\|_{L^2(\Omega)} &\leq C\lambda^p k^{-1} C_{sol}(k) \max\{k,n\}^p \|f_H\|_{L^2(\Omega)} \qquad \forall n \in \mathbb{N}_0
\end{aligned}$$

recalling that $\|f_H\|_{L^2(\Omega)} \leq Ck^{-1}\|f\|_{H^1(\Omega)}$, we see that $u_{H^{s+2}}$ and $\widetilde{u}_{\mathcal{A}}$ have the desired properties. We now turn to u_L. Since f_L and $\partial\Omega$ are analytic, the solution u_L is analytic. For bounds on the derivatives of u_L, we first note that Lemma 3.5.4 yields

$$\|u_L\|_{\mathcal{H}} \leq Ck^{-1}(1 + C_{sol}(k))\|f_L\|_{H^1(\Omega)}. \tag{3.44}$$

For higher order derivatives, we proceed as in the proof of [69, Lemma 4.13]: Upon setting $\varepsilon := 1/k$, we observe that u_L satisfies

$$-\varepsilon^2 \Delta u_L - u_L = \varepsilon^2 f_L \quad \text{in } \Omega, \qquad \varepsilon^2 \partial_n u_L - i\varepsilon u_L = 0 \quad \text{on } \partial\Omega$$

with f_L satisfying the estimates above. Hence, the equation satisfied by u_L has the same structure as in the proof of [72, Lemma 4.13] making [67, Prop. 5.4.5, Rem. 5.4.6] applicable. The result is then

$$\|\nabla^{n+2} u_L\|_{L^2(\Omega)} \leq CK^{n+2} \max\{n,k\}^{n+2}\left[k^{-2}\|f_L\|_{L^2(\Omega)} + k^{-1}\|u_L\|_{\mathcal{H}}\right]$$

for a $K > 0$ independent of k and n. Inserting now the estimate (3.44) for u_L yields

$$\|\nabla^{n+2} u_L\|_{L^2(\Omega)} \leq CK^{n+2} \max\{n,k\}^{n+2} k^{-2} \left\{ \|f_L\|_{L^2(\Omega)} + (1 + C_{sol}(k)) \|f\|_{H^1(\Omega)} \right\}.$$

Using $\|f_L\|_{L^2(\Omega)} \leq C\|f\|_{L^2(\Omega)}$ and setting $u_{\mathcal{A}} := \widetilde{u}_{\mathcal{A}} + u_L$ finishes the proof. □

4. Numerical approximation

4.1. Galerkin discretisation

From [40, Section 4] we repeat here the discussion on abstract stability. For the generic Galerkin discretization we consider a sequence $(V_N)_{N\in\mathbb{N}} \subset H^1(\Omega)$ of finite dimensional subspaces. Furthermore, we assume that $(V_N)_{N\in\mathbb{N}}$ is dense in $H^1(\Omega)$ in the sense that for every $v \in H^1(\Omega)$ we have $\lim_{N\to\infty} \inf_{v_N\in V_N} \|v - v_N\|_{H^1(\Omega)} = 0$. The conforming Galerkin approximation of (2.25) reads then:

$$\text{Find } u_N \in V_N : \quad B(u_N, v) = l(v) \quad \forall v \in V_N. \tag{4.1}$$

Since the sesquilinear form B satisfies a Gårding inequality, the general functional analytic argument show that *asymptotically* the inf-sup condition holds on the discrete level and, the discrete problem (4.1) has a unique solution u_N. In addition, quasi-optimality holds (see, e.g., [86, Thm. 4.2.9], [87]). More precisely, there exists $N_0 > 0$ and $C > 0$ such that for all $N \geq N_0$

$$\|u - u_N\|_{H^1(\Omega)} \leq C \inf_{v\in V_N} \|u - v\|_{H^1(\Omega)}. \tag{4.2}$$

However this asymptotic convergence result does not indicate how C and N_0 depend on discretization parameters of the approximation space and the wave number k. An alternative to such a general convergence analysis faces convergence estimates based on approximation properties of the finite element space that have to be fulfilled. In particular, it is of interest to get a quasi-optimality result which gives an explicit indication (in particular explicit in k) on how to choose an appropriate discretization space. A convergence analysis based on the concept used by Schatz [87] has been developed by Melenk/Sauter [69, 72, 84] and will be discussed in detail in Section 5.1.

The problem (4.1) is solved by the Galerkin method as follows: Let $\{\varphi_n\}_{n=1}^N$ be a basis of the approximation space V_N where the number of degrees of freedom N is the dimension of V_N. While the choice of the basis is immaterial from a theoretical point of view, it is crucial from the point of view of numerics. Inserting the approximation ansatz

$$u_N(x) = \sum_{n=1}^{N} u_n \varphi_n(x)$$

into the Galerkin approximation (4.1) and extracting sums out of the integrals leads to a linear system of equations. The resulting finite element scheme can the be written in matrix form as

$$\text{Find } \mathbf{u} \in \mathbb{C}^N : \quad \mathbf{Bu} = \mathbf{l},$$

where $\mathbf{u} := (u_1, \ldots, u_N)$ is the unknown coefficient vector of u_N, $\mathbf{B} = [B(\varphi_i, \varphi_j)]_{i,j}$ is a $N \times N$-matrix and $\mathbf{l} = [(l, \varphi_i)]_{i=1}^N$. Consequently, we can observe that the choice of basis functions is essential in order to retrieve convenient numerical properties of the matrix $\mathbf{B}$.

A typical example of the Galerkin method is the *Galerkin finite element method* where the domain is divided into elements and the basis functions are of local character and specified in each element. This method will be discussed in detail in the following.

4.2. High-order finite element methods

We review from [69] and [72] the conforming mesh and the high-order basis function and adopt the notations from [29] for our high-order finite element method.

As a first step within the process of the finite element method the domain $\Omega \subset \mathbb{R}^d, d \in \{1, 2, 3\}$ has to be discretisized by a *triangulation* $\mathcal{T}_h$ which has to satisfy the following properties:

- $\bar{\Omega} = \bigcup_{K \in \mathcal{T}_h} K$
- K closed and $K \neq \emptyset \quad \forall K \in \mathcal{T}_h$
- $\forall K \in \mathcal{T}_h : K$ is a Lipschitz domain.
- For each pair $K_1, K_2 \in \mathcal{T}_h$ with $K_1 \neq K_2$ it holds $\mathring{K}_1 \cap \mathring{K}_2 = \emptyset$
- Each element K can be prescribed from the *reference element*

$$\hat{K} := \{x \in \mathbb{R}^d_{\geq 0} : \sum_{i=1}^{d} x_i \leq 1\}$$

  and an associated bijective element mapping $F_K : \hat{K} \to K$ such that $K = F_K(\hat{K})$.

- The *vertices, edges and faces* of each element K are thus the images of vertices, edges and faces of the reference element $\hat{K}$ and the element maps F_K induce the same parametrization on a common edge or face of two adjoining elements.
- For each pair $K_1, K_2 \in \mathcal{T}_h$ with $K_1 \neq K_2$ it holds

$$\overline{K}_1 \cap \overline{K}_2 = \begin{cases} \emptyset \text{ for } d \in \{1, 2, 3\} \\ \text{exactly one common vertex for } d \in \{2, 3\} \\ \text{exactly one common edge for } d \in \{2, 3\} \\ \text{exactly one common face for } d = 3. \end{cases}$$

Other types of reference elements and different strategies of mesh constructions can be seen in [67]. Up to now we considered Lipschitz domains with analytic as well as piecewise (curvi-)linear boundary. To this end we adopt the assumptions made in [72] and [69]:

Assumption 4.2.1 (Triangulation for domains with analytic boundary)**.** *We present here a patch-wise construction for a triangulation in order to preserve the approximation properties of any finite element method. This approach has been studied inter alia in [67, 75].*

1. *Let $\tilde{\mathcal{T}}_{macro}$ be a coarse triangulation of the smooth domain Ω consisting of curvilinear triangles resp. tetrahedral elements of mesh size $O(1)$ and call these elements "patches". For example, this mesh can be constructed by the method of Dubois [35] or Gordon and Hall [49], furnished with the element maps*

$$F_{macro,K} = \tilde{F}_{macro,K} \circ F_{0,K} : \hat{K} \to K_{macro} \to \tilde{K}_{macro}.$$

2. *The finite element mesh with step size h is generated by a standard refinement to N_{loc} elements on each patch leading to the triangulation*

$$\tilde{\mathcal{T}}_h = \{\tilde{K} := \tilde{F}_{macro,K} \circ F_{K,i}(\hat{K}), i = 1, \dots, N_{loc}, K \in \mathcal{T}_{macro}\}.$$

Each element mapping can be written as $F_{\tilde{K}} = D_{\tilde{K}} \circ M_{\tilde{K}} : \hat{K} \to K \to \tilde{K}$, where $M_{\tilde{K}}$ is an affine mapping and $D_{\tilde{K}}$ is an h-independent analytic map which corresponds to the metric distortion at the possibly curved boundary. The maps $D_{\tilde{K}}$ and $M_{\tilde{K}}$ satisfy for constants $C_{affine}, C_{metric}, \gamma > 0$ independent of h:

$$\|M'_{\tilde{K}}\|_{L^\infty(\hat{K})} \leq C_{affine} h, \qquad \|(M'_{\tilde{K}})^{-1}\|_{L^\infty(\hat{K})} \leq C_{affine} h^{-1} \tag{4.3}$$

$$\|(D'_{\tilde{K}})^{-1}\|_{L^\infty(K)} \leq C_{metric}, \qquad \|\nabla^n D_{\tilde{K}}\|_{L^\infty(K)} \leq C_{metric} \gamma^n n! \ \forall n \in \mathbb{N}_0 \tag{4.4}$$

In the case of polygonal domains the regularity analysis suggests that near corners the solution can be better approximated on geometric refined meshes. Therefore, the triangulation away from corners is assembled as in the case of a smooth domain while in an $O(h)$-neighbourhood of the vertices the mesh is refined geometrically.

Assumption 4.2.2 (Triangulation for polygonal domains)**.** *We denote again the vertices of the polygon Ω by $A_j, j = 1, \dots, J$ and the ball with radius ch centered at A_j by $B_{ch}(A_j)$ where c is independent of h. Further let $L \in \mathbb{N}, \sigma \in (0,1)$ be constants independent of h.*

1. *The restriction of $\mathcal{T}_h$ to $\Omega \backslash (\bigcup_{j=1}^J B_{ch}(A_j))$ consists of regular elements with mesh size h. Each element K here can be obtained by an affine mapping $F_K : \hat{K} \to K$ such that $F_K(\hat{K}) = K$. In particular, the element map has the form $F_K(x) = A_K x + d_K$ where $A_K \in \mathbb{R}^{d \times d}$ is an affine scaling and rotation operator and $d_K \in \mathbb{R}^d$ a vector of displacement.*

2. *For each vertex A_j, the restriction of $\mathcal{T}_h$ to $B_{ch}(A_j) \cap \Omega$ is geometrically refined toward A_j by a grading factor $\sigma \in (0,1)$ and $L+1 \in \mathbb{N}$ layers of geometric refinements. Here, for each element K with $dist(K, A_j) > 0$ it holds*

$$C^{-1} dist(K, A_j) \leq diam(K) \leq C dist(K, A_j)$$

and if $dist(K, A_j) = 0$*, then*

$$diam(K) \leq C\sigma^L h.$$

Each element map can be written as $F_K = s_K^{-1} \circ M_K$*, where* $M_K : \hat{K} \to K$ *is an affine element mapping and* s_K *is an affine stretching map* $s_K : \tilde{K} \to K$ *defined by* $s_K(x) := diam(K)x$*. The resulting element map is an analytic diffeomorphism from* $\hat{K}$ *to the stretched element* $\tilde{K}$ *and satisfy for some constants* $C, \gamma > 0$*:*

$$\|\nabla^n (s_K^{-1} \circ M_K)\|_{L^\infty(\hat{K})} \leq C\gamma^n n! \quad \forall n \in \mathbb{N}_0 \tag{4.5}$$

$$\| \left(\left(s_K^{-1} \circ M_K \right)' \right)^{-1} \|_{L^\infty(\hat{K})} \leq C. \tag{4.6}$$

For piece-wise smooth Lipschitz domains such as curvilinear polygons, a more advanced strategy by combining both assumptions would be needed.

Once the appropriate triangulation $\mathcal{T}_h$ is build up, we defined the H^1-conforming finite-element spaces by

$$S^{p,1}(\mathcal{T}_h) := \{u \in H^1(\Omega) \mid \forall K \in \mathcal{T}_h : u|_K \circ F_K \in \mathcal{P}_p\} \tag{4.7}$$

for some $p \in \mathbb{N}$. This space of piecewise polynomials has finite dimension and thus we can imply that there exists at least one basis $\{\varphi_n\}_{n=1}^N$ where the dimension N is called degree of freedom. In particular, the *basis functions* are induced, via the element maps, by the shape function of the reference element $\hat{K}$ which span the approximation space $S^{p,1}(\hat{K})$.

For our numerical computations in 1D we choose the following shape functions base on Legendre polynomials on $\hat{K} = [-1, 1]$:

$$\begin{aligned} N_0(\xi) = \frac{1-\xi}{2}, \quad N_1(\xi) = \frac{1+\xi}{2}, \\ N_i(\xi) = \frac{1}{\sqrt{4i-2}} \big[L_i(\xi) - L_{i-2}(\xi) \big], \quad 2 \leq i \leq p. \end{aligned} \tag{4.8}$$

A detailed definition of the Legendre polynomials can be found in [1]. For our 2D-computations we use the high order FEM software package NETGEN/NGSOLVE by J. Schöberl [88, 89]. For a discussion of the shape functions in higher dimensions see [33, 91].

At last it remains to specify the *quadrature scheme* in order to compute the coefficients of the FEM-matrix $\mathbf{B}$ and the load vector $\mathbf{l}$ as emphasized in Section 4.1. For our one dimensional computations we use the Legendre-Gauss quadrature where the i-th Gauss node ξ_i is the i-th root of Legendre polynomial L_p and its weights are given by $\omega_i = \frac{2}{(1-\xi_i^2)[L_p'(\xi_i)]^2}$. This integration scheme is exact for polynomials of a degree smaller than or equal $2p + 1$.

In order to clarify terminology we speak of *hp*-FEM when both the size of elements as well as the degree of polynomials varies over the domain. In *h*-FEM the degree of polynomials of the basis is fixed while the discretization of the geometry is refined. In *p*-FEM the degree of polynomials increases while the size of elements is fixed.

4.3. Spectral element methods

The spectral elements method (SEM), proposed by Patera in 1984 [81], is similar to the *p*-version of FEM except that it uses different integration points. It combines the advantages of spectral methods [24,50] with the geometrical flexibility of the finite element methods. In one dimension, a typical choice of basis functions would be the high-order Lagrange interpolation polynomials (nodal functions) associated with the local Gauss-Lobatto integration points defined per element. However, for our purpose, we stick to the integrated Legendre polynomials (4.8) together with Gauss-Lobatto quadrature. The main difference to the *p*-FEM is thus the number and location of the integration nodes on each element. Here, the node points are located at the $p+1$ Gauss-Lobatto-Legendre (GLL) points which are the $p+1$ zeros of $(1-\xi^2)L'_p(\xi) = 0$ and the same $p+1$-point GLL quadrature is employed together with the weights $\omega_i = \frac{2}{p(p-1)[L_{p-1}(\xi_i)]^2}$.

Comparing the FEM- and SEM-matrix with basis functions $\varphi_i, \varphi_j \in \mathbb{P}_p$, the matrix coefficients consist of integrals over polynomials $\pi_{ij} = \varphi_i\varphi_j \in \mathbb{P}_{2p}$. We already mentioned that the FEM-coefficients are evaluated exactly as the Legendre-Gauß quadrature is even exact for polynomials with a degree up to $2p+1$. However, the Gauß-Lobatto quadrature is only exact for polynomials of a degree smaller than or equal $2p-1$.

5. Convergence analysis

5.1. Quasi-optimality and adjoint approximability

5.1.1. Abstract Results

Based on the discussion in Section 4.1 we aim to provide convergence estimates for finite element spaces which have to fulfill certain approximation properties. To this end investigations in previous works [9,12,69,70,72,84,87] employed an analytical tool related to a "dual regularity theory". Based on the concept used in [87] and the gerneralisation of the theory in [70] developed in [84] leads to the following result:

Lemma 5.1.1 ([69, Thm. 3.2]). *Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain and B be defined in (2.26). Denote by $S^\star : L^2(\Omega) \to H^1(\Omega)$ the solution operator for the problem*

$$\text{Find } u^\star \in H^1(\Omega) \text{ s.t.} \qquad B(v, u^\star) = (v, f)_{L^2(\Omega)} \qquad \forall v \in H^1(\Omega). \tag{5.1}$$

Define the adjoint approximation property $\eta(V_N)$ *by*

$$\eta(V_N) := \sup_{f \in L^2(\Omega)} \inf_{v \in V_N} \frac{\|S^\star(f) - v\|_{\mathcal{H}}}{\|f\|_{L^2(\Omega)}}.$$

If, for the continuity constant C_B of (2.28), the space V_N satisfies

$$2C_B k \eta(V_N) \leq 1, \tag{5.2}$$

then the solution u_N of (4.1) exists and satisfies

$$\|u - u_N\|_{\mathcal{H}} \leq 2 \inf_{v \in V_N} \|u - v\|_{\mathcal{H}}. \tag{5.3}$$

Remark 5.1.2. *From the proof of quasi-optimality in Lemma 5.1.1 we get as a side result the L^2-estimate*

$$\|u - u_N\|_{L^2} \leq \eta(V_N) \|u - u_N\|_{\mathcal{H}}. \tag{5.4}$$

This can be shown via duality techniques. Denote $e = u - u_N$ and $\psi = S^\star e$. Now let ψ_N

be the best approximation of ψ, then it follows

$$\begin{aligned}
\|e\|_{L^2}^2 &= (e,e)_{L^2} = B(e,\psi) = B(e,\psi-\psi_N) \\
&\lesssim \|e\|_{\mathcal{H}}\|\psi-\psi_N\|_{\mathcal{H}} \\
&\lesssim \|e\|_{\mathcal{H}} \inf_{v\in V_N} \|\psi - v\|_{\mathcal{H}} \\
&\lesssim \|e\|_{\mathcal{H}} \inf_{v\in V_N} \tfrac{\|\psi-v\|_{\mathcal{H}}}{\|e\|_{L^2}} \|e\|_{L^2} \\
&\lesssim \|e\|_{\mathcal{H}} \eta(V_N) \|e\|_{L^2}.
\end{aligned}$$

Thus it remains to focus on estimates for the adjoint approximability that are explicit in the discretization parameters and the wave number. In particular this means we need to get regularity estimates for the solution operator $S^\star$. In fact, we have $S^\star f = \overline{S(\overline{f},0)}$, where an overbar denote complex conjugation. Thus, the regularity theory of Chapter 3 is applicable. The remaining results, here in Chapter 5, are presented by the author and J. M. Melenk in [40] and [39].

Based on the additional considerations in Section 3.5, we obtain the following statement:

Lemma 5.1.3. *Define*

$$\eta_N^{H^1} := \sup_{\substack{f\in H^1(\Omega)\\ f\neq 0}} \inf_{v\in V_N} \frac{\|S^*f - v\|_{\mathcal{H}}}{\|f\|_{H^1(\Omega)}}, \qquad \eta_N^{H_0^1} := \sup_{\substack{f\in H_0^1(\Omega)\\ f\neq 0}} \inf_{v\in V_N} \frac{\|S^*f - v\|_{\mathcal{H}}}{\|f\|_{H^1(\Omega)}}. \tag{5.5}$$

If the solvability condition (5.2) is satisfied, then the Galerkin error $u-u_N$ satisfies

$$\begin{aligned}
\|u-u_N\|_{(H^1(\Omega))'} &\leq 2C_c^2 \eta_N^{H^1} \inf_{v\in V_N} \|u-v\|_{\mathcal{H}} \\
\|u-u_N\|_{H^{-1}(\Omega)} &\leq 2C_c^2 \eta_N^{H_0^1} \inf_{v\in V_N} \|u-v\|_{\mathcal{H}}.
\end{aligned}$$

Proof. We will just prove the estimate for $\|u-u_N\|_{(H^1(\Omega))'}$ using a duality argument and Galerkin orthogonality: For arbitrary $v\in H^1(\Omega)$ and $w_N\in V_N$ we have

$$\left|(u-u_N,v)_{L^2(\Omega)}\right| = |B(u-u_N,S^*v)| = |B(u-u_N,S^*v-w_N)| \leq C_c\|u-u_N\|_{\mathcal{H}}\|S^*v-w_N\|_{\mathcal{H}}.$$

Since w_N is arbitrary, we can conclude

$$\left|(u-u_N,v)_{L^2(\Omega)}\right| \leq C_c\|u-u_N\|_{\mathcal{H}}\|v\|_{H^1(\Omega)}\eta_N^{H^1}.$$

Dividing by $\|v\|_{H^1(\Omega)}$, taking the supremum over $v\in H^1(\Omega)$, and inserting the best approximation result (5.3) yields the claimed bound. □

Lemma 5.1.1 resp. Lemma 5.1.3 decomposes the error analysis of the Galerkin discretization into two separate tasks. On one hand, we are interested in finding bounds for the adjoint approximability η and on the other hand, we need estimates of the best approximation $\inf_{v\in V_N}\|u-v\|$. Through this approach, we will obtain estimates for the FEM-error that are explicit in the wave number and the discretization parameters.

5.1.2. Approximability (for smooth domains)

We study the adjoint approximablity according to *hp*-finite element spaces $V_N = \mathcal{S}^{p,1}(\mathcal{T})$ where $\mathcal{T}$ satisfies the geometric assumptions in Section 4.2. Further we restrict ourselves here to the case of domains with analytic boundary. The regularity assertions of Proposition 3.5.5 and Theorem 3.5.6 allow us to estimate the quantities $\eta_N^{L^2}$ as well as $\eta_N^{H^1}$ and $\eta_N^{H_0^1}$ of (5.5):

Theorem 5.1.4. *Let $\Omega\subset\mathbb{R}^d$, $d\in\{2,3\}$ be a bounded Lipschitz domain with an analytic boundary. Then there exist constants C, $\sigma>0$ independent of h, k, and p such that*

$$\eta_N^{L^2} \leq C\left[\frac{h}{p}+C_{sol}(k)k^{-1}\left\{\left(\frac{h}{h+\sigma}\right)^p+k\left(\frac{kh}{\sigma p}\right)^p\right\}\right], \tag{5.6}$$

$$\eta_N^{H_0^1}\leq\eta_N^{H^1} \leq C\left[\left(\frac{h}{p}\right)^{\min\{2,p\}}+(1+C_{sol}(k))k^{-2}\left\{\left(\frac{h}{h+\sigma}\right)^p+k\left(\frac{kh}{\sigma p}\right)^p\right\}\right] \tag{5.7}$$

Proof. The estimate (5.6) has already been shown in [72, Prop. 5.3]; it follows from the approximation properties of V_N in combination with the regularity assertion in Proposition 3.5.5. The first bound in (5.7) follows directly from the definition. For the second estimate in (5.7), let $f\in H^1(\Omega)$ be arbitrary. Then $u:=S^*f=\overline{S(\overline{f},0)}$ can, according to Theorem 3.5.6 with $s=1$, be written as

$$u=u_{H^3}+u_{\mathcal{A}},$$

where the contributions u_{H^3} and $u_{\mathcal{A}}$ have the regularity properties stated there. Therefore, we get

$$\begin{aligned}\inf_{v\in V_N}\|u_{H^3}-v\|_{\mathcal{H}} &\leq C\left(\frac{h}{p}\right)^{\min\{2,p\}}\|f\|_{H^1(\Omega)},\\ \inf_{v\in V_N}\|u_{\mathcal{A}}-v\|_{\mathcal{H}} &\leq Ck^{-2}(1+C_{sol}(k))\|f\|_{H^1(\Omega)}\left[\left(\frac{h}{h+\sigma}\right)^p+k\left(\frac{hk}{\sigma p}\right)^p\right].\end{aligned}$$

The result now follows. □

Remark 5.1.5. As mentioned in Section 3.2, for our model problem (2.23), the constant $C_{sol}(k)$ satisfies here the polynomial bound (3.5) with $\theta=5/2$. Hence, the crucial condition (5.2) is satisfied if, for a sufficiently small c_1 and a sufficiently large c_2, the following two conditions are satisfied:

$$\frac{kh}{p}\leq c_1 \qquad \text{and} \qquad p\geq c_2\log k. \tag{5.8}$$

5.1.3. Best Approximation (for smooth domains)

By the use of the regularity statements in Section 3.5 we can give the following estimate for the best approximation:

Theorem 5.1.6. *Let $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$ be a bounded Lipschitz domain with an analytic boundary. Let u be the solution of (2.23), then we have:*

(i) If $s \in \mathbb{N}_0$ and $f \in H^s(\Omega)$ and $g \in H^{s+1/2}(\partial\Omega)$, then

$$\begin{aligned}\inf_{v\in V_N} \|u-v\|_{\mathcal{H}} \quad &\leq \quad C\left(\frac{h}{p}\right)^{\min\{s+1,p\}} \left[\|f\|_{H^s(\Omega)} + \|g\|_{H^{s+1/2}(\partial\Omega)}\right] \\ &\quad + C_{sol}(k)k^{-1}\left\{\left(\frac{h}{h+\sigma}\right)^p + k\left(\frac{kh}{\sigma p}\right)^p\right\}\left[\|f\|_{L^2(\Omega)} + \|g\|_{H^{1/2}(\partial\Omega)}\right].\end{aligned}$$

(ii) If $s \in \mathbb{N}$ and $f \in H^s(\Omega)$ and $g = 0$, then

$$\begin{aligned}\inf_{v\in V_N} \|u-v\|_{\mathcal{H}} \quad &\leq \quad C\left(\frac{h}{p}\right)^{\min\{s+1,p\}} \|f\|_{H^s(\Omega)} \\ &\quad + (1+C_{sol}(k))k^{-2}\left\{\left(\frac{h}{h+\sigma}\right)^p + k\left(\frac{kh}{\sigma p}\right)^p\right\}\|f\|_{H^1(\Omega)}.\end{aligned}$$

Proof. (i) follows from Proposition 3.5.5 and the approximation properties of V_N. The estimate in (ii) is shown similarly, but we are able to exploit the improved k-dependence of Theorem 3.5.6. □

5.1.4. Quasi-optimality on polygons

The estimates of Theorem 3.4.6 suggest that the effect of the corner singularities is essentially restricted to an $O(1/k)$-neighborhood of the vertices. This motivates us to consider meshes that are refined in a small neighborhood of the vertices as described in Section 4.2. Then we can show that stability of the *hp*-FEM is ensured if the mesh size h and the polynomial degree p satisfy the scale resolution condition (5.8) and, additionally, $L = O(p)$ layers of geometric refinement are used near the vertices:

Theorem 5.1.7. *Let $\mathcal{T}_{h,L}^{geo}$ denote the geometric meshes on the polygon $\Omega \subset \mathbb{R}^2$ satisfying Assumption 4.2.2. Fix $c_3 > 0$. Then there are constants c_1, $c_2 > 0$ depending solely on Ω and the shape-regularity of the mesh $\mathcal{T}_{h,L}^{geo}$ such that the following is true: If h, p, and L satisfy the conditions*

$$\frac{kh}{p} \leq c_1 \quad \text{and} \quad p \geq c_2 \log k \quad \text{and} \quad L \geq c_3 p \tag{5.9}$$

then the hp-FEM based on the space $S^p(\mathcal{T}_{h,L}^{geo})$ has a unique solution $u_N \in S^p(\mathcal{T}_{h,L}^{geo})$ and

$$\|u-u_N\|_{\mathcal{H}} \leq 2 \inf_{v\in S^p(\mathcal{T}_{h,L}^{geo})} \|u-v\|_{\mathcal{H}} \tag{5.10}$$

Proof. By Lemma 5.1.1, we have to estimate $k\eta(V_N)$ with $V_N = S^p(\mathcal{T}_{h,L}^{geo})$. Recalling the definition of $\eta(V_N)$ we let $f \in L^2(\Omega)$ and observe that we can decompose $S^\star f = u_{H^2} + u_{\mathcal{A}}$, where u_{H^2} and $u_{\mathcal{A}}$ satisfy the bounds

$$\begin{aligned} \|u_{H^2}\|_{H^2(\Omega)} &\leq C\|f\|_{L^2(\Omega)}, \\ \|\Phi_{n,\overrightarrow{\beta},k}\nabla^{n+2}u_{\mathcal{A}}\|_{L^2(\Omega)} &\leq C(C_{sol}(k)+1)k^{-1}\gamma^n \max\{k,n\}^{n+2}\|f\|_{L^2(\Omega)} \qquad \forall n \in \mathbb{N}_0. \end{aligned}$$

Piecewise polynomial approximation on $\mathcal{T}_{h,L}^{geo}$ as discussed in [69, Prop. 5.6] gives under the assumptions $kh/p \leq C$ and $L \geq c_3 p$: (inspection of the proof of [69, Prop. 5.6] shows that only bounds on the derivatives of order ≥ 2 are needed):

$$\begin{aligned} \inf_{v\in V_N} \|u_{H^2} - v\|_{\mathcal{H}} &\leq C\frac{h}{p}\|f\|_{L^2(\Omega)}, \\ \inf_{v\in V_N} \|u_{\mathcal{A}} - v\|_{\mathcal{H}} &\leq C\left[(kh)^{1-\beta_{max}}e^{ckh-bp} + \left(\frac{kh}{\sigma_0 p}\right)^p\right](C_{sol}(k)+1)\|f\|_{L^2(\Omega)}, \end{aligned}$$

where $\beta_{max} = \max_{j=1,\dots,J}\beta_j < 1$, and C, c, $b > 0$ are constants independent of h, p, and k. From this, we can easily infer

$$k\eta(V_N) \leq C\left\{\frac{kh}{p} + k(C_{sol}(k)+1)\left[(kh)^{1-\beta_{max}}e^{ckh-bp} + \left(\frac{kh}{\sigma_0 p}\right)^p\right]\right\}.$$

Noting that Theorem 3.2.1 gives $C_{sol}(k) = O(k^{5/2})$, and selecting c_1 sufficiently small as well as c_2 sufficient large allows us to make $k\eta(V_N)$ so small that the condition (5.2) in Lemma 5.1.1 is satisfied. □

5.2. *hp*-Convergence

For the numerical analysis of the *hp*-FEM the convergence of the error is examined in dependence of both the mesh size h and the polynomial degree p.

5.2.1. Exponential convergence on polygons

Corollary 5.2.1 (exponential convergence on geometric meshes)**.** *Let f be analytic on $\overline{\Omega}$ and g be piecewise analytic, i.e., f, g satisfy (3.23). Given $c_3 > 0$, there exist c_1, $c_2 > 0$ such that under the scale resolution conditions (5.9) of Theorem 5.1.7, the finite-element approximation $u_N \in S^p(\mathcal{T}_{h,L}^{geo})$ exists, and there are constants C, $b > 0$ independent of k such that the error $u - u_N$ satisfies*

$$\|u - u_N\|_{\mathcal{H}} \leq Ce^{-bp}.$$

Proof. In view of Theorem 5.1.7, estimating $\|u - u_N\|_{\mathcal{H}}$ is purely a question of approximability for c_1 sufficiently small and c_2 sufficiently large. Lemma 3.4.1 gives that the

solution $u = S(f,g)$ satisfies the bounds given there and, as in the proof of Theorem 5.1.7, we conclude from [69, Prop. 5.6](more precisely, this follows from its proof)

$$\inf_{v\in V_N} \|u_{\mathcal{A}} - v\|_{\mathcal{H}} \;\leq\; C\left[(kh)^{1-\beta_{max}} e^{ckh-bp} + \left(\frac{kh}{\sigma_0 p}\right)^p\right](C_{sol}(k)+1)(\widetilde{C}_f + \widetilde{C}_g).$$

Theorem 3.2.1 asserts $C_{sol}(k) = O(k^{5/2})$, which implies the result by suitably adjusting c_1 and c_2 if necessary. □

5.2.2. Convergence on smooth domains with higher regularity data

We are now in the position to formulate some *a priori* error estimates. For the reader's convenience we introduce the following shorthand notation:

$$\varepsilon(h,p,k) := \left(\frac{h}{h+\sigma}\right)^p + k\left(\frac{kh}{\sigma p}\right)^p \tag{5.11}$$

Corollary 5.2.2. *Assume the hypotheses of Theorem 5.1.6. Assume in addition that h and p are such that condition (5.2) is satisfied.*

(i) If $s \in \mathbb{N}_0$, $f \in H^s(\Omega)$ and $g \in H^{s+1/2}(\partial\Omega)$, then with $C_{f,g} := \|f\|_{H^s(\Omega)} + \|g\|_{H^{s+1/2}(\partial\Omega)}$

$$\begin{aligned}
\|u-u_N\|_{\mathcal{H}} &\leq CC_{f,g}\left\{\left(\frac{h}{p}\right)^{\min\{s+1,p\}} + k^{-1}C_{sol}(k)\varepsilon(h,p,k)\right\},\\
\|u-u_N\|_{L^2(\Omega)} &\leq C\left\{\left(\frac{h}{p}\right) + k^{-1}C_{sol}(k)\varepsilon(h,p,k)\right\}\|u-u_N\|_{\mathcal{H}},\\
\|u-u_N\|_{H^{-1}(\Omega)} &\leq C\left\{\left(\frac{h}{p}\right)^{\min\{2,p\}} + k^{-2}(1+C_{sol}(k))\varepsilon(h,p,k)\right\}\|u-u_N\|_{\mathcal{H}}.
\end{aligned}$$

(ii) If $s \in \mathbb{N}$ and $f \in H^s(\Omega)$ and $g = 0$, then

$$\begin{aligned}
\|u-u_N\|_{\mathcal{H}} &\leq CC_{f}\left\{\left(\frac{h}{p}\right)^{\min\{s+1,p\}} + k^{-2}(1+C_{sol}(k))\varepsilon(h,p,k)\right\},\\
\|u-u_N\|_{L^2(\Omega)} &\leq C\left\{\left(\frac{h}{p}\right) + k^{-2}(1+C_{sol}(k))\varepsilon(h,p,k)\right\}\|u-u_N\|_{\mathcal{H}},\\
\|u-u_N\|_{H^{-1}(\Omega)} &\leq C\left\{\left(\frac{h}{p}\right)^{\min\{2,p\}} + k^{-2}(1+C_{sol}(k))\varepsilon(h,p,k)\right\}\|u-u_N\|_{\mathcal{H}}.
\end{aligned}$$

Proof. Follows by combining the results on the adjoint approximability and the best approximation in Section 5.1.2 and 5.1.3, respectively. □

5.3. h-Convergence

The above considerations are formulated in a general hp-setting. We will now present some numerical examples for the h-FEM which feature a more pronounced k-dependence. In this case, we can simplify

$$\varepsilon(h,p,k) \leq C_p k^{p+1} h^p.$$

The next corollary follows from Corollary 5.2.2 by fixing p. Additionally, we assume explicitly the condition (3.5) in order to make the k-dependence more visible:

Corollary 5.3.1. *Assume the hypotheses of Corollary 5.2.2 and (3.5). Fix $p \in \mathbb{N}$. Then:*

(i) If $s \in \mathbb{N}_0$, $f \in H^s(\Omega)$, and $g \in H^{s+1/2}(\partial\Omega)$, then with $C_{f,g} = \|f\|_{H^s(\Omega)} + \|g\|_{H^{s+1/2}(\partial\Omega)}$

$$\begin{aligned}
\|u-u_N\|_{\mathcal{H}} &\leq CC_{f,g}\left[h^{\min\{s+1,p\}} + k^{\theta}(kh)^p\right], \\
\|u-u_N\|_{L^2(\Omega)} &\leq CC_{f,g}\left[h^{\min\{s+1,p\}} + k^{\theta}(kh)^p\right] h\left[1 + k^{\theta+1}(kh)^{p-1}\right], \\
\|u-u_N\|_{H^{-1}(\Omega)} &\leq CC_{f,g}\left[h^{\min\{s+1,p\}} + k^{\theta}(kh)^p\right] \begin{cases} h^2(1+k^{\theta+1}(kh)^{p-2}) & \text{if } p \geq 2 \\ hk^{\theta} & \text{if } p = 1. \end{cases}
\end{aligned}$$

(ii) If $s \in \mathbb{N}$ and $f \in H^s(\Omega)$ and $g = 0$, then with $C_f = \|f\|_{H^s(\Omega)}$

$$\begin{aligned}
\|u-u_N\|_{\mathcal{H}} &\leq CC_{f}\left[h^{\min\{s+1,p\}} + k^{\theta-1}(kh)^p\right], \\
\|u-u_N\|_{L^2(\Omega)} &\leq CC_{f}\left[h^{\min\{s+1,p\}} + k^{\theta-1}(kh)^p\right] h\left[1 + k^{\theta+1}(kh)^{p-1}\right], \\
\|u-u_N\|_{H^{-1}(\Omega)} &\leq CC_{f,g}\left[h^{\min\{s+1,p\}} + k^{\theta-1}(kh)^p\right] \begin{cases} h^2(1+k^{\theta+1}(kh)^{p-2}) & \text{if } p \geq 2 \\ hk^{\theta} & \text{if } p = 1. \end{cases}
\end{aligned}$$

Remark 5.3.2. A different way of phrasing the $L^2(\Omega)$-convergence result is as follows: If Ω has an analytic boundary and we assume that the exact solution $u \in H^{m+1}(\Omega)$ of (2.23) satisfies

$$|u|_{H^j(\Omega)} \sim k^j, \qquad j = 0, \ldots, m+1,$$

and the solvability condition (5.2) is satisfied, then

$$\|u-u_N\|_{L^2(\Omega)} \leq C_m\Big(\frac{hk}{p}\Big)^{m+1}\Big\{1 + \Big[1 + \frac{k}{\sigma}\Big(\frac{hk}{\sigma p}\Big)^{p-1}\Big]\big(C_{sol}(k)+1\big)\Big\}. \tag{5.12}$$

This follows by combining the estimate for $\eta_N^{L^2}$ with the a *priori* bound $\inf_{v\in V_N}\|u-v\|_{\mathcal{H}} \leq Ch^m\|u\|_{H^{m+1}(\Omega)} \leq Ck(kh)^m$. The estimate (5.12) illustrates the special nature of the case $p = 1$ when it comes to the k-dependence.

5.4. Numerical examples

5.4.1. Examples for polygons

All calculations reported in this section are performed with the *hp*-FEM code NETGEN/NGSOLVE by J. Schöberl, [88, 89].

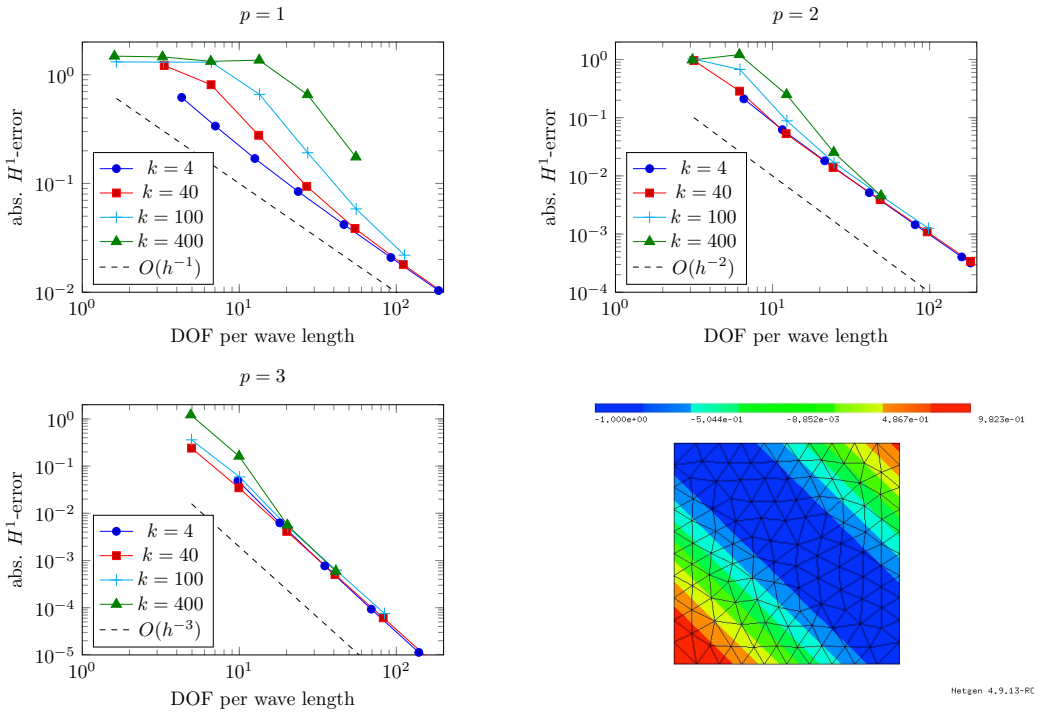


Figure 5.1.: Top: *h*-FEM with $p = 1$ (left) and $p = 2$ (right) as described in Example 5.4.1. Bottom left: *h*-FEM with $p = 3$ as described in Example 5.4.1. Bottom right: Illustration of the solution on the unit square.

Example 5.4.1. We consider the model problem (2.23) with $f = 0$ on the unit square $\Omega = [0,1]^2$. The boundary data g is chosen such that the exact solution is a plane wave $e^{i(k_1 x + k_2 y)}$, where $k_1 = -k_2 = \frac{1}{\sqrt{2}}k$ and $k \in \{4, 40, 100, 400\}$. For fixed $p \in \{1, 2, 3\}$, we show in Fig. 5.1 the performance of the *h*-FEM for $p \in \{1, 2, 3\}$ on quasi-uniform meshes by displaying the relative error in the H^1-seminorm versus the number of degrees of freedom per wavelength $N_\lambda := \frac{2\pi}{k}\sqrt{N/|\Omega|}$. The dotted lines indicate the asymptotic rate of convergence for the *h*-FEM. We observe that higher order methods are less prone to numerical pollution of the FEM scheme. We note that the meshes are quasi-uniform, i.e., no geometric mesh refinement near the vertices is performed in contrast to the requirements of Theorem 5.1.7.

Example 5.4.2. On the *L*-shaped domain $\Omega = [-1,1]^2 \setminus (0,1) \times (-1,0)$ we consider the same problem as in Example 5.4.1. Furthermore, we consider two kinds of meshes,

namely, quasi-uniform meshes $\mathcal{T}_h$ with mesh size h such that $kh \approx 4$ and meshes $\mathcal{T}^{geo}$ that are geometrically refined near the origin. The meshes $\mathcal{T}^{geo}$ are derived from the quasi-uniform mesh $\mathcal{T}_h$ by introducing a geometric grading on the elements abutting the origin; the grading factor is $\sigma = 0.125$ and the number of refinement levels is $L = 10$. Fig. 5.2 shows the relative errors in the H^1-seminorm for the p-version of the FEM where for fixed mesh the approximation order p ranges from 1 to 10. It is particularly noteworthy that the refinement near the origin has hardly any effect on the convergence behavior of the FEM; this is quite in contrast to the stability result Theorem 5.1.7, which requires geometric refinement near all vertices of Ω.

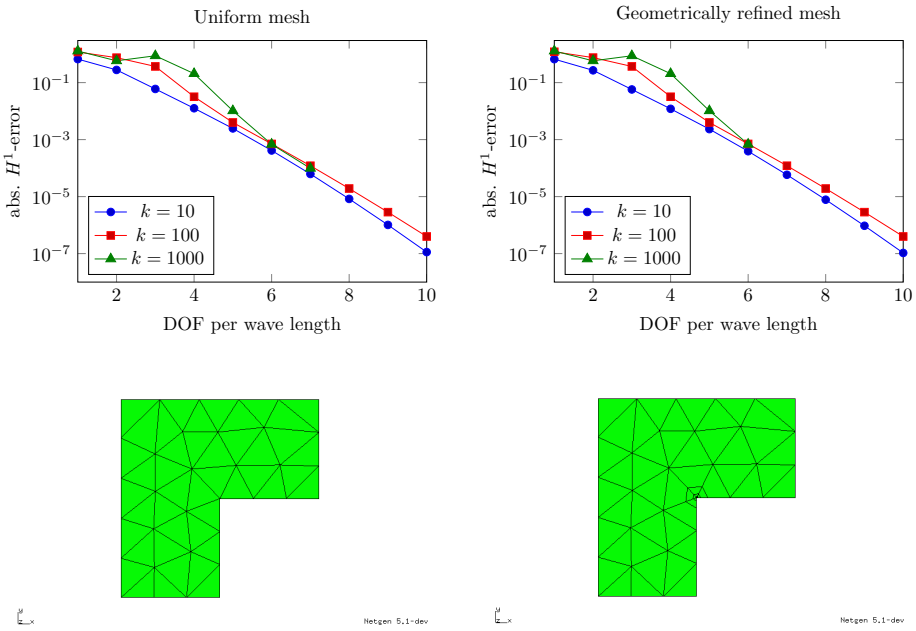


Figure 5.2.: Top: p-FEM for plane wave solution as described in Example 5.4.2 for a quasiuniform mesh $\mathcal{T}_h$ (left) with $kh \approx 4$ and the geometric mesh $\mathcal{T}^{geo}$ (right) obtained from $\mathcal{T}$ by strong geometric refinement near origin. Bottom: Illustration of the uniform (left) and the geometric (right) mesh, generated by NETGEN.

Example 5.4.3. On the sector $\Omega = B_1(0) \setminus (0,1) \times (-1,0)$ with Γ being the union of the two edges meeting at $(0,0)$, we consider the problem

$$-\Delta u - k^2 u = 0 \quad \text{in } \Omega, \quad \partial_n u = 0 \quad \text{on } \Gamma, \quad \partial_n u - iku = g \quad \text{on } \partial\Omega \setminus \Gamma, \tag{5.13}$$

The data g are selected such that the exact solution is $u = J_{2/3}(kr)\cos\frac{2}{3}\varphi$, where (r,φ) denote polar coordinates and J_α is a first kind Bessel function and $k \in \{10, 100, 1000\}$. Our calculations are p-FEMs with $p \in \{1, \dots, 10\}$ on a geometrically refined mesh $\mathcal{T}_h$ as depicted in Figure 5.3. This mesh is a standard one in hp-FEM for elliptic problems,

whose key features are described, for example, in [52, 90]. The results are displayed in Figure 5.3 and show the very fast convergence of the p-FEM on graded meshes.

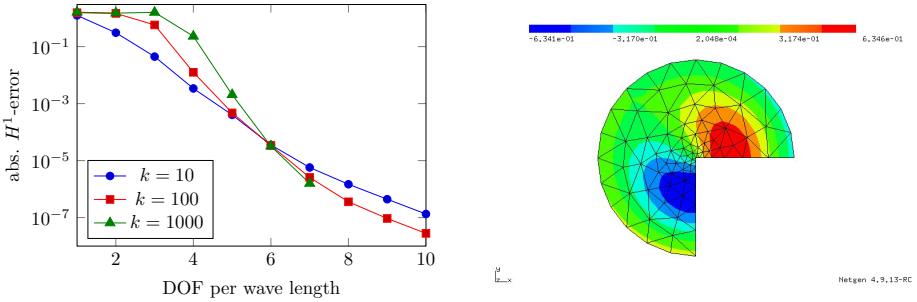


Figure 5.3.: Left: p-FEM for singular solution on a graded mesh as described in Example 5.4.3. Right: Illustration of the solution generated with NETGEN.

5.4.2. Examples for smooth domains in 2D

The following three two-dimensional examples consider

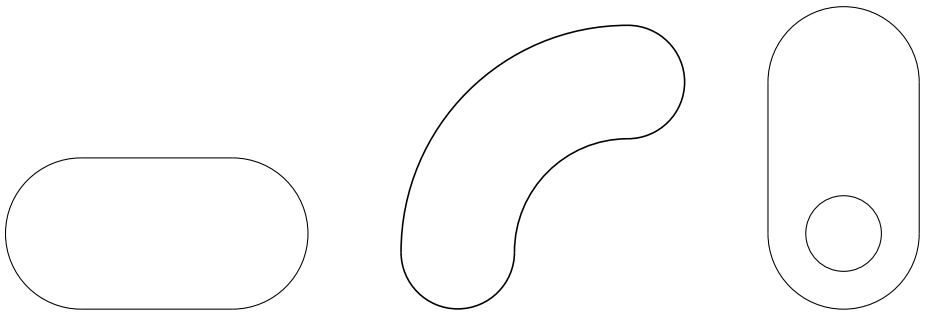
$$\begin{aligned} -\Delta u - |\mathbf{k}|^2 u &= 0 \quad \text{in} \quad \Omega, \\ \partial_n u - i|\mathbf{k}| u &= g \quad \text{on} \quad \partial\Omega, \end{aligned}$$

where g is chosen such that the exact solution u is given by $u(\mathbf{x}) = e^{i\mathbf{x}\cdot\mathbf{k}} = e^{i(k_1 x + k_2 y)}$ with $k_1 = -k_2 = k/\sqrt{2}$ and k ranges from 4 to 80. The first geometry studied in Example 5.4.4 is convex so that the stability bound (3.5) is true with $\theta = 0$; the geometries in Examples 5.4.5, 5.4.6 are non-convex so that only the bound with $\theta = 5/2$ is available. The estimate (3.5) with $\theta = 5/2$ is likely to be pessmistic. Indeed, the numerical results for all three geometries are very similar, and we do not observe a stronger pollution effect in Examples 5.4.5, 5.4.6 than in Example 5.4.4. All three examples are computed with the hp-FEM software package NETGEN/NGSOLVE by J. Schoeberl [88, 89]. The curved geometry is resolved using high order approximations as provided by NETGEN/NGSOLVE.

Example 5.4.4. The domain consists of the square $[-1, 1]^2$ with two semicircular caps attached, i.e., $\overline{\Omega} = [-1, 1]^2 \cup \{(x, y) | (x \pm 1)^2 + y^2 \leq 1\}$ (see Fig. 5.4, (a)), and $k \in \{4, 20, 80\}$. For fixed $p = 1$, 2 we present in Fig. 5.5 the absolute error in L^2 as well as in the H^1-seminorm versus the number of degrees of freedom per wavelength. We observe the same marked difference between the cases $p = 1$ and $p = 2$ that we have seen already emphasized by Remark 5.3.2. and which is explained by Corollary 5.3.1.

Example 5.4.5. The model problem remains the same as in Example 5.4.4. Only the geometry is modified to a non-convex domain, see Fig. 5.4, (b). The domain Ω is a subset of $[-2, 0.5] \times [-0.5, 2]$, thus $\operatorname{diam} \Omega \lesssim 3.5$. The h-FEM with $k = \{4, 20, 80\}$ and $p = 1$, 2 is presented in Fig. 5.6.

Example 5.4.6. We consider the same problem as in Example 5.4.4 on a non-simply connected domain, see Fig. 5.4, (c). The h-FEM has been computed for $k = \{8, 20, 80\}$ and $p = 1$, 2 is shown in Fig. 5.7.



(a) Example 5.4.4 (b) Example 5.4.5 (c) Example 5.4.6

Figure 5.4.: Domain geometries of the concerning examples

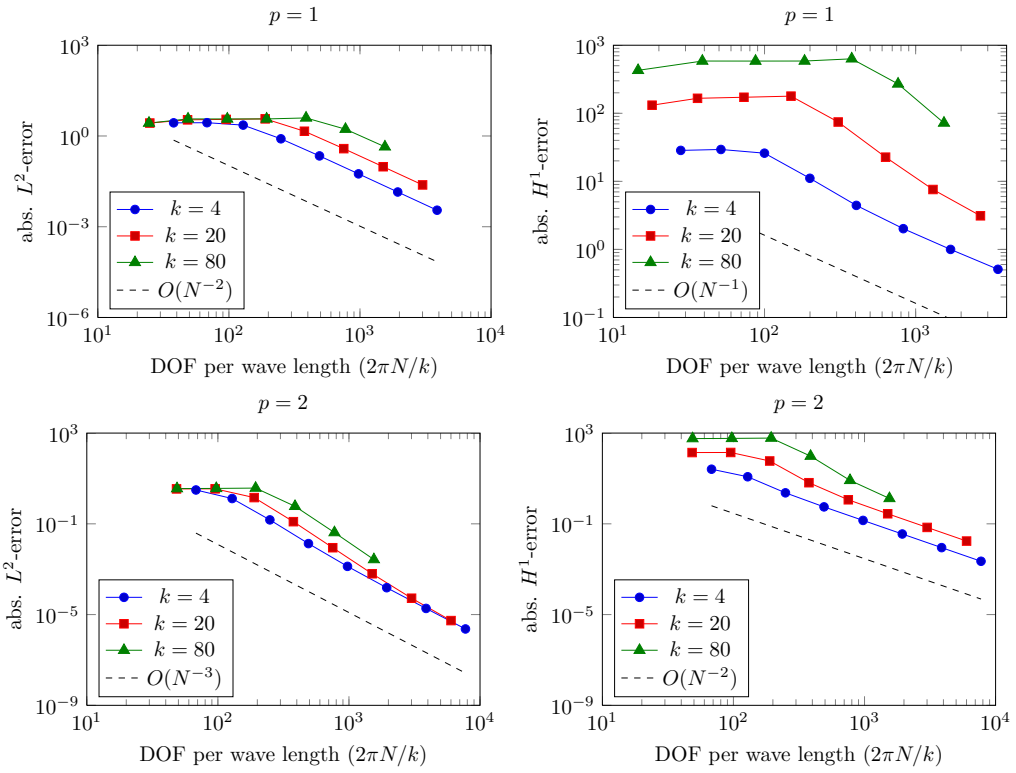


Figure 5.5.: (cf. Example 5.4.4) h-FEM with $p = 1$ and $p = 2$ for the smooth convex domain, see Fig. 5.4(a). Top: L^2-error. Bottom: H^1-seminorm error.

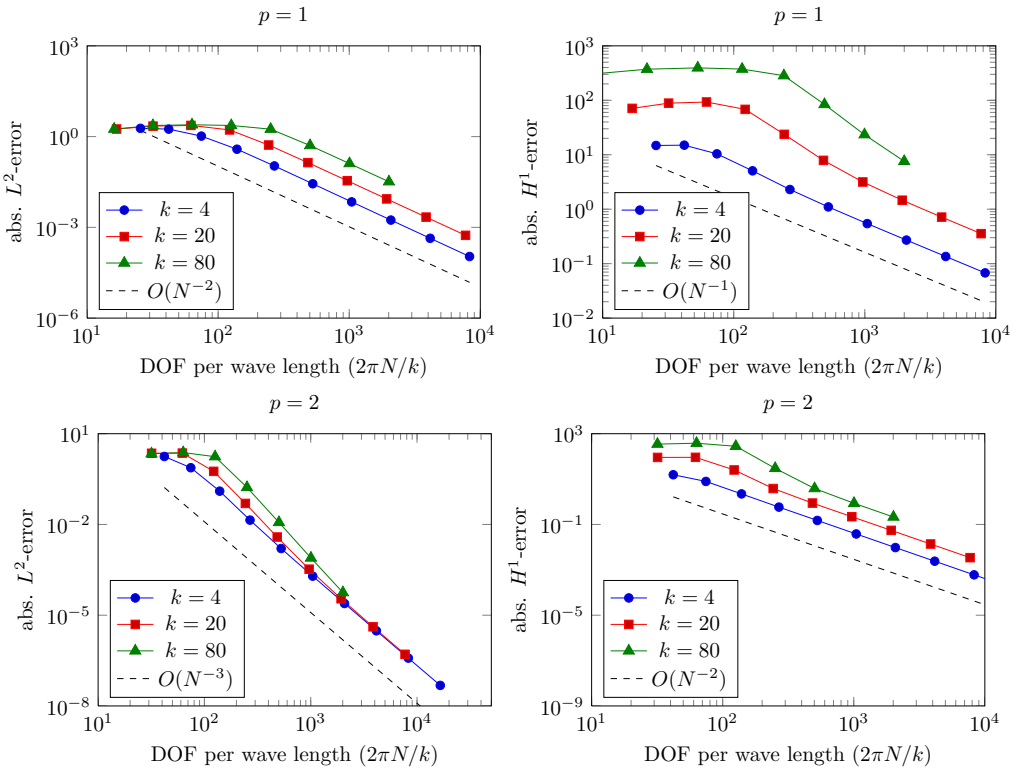


Figure 5.6.: (cf. Example 5.4.5) h-FEM with $p = 1$ and $p = 2$ for the non-convex domain, see Fig. 5.4(b). Top: L^2-error. Bottom: H^1-seminorm error.

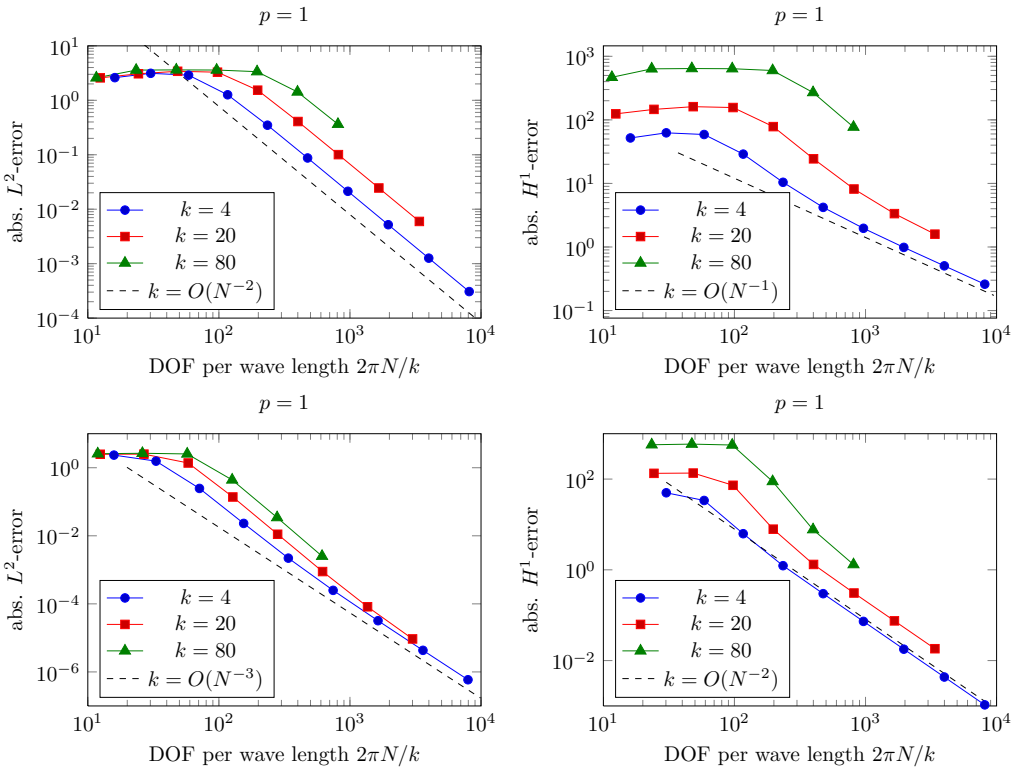


Figure 5.7.: (cf. Example 5.4.6) h-FEM with $p = 1$ and $p = 2$ for the domain with a hole, see Fig. 5.4(c). Top: L^2-error. Bottom: H^1-seminorm error.

6. Dispersion analysis

6.1. Numerical pollution

Differently to the approach in of Chapter 5 we can also carry out an error analysis which focuses on the dispersive behavior of the finite element method. From the discussion on quasi-optimality (Section 4.1, equation (4.2)) we see that the accuracy of the FEM-solution depends on the approximability of the finite element space up to a constant factor which might depend inter alia on the wave number k. This approximability is measured in terms of the minimal error which is only related to the original function space and the according finite element space. If the minimum is reached (uniquely) then this function is called the *best approximation*. Furthermore the error of best approximation is, by definition, smaller than or equal to the interpolation error.

Then the error of the FEM solution can basically be related to the well known error of interpolation [22] by

$$\|u - u_N\| \leq \|u - u_I\| + \|u_I - u_N\|. \tag{6.1}$$

The first term on the right hand side is considered to be the *natural error* (interpolation error, locally determined), while the second term can be interpreted as the numerical *pollution error* related to the computational scheme; this contribution of the FEM-error is of non-local character. This effect comes essentially from the *numerical dispersion*, that is the wave number of the FEM solution is different from the wave number of the exact solution.

In order to illustrate the dispersion effect in a more explicit way we consider the homogeneous Helmholtz problem with the Robin boundary condition in one dimension. There the exact solution on the unbounded domain $\mathbb{R}$ is $u(x) = e^{ikx}$. Considering the low order FEM discretization with $P = 1$ and uniform mesh size h the resulting FEM matrix can be expressed by

$$B = tridiag(R(h,k), 2S(h,k), R(h,k))$$

with R, S being some algebraic functions of h and k. A detailed definition can be found in [58]. Imaging an infinite regular grid the linear system reads line by line:

$$R(h,k)u_{j-1} + 2S(h,k)u_j + R(h,k)u_{j+1} = 0, j \in \mathbb{Z}$$

With the ansatz $u_N := e^{i\tilde{k}\tilde{x}}$(in analogy to the exact solution) where $\tilde{x}_j := jh$ and $\tilde{k}$ the corresponding *discrete wave number*, this leads to

$$R(h,k)e^{i\tilde{k}h(j-1)} + 2S(h,k)e^{i\tilde{k}hj} + R(h,k)e^{i\tilde{k}h(j+1)} = 0.$$

Using the notation $\lambda := e^{i\tilde{k}h}$ transforms the previous equation into a quadratic equation with solutions

$$\begin{array}{rcl} \lambda_{1,2} & = & -\frac{S(h,k)}{R(h,k)} \pm \sqrt{\frac{S^2(h,k)}{R^2(h,k)} - 1} \\ \| & & \\ e^{\pm i\tilde{k}h} & = & \cos(h\tilde{k}) \pm i\sin(h\tilde{k}) \end{array}$$

Note that we assume $\tilde{k}$ to be real valued for a proper description of propagating waves. In consequence $\lambda_{1,2}$ are complex which means that $\left|\frac{S(h,k)}{R(h,k)}\right| < 1$ has to hold. Identifying the real and imaginary part gives then $\tilde{k} = \frac{1}{h}\arccos(-\frac{S(h,k)}{R(h,k)})$, thus via Taylor expansion we find that

$$\tilde{k} = k(1 + O(hk)^2).$$

For high-order FEM, the matrix can again be transformed to a tridiagonal matrix via condensation. The general *dispersion relation* can be written as [5]

$$\tilde{k} = k(1 + O(hk)^{2p}), \quad p \geq 1.$$

Hence the error of the FEM solution is mainly subject to a phase shift in the corresponding discrete wave number $\tilde{k}$. As the polynomial interpolant is obviously in phase with the exact solution the pollution term results from the evaluation of the second term on the right hand side in (6.1). This dispersive character is thus a global property. In 1D this can be avoided with a special FEM, however it can be shown that this effect is inevitable in higher dimensions [10].

6.2. The optimally blended FE-SE scheme

A more detailed investigation of the dispersive behavior of the finite element approximation shows that this computational scheme provides a *phase lead* for the corresponding discrete wave number. In contrast, spectral element methods as introduced in Section 4.3 exhibits a *phase lag*. This motivates to design a new scheme that exploits the different dispersive behaviors of two distinct schemes in order to minimize the numerical pollution. This idea is worked out in [5] by blending the finite element method with the spectral element method giving the two terms the relative weight $1 : p$. That is, the bilinear form of the so-called "optimally blended scheme" is given by

$$B_{FE-SE}(u,v) = \frac{1}{p+1} B_{FEM}(u,v) + \left(1 - \frac{1}{p+1}\right) B_{SEM}(u,v),$$

where p is the polynomial degree employed and $B_{FEM}(\cdot,\cdot)$ and $B_{SEM}(\cdot,\cdot)$ are bilinear forms of the Galerkin method and the spectral element method respectively. Seeking a solution in $S^{p,1}(\mathcal{T}_h)$ the coefficients of the bilinearform consists of integrals of polynomials with maximal degree $2p$. In comparison, using p quadrature nodes, B_{FEM} is thus evaluated exactly as the Gauß-Legendre quadrature is exact for polynomials with degree small than or equal to $2p+1$, while B_{SEM} with the Gauß-Lobatto rule is only exact for

polynomials with degree small than or equal to $2p-1$.

Published in [39], we show in Theorem 6.3.1 below that estimates for the dispersion error translate into actually error bounds. We will do this specifically for the lowest order case $p=1$ and errors in the L^2-norm for the model problem (3.32), since the lowest order case is particularly striking in that the difference between the Galerkin method and the optimally blended method is most pronounced in this case. To that end, we introduce the following general notation for the discrete formulation of the Helmholtz problem: Find $u_h^\alpha \in V_N$ such that

$$B_h^\alpha(u_h^\alpha, v) = l(v) := \int_I f\varphi \qquad \forall \varphi \in V_N, \tag{6.2}$$

where $V_N \subset \{v \in H^1(I) \,|\, v(0)=0\}$ consists of the classical piecewise linear functions on a regular mesh of mesh size h. Here, the parameter α indicates the dispersion order of the approximation scheme providing the following property of the discrete wave number $\tilde{k}$:

$$|\tilde{k}-k| = kO(hk)^{2\alpha}, \quad \alpha \geq 1 \tag{6.3}$$

This assumption covers the classical Galerkin method with $\alpha=1$ (this case has been analyzed previously in [58, Sec. 4.6.4]) and the optimally blended scheme with $\alpha=2$. Moreover it was shown in [3–5] that

$$\begin{array}{rcll} |k-\tilde{k}| & = & kO((kh)^{2p}) & \text{for the FEM and SEM approximation with } \alpha = p, \\ |k-\tilde{k}| & = & kO((kh)^{2p+2}) & \text{for the optimally blended FE-SE method and } \alpha = p+1. \end{array}$$

Thus one can see that this method leads to a better accuracy of additional two orders compared to FEM or SEM approximation.

6.3. "Phase shift"-explicit convergence theory

6.3.1. Setting and main result

As Ainsworth and Wajid [5] restrict their work to present a constructive numerical scheme and the dispersion analysis, we focus on the numerical error analysis in dependence on the dispersion rate. In order to point this out we consider the following Helmholtz model in one dimension:

$$\begin{array}{rl} -\Delta u - k^2 u = f \quad \text{in} \quad \Omega = (0,1) \\ u(0) = 0 \\ u'(1) - iku(1) = 0. \end{array} \tag{6.4}$$

For $f \in L^2(\Omega)$ the solution can be written in the form

$$u(x) = \int_\Omega G_k(x,y) f(y) dy$$

with the Green's function G_k introduced in Section 2.1. Likewise the discrete Green's function as described, for example, in [58], is given in terms of the discrete wavenumber $\tilde{k}(\alpha)$ by

$$G_h^\alpha(x,y) = \frac{1}{h\sin(h\tilde{k})} \begin{cases} \sin(\tilde{k}x)\big(A\sin(\tilde{k}y) + \cos(\tilde{k}y)\big) & 0 \le x \le y \le 1 \\ \sin(\tilde{k}y)\big(A\sin(\tilde{k}x) + \cos(\tilde{k}x)\big) & 0 \le y \le x \le 1 \end{cases}$$

with

$$A := A(k,\tilde{k}) = \frac{(hk)^2\sin(\tilde{k})\cos(\tilde{k}) + i\sqrt{12}\sqrt{12-(hk)^2}}{12-(hk)^2\cos^2(\tilde{k})}.$$

Then the discrete solution u_h^α can be expressed by

$$u_h^\alpha(x_i) = h\sum_j G_h^\alpha(x_i,x_j)r_h(x_j).$$

Here, the nodes $x_i = ih$, $i = 0,\dots,N$, represent the mesh, and the functions φ_i, $i = 0,\dots,N$ are the classical hat functions associated with the nodes x_i. Furthermore the discrete right-hand side is defined by the scalar product of the data f and the nodal shape functions, i.e.:

$$r_h(x_j) = h\int_\Omega f(y)\varphi_j(y)dy.$$

In particular we can see via Taylor expansion at $hk = 0$ that $A = i + O(hk)^2$ and note that $\tilde{\alpha}_i := A(k,\tilde{k})\sin(\tilde{k}x_i) + \cos(\tilde{k}x_i) = e^{i\tilde{k}x_i} + O(hk)^2$.

Then we can state the following result:

Theorem 6.3.1. *Let u be the exact solution of the 1D-Helmholtz problem (3.32) with $g = 0$. Let u_h^α be the piecewise linear function solving (6.2). Then, for sufficiently smooth f and $kh = O(1)$ the error $u - u_h^\alpha$ can be estimated by*

$$\|u - u_h^\alpha\|_{L^2} \lesssim h^2\big(1 + k(hk)^{2(\alpha-1)}\big)C_f,$$

where C_f depends only on the data f.

Proof. Let $u_I \in \mathcal{S}^{1,p}(\mathcal{T})$ be the linear interpolation of the solution u and apply the triangle inequality:

$$\|u - u_h^\alpha\|_{L^2} \le \|u - u_I\|_{L^2} + \|u_I - u_h^\alpha\|_{L^2}$$

From approximation theory we know

$$\|u - u_I\|_{L^2} \le Ch^2\big(\|f\|_\infty + \|f'\|_\infty\big) \lesssim h^2 C_f.$$

Hence it remains to estimate

$$\begin{aligned} \|u_I - u_h^\alpha\|_{L^2}^2 &= \textstyle\int_\Omega |(u_I - u_h^\alpha)(x)|^2 dx \le h\sum_i |u(x_i) - u_h^\alpha(x_i)|^2 \\ &\le h\sum_i \Big|\int_\Omega G(x_i,y)f(y)dy - h\sum_j G_h^\alpha(x_i,x_j)r_h(x_j)\Big|^2 \\ &\le h\sum_i \Big|A_i + B_i\Big|^2 \end{aligned} \tag{6.5}$$

with

$$A_i := \int_0^{x_i} G(x_i,y)f(y)dy - \Big(h\sum_{j=0}^{i-1} G_h^\alpha(x_i,x_j)r_h(x_j) + \frac{h}{2}G_h^\alpha(x_i,x_i)r_h(x_i)\Big)$$

and

$$B_i := \int_{x_i}^1 G(x_i,y)f(y)dy - \Big(h\sum_{j=i+1}^{N} G_h^\alpha(x_i,x_j)r_h(x_j) + \frac{h}{2}G_h^\alpha(x_i,x_i)r_h(x_i)\Big).$$

Under the assumption that kh (and thus $\tilde{k}h$) is small the following estimate holds

$$|A_i + B_i| \;\lesssim\; \Big\{\frac{1}{k^2}(hk)^2\big(1+O(hk)^2\big) + \varepsilon\frac{1}{k}\big(1+O(hk)^2\big)\Big\}\big(\|f\|_\infty + \|f'\|_\infty\big)$$

where $\varepsilon = (hk)^{2\alpha}, \tilde{k} = k(1+\varepsilon)$ and $\alpha \geq 1$. These estimates will be discussed in more detail in Section 6.3.2. Hence

$$\|u_I - u_h\|_{L^2} \lesssim \sqrt{h\sum_i |A_i+B_i|^2} \lesssim \Big\{\frac{1}{k^2}(hk)^2\big(1+O(hk)^2\big) + \varepsilon\frac{1}{k}\big(1+O(hk)^2\big)\Big\}\big(\|f\|_\infty + \|f'\|_\infty\big)$$

and

$$\|u-u_h\|_{L^2} \lesssim \big(h^2 + k^{-1}\varepsilon\big)C_f \lesssim h^2\big(1+k(hk)^{2(\alpha-1)}\big)C_f.$$

□

Remark 6.3.2. *We assume that the solution behaves like $\|u\|_{L^2} \sim k^{-2}$ (as can be acertained for smooth f using the solution formula). For the case of $\alpha = 1$, which represents the FEM approximation for the discrete wave number $\tilde{k}$, it follows:*

$$\|u-u_h\|_{L^2} \lesssim kh^2 C_f$$

which leads us to the estimate

$$\frac{\|u-u_h\|_{L^2}}{\|u\|_{L^2}} \lesssim k(hk)^2. \tag{6.6}$$

This means that the convergence of the relative error of the FEM approximation is explicitly dependent on the wave number. And the case $\alpha = 2$ representing the optimally blended scheme gives

$$\|u_I - u_h\|_{L^2} \lesssim h^2(1+k(hk)^2)C_f.$$

Thus in this case we arrive at

$$\frac{\|u-u_h\|_{L^2}}{\|u\|_{L^2}} \lesssim (hk)^2 \tag{6.7}$$

which shows that the convergence of the relative error of the optimally blended scheme is independent of the wave number.

The statements in Theorem 6.3.1 and the auxiliary results in the following Section have been published in [39].

6.3.2. Auxiliary results

As discussed in Theorem 6.3.1 we need to estimate the pollution error $e_{poll} = u_I - u_h^\alpha$ in the L^2-norm. This will be done in (6.3.2). In particular from (6.5) we have to estimate

$$\|u_I - u_h^\alpha\|_{L^2}^2 \lesssim h \sum_i |A_i + B_i|^2.$$

Our strategy now is to estimate both terms A_i and B_i by the same expression (independent of i). The basic mechanism to obtain additional powers of k^{-1} on the continuous level is an integration by parts. On the discrete level, this role is taken by the summation by parts formula:

$$\sum_{j=0}^{N} x_j y_j = y_N \sum_{j=0}^{N} x_j - \sum_{j=0}^{N-1} \sum_{l=0}^{j} x_l \big(y_{j+1} - y_j\big) \tag{6.8}$$

and we will also require the identity

$$\sum_{l=0}^{j} \sin \tilde{k} x_l = \frac{\sin \frac{j\tilde{k}h}{2} \sin(\frac{j+1}{2}\tilde{k}h)}{\sin \frac{\tilde{k}h}{2}} = \frac{\cos \frac{\tilde{k}h}{2} - \cos(j + \frac{1}{2})\tilde{k}h}{2 \sin \frac{\tilde{k}h}{2}}. \tag{6.9}$$

In addition we will use the abbreviation

$$\tilde{\alpha}_i := A(k, \tilde{k}) \sin(\tilde{k} x_i) + \cos(\tilde{k} x_i). \tag{6.10}$$

Estimations of the right hand side

We start by studying the discrete right-hand side defined by

$$r_h(x_j) = h \int_\Omega f(y) \varphi_j(y) dy$$

which has the following properties:

Lemma 6.3.3. *With the abbreviation $x_{j+1/2} := (j + 1/2)h$, we have*

(i) $r_h(0) = 0$

(ii) $r_h(x_N) = \frac{h^2}{2} f(x_N) - \frac{h^3}{6} f'(x_N) + O(h^4 \|f''\|_\infty)$

(iii) $r_h(x_j) = h^2 f(x_j) + O(h^4 \|f''\|_\infty), \quad$ *for* $0 < i < N$

(iv) $r_h(x_1) - r_h(0) = r_h(x_1)$

(v) $r_h(x_N) - r_h(x_{N-1}) = -\frac{h^2}{2} f(x_{N-1/2}) + f'(x_{N-1/2}) \frac{7h^3}{12} + O(h^4 \|f''\|_\infty)$

(vi) $r_h(x_{j+1}) - r_h(x_j) = h^3 f'(x_{j+1/2}) + \frac{h^5}{8} f'''(x_{j+1/2}) + O(h^6 \|f^{(4)}\|_\infty)$ *for* $0 < j < N-2$

Proof. In more detail, we have due to the Dirichlet boundary conditions on the left boundary that $r_h(0) = 0$. Further we can calculate

$$\begin{aligned}
r_h(N) &= h\int_0^1 f(y)\varphi_N(y)dy = h\int_{x_{N-1}}^{x_N} f(y)\frac{y-x_{N-1}}{h}dy \\
&= \int_{x_{N-1}}^{x_N} \big(f(x_N) + f'(x_N)(y-x_N) + f''(x_N)\frac{(y-x_N)^2}{2} + O((y-x_N)^3)\big)(y-x_{N-1})dy \\
&= f(x_N)\int_{x_{N-1}}^{x_N}(y-x_{N-1})dy + f'(x_N)\int_{x_{N-1}}^{x_N}(y-x_N)(y-x_{N-1})dy \\
&\quad + f''(x_N)\int_{x_{N-1}}^{x_N}\frac{(y-x_N)^2}{2}(y-x_{N-1}) + \int_{x_{N-1}}^{x_N} O((y-x_N)^3)(y-x_{N-1})dy \\
&= f(x_N)\frac{h^2}{2} - \frac{h^3}{6}f'(x_N) + O(h^4)
\end{aligned}$$

and

$$\begin{aligned}
r_h(j) &= h\int_0^1 f(y)\varphi_i(y)dy = h\int_{x_{j-1}}^{x_j} f(y)\frac{y-x_{j-1}}{h}dy + h\int_{x_j}^{x_{j+1}} f(y)\frac{x_{i+1}-y}{h}dy \\
&= \int_{x_{j-1}}^{x_j} \big(f(x_j) + f'(x_j)(y-x_j) + f''(x_j)\frac{(y-x_j)^2}{2} + O((y-x_j)^3)\big)(y-x_{j-1})dy \\
&\quad + \int_{x_j}^{x_{j+1}} \big(f(x_j) + f'(x_j)(y-x_j) + f''(x_j)\frac{(y-x_j)^2}{2} + O((y-x_j)^3)\big)(x_{j+1}-y)dy \\
&= f(x_j)\Big(\int_{x_{j-1}}^{x_j}(y-x_{j-1})dy + \int_{x_j}^{x_{j+1}}(x_{j+1}-y)dy\Big) \\
&\quad + f'(x_j)\Big(\int_{x_{j-1}}^{x_j}(y-x_j)(y-x_{j-1})dy + \int_{x_j}^{x_{j+1}}(y-x_j)(x_{j+1}-y)dy\Big) \\
&\quad + f''(x_j)/2\Big(\int_{x_{j-1}}^{x_j}(y-x_j)^2(y-x_{j-1})dy + \int_{x_j}^{x_{j+1}}(y-x_j)^2(x_{j+1}-y)dy\Big) + \dots \\
&= f(x_j)h^2 + O(h^4),
\end{aligned}$$

for $0 < i < N$. Property (iv) follows obviously from (i). Next,

$$\begin{aligned}
&r_h(x_N) - r_h(x_{N-1}) = \\
&= h\int_0^1 f(y)\varphi_N(y)dy - h\int_0^1 f(y)\varphi_{N-1}(y)dy \\
&= -h\int_{x_{N-2}}^{x_{N-1}} f(y)\varphi_{N-1}(y)dy + h\int_{x_{N-1}}^{x_N} f(y)\big(\varphi_N(y) - \varphi_{N-1}(y)\big)dy \\
&= -\int_{x_{N-2}}^{x_{N-1}} \big(f(x_{N-1/2}) + f'(x_{N-1/2})(y-x_{N-1/2}) + O((y-x_{N-1/2})^2)\big)(y-x_{N-2})dy \\
&\quad + \int_{x_{N-1}}^{x_N} \big(f(x_{N-1/2}) + f'(x_{N-1/2})(y-x_{N-1/2}) + O((y-x_{N-1/2})^2)\big)(2y-(x_N+x_{N-1}))dy
\end{aligned}$$

$$
\begin{aligned}
&= f(x_{N-1/2})\Big(- \int_{x_{N-2}}^{x_{N-1}} (y - x_{N-2})dy + \int_{x_{N-1}}^{x_N} (2y - (x_N + x_{N-1}))dy\Big) \\
&\quad + f'(x_{N-1/2})\Big(- \int_{x_{N-2}}^{x_{N-1}} (y - x_{N-1/2})(y - x_{N-2})dy \\
&\qquad\qquad + \int_{x_{N-1}}^{x_N} (y - x_{N-1/2})(2y - (x_N + x_{N-1}))dy\Big) \\
&\quad + \Big(\int_{x_{N-2}}^{x_{N-1}} O((y - x_{N-1/2})^2)(y - x_{N-2})dy \\
&\qquad + \int_{x_{N-1}}^{x_N} O((y - x_{N-1/2})^2)(2y - (x_N + x_{N-1}))dy\Big) \\
&= -f(x_{N-1/2})\frac{h^2}{2} + f'(x_{N-1/2})\frac{7h^3}{12} + O(h^4).
\end{aligned}
$$

Finally

$$
\begin{aligned}
r_h(x_{j+1}) - r_h(x_j) &= \\
&= h\int_0^1 f(y)\varphi_{j+1}(y)dy - h\int_0^1 f(y)\varphi_j(y)dy = h\int_0^1 f(y)\left(\varphi_{j+1}(y) - \varphi_j(y)\right)dy \\
&= h\int_0^1 \Big(f(x_{j+1/2}) + f'(x_{j+1/2})(y - x_{j+1/2}) + f''(x_{j+1/2})\ \frac{(y - x_{j+1/2})^2}{2} \\
&\quad + f'''(x_{j+1/2})\ \frac{(y - x_{j+1/2})^3}{6} + O((y - x_{j+1/2})^4)\Big)\big(\varphi_{j+1}(y) - \varphi_j(y)\big)dy \\
&= hf(x_{j+1/2})\int_0^1 \varphi_{j+1}(y) - \varphi_j(y)dy \\
&\quad + hf'(x_{j+1/2})\int_0^1 (y - x_{j+1/2})(\varphi_{j+1}(y) - \varphi_j(y))dy \\
&\quad + hf''(x_{j+1/2})\int_0^1 \frac{(y - x_{j+1/2})^2}{2}(\varphi_{j+1}(y) - \varphi_j(y))dy \\
&\quad + hf'''(x_{j+1/2})\int_0^1 \frac{(y - x_{j+1/2})^3}{6}(\varphi_{j+1}(y) - \varphi_j(y))dy \\
&\quad + h\int_0^1 O((y - x_{j+1/2})^4)(\varphi_{j+1}(y) - \varphi_j(y))dy \\
&= h^3 f'(x_{j+1/2}) + \frac{h^5}{8}f'''(x_{j+1/2}) + h\int_0^1 O((y - x_{j+1/2})^4)(\varphi_{j+1}(y) - \varphi_j(y))dy \\
&= h^3 f'(x_{j+1/2}) + \frac{h^5}{8}f'''(x_{j+1/2}) + O(h^6)
\end{aligned}
$$

for $0 < j < N - 2$. □

Estimations of the term A_i

In order to estimate the term A_i, we will need the following result:

Lemma 6.3.4.

$$\sum_{j=0}^{i-1}\sin(\tilde{k}x_j)r_h(x_j)+\frac{1}{2}\sin(\tilde{k}x_i)r_h(x_i)=$$

$$=\frac{h^2\cos(\frac{h\tilde{k}}{2})}{2\sin(\frac{h\tilde{k}}{2})}\big[f(0)+\cos(\tilde{k}x_i)f(x_i)+O(h^2\|f''\|_\infty)\big]$$

$$+\frac{h^2}{2\sin(\frac{h\tilde{k}}{2})}\sum_{j=0}^{i-1}\cos(\tilde{k}x_{j+1/2})\big(hf'(x_{j+1/2})+O(h^3\|f'''\|_\infty)\big)$$

Proof. The essential ingredient of the proof is a summation by parts given by (6.8). Together with (6.9) we get

$$\sum_{j=0}^{i-1}\sin(\tilde{k}x_j)r_h(x_j)+\frac{1}{2}\sin(\tilde{k}x_i)r_h(x_i)=$$

$$=\left(\sum_{j=0}^{i-1}\sin(\tilde{k}x_j)\right)r_h(x_{i-1})-\sum_{j=0}^{i-2}\left(\sum_{l=0}^{j}\sin(\tilde{k}x_l)\right)\big(r_h(x_{j+1})-r_h(x_j)\big)+\frac{1}{2}\sin(\tilde{k}x_i)r_h(x_i)$$

$$=\frac{1}{2\sin(\frac{h\tilde{k}}{2})}\big(\cos(\frac{h\tilde{k}}{2})-\cos(\tilde{k}x_i-\frac{h\tilde{k}}{2})\big)r_h(x_{i-1})$$

$$-\frac{1}{2\sin(\frac{h\tilde{k}}{2})}\sum_{j=0}^{i-2}\big(\cos(\frac{h\tilde{k}}{2})-\cos(\tilde{k}x_j+\frac{h\tilde{k}}{2})\big)\big(r_h(x_{j+1})-r_h(x_j)\big)+\frac{1}{2}\sin(\tilde{k}x_i)r_h(x_i)$$

$$=\frac{1}{2\sin(\frac{h\tilde{k}}{2})}\Big[\big(\cos(\frac{h\tilde{k}}{2})-\cos(\tilde{k}x_i-\frac{h\tilde{k}}{2})\big)r_h(x_{i-1})-\cos(\frac{h\tilde{k}}{2})\sum_{j=0}^{i-2}\big(r_h(x_{j+1})-r_h(x_j)\big)$$

$$+\sum_{j=0}^{i-2}\cos(\tilde{k}x_j+\frac{h\tilde{k}}{2})\big(r_h(x_{j+1})-r_h(x_j)\big)+\sin(\frac{h\tilde{k}}{2})\sin(\tilde{k}x_i)r_h(x_i)\Big]$$

$$=\frac{1}{2\sin(\frac{h\tilde{k}}{2})}\Big[\big(\cos(\frac{h\tilde{k}}{2})-\cos(\tilde{k}x_i-\frac{h\tilde{k}}{2})\big)r_h(x_{i-1})-\cos(\frac{h\tilde{k}}{2})\big(r_h(x_{i-1})-r_h(0)\big)$$

$$+\sum_{j=1}^{i-1}\cos(\tilde{k}x_j+\frac{h\tilde{k}}{2})\big(r_h(x_{j+1})-r_h(x_j)\big)-\cos(\tilde{k}x_i-\frac{h\tilde{k}}{2})\big(r_h(x_i)-r_h(x_{i-1})\big)$$

$$+\cos(\frac{h\tilde{k}}{2})\big(r_h(x_1)-r_h(0)+\sin(\frac{h\tilde{k}}{2})\sin(\tilde{k}x_i)r_h(x_i)\Big]$$

$$=\frac{\cos(\frac{h\tilde{k}}{2})}{2\sin(\frac{h\tilde{k}}{2})}r_h(x_1)+\frac{\cos(\frac{h\tilde{k}}{2})}{2\sin(\frac{h\tilde{k}}{2})}\cos(\tilde{k}x_i)r_h(x_i)+\frac{1}{2\sin(\frac{h\tilde{k}}{2})}\sum_{j=1}^{i-1}\cos(\tilde{k}x_j+\frac{h\tilde{k}}{2})\big(r_h(x_{j+1})-r_h(x_j)\big)$$

Then, after applying the properties of the discrete right-hand side given in Lemma 6.3.3,

we get

$$
\begin{aligned}
&\sum_{j=0}^{i-1}\sin(\tilde{k}x_j)r_h(x_j)+\frac{1}{2}\sin(\tilde{k}x_i)r_h(x_i)= \\
&\quad=\frac{h^2\cos(\frac{h\tilde{k}}{2})}{2\sin(\frac{h\tilde{k}}{2})}\big(f(x_1)+O(h^2\|f''\|_\infty)\big)+\frac{h^2\cos(\frac{h\tilde{k}}{2})}{2\sin(\frac{h\tilde{k}}{2})}\cos(\tilde{k}x_i)\big(f(x_i)+O(h^2\|f''\|_\infty)\big) \\
&\qquad+\frac{1}{2\sin(\frac{h\tilde{k}}{2})}\sum_{j=1}^{i-1}\cos(\tilde{k}x_j+\frac{h\tilde{k}}{2})\big(h^3f'(x_{j+1/2})+O(h^5\|f'''\|_\infty)\big) \\
&\quad=\frac{h^2\cos(\frac{h\tilde{k}}{2})}{2\sin(\frac{h\tilde{k}}{2})}\big(f(x_1)+O(h^2\|f''\|_\infty)-hf'(x_{1/2})+O(h^3\|f'''\|_\infty)\big) \\
&\qquad+\frac{h^2\cos(\frac{h\tilde{k}}{2})}{2\sin(\frac{h\tilde{k}}{2})}\cos(\tilde{k}x_i)\big(f(x_i)+O(h^2\|f''\|_\infty)\big) \\
&\qquad+\frac{h^2}{2\sin(\frac{h\tilde{k}}{2})}\sum_{j=0}^{i-1}\cos(\tilde{k}x_j+\frac{h\tilde{k}}{2})\big(hf'(x_{j+1/2})+O(h^3\|f'''\|_\infty)\big) \\
&\quad=\frac{h^2\cos(\frac{h\tilde{k}}{2})}{2\sin(\frac{h\tilde{k}}{2})}\big(f(0)+hf'(0)+O(h^2\|f''\|_\infty)-h(f'(0)+\frac{h}{2}f''(0)+O(h^2\|f'''\|_\infty)\big) \\
&\qquad+\frac{h^2\cos(\frac{h\tilde{k}}{2})}{2\sin(\frac{h\tilde{k}}{2})}\cos(\tilde{k}x_i)\big(f(x_i)+O(h^2\|f''\|_\infty)\big) \\
&\qquad+\frac{h^2}{2\sin(\frac{h\tilde{k}}{2})}\sum_{j=0}^{i-1}\cos(\tilde{k}x_j+\frac{h\tilde{k}}{2})\big(hf'(x_{j+1/2})+O(h^3\|f'''\|_\infty)\big) \\
&\quad=\frac{h^2\cos(\frac{h\tilde{k}}{2})}{2\sin(\frac{h\tilde{k}}{2})}\big(f(0)+O(h^2\|f''\|_\infty)\big)+\frac{h^2\cos(\frac{h\tilde{k}}{2})}{2\sin(\frac{h\tilde{k}}{2})}\cos(\tilde{k}x_i)\big(f(x_i)+O(h^2\|f''\|_\infty)\big) \\
&\qquad+\frac{h^2}{2\sin(\frac{h\tilde{k}}{2})}\sum_{j=0}^{i-1}\cos(\tilde{k}x_j+\frac{h\tilde{k}}{2})\big(hf'(x_{j+1/2})+O(h^3\|f'''\|_\infty)\big)
\end{aligned}
$$

□

Using Lemma 6.3.4, we obtain with the definition of $\tilde{\alpha}_i$ in (6.10):

$$
\begin{aligned}
A_i &:= \int_0^{x_i} G(x_i,y)f(y)dy-\Big(h\sum_{j=0}^{i-1}G_h^\alpha(x_i,x_j)r_h(x_j)+\frac{1}{2}hG_h^\alpha(x_i,x_i)r_h(x_i)\Big) \\
&=\frac{1}{k}e^{\mathbf{i}kx_i}\int_0^{x_i}\sin(ky)f(y)dy \\
&\quad-\Big(h\frac{1}{h\sin(h\tilde{k})}\tilde{\alpha}_i\sum_{j=0}^{i-1}\sin(\tilde{k}x_j)r_h(x_j)+\frac{h}{2h\sin(h\tilde{k})}\tilde{\alpha}_i\sin(\tilde{k}x_i)r_h(x_i)\Big)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{k}e^{\mathbf{i}kx_i}\Big[-\frac{1}{k}\cos(ky)f(y)\Big|_0^{x_i}+\frac{1}{k}\int_0^{x_i}\cos(ky)f'(y)dy\Big]\\
&\quad-\frac{\tilde\alpha_i}{\sin(h\tilde k)}\Big[\sum_{j=0}^{i-1}\sin(\tilde k x_j)r_h(x_j)+\frac{1}{2}\sin(\tilde k x_i)r_h(x_i)\Big]\\
&= \frac{1}{k^2}e^{\mathbf{i}kx_i}\Big[f(0)-\cos(kx_i)f(x_i)+\int_0^{x_i}\cos(ky)f'(y)dy\Big]\\
&\quad-\frac{\tilde\alpha_i}{\sin(h\tilde k)}\Big[\frac{h^2\cos(\frac{h\tilde k}{2})}{2\sin(\frac{h\tilde k}{2})}\{f(0)+\cos(\tilde k x_i)f(x_i)+O(h^2\|f''\|_\infty)\}+\\
&\quad\frac{h^2}{2\sin(\frac{h\tilde k}{2})}\sum_{j=0}^{i-1}\cos(\tilde k x_{j+1/2})\big\{hf'(x_{j+1/2})+O(h^3\|f'''\|_\infty)\big\}\Big]\\
&= \frac{1}{k^2}e^{\mathbf{i}kx_i}f(0)-\frac{h^2}{4\sin^2(\frac{h\tilde k}{2})}\tilde\alpha_i\big(f(0)+O(h^2\|f''\|_\infty)\big) && \}\text{ⓐ}\\
&\quad-\Big[\frac{1}{k^2}e^{\mathbf{i}kx_i}\cos(kx_i)f(x_i)-\frac{h^2}{4\sin^2(\frac{h\tilde k}{2})}\tilde\alpha_i\cos(\tilde k x_i)\big(f(x_i)+O(h^2\|f''\|_\infty)\big)\Big] && \}\text{ⓑ}\\
&\quad+\frac{1}{k^2}e^{\mathbf{i}kx_i}\int_0^{x_i}\cos(ky)f'(y)dy-\frac{h^2}{2\sin(h\tilde k)\sin(\frac{h\tilde k}{2})}\tilde\alpha_i h\sum_{j=0}^{i-1}\cos(\tilde k x_j+\frac{h\tilde k}{2})\big(f'(x_{j+1/2})+O(h^2\|f'''\|_\infty)\big) && \}\text{ⓒ}
\end{aligned}
$$

Next, we aim to estimate each of the terms ⓐ, ⓑ, and ⓒ. Therefore we will need the following two lemmata:

Lemma 6.3.5. *As $kh\to 0$ (and thus $\varepsilon\to 0$) we have (uniformly in x_i, $x_j\in[0,1]$)*

$$e^{\mathbf{i}\tilde k x}-e^{\mathbf{i}kx}=O(k\varepsilon) \tag{6.11}$$

$$e^{\mathbf{i}\tilde k x}\cos(\tilde k y)-e^{\mathbf{i}kx}\cos(ky)=O(k\varepsilon) \tag{6.12}$$

$$\frac{(hk)^2}{4\sin^2(\frac{h\tilde k}{2})}e^{\mathbf{i}(\tilde k-k)x_i}=(1+O((kh)^2)+O(\varepsilon))(1+O(k\varepsilon)). \tag{6.13}$$

Proof. We start with the proof of (6.11):

$$e^{\mathbf{i}\tilde k x}-e^{\mathbf{i}kx} \;=\; e^{\mathbf{i}kx}\left(e^{\mathbf{i}(\tilde k-k)x}-1\right)=e^{\mathbf{i}kx}\left(e^{\mathbf{i}k\varepsilon x}-1\right)=O(k\varepsilon),\qquad \forall\varepsilon\to 0. \tag{6.14}$$

(If $k\varepsilon$ is small, then the statement is shown by Taylor expansion; if $k\varepsilon$ is *not* small, then $e^{\mathbf{i}k\varepsilon x}-1$ is $O(1)$ since k, ε, x are real).

Next, we observe that (6.14) implies

$$\cos(kx)-\cos(\tilde k x) \;=\; \frac{1}{2}\left(e^{\mathbf{i}kx}-e^{\mathbf{i}\tilde k x}+e^{-\mathbf{i}kx}-e^{\mathbf{i}\tilde k x}\right)=O(k\varepsilon). \tag{6.15}$$

The bounds (6.12) is shown similarly using (6.11), (6.15):

$$e^{\mathbf{i}\tilde k x}\cos(ky)-e^{\mathbf{i}\tilde k x}\cos(\tilde k y)=$$

$$
\begin{aligned}
&= e^{\mathbf{i}\tilde{k}x}\left[\cos(ky) - e^{\mathbf{i}(\tilde{k}-k)x}\left(\cos(\tilde{k}y) - \cos(ky) + \cos(ky)\right)\right] \\
&= e^{\mathbf{i}\tilde{k}x}\left[\cos(ky) - (1+O(k\varepsilon))(\cos(ky) + O(k\varepsilon))\right] \\
&= O(k\varepsilon).
\end{aligned}
$$

(Here, we ignore the term $O(k^2\varepsilon^2)$ that would formally arise since if $k\varepsilon = O(1)$, then the left-hand side is also $O(1)$).

Next, we show (6.13):

$$
\begin{aligned}
&\frac{(hk)^2}{4\sin^2(\frac{h\tilde{k}}{2})} e^{\mathbf{i}(\tilde{k}-k)x_i} = \\
&\qquad = \left(\frac{k}{\tilde{k}}\right)^2 \frac{(h\tilde{k})^2}{4\sin^2(\frac{h\tilde{k}}{2})} e^{\mathbf{i}(\tilde{k}-k)x_i} = \left(\frac{1}{1+\varepsilon}\right)^2 \frac{(h\tilde{k})^2}{4\sin^2(\frac{h\tilde{k}}{2})} e^{\mathbf{i}\varepsilon k x_i} \\
&\qquad = (1+O(\varepsilon))\left(1+O(h\tilde{k})^2\right)(1+O(k\varepsilon)) \\
&\qquad = \left(1+O((kh)^2)+O(\varepsilon)\right)(1+O(k\varepsilon)).
\end{aligned}
$$

□

Lemma 6.3.6. *As $kh \to 0$ (and thus $\varepsilon \to 0$) we have (uniformly in x, $y \in [0,1]$)*

$$
\frac{1}{k^2}\left| e^{\mathbf{i}kx_i} - \frac{(kh)^2}{4\sin^2(\frac{h\tilde{k}}{2})}\tilde{\alpha}_i \right| = O(h^2) + O(k^{-1}\varepsilon) \tag{6.16}
$$

$$
\frac{1}{k^2}\left| e^{\mathbf{i}kx_i}\cos(kx_j) - \frac{(kh)^2}{4\sin^2(\frac{h\tilde{k}}{2})}\tilde{\alpha}_i \cos(\tilde{k}x_j) \right| = O(h^2) + O(k^{-1}\varepsilon) \tag{6.17}
$$

Proof. In view of (6.10) and (6.13) we get that (as $kh \to 0$)

$$
\begin{aligned}
&\Big|\frac{1}{k^2}e^{\mathbf{i}kx_i} - \frac{(kh)^2}{4\sin^2(\frac{h\tilde{k}}{2})}\tilde{\alpha}_i\Big| = \\
&\quad = \frac{1}{k^2}\Big|1 - \frac{(hk)^2}{4\sin^2(\frac{h\tilde{k}}{2})} e^{\mathbf{i}(\tilde{k}-k)x_i}(1+O(kh)^2)\Big| \\
&\quad = \frac{1}{k^2}\Big|1 - \frac{(hk)^2}{4\sin^2(\frac{h\tilde{k}}{2})} e^{\mathbf{i}(\tilde{k}-k)x_i}\Big| + O(h^2) \\
&\quad = k^{-2}\Big\{O((kh)^2) + O(\varepsilon) + (1+O((kh)^2)+O(\varepsilon))O(k\varepsilon) + k^2 O(\varepsilon^2)\Big\} + O(h^2) \\
&\quad = k^{-2}\left\{O((kh)^2) + O(k\varepsilon)\right\} + O(h^2) = O(h^2) + O(k^{-1}\varepsilon)
\end{aligned}
$$

We finally show (6.17):

$$
\frac{1}{k^2}\left| e^{\mathbf{i}kx_i}\cos(kx_j) - \frac{(kh)^2}{4\sin^2(\frac{h\tilde{k}}{2})}\tilde{\alpha}_i\cos(\tilde{k}x_j) \right| =
$$

$$
\begin{aligned}
&= \frac{1}{k^2}\left| e^{\mathbf{i}kx_i}\cos(kx_j) - \frac{(kh)^2}{4\sin^2(\frac{h\tilde{k}}{2})}(e^{\mathbf{i}\tilde{k}x_j} + O(kh)^2)\cos(\tilde{k}x_j)\right| \\
&= \frac{1}{k^2}\left| e^{\mathbf{i}kx_i}\cos(kx_j) - \frac{(kh)^2}{4\sin^2(\frac{h\tilde{k}}{2})}e^{\mathbf{i}\tilde{k}x_j}\cos(\tilde{k}x_j)\right| + O(h^2) \\
&= O(h^2) + O(k^{-1}\varepsilon).
\end{aligned}
$$

□

With these estimates, we can analyze ⓐ, ⓑ, ⓒ. From (6.16) we get

$$
\begin{aligned}
ⓐ \;&:=\; \frac{1}{k^2}e^{\mathbf{i}kx_i}f(0) - \frac{h^2}{4\sin(\frac{h\tilde{k}}{2})^2}\tilde{\alpha}_i\big(f(0) + O(h^2\|f''\|_\infty)\big) \\
&=\; \frac{1}{k^2}\left(e^{\mathbf{i}kx_i} - \frac{(kh)^2}{4\sin(\frac{h\tilde{k}}{2})^2}\tilde{\alpha}_i\right)f(0) + \frac{1}{k^2}\frac{(kh)^2}{4\sin^2(h\tilde{k}/2)}O(h^2\|f''\|_\infty) \\
&\lesssim\; (h^2 + k^{-1}\varepsilon)\|f\|_\infty + h^2k^{-2}\|f''\|_\infty.
\end{aligned}
$$

Similarly we get with (6.17):

$$
\begin{aligned}
ⓑ \;&:=\; -\frac{1}{k^2}e^{\mathbf{i}kx_i}\cos(kx_i)f(x_i) + \frac{h^2}{4\sin\left(\frac{h\tilde{k}}{2}\right)^2}\tilde{\alpha}_i\cos(\tilde{k}x_i)\Big[f(x_i) + O(h^2\|f''\|_\infty)\Big] \\
&=\; \left[-\frac{1}{k^2}e^{\mathbf{i}kx_i}\cos(kx_i) + \frac{h^2}{4\sin\left(\frac{h\tilde{k}}{2}\right)^2}\tilde{\alpha}_i\cos(\tilde{k}x_i)\right]f(x_i) + O(k^{-2}h^2\|f''\|_\infty) \\
&\lesssim\; (h^2 + k^{-1}\varepsilon)\|f\|_\infty + k^{-2}h^2\|f''\|_\infty.
\end{aligned}
$$

For the third term, ⓒ, we start with the observation

$$
\frac{1}{2\sin(h\tilde{k})\sin(\frac{h\tilde{k}}{2})} = \frac{1}{4\sin^2(\frac{h\tilde{k}}{2})}\frac{1}{\cos(\frac{h\tilde{k}}{2})} = \frac{1}{4\sin^2(\frac{h\tilde{k}}{2})}(1 + O(kh)^2), \qquad kh \to 0. \quad (6.18)
$$

and use the midpoint rule to discretize the integral:

$$
\frac{1}{k^2}e^{\mathbf{i}kx_i}\int_0^{x_i}\cos(ky)f'(y)\,dy = \frac{1}{k^2}e^{\mathbf{i}kx_i}h\sum_{j=0}^{i-1}\cos(kx_{j+1/2})f'(x_{j+1/2}) + k^{-2}h^2O(k^2\|f'\|_\infty + k\|f''\|_\infty + \|f'''\|_\infty) \quad (6.19)
$$

Using (6.18) and (6.19), (6.17) we get thus:

$$
\begin{aligned}
ⓒ \;&:=\; \frac{1}{k^2}e^{\mathbf{i}kx_i}\int_0^{x_i}\cos(ky)f'(y)dy - \Big(\frac{1}{2\sin(h\tilde{k})\sin(\frac{h\tilde{k}}{2})}\tilde{\alpha}_i\sum_{j=0}^{i-1}\cos(\tilde{k}x_{j+1/2})\big(h^3f'(x_{j+1/2}) + O(\frac{h^5}{8}\|f'''\|_\infty)\Big) \\
&=\; \frac{1}{k^2}e^{\mathbf{i}kx_i}h\sum_{j=0}^{i-1}\cos(kx_{j+1/2})f'(x_{j+1/2}) \quad + h^2O(\|f'\|_\infty + k^{-1}\|f''\|_\infty + k^{-2}\|f'''\|_\infty)
\end{aligned}
$$

$$
\begin{aligned}
& -\frac{h^2}{2\sin(h\tilde{k})\sin(\frac{h\tilde{k}}{2})}\tilde{\alpha}_i h\sum_{j=0}^{i-1}\cos(\tilde{k}x_{j+1/2})\big(f'(x_{j+1/2})+O(\frac{h^2}{8}\|f'''\|_\infty)\big)\big) \\
= \;& \frac{1}{k^2}e^{\mathbf{i}kx_i}h\sum_{j=0}^{i-1}\cos(kx_{j+1/2})f'(x_{j+1/2}) - \frac{h^2}{2\sin(h\tilde{k})\sin(\frac{h\tilde{k}}{2})}\tilde{\alpha}_i h\sum_{j=0}^{i-1}\cos(\tilde{k}x_{j+1/2})f'(x_{j+1/2}) \\
& + h^2O(\|f'\|_\infty + k^{-1}\|f''\|_\infty + k^{-2}\|f'''\|_\infty) \\
= \;& h\sum_{j=0}^{i-1}\Big[\frac{1}{k^2}e^{\mathbf{i}kx_i}\cos(kx_{j+1/2}) - \frac{h^2}{4\sin^2(\frac{h\tilde{k}}{2})}(1+O(kh)^2)\tilde{\alpha}_i\cos(\tilde{k}x_{j+1/2})\Big]f'(x_{j+1/2}) \\
& + h^2O(\|f'\|_\infty + k^{-1}\|f''\|_\infty + k^{-2}\|f'''\|_\infty) \\
\lesssim \;& (h^2+k^{-1}\varepsilon)\|f'\|_\infty + h^2k^{-1}\|f''\|_\infty + h^2k^{-2}\|f'''\|_\infty
\end{aligned}
$$

In summary, this leads us to

$$
A_i = ⓐ + ⓑ + ⓒ \lesssim \big(h^2+k^{-1}\varepsilon\big)\big(\|f\|_\infty + \|f'\|_\infty\big) + k^{-1}h^2\|f''\|_\infty + k^{-2}h^2\|f'''\|_\infty\big) \quad (6.20)
$$

Estimation of the term B_i

For the second term B_i we will need

Lemma 6.3.7. *For $kh \to 0$ (and thus $\tilde{k}h \to 0$) we have*

$$
\begin{aligned}
& \sum_{j=i+1}^{N} e^{\mathbf{i}\tilde{k}x_j}r_h(x_j) + \frac{1}{2}e^{\mathbf{i}\tilde{k}x_i}r_h(x_i) = \\
& \qquad = -\frac{h^2(1+e^{\mathbf{i}\tilde{k}h})}{2(1-e^{\mathbf{i}\tilde{k}h})}e^{\mathbf{i}\tilde{k}}\big(f(1)+O(h^2\|f''\|_\infty)\big) \\
& \qquad\quad + \frac{h^2(1+e^{\mathbf{i}\tilde{k}h})}{2(1-e^{\mathbf{i}\tilde{k}h})}e^{\mathbf{i}\tilde{k}x_i}\big(f(x_i)+O(h^2\|f''\|_\infty)\big) \\
& \qquad\quad + \frac{h^2e^{\frac{i\tilde{k}h}{2}}}{(1-e^{\mathbf{i}\tilde{k}h})}h\sum_{j=i}^{N-1}e^{\mathbf{i}\tilde{k}x_{j+1/2}}\big(f'(x_{j+1/2})+O(h^2\|f'''\|_\infty)\big) \\
& \sum_{j=i+1}^{N}\sin(\tilde{k}x_j)r_h(x_j) + \frac{1}{2}\sin(\tilde{k}x_i)r_h(x_i) = \\
& \qquad = -\frac{h^2\cos(\frac{h\tilde{k}}{2})}{2\sin(\frac{h\tilde{k}}{2})}\cos(\tilde{k}x_N)f(1) + O(h^3\|f'\|_\infty + \tilde{k}^{-1}h^3\|f''\|_\infty) \\
& \qquad\quad + \frac{h^2\cos(\frac{h\tilde{k}}{2})}{2\sin(\frac{h\tilde{k}}{2})}\cos(\tilde{k}x_i)\big(f(x_i)+O(h^2\|f''\|_\infty)\big) \\
& \qquad\quad + \frac{h^2}{2\sin(\frac{h\tilde{k}}{2})}h\sum_{j=i}^{N-1}\cos(\tilde{k}x_{j+1/2})\big(f'(x_{j+1/2})+O(h^2\|f'''\|_\infty)\big)
\end{aligned}
$$

Proof. These two identities are again shown via summation by parts:

$$\begin{aligned}
&\sum_{j=i+1}^{N} e^{\mathbf{i}\tilde{k}x_j} r_h(x_j) + \frac{1}{2} e^{\mathbf{i}\tilde{k}x_i} r_h(x_i) = \\
&= \left(\sum_{j=i+1}^{N} e^{\mathbf{i}\tilde{k}x_j} \right) r_h(x_N) - \sum_{j=i+1}^{N-1} \left(\sum_{l=i+1}^{j} e^{\mathbf{i}\tilde{k}x_l} \right) \big(r_h(x_{j+1}) - r_h(x_j) \big) + \frac{1}{2} e^{\mathbf{i}\tilde{k}x_i} r_h(x_i) \\
&= \frac{e^{\mathbf{i}\tilde{k}x_{i+1}} - e^{\mathbf{i}\tilde{k}x_{N+1}}}{1 - e^{\mathbf{i}\tilde{k}h}} r_h(x_N) - \sum_{j=i+1}^{N-1} \frac{e^{\mathbf{i}\tilde{k}x_{i+1}} - e^{\mathbf{i}\tilde{k}x_{j+1}}}{1 - e^{\mathbf{i}\tilde{k}h}} \big(r_h(x_{j+1}) - r_h(x_j) \big) + \frac{1}{2} e^{\mathbf{i}\tilde{k}x_i} r_h(x_i) \\
&= \frac{1}{(1 - e^{\mathbf{i}\tilde{k}h})} \Big[(e^{\mathbf{i}\tilde{k}x_{i+1}} - e^{\mathbf{i}\tilde{k}x_{N+1}}) r_h(x_N) - e^{\mathbf{i}\tilde{k}x_{i+1}} \sum_{j=i+1}^{N-1} \big(r_h(x_{j+1}) - r_h(x_j) \big) \\
&\quad + \sum_{j=i+1}^{N-1} e^{\mathbf{i}\tilde{k}x_{j+1}} \big(r_h(x_{j+1}) - r_h(x_j) \big) + \frac{1}{2} (1 - e^{\mathbf{i}\tilde{k}h}) e^{\mathbf{i}\tilde{k}x_i} r_h(x_i) \Big] \\
&= \frac{1}{(1 - e^{\mathbf{i}\tilde{k}h})} \Big[(e^{\mathbf{i}\tilde{k}x_{i+1}} - e^{\mathbf{i}\tilde{k}x_{N+1}}) r_h(x_N) - e^{\mathbf{i}\tilde{k}x_{i+1}} \big(r_h(x_N) - r_h(x_{i+1}) \big) \\
&\quad + \sum_{j=i}^{N-1} e^{\mathbf{i}\tilde{k}x_{j+1}} \big(r_h(x_{j+1}) - r_h(x_j) \big) - e^{\mathbf{i}\tilde{k}x_{i+1}} \big(r_h(x_{i+1}) - r_h(x_i) \big) + \frac{1}{2} (1 - e^{\mathbf{i}\tilde{k}h}) e^{\mathbf{i}\tilde{k}x_i} r_h(x_i) \Big] \\
&= \frac{1}{(1 - e^{\mathbf{i}\tilde{k}h})} \Big[-e^{\mathbf{i}\tilde{k}x_{N+1}} r_h(x_N) + e^{\mathbf{i}\tilde{k}x_{i+1}} r_h(x_i) + \frac{1}{2} (1 - e^{\mathbf{i}\tilde{k}h}) e^{\mathbf{i}\tilde{k}x_i} r_h(x_i) \\
&\quad + \sum_{j=i}^{N-2} e^{\mathbf{i}\tilde{k}x_{j+1}} \big(r_h(x_{j+1}) - r_h(x_j) \big) + e^{\mathbf{i}\tilde{k}x_N} \big(r_h(x_N) - r_h(x_{N-1}) \big) \Big] \\
&= \frac{1}{(1 - e^{\mathbf{i}\tilde{k}h})} \Big[e^{\mathbf{i}\tilde{k}} \Big(\big(r_h(x_N) - r_h(x_{N-1}) \big) - e^{\mathbf{i}\tilde{k}h} r_h(x_N) \Big) + \frac{1}{2} (1 + e^{\mathbf{i}\tilde{k}h}) e^{\mathbf{i}\tilde{k}x_i} r_h(x_i) \\
&\quad + \sum_{j=i}^{N-2} e^{\mathbf{i}\tilde{k}x_{j+1}} \big(r_h(x_{j+1}) - r_h(x_j) \big) \Big]
\end{aligned}$$

Applying the properties of r_h shown in Lemma 6.3.3 produces

$$\begin{aligned}
&\sum_{j=i+1}^{N} e^{\mathbf{i}\tilde{k}x_j} r_h(x_j) + \frac{1}{2} e^{\mathbf{i}\tilde{k}x_i} r_h(x_i) = \\
&= \frac{1}{(1 - e^{\mathbf{i}\tilde{k}h})} \Big[e^{\mathbf{i}\tilde{k}} \Big(\big(- f(x_{N-1/2}) \frac{h^2}{2} + f'(x_{N-1/2}) \frac{7h^3}{12} + O(h^4 \|f''\|_\infty) \big) \\
&- e^{\mathbf{i}\tilde{k}h} \big(\frac{h^2}{2} f(x_N) - \frac{h^3}{6} f'(x_N) + O(h^4 \|f''\|_\infty) \big) \Big) + \frac{1}{2} (1 + e^{\mathbf{i}\tilde{k}h}) e^{\mathbf{i}\tilde{k}x_i} \big(f(x_i) h^2 + O(h^4 \|f''\|_\infty) \big) \\
&+ \sum_{j=i}^{N-1} e^{\mathbf{i}\tilde{k}x_{j+1}} \big(h^3 f'(x_{j+1/2}) + O(h^5 \|f'''\|_\infty) \big) - e^{\mathbf{i}\tilde{k}} \big(h^3 f'(x_{N-1/2}) + O(h^5 \|f'''\|_\infty) \big) \Big]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{e^{\mathbf{i}\tilde{k}}}{(1-e^{\mathbf{i}\tilde{k}h})}\Big[-\frac{h^2}{2}\Big(f(x_N)-\frac{h}{2}f'(x_N)\Big)+\frac{7h^3}{12}\Big(f'(x_N)\Big)+O(h^4\|f''\|_\infty)\\
&- e^{\mathbf{i}\tilde{k}h}\big(\frac{h^2}{2}f(x_N)-\frac{h^3}{6}f'(x_N)\big)-h^3\Big(f'(x_N)\Big)+O(h^4\|f'''\|_\infty)\Big]\\
&+\frac{1+e^{\mathbf{i}\tilde{k}h}}{2(1-e^{\mathbf{i}\tilde{k}h})}e^{\mathbf{i}\tilde{k}x_i}\big(f(x_i)h^2+O(h^4\|f''\|_\infty)\big)+\frac{1}{(1-e^{\mathbf{i}\tilde{k}h})}\sum_{j=i}^{N-1}e^{\mathbf{i}\tilde{k}x_{j+1}}\big(h^3f'(x_{j+1/2})+O(h^5\|f'''\|_\infty)\big)\\
&= \frac{e^{\mathbf{i}\tilde{k}}}{(1-e^{\mathbf{i}\tilde{k}h})}\Big[-\frac{h^2}{2}(1+e^{\mathbf{i}\tilde{k}h})f(x_N)-\frac{h^3}{6}(1-e^{\mathbf{i}\tilde{k}h})f'(x_N)+O(h^4\|f''\|_\infty)\Big]\\
&+\frac{1+e^{\mathbf{i}\tilde{k}h}}{2(1-e^{\mathbf{i}\tilde{k}h})}e^{\mathbf{i}\tilde{k}x_i}\big(f(x_i)h^2+O(h^4\|f''\|_\infty)\big)+\frac{1}{(1-e^{\mathbf{i}\tilde{k}h})}\sum_{j=i}^{N-1}e^{\mathbf{i}\tilde{k}x_{j+1}}\big(h^3f'(x_{j+1/2})+O(h^5\|f'''\|_\infty)\big)\\
&= -\frac{h^2(1+e^{\mathbf{i}\tilde{k}h})}{2(1-e^{\mathbf{i}\tilde{k}h})}e^{\mathbf{i}\tilde{k}}\big(f(1)+O(h^2\|f''\|_\infty)\big)-\frac{h^3}{6}f'(1)e^{\mathbf{i}\tilde{k}}+\frac{h^2(1+e^{\mathbf{i}\tilde{k}h})}{2(1-e^{\mathbf{i}\tilde{k}h})}e^{\mathbf{i}\tilde{k}x_i}\big(f(x_i)+O(h^2\|f''\|_\infty)\big)\\
&+\frac{e^{\frac{i\tilde{k}h}{2}}}{(1-e^{\mathbf{i}\tilde{k}h})}\sum_{j=i}^{N-1}e^{\mathbf{i}\tilde{k}x_{j+1/2}}\big(h^3f'(x_{j+1/2})+O(h^5\|f'''\|_\infty)\big)\\
&= -\frac{h^2(1+e^{\mathbf{i}\tilde{k}h})}{2(1-e^{\mathbf{i}\tilde{k}h})}e^{\mathbf{i}\tilde{k}}\big(f(1)+O(h^2\|f''\|_\infty)\big)+\frac{h^2(1+e^{\mathbf{i}\tilde{k}h})}{2(1-e^{\mathbf{i}\tilde{k}h})}e^{\mathbf{i}\tilde{k}x_i}\big(f(x_i)+O(h^2\|f''\|_\infty)\big)\\
&+\frac{h^2e^{\frac{i\tilde{k}h}{2}}}{(1-e^{\mathbf{i}\tilde{k}h})}h\sum_{j=i}^{N-1}e^{\mathbf{i}\tilde{k}x_{j+1/2}}\big(f'(x_{j+1/2})+O(h^2\|f''\|_\infty)\big)
\end{aligned}
$$

For the second identity of the lemma, we calculate with the summation by parts formula (6.8) and the trigonometric identity (6.9):

$$
\begin{aligned}
&\sum_{j=i+1}^{N}\sin(\tilde{k}x_j)r_h(x_j)+\frac{1}{2}\sin(\tilde{k}x_i)r_h(x_i)=\\
&=\left(\sum_{j=i+1}^{N}\sin(\tilde{k}x_j)\right)r_h(x_N)-\sum_{j=i+1}^{N-1}\left(\sum_{l=i+1}^{j}\sin(\tilde{k}x_l)\right)\big(r_h(x_{j+1})-r_h(x_j)\big)+\frac{1}{2}\sin(\tilde{k}x_i)r_h(x_i)\\
&=\frac{1}{2\sin(\frac{h\tilde{k}}{2})}\big(\cos(\tilde{k}x_{i+1/2})-\cos(\tilde{k}x_{N+1/2})\big)r_h(x_N)\\
&-\frac{1}{2\sin(\frac{h\tilde{k}}{2})}\sum_{j=i+1}^{N-1}\big(\cos(\tilde{k}x_{i+1/2})-\cos(\tilde{k}x_{j+1/2})\big)\big(r_h(x_{j+1})-r_h(x_j)\big)+\frac{1}{2}\sin(\tilde{k}x_i)r_h(x_i)\\
&=\frac{1}{2\sin(\frac{h\tilde{k}}{2})}\Big[\big(\cos(\tilde{k}x_{i+1/2})-\cos(\tilde{k}x_{N+1/2})\big)r_h(x_N)-\cos(\tilde{k}x_{i+1/2})\sum_{j=i+1}^{N-1}\big(r_h(x_{j+1})-r_h(x_j)\big)\\
&+\sum_{j=i+1}^{N-1}\cos(\tilde{k}x_{j+1/2})\big(r_h(x_{j+1})-r_h(x_j)\big)+\sin(\frac{h\tilde{k}}{2})\sin(\tilde{k}x_i)r_h(x_i)\Big]
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{2\sin(\frac{h\tilde{k}}{2})}\Big[\big(\cos(\tilde{k}x_{i+1/2}) - \cos(\tilde{k}x_{N+1/2})\big)r_h(x_N) - \cos(\tilde{k}x_{i+1/2})\big(r_h(x_N) - r_h(x_{i+1})\big)\\
&+ \sum_{j=i}^{N-2} \cos(\tilde{k}x_{j+1/2})\big(r_h(x_{j+1}) - r_h(x_j)\big) - \cos(\tilde{k}x_{i+1/2})\big(r_h(x_{i+1}) - r_h(x_i)\big)\\
&+ \cos(\tilde{k}x_{N-1/2})\big(r_h(x_N) - r_h(x_{N-1})\big) + \sin(\frac{h\tilde{k}}{2})\sin(\tilde{k}x_i)r_h(x_i)\Big]\\
&= \frac{1}{2\sin(\frac{h\tilde{k}}{2})}\Big[\big(\cos(\tilde{k}x_{N-1/2}) - \cos(\tilde{k}x_{N+1/2})\big)r_h(x_N) - \cos(\tilde{k}x_{N-1/2})r_h(x_{N-1})\big)\\
&+ \sum_{j=i}^{N-2} \cos(\tilde{k}x_{j+1/2})\big(r_h(x_{j+1}) - r_h(x_j)\big) + \Big(\cos(\tilde{k}x_{i+1/2}) + \sin(\frac{h\tilde{k}}{2})\sin(\tilde{k}x_i)\Big)r_h(x_i)\Big]\\
&= \sin(\tilde{k}x_N)r_h(x_N) - \frac{1}{2\sin(\frac{h\tilde{k}}{2})}\cos(\tilde{k}x_{N-1/2})r_h(x_{N-1}) + \frac{\cos(\frac{h\tilde{k}}{2})}{2\sin(\frac{h\tilde{k}}{2})}\cos(\tilde{k}x_i)r_h(x_i)\\
&+ \frac{1}{2\sin(\frac{h\tilde{k}}{2})}\sum_{j=i}^{N-2}\cos(\tilde{k}x_{j+1/2})\big(r_h(x_{j+1}) - r_h(x_j)\big)\\
&= \sin(\tilde{k}x_N)\big(\frac{h^2}{2}f(x_N) - \frac{h^3}{6}f'(x_N) + O(h^4\|f''\|_\infty)\big) - \frac{1}{2\sin(\frac{h\tilde{k}}{2})}\cos(\tilde{k}x_{N-1/2})\big(h^2 f(x_{N-1}) + O(h^4\|f''\|_\infty)\big)\\
&+ \frac{\cos(\frac{h\tilde{k}}{2})}{2\sin(\frac{h\tilde{k}}{2})}\cos(\tilde{k}x_i)\big(h^2 f(x_i) + O(h^4\|f''\|_\infty)\big) + \frac{1}{2\sin(\frac{h\tilde{k}}{2})}\sum_{j=i}^{N-1}\cos(\tilde{k}x_{j+1/2})\big(h^3 f'(x_{j+1/2}) + O(h^5\|f'''\|_\infty)\big)\\
&- \frac{1}{2\sin(\frac{h\tilde{k}}{2})}\cos(\tilde{k}x_{N-1/2})\big(h^3 f'(x_{N-1/2}) + O(h^5\|f'''\|_\infty)\big)\\
&= \sin(\tilde{k}x_N)\big(\frac{h^2}{2}f(x_N) - \frac{h^3}{6}f'(x_N) + O(h^4\|f''\|_\infty)\big)\\
&- \frac{1}{2\sin(\frac{h\tilde{k}}{2})}\cos(\tilde{k}x_{N-1/2})\big(h^2(f(x_N) - hf'(x_N) + O(h^2\|f''\|_\infty)) + O(h^4\|f''\|_\infty)\big)\\
&+ \frac{\cos(\frac{h\tilde{k}}{2})}{2\sin(\frac{h\tilde{k}}{2})}\cos(\tilde{k}x_i)\big(h^2 f(x_i) + O(h^4\|f''\|_\infty)\big) + \frac{1}{2\sin(\frac{h\tilde{k}}{2})}\sum_{j=i}^{N-1}\cos(\tilde{k}x_{j+1/2})\big(h^3 f'(x_{j+1/2}) + O(h^5\|f'''\|_\infty)\big)\\
&- \frac{1}{2\sin(\frac{h\tilde{k}}{2})}\cos(\tilde{k}x_{N-1/2})\big(h^3(f'(x_N) - \frac{h}{2}f''(x_N) + O(h^2\|f'''\|_\infty))) + O(h^5\|f'''\|_\infty)\big)\\
&= -\frac{h^2\cos(\frac{h\tilde{k}}{2})}{2\sin(\frac{h\tilde{k}}{2})}\cos(\tilde{k}x_N)f(x_N) - \frac{h^3}{6}f'(1)\sin(\tilde{k}) + \sin(\tilde{k})O(\tilde{k}^{-1}h^3\|f''\|_\infty)\\
&+ \frac{h^2\cos(\frac{h\tilde{k}}{2})}{2\sin(\frac{h\tilde{k}}{2})}\cos(\tilde{k}x_i)\big(f(x_i) + O(h^2\|f''\|_\infty)\big) + \frac{1}{2\sin(\frac{h\tilde{k}}{2})}\sum_{j=i}^{N-1}\cos(\tilde{k}x_{j+1/2})\big(h^3 f'(x_{j+1/2}) + O(h^5\|f'''\|_\infty)\big)
\end{aligned}
$$

$$= -\frac{h^2\cos(\frac{h\tilde{k}}{2})}{2\sin(\frac{h\tilde{k}}{2})}\cos(\tilde{k}x_N)f(x_N) + O(h^3\|f'\|_\infty + \tilde{k}^{-1}h^3\|f''\|_\infty) + \frac{h^2\cos(\frac{h\tilde{k}}{2})}{2\sin(\frac{h\tilde{k}}{2})}\cos(\tilde{k}x_i)\big(f(x_i) + O(h^2\|f''\|_\infty)\big)$$
$$+ \frac{h^2}{2\sin(\frac{h\tilde{k}}{2})}h\sum_{j=i}^{N-1}\cos(\tilde{k}x_{j+1/2})\big(f'(x_{j+1/2}) + O(h^2\|f'''\|_\infty)\big)$$

□

Thus, using the structure of $\tilde{\alpha}_j$ given in (6.10) and Lemma 6.3.7 we receive

$$\begin{aligned}
B_i &:= \int_{x_i}^1 G(x_i,y)f(y)dy - \Big(h\sum_{j=i+1}^N G_h^\alpha(x_i,x_j)r_h(x_j) + \frac{h}{2}G_h^\alpha(x_i,x_i)r_h(x_i)\Big)\\
&= \frac{1}{k}\sin(kx_i)\int_{x_i}^1 e^{\mathbf{i}ky}f(y)dy - \Big(\frac{h}{h\sin(h\tilde{k})}\sin(\tilde{k}x_i)\sum_{j=i+1}^N\tilde{\alpha}_j r_h(x_j) + \frac{h}{2h\sin(h\tilde{k})}\sin(\tilde{k}x_i)\tilde{\alpha}_i r_h(x_i)\Big)\\
&= \frac{1}{k}\sin(kx_i)\Big[\frac{1}{\mathbf{i}k}e^{\mathbf{i}ky}f(y)\Big|_{x_i}^1 - \frac{1}{\mathbf{i}k}\int_{x_i}^1 e^{\mathbf{i}ky}f'(y)dy\Big]\\
&\quad -\frac{1}{\sin(h\tilde{k})}\sin(\tilde{k}x_i)\Big[\sum_{j=i+1}^N\big(e^{\mathbf{i}\tilde{k}x_j} + \sin(\tilde{k}x_j)(A(k,\tilde{k})-\mathbf{i})\big)r_h(x_j)\\
&\qquad\qquad +\frac{1}{2}\big(e^{\mathbf{i}\tilde{k}x_i} + \sin(\tilde{k}x_i)(A(k,\tilde{k})-\mathbf{i})\big)r_h(x_i)\Big]\\
&= \frac{1}{\mathbf{i}k^2}\sin(kx_i)\Big[e^{\mathbf{i}k}f(1) - e^{\mathbf{i}kx_i}f(x_i) - \int_{x_i}^1 e^{\mathbf{i}ky}f'(y)dy\Big]\\
&\quad -\frac{1}{\sin(h\tilde{k})}\sin(\tilde{k}x_i)\Big[\sum_{j=i+1}^N e^{\mathbf{i}\tilde{k}x_j}r_h(x_j) + \frac{1}{2}e^{\mathbf{i}\tilde{k}x_i}r_h(x_i) +\\
&\qquad\qquad \Big(\sum_{j=i+1}^N\sin(\tilde{k}x_j)r_h(x_j) + \frac{1}{2}\sin(\tilde{k}x_i)r_h(x_i)\Big)(A(k,\tilde{k})-\mathbf{i})\Big]\\
&= \frac{1}{\mathbf{i}k^2}\sin(kx_i)\Big[e^{\mathbf{i}k}f(1) - e^{\mathbf{i}kx_i}f(x_i) - \int_{x_i}^1 e^{\mathbf{i}ky}f'(y)dy\Big]\\
&\quad -\frac{1}{\sin(h\tilde{k})}\sin(\tilde{k}x_i)\Big(-\frac{h^2(1+e^{\mathbf{i}\tilde{k}h})}{2(1-e^{\mathbf{i}\tilde{k}h})}\big(e^{\mathbf{i}\tilde{k}}f(1) - e^{\mathbf{i}\tilde{k}x_i}f(x_i) + O(h^2\|f''\|_\infty)\big)\\
&\quad +\frac{h^2e^{\frac{\mathbf{i}\tilde{k}h}{2}}}{(1-e^{\mathbf{i}\tilde{k}h})}h\sum_{j=i}^{N-1}e^{\mathbf{i}\tilde{k}x_{j+1/2}}\big(f'(x_{j+1/2}) + O(h^2\|f''\|_\infty)\big)\Big)\\
&\quad -\frac{1}{\sin(h\tilde{k})}\sin(\tilde{k}x_i)(A(k,\tilde{k})-\mathbf{i})\Big\{-\frac{h^2\cos(\frac{h\tilde{k}}{2})}{2\sin(\frac{h\tilde{k}}{2})}\big(\cos(\tilde{k})f(1) - \cos(\tilde{k}x_i)f(x_i)\big)\\
&\qquad\qquad +O(h^3\|f'\|_\infty + \tilde{k}^{-1}h^3\|f''\|_\infty)
\end{aligned}$$

$$+\frac{h^2}{2\sin(\frac{h\tilde{k}}{2})}h\sum_{j=i}^{N-1}\cos(\tilde{k}x_{j+1/2})\big(f'(x_{j+1/2})+O(h^2\|f'''\|_\infty)\big)\Big\}$$

$$= \Big\{\frac{1}{\mathbf{i}k^2}\sin(kx_i)e^{\mathbf{i}k}f(1)+\frac{h^2(1+e^{\mathbf{i}\tilde{k}h})}{2\sin(h\tilde{k})(1-e^{\mathbf{i}\tilde{k}h})}\sin(\tilde{k}x_i)e^{\mathbf{i}\tilde{k}}f(1)\Big\} \quad ⓓ$$

$$+\Big\{\frac{h^2}{4\sin^2(\frac{\tilde{k}h}{2})}\sin(\tilde{k}x_i)\cos(\tilde{k})f(1)(A(k,\tilde{k})-\mathbf{i})$$

$$-\frac{1}{\mathbf{i}k^2}\sin(kx_i)e^{\mathbf{i}kx_i}f(x_i)-\frac{h^2(1+e^{\mathbf{i}\tilde{k}h})}{2\sin(h\tilde{k})(1-e^{\mathbf{i}\tilde{k}h})}\sin(\tilde{k}x_i)e^{\mathbf{i}\tilde{k}x_i}f(x_i)\Big\} \quad ⓔ$$

$$\Big\{-\frac{h^2}{4\sin^2(\frac{\tilde{k}h}{2})}\sin(\tilde{k}x_i)\cos(\tilde{k}x_i)f(x_i)(A(k,\tilde{k})-\mathbf{i})$$

$$-\frac{1}{\mathbf{i}k^2}\sin(kx_i)\int_{x_i}^1 e^{\mathbf{i}ky}f'(y)dy-\frac{h^2e^{\frac{\mathbf{i}\tilde{k}h}{2}}}{\sin(h\tilde{k})(1-e^{\mathbf{i}\tilde{k}h})}\sin(\tilde{k}x_i)$$

$$h\sum_{j=i}^{N-1}e^{\mathbf{i}\tilde{k}x_{j+1/2}}\big(f'(x_{j+1/2})+O(h^2\|f'''\|_\infty)\big)$$

$$+\frac{h^2}{2\sin(\frac{h\tilde{k}}{2})}h\sum_{j=i}^{N-1}\cos(\tilde{k}x_{j+1/2})\big(f'(x_{j+1/2})+O(h^2\|f'''\|_\infty)\big)(A(k,\tilde{k})-\mathbf{i})\Big\} \quad ⓕ$$

In order to simplify these three terms, we need a lemma:

Lemma 6.3.8. *For $kh\to 0$ (and thus $\tilde{k}h\to 0$)*

$$A(k,\tilde{k})-\mathbf{i} = O((kh)^2) \tag{6.21}$$

$$\frac{1}{\mathbf{i}k^2}\sin(kx_i)e^{\mathbf{i}k}+\frac{h^2(1+e^{\mathbf{i}\tilde{k}h})}{2\sin(h\tilde{k})(1-e^{\mathbf{i}\tilde{k}h})}\sin(\tilde{k}x_i)e^{\mathbf{i}\tilde{k}} = O(k\varepsilon)+O(h^2) \tag{6.22}$$

$$\frac{1}{\mathbf{i}k^2}\sin(kx_i)e^{\mathbf{i}ky}+\frac{h^2e^{\mathbf{i}\tilde{k}h/2}}{\sin(h\tilde{k})(1-e^{\mathbf{i}h\tilde{k}})}\sin(\tilde{k}x_i)e^{\mathbf{i}\tilde{k}y} = O(h^2)+O(k^{-1}\varepsilon). \tag{6.23}$$

Proof. (6.21) follows from the definition of $A(k,\tilde{k})$ and a straight forward Taylor expansion.

For (6.22), we first note that the estimate is trivial if $k\varepsilon=O(1)$ (and kh is small). We may therefore assume that additionally $k\varepsilon$ is small. With $\tilde{k}=k(1+\varepsilon)$ we then have

$$\frac{1}{\mathbf{i}k^2}\sin(kx_i)e^{\mathbf{i}k}+\frac{h^2(1+e^{\mathbf{i}\tilde{k}h})}{2\sin(h\tilde{k})(1-e^{\mathbf{i}\tilde{k}h})}\sin(\tilde{k}x_i)e^{\mathbf{i}\tilde{k}} =$$

$$\frac{1}{\mathbf{i}k^2}e^{\mathbf{i}k}\left(\sin(kx_i)+\mathbf{i}\frac{(kh)^2(1+e^{\mathbf{i}kh(1+\varepsilon)})}{2\sin(kh(1+\varepsilon))(1-e^{\mathbf{i}kh(1+\varepsilon)})}\sin(k(1+\varepsilon)x_i)e^{\mathbf{i}k\varepsilon)}\right)$$

We set $\delta = kh$ and perform a Taylor expansions (assuming δ and ε to be small) to get

$$\begin{aligned}
\frac{(kh)^2(1+e^{\mathbf{i}kh(1+\varepsilon)})}{2\sin(kh(1+\varepsilon))(1-e^{\mathbf{i}kh(1+\varepsilon)})} &= \frac{\delta^2(2+\mathbf{i}\delta(1+\varepsilon)+O(\delta^2))}{2(\delta(1+\varepsilon)+O(\delta^3))(1-(1+\mathbf{i}\delta(1+\varepsilon)-\frac{1}{2}(\delta(1+\varepsilon))^2+O(\delta^3))} \\
&= \frac{1+\frac{\mathbf{i}\delta(1+\varepsilon)}{2}+O(\delta^2)}{(1+\varepsilon+O(\delta^2))(-\mathbf{i}(1+\varepsilon)+\frac{1}{2}\delta(1+\varepsilon)^2+O(\delta^2))} \\
&= \frac{1}{-\mathbf{i}(1+\varepsilon)}\frac{1+\frac{\mathbf{i}\delta(1+\varepsilon)}{2}+O(\delta^2)}{(1+\varepsilon+O(\delta^2))(1+\mathbf{i}\frac{1}{2}\delta(1+\varepsilon)+O(\delta^2))} \\
&= \frac{1}{-\mathbf{i}(1+\varepsilon)^2}(1+O(\delta^2)) = \frac{1}{-\mathbf{i}}(1+O(\delta^2)+O(\varepsilon))
\end{aligned}$$

Therefore, we get

$$\begin{aligned}
&\frac{1}{\mathbf{i}k^2}e^{\mathbf{i}k}\left(\sin(kx_i)+\mathbf{i}\frac{(kh)^2(1+e^{\mathbf{i}kh(1+\varepsilon)})}{2\sin(kh(1+\varepsilon))(1-e^{\mathbf{i}kh(1+\varepsilon)})}\sin(k(1+\varepsilon)x_i)e^{\mathbf{i}k\varepsilon)}\right) \\
&= \frac{1}{\mathbf{i}k^2}e^{\mathbf{i}k}\left(\sin(kx_i)-(1+O(\delta^2)+O(\varepsilon))\sin(kx_i(1+\varepsilon))e^{\mathbf{i}k\varepsilon}\right) \\
&= \frac{1}{\mathbf{i}k^2}e^{\mathbf{i}k}\left(\sin(kx_i)-(1+O(\delta^2)+O(\varepsilon))(\sin(kx_i)+O(k\varepsilon))(1+O(k\varepsilon))\right) \\
&k^{-2}\left(O(\delta^2)+O(k\varepsilon)\right)
\end{aligned}$$

Recalling that $\delta = kh$ finishes the proof of (6.22).

We now show (6.23). Taylor expansion gives (for small δ and ε)

$$\frac{\delta^2 e^{\mathbf{i}\delta(1+\varepsilon)/2}}{\sin(\delta(1+\varepsilon))(1-e^{\mathbf{i}\delta(1+\varepsilon)})} = -1+O(\delta^2)+O(\varepsilon)$$

Hence, we get with the notation $\delta = kh$

$$\begin{aligned}
&\frac{1}{\mathbf{i}k^2}\sin(kx_i)e^{\mathbf{i}ky}+\frac{h^2 e^{\mathbf{i}\tilde{k}h/2}}{\sin(h\tilde{k})(1-e^{\mathbf{i}h\tilde{k}})}\sin(\tilde{k}x_i)e^{\mathbf{i}\tilde{k}y} \\
&= \frac{1}{\mathbf{i}k^2}e^{\mathbf{i}ky}\left(\sin(kx_i)+(-1+O(\delta^2)+O(\varepsilon))\sin(k(1+\varepsilon)x_i)e^{\mathbf{i}k\varepsilon y}\right) \\
&= \frac{1}{\mathbf{i}k^2}e^{\mathbf{i}ky}\left(\sin(kx_i)+(-1+O(\delta^2)+O(\varepsilon))(\sin(kx_i)+O(k\varepsilon))(1+O(k\varepsilon))\right) \\
&= k^{-2}\left(O(\delta^2)+O(k\varepsilon)\right),
\end{aligned}$$

which concludes the proof of (6.23). □

With Lemma 6.3.8 in hand, we can bounds the terms ⓓ, ⓔ, and ⓕ.

From (6.22), we get

$$|ⓓ| \leq C|f(1)|\left(h^2+k^{-1}\varepsilon\right).$$

Combining (6.21) and (6.22) yields

$$|ⓔ| \quad \leq \quad C|f(x_i)|\left(h^2+k^{-1}\varepsilon\right)+Ch^2|f(1)|.$$

The term ⓕ consists of three terms

$$ⓕ = ⓕ_1 + ⓕ_2 + ⓕ_3.$$

The terms $ⓕ_1$ and $ⓕ_2$ can be estimated using (6.21) by

$$|ⓕ_1| + |ⓕ_3| \leq Ch^2|f(x_i)| + Ch^2\left(\|f'\|_\infty + h^2\|f'''\|_\infty\right).$$

The term $ⓕ_2$ requires more care. Discretizing the integral in the term $ⓕ_2$ with the midpoint rule we get

$$-\frac{1}{\mathbf{i}k^2}\sin(kx_i)\int_{x_i}^1 e^{\mathbf{i}ky}f'(y)\,dy = -\frac{1}{\mathbf{i}k^2}\sin(kx_i)h\sum_{j=i}^{N-1}e^{\mathbf{i}kx_{j+1/2}}f'(x_{j+1/2})+h^2O(k^2\|f'\|_\infty+k\|f''\|_\infty+\|f'''\|_\infty). \tag{6.24}$$

With the aid of (6.23) and (6.24) we get for ⓒ:

$$\begin{aligned}
ⓕ \quad := \quad & -\frac{1}{\mathbf{i}k^2}\sin(kx_i)\int_{x_i}^1 e^{\mathbf{i}ky}f'(y)dy - \frac{h^2e^{\frac{i\tilde{k}h}{2}}}{\sin(h\tilde{k})(1-e^{\mathbf{i}\tilde{k}h})}\sin(\tilde{k}x_i)h\sum_{j=i}^{N-1}e^{\mathbf{i}\tilde{k}x_{j+1/2}}\left(f'(x_{j+1/2})+O(h^2\|f'''\|_\infty)\right)\\
= \quad & -\frac{1}{\mathbf{i}k^2}\sin(kx_i)h\sum_{j=i}^{N}e^{\mathbf{i}kx_{j+1/2}}f'(x_{j+1/2}) \quad +k^{-2}h^2O(k^2\|f'\|_\infty+k\|f''\|_\infty+\|f'''\|_\infty)\\
& -\frac{h^2e^{\frac{i\tilde{k}h}{2}}}{\sin(h\tilde{k})(1-e^{\mathbf{i}\tilde{k}h})}\sin(\tilde{k}x_i)h\sum_{j=i}^{N-1}e^{\mathbf{i}\tilde{k}x_{j+1/2}}\left(f'(x_{j+1/2})+O(h^2\|f'''\|_\infty)\right)\\
= \quad & -h\sum_{j=i}^{N-1}\Big[\frac{1}{\mathbf{i}k^2}\sin(kx_i)e^{\mathbf{i}kx_{j+1/2}}+\frac{h^2e^{\frac{\mathbf{i}\tilde{k}h}{2}}}{\sin(h\tilde{k})(1-e^{\mathbf{i}\tilde{k}h})}\sin(\tilde{k}x_i)e^{\mathbf{i}\tilde{k}x_{j+1/2}}\Big]f'(x_{j+1/2})\\
& +k^{-2}h^2O(k^2\|f'\|_\infty+k\|f''\|_\infty+\|f'''\|_\infty)\\
= \quad & (O(h^2)+O(k^{-1}\varepsilon))\|f'\|_\infty+k^{-2}h^2O(k^2\|f'\|_\infty+k\|f''\|_\infty+\|f'''\|_\infty).
\end{aligned}$$

This leads us to

$$B_i = ⓓ+ⓔ+ⓕ \lesssim \left(h^2+k^{-1}\varepsilon\right)\left(\|f\|_\infty+\|f'\|_\infty\right)+h^2(k^{-1}\|f''\|_\infty+k^{-2}\|f'''\|_\infty). \tag{6.25}$$

The final estimate

Combining (6.20) and (6.25) we arrive at

$$\|e_{poll}\|_{L^2} \lesssim \sqrt{h\sum_i|A_i+B_i|^2} \lesssim \left(h^2+k^{-1}\varepsilon\right)\left(\|f\|_\infty+\|f'\|_\infty\right)+h^2k^{-1}\|f''\|_\infty+h^2k^{-2}\|f'''\|_\infty,$$

leading to the required estimate used in the proof of Theorem 6.3.1.

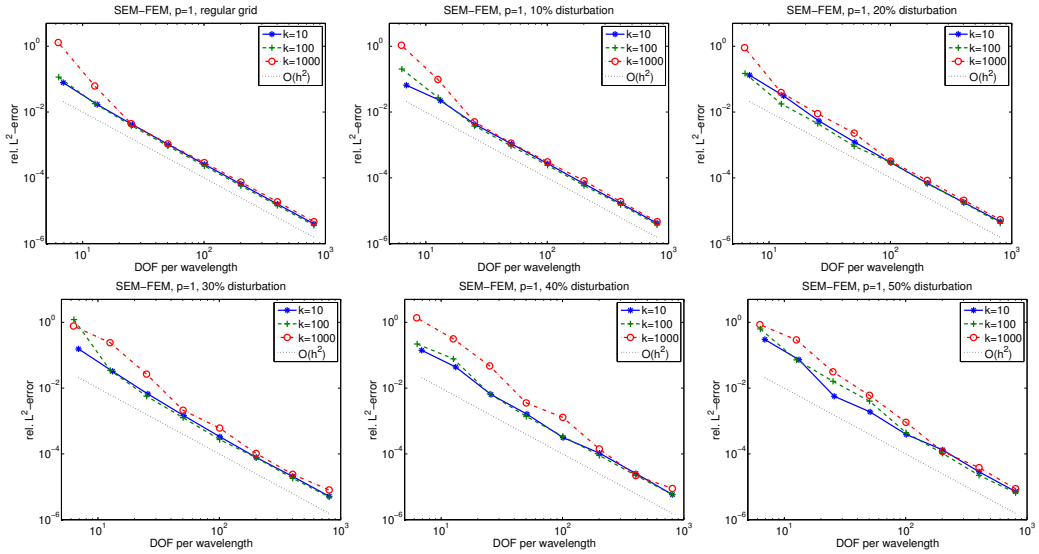


Figure 6.1.: (cf. Example 6.4.1) h-method for the 1D blended SEM-FEM on uniform and randomly perturbed meshes (mesh points of regular mesh perturbed by $n\%$ for $n \in \{0, 10, 20, 30, 40, 50\}$).

6.4. Numerical examples

In addition we implemented the blended spectral-low order finite element scheme for a 1D-Helmholtz problem whose results were published by the author and J. M. Melenk in [39]. In contrast to numerical computations described in Section 5.4 this leads to a distinct improvement in the L^2-error, see Fig. 6.1. In particular we run the scheme on regular mesh as well as on irregular meshes with random perturbations from 10-50%. Thus from the picture in Fig. 6.1 we can infer that the statement in Section 6.3 can't be extended from regular meshes to arbitrary meshes.

The observation of Remark 6.3.2 is illustrated in the following numerical example.

Example 6.4.1. We consider (3.32) with $f = 1$ and $g = 0$. The exact solution u is given by

$$u(x) = \frac{1}{k^2}(e^{ikx} - ie^{ik}\sin(kx) - 1). \tag{6.26}$$

The top leftmost plot in Fig. 6.1 shows the performance of the optimally blended scheme. The relative error in L^2 is plotted versus the number of degrees of freedom per wavelength N_λ. We note the good agreement with the *a priori* estimate (6.7). The remaining plots in Fig. 6.1 show the performance of the optimally blended scheme on non-uniform meshes. The mesh points of a regular mesh were randomly perturbed by n percent, where $n \in \{10, 20, 30, 40, 50\}$. Although the favorable properties of the optimally blended scheme are not proved under these circumstances, the numerical results indicate a certain robustness of the method under mesh perturbation.

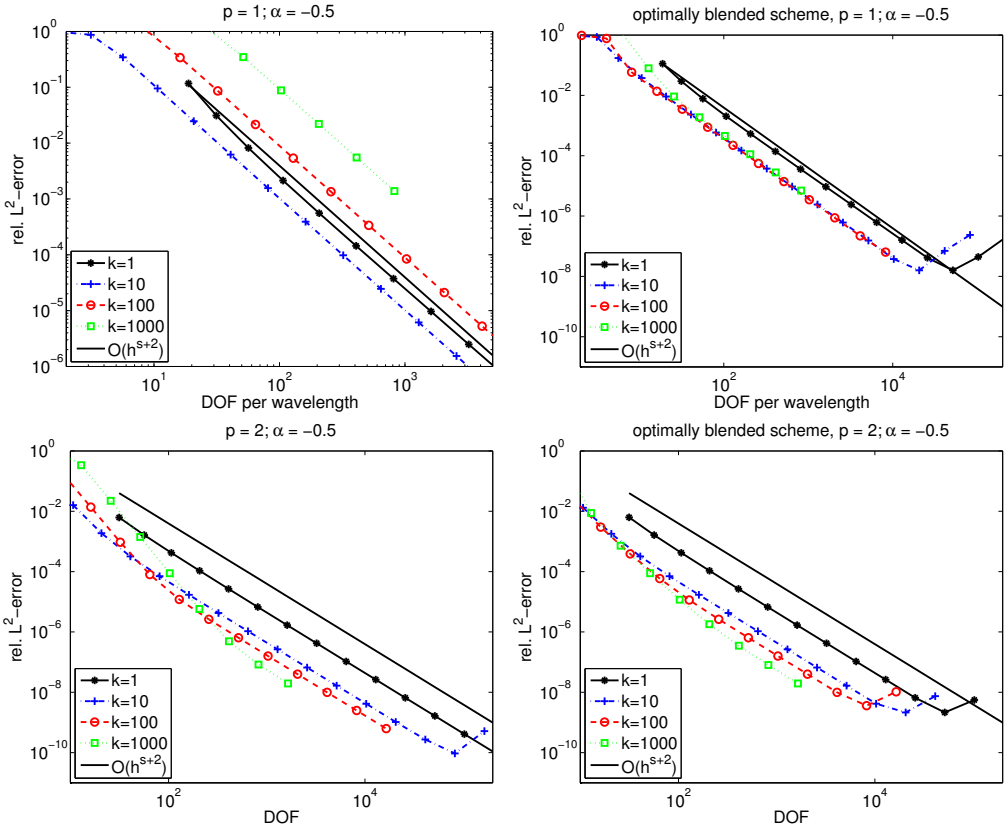


Figure 6.2.: (cf. Example 6.4.2) Galerkin FEM and optimally blended scheme for non-smooth right-hand side. top: $p = 1$ (left: Galerkin, right: optimally blended), bottom: $p = 2$ (left: Galerkin, right: optimally blended).

Example 6.4.2. We consider (3.32) with $f = x^\alpha$ and $g = 0$. For the case $\alpha = -1/2$ we present the relative error in L^2 versus the number of degrees of freedom per wavelength N_λ. We compare, for $p = 1$ and $p = 2$ the Galerkin method with the optimally blended scheme. Fig. 6.2. We remark in passing that the L^2-norm of the exact solution is observed numerically to scale like $O(k^{-3/2})$. Although this examples is not covered by Theorem 6.3.1, the optimally blended scheme is, in particular for the lowest order case, superior to the Galerkin method.

Part II.

Application to a nonlinear Helmholtz model in laser physics

7. The Steady-state Ab-initio Laser Theory

7.1. Introduction

7.1.1. Laser terminology

To begin with, we introduce and clarify some physical terms. For that purpose we start by briefly motivating a classical laser setting and the origin of laser light as a special case of electromagnetic waves [54]. Furthermore, we introduce the concept of active and inactive modes. Then we go into more detail of the properties of the cavity, the gain medium and the type of energy pumping. It is only the knowledge of these features that is required as input, as suggested by the term *ab initio* in the name SALT, which makes this theory so reliable and efficient.

Multimode lasers and nonlinear interaction

A basic laser consists of a gain medium with properties that allow it to amplify light, an optical resonator (cavity), typically a confining mirror arrangement around the gain medium, and an external energy pump source, see Figure 7.1.

The laser process is induced by excitation of the atoms of the gain material which initially only leads to spontaneous emission of electromagnetic radiation. From a quantum mechanical point of view this means that an electron absorbs the induced energy and jumps from one energy level to a higher one and returns back (randomly) to the original energy level, emitting a photon of the corresponding energy difference. The establishment of a population inversion between two energy levels (i.e. the population of electrons in the upper level is higher as in the lower one) by increasing the external pump leads to a stimulated emission of photons such that the power of the incoming radiation is amplified.

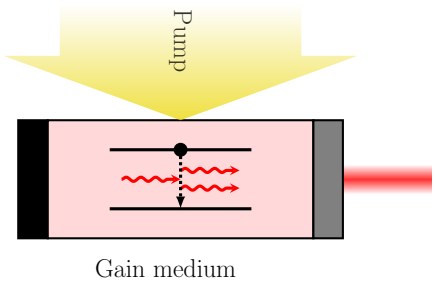


Figure 7.1.: Schematic illustration of a laser

In the presence of perfectly reflecting mirrors, this procedure results in a discrete set of infinitely many eigenstates which are coherent in space and time. These resonances are harmonic in time with real-valued frequencies. In order to couple out these modes, at least one mirror has to be partially transparent. This again induces an energy loss of the system which manifests itself in complex-valued frequencies with a negative imaginary

part and thus make these resonances decaying. However, applying an external energy pump, the loss and the gain of the system are balanced with some modes experiencing a positive shift in the imaginary direction. For a sufficiently large pump strength, the negative imaginary part in theses frequencies vanishes again and an equilibrium value of laser power is produced.

These out coupled resonances with real-valued frequencies are referred to as the laser modes that contribute to the overall laser output and will be denoted by Ψ_μ. In the further course of this work, we will also refer to these modes as the active laser modes or simply **active modes**. This is in contrast to the **inactive modes** which correspond to the resonances with frequencies that have an negative imaginary part, as they can be understood, under sufficient excitation, as possible candidates for active laser modes.

In the further course of amplification, the population inversion is naturally saturated when the density of atoms in the ground level is depleted. This effect, called **spatial hole burning**, causes a distortion of the gain shape and prevents the lasing modes of experiencing an "overload". That is, instead of the lasing modes to explode in a non-physical way, exhibited by getting frequencies with an additional positive imaginary part, their frequencies stay real-valued and the modes remain stable.

As several modes share the same amplification in the gain medium this leads in addition to cross-saturation effects. The effect of self- and cross-saturation manifests itself in a **nonlinear coupling** between the laser modes. In general, the mode with the highest power will saturate the gain, while any other mode might even experience a negative amplification, which causes its power to fade away. However, through an elaborate manipulation of loss and gain by changing the material properties or the pump distribution, many other phenomena can occur.

The cavity and pump configuration

Beside the discussed basic laser set-up a vast number of more elaborate laser processes have been achieved without a feedback mechanism as realized by standard mirrors. Modern laser cavities of that kind include dielectric micro-disk or micro-sphere lasers. In these cases the difference between the **index of refraction** of the cavity medium and the surrounding medium is very big which leads to an extremely high internal reflection of the cavity. This index of refraction enters the describing model as a function $n(x)$ which governs essentially the passive contribution of the **dielectric function** $\varepsilon(x)$ in the system. From a mathematical point of view we should mention here that this function might be piece-wise constant/continuous in space and thus one has to pay attention to the discretization for our numerical computations. In general we assume here that this is a real function which only varies in space. However, it might also be complex valued when the material has some inherent absorption properties.

So far, it seems that the laser output energy is composed of an infinite number of laser modes. But in fact, the range of frequencies is restricted by the **gain curve** Γ of the lasing medium. The explicit shape of this distribution is assumed to be of Lorentz-type, centered at the atomic transition frequency k_a of the active medium and a width at the half maximum which is twice the decay rate $\gamma_\perp$. This feature leads to a "favouring" of

laser modes with frequencies that lie near the maximum of that gain curve.

As emphasized before, sufficiently many atoms have to be excited in order to obtain laser action. To do that the atoms have to be pumped energetically from an external energy source. Many different ways of pumping a laser can be realized in practice. In particular there are no limitations to the design of the pump configuration with respect to space or pump amplitude. This motivates us to describe the pump configuration as a **pump trajectory** $D_0(\mathbf{x}, d)$ in dependence of the space variable $\mathbf{x}$ and a pump parameter d, similarly as it has been done in [65]. In previous works [45,96,98], pump configurations were considered with a global parametrization $D_0(\mathbf{x}, d) = dF(\mathbf{x})$ where $F(\mathbf{x})$ is a fixed pump profile that gets amplified via the global pump parameter d. The simplest case is $F(\mathbf{x}) \equiv 1$ which describes uniform pumping of the cavity. But even more elaborate configurations where different areas are pumped differently, for instance having two spatially separated cavities which are pumped differently as it has been done in [65], lie within the scope of our new solution method.

7.1.2. The model problem

From Maxwell-Bloch to SALT equations

The origin of the SALT equations are the Maxwell-Bloch (MB) equations which are at the heart of semi-classical laser theory [45, 96, 98]. There the gain medium is treated as an ensemble of two-level atoms embedded in a host medium. In particular they are prescribed by three time-dependent equations which are nonlinearly coupled. In the rotating wave approximation (RWA) they are given by

$$\nabla^2 E^+ - \tfrac{1}{c^2}\varepsilon_c(\mathbf{x})\ddot{E}^+ = \tfrac{1}{\varepsilon_0 c^2}\ddot{P}^+, \tag{7.1}$$

$$\dot{P}^+ = -i(k_a - i\gamma_\perp)P^+ + \tfrac{g^2}{i\hbar}E^+ D, \tag{7.2}$$

$$\dot{D} = \gamma_\parallel(D_0 - D) - \tfrac{2}{i\hbar}[E^+(P^+)^* - P^+(E^+)^*], \tag{7.3}$$

where $E^+(\mathbf{x}, t)$ and $P^+(\mathbf{x}, t)$ are the positive frequency components of the electric field and the polarization and $D(\mathbf{x}, t)$ is the population inversion. They include further physical parameters as the relaxation rates of the polarization and inversion $\gamma_\perp$ and $\gamma_\parallel$ respectively, the transition frequency of the two level atoms k_a, the so called dipole matrix element g, the dielectric function of the passive resonator denoted by $\varepsilon_c(\mathbf{x})$ and the pump with D_0. Note that this formulation describes the SALT model in one dimension as well as for TM modes in two dimensions. The key assumption of the SALT is that both the electric field as well as the polarization are assumed to be multi-periodic in time, i.e.,

$$E^+(\mathbf{x}, t) = \sum_{\mu=1}^{M} \Psi_\mu(\mathbf{x}, t)e^{-ik_\mu t} \tag{7.4}$$

$$P^+(\mathbf{x}, t) = \sum_{\mu=1}^{M} p_\mu(\mathbf{x}, t)e^{-ik_\mu t}. \tag{7.5}$$

Here, Ψ_μ and k_μ are the a priori unknown laser modes and their corresponding *real* frequencies, and M is the number of active laser modes, which is self-consistently determined at each pump step when solving the SALT equations. Furthermore, the stationary inversion approximation (SIA) is required in the case of multi-mode lasing and states that the condition $\gamma_\parallel \ll |k_\mu - k_\nu|$ must be satisfied [46]. This condition is always fulfilled in the case of single mode lasing and models multi-mode lasing for lasers which exhibits no chaotic or pulsed, but only time-harmonic behaviour [80].

By inserting the ansatz Eq. (7.5) into the MB equations, making use of the stationary inversion approximation ($\dot{D} = 0$) and taking the Fourier transformation we can formulate the SALT equations as a system of nonlinear Helmholtz-type equations:

Find $(\Psi_\mu, k_\mu), \mu = 1, \ldots, M$ such that

$$\begin{aligned} \left[\Delta + k_\mu^2 \varepsilon_\mu \left(\mathbf{x}, \{\Psi_\nu, k_\nu\}_{\nu=1}^M\right)\right] \Psi_\mu(\mathbf{x}) &= 0 \\ \lim_{r\to\infty} r^{\frac{d-1}{2}} \left(\partial_r \Psi_\mu - ik\Psi_\mu\right) &= 0 \\ \Im(k_\mu) &= 0 \end{aligned} \tag{7.6}$$

A detailed derivation has been done to a various extent in [44, 45, 64]. The nonlinear contribution in (7.6) can be split into a passive non-interacting part and a complex valued coupling term:

$$\varepsilon_\mu \left(\mathbf{x}, \{\Psi_\nu, k_\nu\}\right) = \varepsilon_c(\mathbf{x}) + \varepsilon_g^\mu \left(\mathbf{x}, \{\Psi_\nu, k_\nu\}\right),$$

where $\varepsilon_c(\mathbf{x}) : \mathbb{R}^d \to \mathbb{C}$ is a piece-wise continuous, possibly complex valued function, representing the dielectric function of the cavity. This function can basically be related to the index of refraction by $\varepsilon_c = n^2$ inside the cavity. Outside the cavity, the dielectric function is assumed to be equal to 1, representing the air or a vacuum as surrounding medium. The interaction term explicitly depends on the wave number k_μ and has the form

$$\varepsilon_g^\mu \left(\mathbf{x}, \{\Psi_\nu, k_\nu\}\right) = \frac{\gamma_\perp}{k_\mu - k_a + i\gamma_\perp} D\left(\mathbf{x}, \{\Psi_\nu, k_\nu\}\right). \tag{7.7}$$

Here, $D\left(\mathbf{x}, \{\Psi_\nu, k_\nu\}\right)$ originates from the characteristics of the population inversion. It induces the nonlinear interaction between the laser modes given by

$$D\left(\mathbf{x}, \{\Psi_\nu, k_\nu\}\right) = D_0(\mathbf{x}, d)\left[1 + \sum_{\nu=1}^{M} \Gamma(k_\nu)|\Psi_\nu(\mathbf{x})|^2\right]^{-1}. \tag{7.8}$$

Next to the modal interaction, this comprises the external pump $D_0(\mathbf{x}, d) \geq 0$. For convenience we will use the notation $\gamma(k_\mu) := \frac{\gamma_\perp}{k_\mu - k_a + i\gamma_\perp}$ with $\gamma_\perp > 0$ being the gain width and $k_a > 0$ being the frequency at the center of the gain curve Γ imposed by the cavity. The non-linearity also includes the Lorentzian gain curve $\Gamma(k_\nu)$ evaluated at k_ν and satisfies the property $\Gamma(k) = |\gamma(k)|^2$.

Previous solution method

So far the SALT equation (7.6) has only been solved in its integral form

$$\Psi_\mu(\mathbf{x}) = -\gamma(k_\mu)k_\mu^2 \int_{\mathcal{C}} D(\mathbf{x}')G(\mathbf{x},\mathbf{x}';k_\mu)\Psi_\mu(\mathbf{x}')\mathrm{d}\mathbf{x}', \tag{7.9}$$

which can be obtained using similar arguments as in Section 2.1. Then both the Green's function G and the modes Ψ_μ were expanded in terms of *constant flux* (CF) states [44] or alternatively by the *threshold constant flux* (TCF) states. Here, we just recall the approach using the TCF states. Within that approach, the external pump is assumed to be of the form $D = D_0 f(x)$. Then the TCF states are defined by the eigenvalue problem

$$\begin{aligned} \left[\nabla^2 + k^2(\varepsilon_c(\mathbf{x}) + \eta_n(k)f(\mathbf{x}))\right]\psi_n(\mathbf{x};k) &= 0 \\ \lim_{r\to\infty} r^{\frac{d-1}{2}}\left(\partial_r\psi_n(\mathbf{x};k) - ik\psi_n(\mathbf{x};k)\right) &= 0, \end{aligned}$$

where $\eta_n(k)$ are the parametrized eigenvalues and u_n the corresponding parametrized eigenstates. The domain of the cavity is given by the support of the pump function $f(\mathbf{x})$. Note that the eigenstates and their respective eigenvalues both depend on the real-valued frequency k, which is considered as an input parameter to the CF states. Consequently, these states always fulfill the SALT conditions. Using these states the expansion of laser modes reads

$$\Psi_\mu(\mathbf{x}) = \sum_n a_n^\mu \psi_n(\mathbf{x};k_\mu).$$

Substituting the Green's function by its spectral representation

$$G(\mathbf{x},\mathbf{x}';k) = -\frac{1}{k^2}\sum_i \frac{\psi_i(\mathbf{x};k)\psi_i(\mathbf{x}';k)}{\eta_i}.$$

in equation (7.9), the SALT equations (7.6) can be rewritten in terms of the expansion coefficients a_n^μ and basis states $\psi_n^\mu(\mathbf{x}) = \psi_n(\mathbf{x};k_\mu)$ as

$$a_n^\mu = \frac{\gamma(k_\mu)}{\eta_n^\mu}\sum_{n'}\int_{\mathcal{C}} \frac{f(\mathbf{x}',d)\psi_n^\mu(\mathbf{x}')\psi_{n'}^\mu(\mathbf{x}')}{1+h(\mathbf{x}')}a_{n'}^\mu \mathrm{d}\mathbf{x}', \tag{7.10}$$

where $h(x') = \sum_{\nu=1}^N \Gamma_\nu |\sum_n a_n^\nu \psi_n(\mathbf{x}';k_\nu)|^2$ contains the spatial hole burning term of equation (7.8).

Using this approach the TCF basis is ideal for studying laser systems close above threshold. The advantage of the parametrization is, that the basis set is independent of the actual gain curve of the system and can be used to easily extract general information about a laser system as used, e.g. by the single pole approximation of the SALT [44] or to make a general statement about the threshold of a coupled laser system [65]. However, the parametrized basis is of disadvantage with regard to its size when used in an actual numeric simulation.

7.2. Mathematical framework

In this section we want to put the system of PDEs as introduced in the previous section into a rigorous mathematical framework. For that purpose we denote the unknown wave functions by $u_1, \ldots, u_M$ and assume them to lie all in the same function space which we denote by $V(= H^1(\Omega))$. Then we can rewrite the governing equations in an abstract form as

$$\begin{aligned} \mathcal{L}_1(u_1, \ldots, u_M) &= 0, \\ &\vdots \\ \mathcal{L}_M(u_1, \ldots, u_M) &= 0, \end{aligned}$$

where $\mathcal{L}_\mu$ is the differential operator defining the μth differential equation in (7.6).

For the corresponding weak form there are basically two equivalent ways of description. One approach treats each equation independently as a scalar equation while the other recognizes the total system as one vector-valued equation in a global divergence form.

In the first case, one multiplies each equation by a test function $v \in V$ and integrates over the domain:

$$\begin{aligned} \int_\Omega \mathcal{L}_1(u_1, \ldots, u_M) v \mathrm{dx} &= 0, \\ &\vdots \\ \int_\Omega \mathcal{L}_M(u_1, \ldots, u_M) v \mathrm{dx} &= 0\,, \end{aligned} \tag{7.11}$$

The the second approach uses the following description in terms of a vector-valued function

$$\mathbf{u} = (u_1, \ldots, u_M) \in \mathcal{V} := V \times \cdots \times V\,.$$

Then the governing system of differential equations can then be written as

$$\boldsymbol{\mathcal{L}}(\mathbf{u}) = 0,$$

where

$$\boldsymbol{\mathcal{L}}(\mathbf{u}) = (\mathcal{L}_1(\mathbf{u}), \ldots, \mathcal{L}_M(\mathbf{u}))\,.$$

The corresponding weak formulation is obtained by taking the inner product of the vector of equations and a test function vector $\mathbf{v} \in \mathcal{V}$:

$$\int_\Omega \boldsymbol{\mathcal{L}}(\mathbf{u}) \cdot \mathbf{v} = 0 \quad \forall \mathbf{v} \in \mathcal{V}. \tag{7.12}$$

Note that (7.12) is one scalar equation. In order to generate the M independent variational equations one has to choose M linearly independent test function vectors $\mathbf{v} \in \mathcal{V}$. In particular, $\mathbf{v} = (0, \ldots, 0, v^{(i)}, 0, \ldots, 0)$ recovers the ith variational form $\int_\Omega \mathcal{L}^{(i)} v^{(i)} \mathrm{dx} = 0$

in (7.11) for $i = 1, \ldots, M$.

For a closer inspection we have to extend to the function space $\mathcal{V} = H^1(\Omega, \mathbb{C}^M) := \left(H^1(\Omega)\right)^M$, provided with the inner product

$$(\mathbf{u}, \mathbf{v}) := (\mathbf{u}, \mathbf{v})_{H^1(\Omega, \mathbb{R}^M)} = \sum_{i=1}^{M} \sum_{|\alpha| \leq 1} (\partial^\alpha u^i, \partial^\alpha v^i)_{L^2(\Omega)}$$

and the norm

$$\|\mathbf{u}\|_{H^1(\Omega)} := \|\mathbf{u}\|_{H^1(\Omega, \mathbb{R}^M)} = \left(\sum_{i=1}^{M} \sum_{|\alpha| \leq 1} \|\partial^\alpha u_i\|^2_{L^2(\Omega)} \right)^{1/2} \tag{7.13}$$

A detailed structure of the quasi-linear system in divergence form can be written as

$$\boldsymbol{\mathcal{L}}(\mathbf{u}) = -\text{Div } A(D\mathbf{u}) + \mathbf{b}(x, \mathbf{u}) \tag{7.14}$$

where $D\mathbf{u}$ denotes the Jacobian matrix of the map $\mathbf{u} : \mathbb{R}^d \supset \Omega \to \mathbb{C}^{M \times d}$ defined by

$$D\mathbf{u} := (D\mathbf{u})^i_\alpha = D_\alpha u^i(x) = \partial u^i(x)/\partial x_\alpha, \quad i = 1, \ldots, M, \alpha = 1, \ldots, d$$

and $A : \mathbb{C}^{M \times d} \to \mathbb{C}^{M \times d}$ defined element-wise by

$$A^i_\alpha(\xi^i_\beta) = \sum_{j=1}^{M} \sum_{\beta=1}^{d} A^{i,j}_{\alpha,\beta} \xi^j_\beta \quad \text{and} \quad A^{i,j}_{\alpha,\beta} := \delta_{i,j}\delta_{\alpha,\beta}.$$

Note that A is completely linear and decoupled. Furthermore, the elliptic system is in fact semi-linear as it is only linear in the second (leading) order term. However, the system is coupled non-linearly through the lower order term bfb in (7.14).

Two major techniques for the solvability of such semi-linear systems are the variational method and the monotone operator method [27]. The special case of a system with two equations written as

$$\vec{\Delta}\mathbf{u} + K\mathbf{u} + Q(\mathbf{u}) = 0$$

with $\mathbf{u} = (u^1, u^2), \vec{\Delta} = \begin{pmatrix} \Delta & 0 \\ 0 & \Delta \end{pmatrix}, K = \begin{pmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{pmatrix}$ and $Q = \begin{pmatrix} q_1(\mathbf{u}) & 0 \\ 0 & q_2(\mathbf{u}) \end{pmatrix}$ where q_1, q_2 being bounded, has also been investigated by Zuluaga [103] including some bifurcation analysis at resonance. However, the investigation of systems is rather poorly developed compared to the analysis of one equation. The single equation for one unknown function is basically of Landesman-Lazer type, see Zeidler, [102, §29.9] and

more generally [101, §7.16]. A detailed examination has been done by Ambrosetti and Mancini [6] of the nonlinear elliptic equation

$$\Delta u + \lambda u + \frac{u}{1+u^2} = g$$

for smooth right hand sides $g \in H_0^1(\Omega)$. Further analysis of bifurcation on a multi-parameter system of elliptic equations

$$-\Delta u = \lambda u + \Upsilon(u, \lambda)$$

with Dirichlet boundary conditions and $\Upsilon(u, \lambda) = o(\|u\|)$ locally uniform in λ, has been done by Gawrycka, Rybick in [43]. For solvability results of the semi-linear equation one usually aims to exploit the results on linear elliptic PDEs. This requires to have the corresponding linear differential equation under functional-analytic control and provides us with a legitimate nexus to Part I of this work.

7.3. The FEM discretization

Next we used the high order FEM as introduced in Section 4.2 for numerical computations. To this end, we repeat the weak formulation of the coupled system of nonlinear PDEs from Section 7.1:

Find $(u_\mu, k_\mu) \in V \times \mathbb{R},\ \mu = 1, \ldots, M$ such that

$$\begin{aligned} -\int_\Omega \nabla u_\mu \nabla v_\mu + ik_\mu \int_{\partial\Omega} u_\mu v_\mu + k_\mu^2 \int_\Omega \varepsilon_c(x) u_\mu v_\mu \\ + k_\mu^2 \gamma(k_\mu) \int_\Omega \frac{D_0(x,d)}{1 + \sum_\nu \Gamma(k_\nu)|u_\nu(x)|^2} u_\mu v_\mu = 0 \end{aligned} \tag{7.15}$$

for arbitrary $v_\mu \in V$ and M the number of active modes at pump strength d. Note that we discuss here the discretization scheme using the "Robin-notation" from the one dimensional case. This is basically also applicable in higher dimensions, better approximation schemes have been discussed in Section 2.2.

To complete the discretization, we recall some notations from Section 4.2:
Let $\overline{\Omega}_h := \bigcup_{K \in \mathcal{T}_h} K$ be the polygonal approximation of the cavity domain Ω, constructed by partition $\mathcal{T}_h$ of finitely many closed triangles K. Further let $\mathcal{V}_N$ be a finite dimensional subspace of the function space $\mathcal{V}$ and $\{\varphi_i\}_{n=1}^N$ a basis of $\mathcal{V}_N$ and replace the exact solution $u_\mu \in \mathcal{V}$ in (7.15) by the discrete approximation

$$u_N^\mu := \sum_i u_i^\mu \varphi_i \in \mathcal{V}_h.$$

Details on the choice of the domain triangulation and the basis function were discussed in Section 4.2. In addition, we denote the complex coefficient vector as $\mathbf{u}_\mu := \left(u_1^\mu, \ldots, u_M^\mu\right)$ and $X := (X_1, \ldots, X_M), X_\mu := (\mathbf{u}_\mu, k_\mu)$. Then, by extracting sums out of the integrals

and assembling the contributions for all elements we arrive at the following finite element scheme in matrix form:

Find $X_\mu \in \mathbb{C}^N \times \mathbb{R}, \mu = 1, \ldots, M$ such that

$$F_\mu(X) := \big[-\mathbf{L} + ik_\mu \mathbf{R} + k_\mu^2 \mathbf{M}^{\varepsilon_c} + k_\mu^2 \gamma(k_\mu)\mathbf{Q}(X)\big]\mathbf{u}_\mu = 0. \tag{7.16}$$

Here, the sparse $N \times N$-matrices $\mathbf{L}, \mathbf{R}, \mathbf{M}^{\varepsilon_c}, \mathbf{Q}(X)$ are the stiffness matrix

$$\mathbf{L} := \left(\int_\Omega \nabla\varphi_i \cdot \nabla\varphi_j dx \right)_{i,j}$$

corresponding to the Laplacian term, the mass matrix

$$\mathbf{M}^{\varepsilon_c} := \left(\int_\Omega \varepsilon_c(x)\varphi_i\varphi_j dx \right)_{i,j}$$

containing the passive dielectric function, the matrix

$$\mathbf{R} := \left(\int_{\partial\Omega} \varphi_i\varphi_j d\sigma_x \right)_{i,j}$$

which only involves the boundary elements and reflects the outgoing boundary condition, and the nonlinear contribution

$$\mathbf{Q}(X) := \left(\int_\Omega \frac{D_0(x,d)\varphi_i(x)\varphi_j(x)}{1 + \sum_\nu \Gamma(k_\nu)|\sum_l b_l^\nu \varphi_l(x)|^2} dx \right)_{i,j}$$

which accounts for the nonlinear coupling including the spatial hole burning effect.

.

8. The nonlinear SALT-Algorithm

In this chapter, we describe the direct solution approach of the coupled system of the nonlinear SALT equations introduced in Chapter 7. In particular, we present two different concepts of a direct solution method for the SALT equations.

Following the idea of a pump trajectory, it is possible to use our new solution method as a path-following algorithm which can be understood as intuitively as a real-life experimental set-up where scientists increase the pump strength systematically in order to amplify the laser emission. The algorithm solves directly the nonlinear system (7.6) while successively increasing the pump parameter. In this manner we directly track the whole laser mode information according to the imposed pump history.

In addition, we present a further solution method to compute the SALT solutions for one single pump strength independent of any foregoing pump configuration. In principle this uses the same algorithmic components as used in the algorithm that follows the pump trajectory.

For both concepts we will see that the solution is essentially reduced to the solution of a coupled system of nonlinear PDEs via Newton method and the computation of a nonlinear eigenvalue problem.

8.1. The consecutive pump algorithm

Following the concept of a pump trajectory as discussed in Section 7.1, we construct a path-following algorithm which computes directly all laser modes and frequencies at each pump step. In this manner, our algorithm simulates the process of multi-mode lasing for a general pump configuration with pump parameter d.

We start with $d = 0$ and increase successively the pump parameter as the experimentalists would do when amplifying the laser cavity. The pump strength is thus below the first threshold where the laser/first mode starts to emit/activate and no nonlinear interaction takes place. Thereby, the system is reduced then to a single equation which obviously has no solution (Ψ, k) with $\Im(k) = 0$ as requested in the SALT equations. Nevertheless, in order to find the first pump threshold d_1 and the corresponding solution $(\Psi_1, k_1)(d_1)$ we solve the resulting equation as a rational eigenvalue problem

$$T_{d,0}(k)\Psi(x) := \Delta\psi(x) + k^2\big(\varepsilon_c(x) + \gamma(k)D_0(x,d)\big)\Psi(x) = 0 \quad + \ b.c. \tag{8.1}$$

for $0 \leq d \leq d_1$. So far the system has initially only eigenpairs (Ψ_n, k_n) with $\Im(k_n) < 0$, but for each pump step the eigenvalues in the negative imaginary plane will travel towards the real axis, see Fig. 8.1. Thus at $d = d_1$ the first laser mode activates and nonlinear interaction sets in.

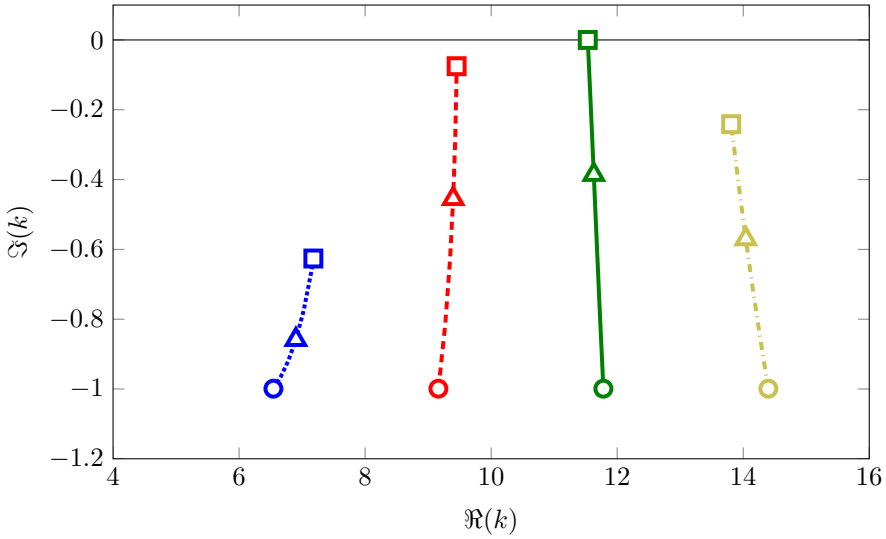


Figure 8.1.: Trajectories of the eigenvalues in the complex plane with a uniformly increasing pump strength $d \in [0, 0.267]$ for a 1D slab cavity laser of length one. The gain medium has a spatially uniform index of refraction $n = 1.2$ and the gain paramters $k_a = 10$ and $\gamma_\perp = 4$.

As soon as the first threshold pump strength d_1 and the corresponding solution $(\Psi_1, k_1)(d_1)$ are determined we continue to increase the pump parameter and solve the nonlinear problem

$$F_{d,1}(\Psi, k) = \Delta\Psi(x) + k^2\big(\varepsilon_c(x) + \frac{\gamma(k)D_0(x,d)}{1+\Gamma(k)|\Psi(x)|^2}\big)\Psi(x) = 0 \quad + \ b.c. \tag{8.2}$$

for $d_1 < d < d_2$ with d_2 being the pump strength where the second laser mode activates. We solve this problem via the Newton scheme where the Jacobian is known analytically and use the solution of the previous pump step as initial guess.

In order to verify if a new laser mode activates we additionally need to check the other eigenpairs of that system at pump strength d. Therefore we insert the previously computed solution $(\Psi_1^*, k_1^*)(d)$ into the denominator in (8.2) which turns the nonlinear problem again into a rational eigenvalue problem

$$T_{d,1}(k)\Psi(x) := \Delta\psi(x) + k^2\Big(\varepsilon_c(x) + \gamma(k)\frac{D_0(x,d)}{1+\Gamma(k_1^*)|\Psi_1^*(x)|^2}\Big)\Psi(x) = 0 \quad + \ b.c.$$

As soon as the imaginary part of one of the eigenvalues becomes positive, we know that a new laser mode must have become active. Thus from the next pump step on we have to solve the system with an additional mode and the size of the nonlinear problem is increased by one.

This procedure can now be continued with the increasing pump strength such that for $d_{M-1} \le d \le d_M$ we search for $X := \{(\Psi_\mu, k_\mu)\}_{\mu=1}^M$ satisfying the nonlinear coupled

system:

$$F_{d,M}(X) := (F_1(X), \dots, F_M(X)) = 0 \quad + \ b.c. \tag{8.3}$$

with

$$F_{d,\mu}(X) = \Delta\Psi_\mu(x) + k_\mu^2\big(\varepsilon_x(x) + \frac{\gamma(k_\mu)D_0(x,d)}{1+\sum_{\nu=1}^{M}\Gamma(k_\nu)|\Psi_\nu(x)|^2}\big)\Psi_\mu(x)$$

for $\mu = 1, \dots, M$ and M being the number of active modes. The corresponding rational eigenvalue problem is of the form

$$T_{d,M}(k)\Psi(x) := \Delta\psi(x) + k^2\big(\varepsilon_c(x) + \frac{\gamma(k)D_0(x,d)}{1+\sum_{\nu=1}^{M}\Gamma(k_\nu^*)|\Psi_\nu^*(x)|^2}\big)\Psi(x) = 0 \quad + \ b.c. \tag{8.4}$$

Thus within each pump step along the pump trajectory we have to compute a rational EVP and after the first pump threshold we can solve the coupled system of nonlinear equations directly. For the computation of the nonlinear problem we use the Newton iteration and choose the previously computed solution as the initial guess for the next pump step. In case of the activation of an additional mode, detected during the computation of the EVP, we append the eigenfunction of the corresponding eigenvalue with the detected sign change in the imaginary part and join only the real part of this eigenvalue.

A schematic structure of this solution algorithm can be seen in Algorithm **??**:
For some laser configurations there might exist a specific pump configuration which causes the laser modes to shut down again [65]. Even this phenomenon can be handled by our solution method without any notable effort. In a situation when the laser mode turns off again, the Newton solver of the current number of active modes M will not converge as we suppress the zero solution for stability reasons, see Section 9. Thus by detecting the mode with minimal intensity (which is close to zero) and omitting it, the procedure can be continued for the $M-1$ remaining modes. Thus we reduce the number of active modes, but the solution strategy remains the same.

In summary, our problem reduces essentially to the solution of a coupled system of nonlinear PDEs via Newton's method and the computation of a rational eigenvalue problem which will be discussed in more detail in Section 9 and Section 10 respectively.

8.2. The instant pump algorithm

In contrast to the consecutive pump algorithm it is possible to speed up the calculations (of the direct solver) when the laser information is only requested at a specific pump. This means that we are also able to solve the SALT system for a single pump strength independently of any foregoing pump configuration. We will see that this method uses in principle the same algorithmic components as within the algorithm that follows successively the pump trajectory.

We start again by solving the linear problem (8.1) for the distinct pump d^* as a nonlinear EVP of the form

$$T_0(k)\Psi(x) := \Delta\Psi(x) + k^2\big(\varepsilon_c(x) + k^2\gamma(k)D_0(x,d^*)\big)\Psi(x) = 0 \quad + \ b.c. \tag{8.5}$$

In case of d^* being bigger than the first threshold pump this provides non-physical solutions with eigenfrequencies with positive imaginary part. In order to find the right number of active modes we continue by solving the nonlinear SALT system for a single lasing mode

$$F_1(X) := \Delta\Psi + k^2\big(\varepsilon_c(x) + k^2\frac{\gamma(k)D_0(x,d^*)}{1+\Gamma(k)|\Psi|^2}\big)\Psi(x) = 0 \quad + \ b.c. \tag{8.6}$$

via the Newton scheme and use the laser mode of the resonance with the highest positive imaginary part and the corresponding real part of this eigenfrequency as initial data. Following the same argumentation as in Sec. 8.1, we insert then the solution $\{\Psi^*, k^*\}$ of (8.6) into the denominator of the same equation and solve again the resulting EVP

$$T_1(k)\Psi(x) := \Delta\Psi(x) + k^2\big(\varepsilon_c(x) + \frac{\gamma(k)D_0(x,d^*)}{1+\Gamma(k^*)|\Psi^*|^2}\big)\Psi(x) = 0.$$

The SALT solution $\{u^*, k^*\}$ solves again this EVP, but further non-physical resonances might still exist. In that case we extend the SALT system for an additional lasing mode and use again the resonance with the highest positive imaginary part as initial data. This procedure continues until no solution with eigenfrequencies with positive imaginary part remains.

In order to get a better insight of how the solution strategy for an instant pump proceeds we illustrate this with the following example:

Example 8.2.1. We consider a 1D model of a slab cavity $\Omega = [0, 1]$ with perfect mirror on the left side, that is $\Psi(0) = 0$, and open boundary on the right side, i.e. $\Psi'(1) - ik\Psi(1) = 0$. Furthermore, we choose $\varepsilon(x) \equiv 1.2, k_a = 10, \gamma_\perp = 4$ and $D_0(\mathbf{x}, d) \equiv 1$.

For solving the eigenvalue problem we use here the contour integral method which we will explain in detail in Section 10.3. Thus initially solving the EVP $T_0(k)u = 0$ produces 8 nonphysical eigenvalues in the positive imaginary half-plane, see Figure 8.2(a). Again we consider the corresponding eigenvalues as possible candidates for active lasing modes satisfying the SALT equations. In particular the mode to the eigenvalue with the highest imaginary part is most likely a mode that is lasing within the fully interacting regime.

Without knowing the final number of active modes at the desired pump d we start with "testing" for one laser mode and solve the nonlinear system (which is in this case only a single equation) by using the mode with the highest imaginary part and the corresponding real part of that frequency as initial guess. If the nonlinear iteration converges, we have found an active mode with a frequency that is "forced" to be real. We then have to include this particular mode into the coupling term in order to check the remaining eigenvalues. As the inclusion of the current data into the nonlinear term reduces the pump within the system, all other eigenvalues have to move downwards in the complex plane. This can be seen in Figure 8.2(b).

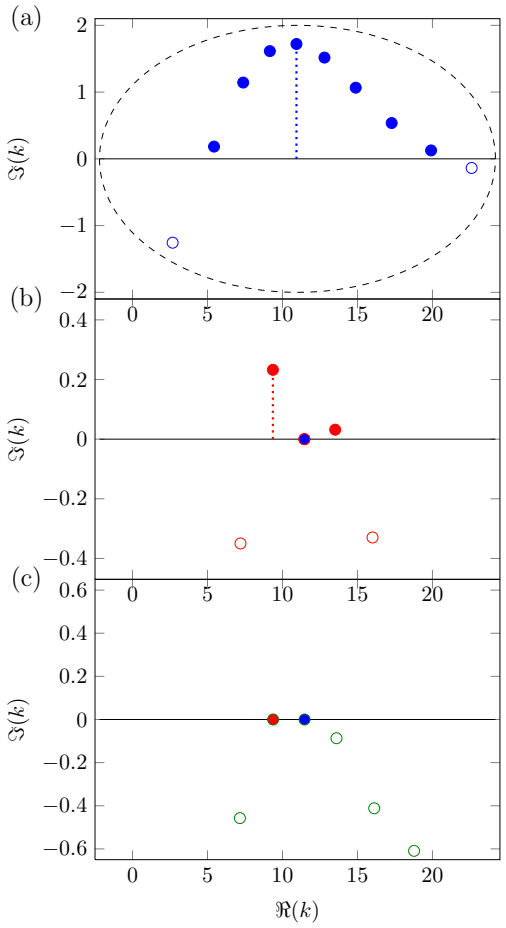


Figure 8.2.: Fast solution algorithm

While some eigenvalues have been suppressed such that they already lie far below the real axis, there are still two eigenvalues above the real axis. Thus there is obviously more than one mode active at that specific pump strength. Hence we choose again the mode with the highest imaginary part as the additional initial guess and solve the SALT equations, but this time for two modes. Finally the Newton solver converges again providing two laser modes of purely real valued frequencies. Verifying again the remaining inactive modes by solving the corresponding nonlinear EVP shows that we have found a correct solution for the SALT system at the pump d, see Figure 8.2(c).

Of course one could start with any other non-physical mode from Figure 8.2(a) and test all possible combinations of initial guesses. We have experienced that either a different combination does not converge at all or it finds the same final pair of active laser modes at the end. However, a rigorous mathematical investigation of the subject of multiple combinations of active modes solving the SALT equations for the same distinct pump hasn't been carried out yet.

A scheme of this instant pump algorithm can be seen in Algorithm 8.1 below:

Algorithm 8.1 Scheme of the instant pump algorithm

Input: Laser geometry, FEM parameter, gain profile, pump strength d^*
  Solve EVP $T_0(k)U = 0$
  Determine M_+, the number of eigenvalues with pos. imaginary part
  Determine initial data (U_1, k_1)
  Initialize number of active modes $M_0 = 0$
  while $M_+ > 0$ **do**
    $M_0 = M_0 + 1$
    Solve Newton $F(X) = 0, F = (F_1, \ldots, F_{M_0}),\ X = (U, \vec{k})$;
    Solve EVP $T_{M_0}(k)U = 0$
    Determine M_+, number of eigenvalues with pos. imaginary part
    Determine initial data (U_{M_0}, k_{M_0}) and append to X
  end while
Output: $X = (U, \vec{k}), M_0$

9. Solving the nonlinear system

In this chapter we concentrate on the computation of the fully nonlinear SALT equations. A straight forward approach to solve the nonlinear system is to apply a Newton scheme and search for all modes and frequencies $\{u_\mu, k_\mu\}$ simultaneously.

9.1. Newton's method

As emphasized in the two previous sections, one of the crucial steps in the computations is the direct solution of the coupled system of nonlinear equations in (7.6). Following the discussions in Section 7.3 the discretized system can be written as a set of algebraic equations,

$$F(\{\mathbf{b}_\mu, k_\mu\}) = 0, \tag{9.1}$$

where μ ranges from 1 to M, the number of currently active laser modes. Since we have to solve the system of equations, Eq. (9.1) has to be understood as an M-tuple $F(\{\mathbf{b}_\mu, k_\mu\}) = \big(F_1(\{\mathbf{b}_\mu, k_\mu\}), \ldots, F_M(\{\mathbf{b}_\mu, k_\mu\})\big)$.

Assuming that we have an appropriate initial guess X_0, this problem is solved by using the Newton scheme

$$X_{n+1} = X_n - J(X_n)^{-1} F(X_n)\,. \tag{9.2}$$

From equations (9.1) rand (7.16) we observe that $F : \mathbb{C}^{N\times M} \times \mathbb{R}^M \to \mathbb{C}^{N\times M}$ and $F_\mu : \mathbb{C}^N \times \mathbb{R} \to \mathbb{C}^N$. Thus our discrete system is so far under-determined. Formally we would have to add a supplementary condition $f(X) = 0$ with $f : \mathbb{C}^{N\times M} \times \mathbb{R}^M \to \mathbb{R}^M$ in order to establish an appropriate linearization scheme. This leads then to a completed nonlinear system $\tilde{F}(X) = \big(F(X), f(X)\big) = 0$ where $\tilde{F} : \mathbb{C}^{N\times M} \times \mathbb{R}^M \to \mathbb{C}^{N\times M} \times \mathbb{R}^M$.

In effect the solutions of (7.6) are phase invariant. Thus one has to fix the phase of the problem in order to obtain a well determined system such that the Newton scheme can be applied and the Jacobian becomes invertible. The concrete treatment of this subject will be discussed in Section 9.2. The Jacobian $J(X)$ in (9.2) is computed explicitly by "symbolic differentiation" and discussed in detail in Section 9.3.

9.2. Stability conditions

As mentioned in the previous section, we observe that the solutions of (7.6) are phase invariant. In order to fix this lack of uniqueness for our Newton scheme, we have to stipulate further phase conditions. That is, we choose for each coefficient vector $\mathbf{u}_\mu$ of the initial data $X_0 = (U, \vec{k})$ one index n^*_μ such that $u^*_\mu := u^\mu_{n^*_\mu} \neq 0$ and rotate the initial coefficient vectors by the phase angle $\alpha^*_\mu := \Phi(u^*_\mu)$ to $\tilde{\mathbf{u}}_\mu := e^{-i\alpha^*_\mu}\mathbf{u}_\mu$ such that the

corresponding coefficient becomes real, i.e. $\Im(u_\mu^*) = 0$. Here, $\Phi : \mathbb{C} \to [-\pi, \pi), \Phi(z) := \Im(\log(z))$ returns the phase angle in radians to each point in the complex plane. In order to keep these phase conditions during Newton's iterations, we append the M additional equations

$$f_\mu(X) := \Im(u_\mu^*) = 0$$

where $f_\mu : \mathbb{C}^N \times \mathbb{R} \to \mathbb{R}$. In consequence, this modification turns our Newton iteration into a feasible procedure.

Additionally, one can see that the trivial zero solution always satisfies the system. This causes a second issue at the points when a new lasing mode appears. At that moment, its amplitude is small and difficult to distinguish from the trivial zero solution. Therefore, we have to exclude the trivial solution for further stability of the Newton scheme. In our scheme we scale the modified initial coefficient vectors by a factor s_μ to $\hat{\mathbf{u}}_\mu := \tilde{\mathbf{u}}_\mu / s_\mu$ such that the n_μ^*-th entry of $\hat{\mathbf{u}}_\mu$ is normalized to one, i.e. $\hat{u}_\mu^* = 1$. However these scaling factors have to be determined as M additional unknown variables. This extends the overall unknown variable to $\hat{X} := (X, s)$ with $s = (s_1, \ldots, s_M)$. Keeping these conditions as M further equations

$$g_\mu(\hat{X}) := \Re(u_{n^*}^\mu) - 1$$

during Newton's iteration causes then the resulting solution never to attain the trivial zero solution. In summary, our stable Newton problem reads:

Find $\hat{X}_\mu := (X_\mu, s_\mu) \in \mathbb{C}^N \times \mathbb{R} \times \mathbb{R}, 1 \leq \mu \leq M$ such that

$$\hat{F}(\hat{X}) := \big(\hat{F}_1(\hat{X}), \ldots, \hat{F}_N(\hat{X})\big) = 0$$

with

$$F_\mu(\hat{X}) := \big(F_\mu(\hat{X}), f_\mu(\hat{X}), g_\mu(\hat{X})\big) \tag{9.3}$$

for $1 \leq \mu \leq M$.

Note that from the point of view of implementation the FEM structure remains the same in all linear components of

$$F_\mu(\hat{X}) = \big[-\mathbf{L} + ik_\mu \mathbf{R} + k_\mu^2 \mathbf{M}^{\varepsilon_c} + k_\mu^2 \gamma(k_\mu) \hat{\mathbf{Q}}(\hat{X}) \big] \mathbf{u}^\mu$$

as the additional unknown variables s_μ only enter in the nonlinear term by

$$\hat{\mathbf{Q}}(\hat{X}) := \Big(\int_\Omega \frac{F(x)}{1 + \sum_\nu s_\nu^2 \Gamma(k_\nu) |\sum_l u_l^\nu \varphi_l(x)|^2} \varphi_i(x) \varphi_j(x) dx \Big)_{i,j}.$$

The schematic structure of the stabilized Newton scheme which fixes one specific phase and extracts the auxiliary unknowns s_μ during the iteration process can be seen in Algorithm 9.1.

Algorithm 9.1 Stable Newton scheme

Input: $X_0 = (U, \vec{k})$ where $U = (\mathbf{u}_1, \ldots, \mathbf{u}_M)^\top, \mathbf{u}_\mu = (u_1^\mu, \ldots, u_N^\mu), \vec{k} = (k_1, \ldots, k_M)$
  for $\mu = 1, \ldots, M$ **do**
    Choose index $n_\mu^* \in \{1, \ldots, N\}$ such that $u_\mu^* := u_{n_\mu^*}^\mu \neq 0$
    Determine $[\varphi_\mu, s_\mu] = \texttt{cat2pol}(u_\mu^*)$
    Transform $\mathbf{u}_\mu = 1/s_\mu e^{i\varphi_\mu}\mathbf{u}_\mu$
  end for
  $Y_0 := (X_0, \vec{s})$ where $\vec{s} := (s_1, \ldots, s_M)$
  $err := \|\hat{F}(Y_0)\|,\ tol := 10^{-14}$
  $Y = Y_0$
  while $err \geq tol$ **do**
    Compute $\hat{J}(Y), \hat{F}(Y)$
    Solve $\hat{J}(Y)x = \hat{F}(Y)$
    $Y = Y - x$
    $err = \|\hat{F}(Y)\|$
  end while
  for $\mu = 1, \ldots, M$ **do**
    Back transform $\mathbf{u}_\mu = s_\mu \mathbf{u}_\mu$
  end for
Output: $X = (U, \vec{k})$

9.3. The explicit Jacobi matrix

As one can see from previous discussions in Section 7.3, 9.1 and 9.2, the stabilized nonlinear function $\hat{F}$ defined on the discrete FEM space produces a squared Jacobi matrix J of large extent. For efficiency reasons, it is thus recommendable to understand and exploit the underlying sparsity structure. Assembling the Jacobian matrix through its explicitly analytic expression becomes much more efficient than computing the Jacobian matrix by numerical perturbation. Furthermore the implementation of the analytic Jacobian is also much easier as commonly thought.

In order to derive the Jacobi matrix explicitly, we should first mention that it is convenient to reconsider the complex-valued coefficients in $\mathbb{C}$ as a real-valued vector $u_j^\mu = (v_j^\mu, w_j^\mu)$ in $\mathbb{R}^2$. To this end, we introduce the notation $\mathbf{u}_\mu := \mathbf{v}_\mu + i\mathbf{w}_\mu$. Consequently (7.16) has to be split into its real and imaginary contribution. Therefore we note first

$$\gamma(k) = \frac{\gamma_\perp}{k - k_a + i\gamma_\perp} = \frac{\gamma_\perp(k - k_a)}{(k - k_a)^2 + \gamma_\perp^2} - i\frac{\gamma_\perp^2}{(k - k_a)^2 + \gamma_\perp^2} = \frac{k - k_a}{\gamma_\perp}\Gamma(k) - i\Gamma(k).$$

Then we get

$$\begin{aligned} F_\mu(\hat{X}) &= \big[-\mathbf{L} + ik_\mu\mathbf{R} + k_\mu^2\mathbf{M}^{\varepsilon_c} + k_\mu^2\gamma(k_\mu)\mathbf{Q}(\hat{X})\big](\mathbf{v}_\mu + i\mathbf{w}_\mu) \\ &= \big[-\mathbf{L} + ik_\mu\mathbf{R} + k_\mu^2\mathbf{M}^{\varepsilon_c} + k_\mu^2\big(\frac{k_\mu - k_a}{\gamma_\perp}\Gamma(k_\mu) - i\Gamma(k_\mu)\big)\mathbf{Q}(\hat{X})\big]\mathbf{v}_\mu \end{aligned}$$

$$
\begin{aligned}
&+ i\big[-\mathbf{L} + ik_\mu\mathbf{R} + k_\mu^2\mathbf{M}^{\varepsilon_c} + k_\mu^2\big(\frac{k_\mu - k_a}{\gamma_\perp}\Gamma(k_\mu) - i\Gamma(k_\mu)\big)\mathbf{Q}(\hat{X})\big]\mathbf{w}_\mu \\
&= -\mathbf{L}\mathbf{v}_\mu - k_\mu\mathbf{R}\mathbf{w}_\mu + k_\mu^2\mathbf{M}^{\varepsilon_c}\mathbf{v}_\mu + k_\mu^2\Gamma(k_\mu)\big(\frac{k_\mu - k_a}{\gamma_\perp}\mathbf{Q}(\hat{X})\mathbf{v}_\mu + \mathbf{Q}(\hat{X})\mathbf{w}_\mu\big) + \\
&+ i\big(-\mathbf{L}\mathbf{w}_\mu + k_\mu\mathbf{R}\mathbf{v}_\mu + k_\mu^2\mathbf{M}^{\varepsilon_c}\mathbf{w}_\mu + k_\mu^2\Gamma(k_\mu)\big(\frac{k_\mu - k_a}{\gamma_\perp}\mathbf{Q}(\hat{X})\mathbf{w}_\mu - \mathbf{Q}(\hat{X})\mathbf{v}_\mu\big)\big) \\
&=: F_\mu^{\Re}(\hat{X}) + iF_\mu^{\Im}(\hat{X})
\end{aligned}
$$

Thus, identifiying the complex plane $\mathbb{C}$ with $\mathbb{R}^2$, (7.16) can also be written in the matrix form as

$$
\begin{pmatrix} -\mathbf{L} + k_\mu^2\mathbf{M}^{\varepsilon_c} & -k_\mu\mathbf{R} \\ k_\mu\mathbf{R} & -\mathbf{L} + k_\mu^2\mathbf{M}^{\varepsilon_c} \end{pmatrix} \begin{pmatrix} \mathbf{v}_\mu \\ \mathbf{w}_\mu \end{pmatrix} + k_\mu^2\Gamma(k_\mu) \begin{pmatrix} \frac{k_\mu - k_a}{\gamma_\perp}\mathbf{Q} & \mathbf{Q} \\ -\mathbf{Q} & \frac{k_\mu - k_a}{\gamma_\perp}\mathbf{Q} \end{pmatrix} \begin{pmatrix} \mathbf{v}_\mu \\ \mathbf{w}_\mu \end{pmatrix} = 0
$$

We mentioned before that $\mathbf{M}^{\varepsilon_c}$ might be a complex matrix in case of ε_c being a complex-valued function. Then we also have to split $\mathbf{M}^{\varepsilon_c} = \Re(\mathbf{M}^{\varepsilon_c}) + i\Im(\mathbf{M}^{\varepsilon_c})$ and the first matrix has to be rearranged to

$$
\begin{pmatrix} -\mathbf{L} + k_\mu^2\Re(\mathbf{M}^{\varepsilon_c}) & -k_\mu\mathbf{R} - k_\mu^2\Im(\mathbf{M}^{\varepsilon_c}) \\ k_\mu\mathbf{R} + k_\mu^2\Im(\mathbf{M}^{\varepsilon_c}) & -\mathbf{L} + k_\mu^2\Re(\mathbf{M}^{\varepsilon_c}) \end{pmatrix}.
$$

This leads to the following real-valued formulation of the discrete nonlinear problem (9.1):

$$
F(\hat{X}) = \begin{pmatrix} F_1^{\Re}(\hat{X}), \ldots, F_M^{\Re}(\hat{X}) \\ F_1^{\Im}(\hat{X}), \ldots, F_M^{\Im}(\hat{X}) \end{pmatrix} = 0 \tag{9.4}
$$

For the corresponding symbolic derivation of the Jacobi matrix we use again the following common notation:

$$
\hat{X} = (\hat{X}_1, \ldots, \hat{X}_M),\ \hat{X}_\mu = (\mathbf{v}_\mu, \mathbf{w}_\mu, k_\mu, s_\mu),\ \mu = 1, \ldots, M
$$

and

$$
F(\hat{X}) = \left(F_1\left(\hat{X}\right), \ldots, F_M\left(\hat{X}\right)\right),\ J(\hat{X}) = \left(\frac{\partial F_\mu(\hat{X})}{\partial \hat{X}_\nu}\right)_{\mu,\nu=1}^{M}
$$

and each Jacobi block has the structure

$$
\frac{\partial F_\mu(\hat{X})}{\partial \hat{X}_\nu} = \begin{pmatrix} \frac{\partial F_\mu^{\Re}}{\partial \mathbf{v}_\nu} & \frac{\partial F_\mu^{\Re}}{\partial \mathbf{w}_\nu} & \frac{\partial F_\mu^{\Re}}{\partial k_\nu} & \frac{\partial F_\mu^{\Re}}{\partial s_\nu} \\ \frac{\partial F_\mu^{\Im}}{\partial \mathbf{v}_\nu} & \frac{\partial F_\mu^{\Im}}{\partial \mathbf{w}_\nu} & \frac{\partial F_\mu^{\Im}}{\partial k_\nu} & \frac{\partial F_\mu^{\Im}}{\partial s_\nu} \\ \frac{\partial f_\mu}{\partial \mathbf{v}_\nu} & 0 & 0 & 0 \\ 0 & \frac{\partial g_\mu}{\partial \mathbf{w}_\nu} & 0 & 0 \end{pmatrix} (\hat{X}).
$$

Thus the Jacobi matrix doesn't seem to be that sparse in the first instance. However we have

$$\frac{\partial f_\mu}{\partial \mathbf{v}_\nu}, \frac{\partial g_\mu}{\partial \mathbf{w}_\nu} = (0, \ldots, 0, \delta_{\mu\nu}, 0, \ldots, 0)$$

Then we consider first the case of $\mu = \nu$ and get in detail:

$$\begin{aligned}
\frac{\partial F_\mu^{\Re}}{\partial \mathbf{v}_\mu} &= -\mathbf{L} + k_\mu^2 \mathbf{M}^{\varepsilon_c} + k_\mu^2 \Gamma(k_\mu) \Big(\frac{k_\mu - k_a}{\gamma_\perp} \big(\frac{\partial}{\partial \mathbf{v}_\mu} \big[\mathbf{Q}(X) \big] \mathbf{v}_\mu + \mathbf{Q}(X) \big) + \frac{\partial}{\partial \mathbf{v}_\mu} \big[\mathbf{Q}(X) \big] \mathbf{w}_\mu \Big) \\
\frac{\partial F_\mu^{\Re}}{\partial \mathbf{w}_\mu} &= -k_\mu \mathbf{R} + k_\mu^2 \Gamma(k_\mu) \Big(\frac{k_\mu - k_a}{\gamma_\perp} \frac{\partial}{\partial \mathbf{w}_\mu} \big[\mathbf{Q}(X) \big] \mathbf{v}_\mu + \frac{\partial}{\partial \mathbf{w}_\mu} \big[\mathbf{Q}(X) \big] \mathbf{w}_\mu + \mathbf{Q}(X) \Big) \\
\frac{\partial F_\mu^{\Re}}{\partial k_\mu} &= -\mathbf{R}\mathbf{w}_\mu + 2k_\mu \mathbf{M}^{\varepsilon_c} \mathbf{v}_\mu + \Big(\frac{\partial}{\partial k_\mu} \big[k_\mu^2 \frac{k_\mu - k_a}{\gamma_\perp} \Gamma(k_\mu) \mathbf{Q}(X) \mathbf{v}_\mu \big] + \frac{\partial}{\partial k_\mu} \big[k_\mu^2 \Gamma(k_\mu) \mathbf{Q}(X) \mathbf{w}_\mu \big] \Big) \\
&= -\mathbf{R}\mathbf{w}_\mu + 2k_\mu \mathbf{M}^{\varepsilon_c} \mathbf{v}_\mu + \Big(k_\mu^2 \frac{k_\mu - k_a}{\gamma_\perp} \Gamma(k_\mu) \frac{\partial}{\partial k_\mu} \big[\mathbf{Q}(X) \big] \mathbf{v}_\mu + \frac{\partial}{\partial k_\mu} \big[k_\mu^2 \frac{k_\mu - k_a}{\gamma_\perp} \Gamma(k_\mu) \big] \mathbf{Q}(X) \mathbf{v}_\mu \\
&\quad + k_\mu^2 \Gamma(k_\mu) \frac{\partial}{\partial k_\mu} \big[\mathbf{Q}(X) \big] \mathbf{w}_\mu + \frac{\partial}{\partial k_\mu} \big[k_\mu^2 \Gamma(k_\mu) \big] \mathbf{Q}(X) \mathbf{w}_\mu \Big) \\
\frac{\partial F_\mu^{\Re}}{\partial s_\mu} &= k_\mu^2 \frac{k_\mu - k_a}{\gamma_\perp} \Gamma(k_\mu) \frac{\partial}{\partial s_\mu} \big[\mathbf{Q}(X) \mathbf{v}_\mu + k_\mu^2 \Gamma(k_\mu) \frac{\partial}{\partial s_\mu} \mathbf{Q}(X) \mathbf{w}_\mu
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial F_\mu^{\Im}}{\partial \mathbf{v}_\mu} &= k_\mu \mathbf{R} + k_\mu^2 \Gamma(k_\mu) \Big(\frac{k_\mu - k_a}{\gamma_\perp} \frac{\partial}{\partial \mathbf{v}_\mu} \big[\mathbf{Q}(X) \big] \mathbf{w}_\mu - \big(\frac{\partial}{\partial \mathbf{v}_\mu} \big[\mathbf{Q}(X) \big] \mathbf{v}_\mu + \mathbf{Q}(X) \big) \Big) \\
\frac{\partial F_\mu^{\Im}}{\partial \mathbf{w}_\mu} &= -\mathbf{L} + k_\mu^2 \mathbf{M}^{\varepsilon_c} + k_\mu^2 \Gamma(k_\mu) \Big(\frac{k_\mu - k_a}{\gamma_\perp} \big(\frac{\partial}{\partial \mathbf{w}_\mu} \big[\mathbf{Q}(X) \big] \mathbf{w}_\mu + \mathbf{Q}(X) \big) - \frac{\partial}{\partial \mathbf{w}_\mu} \big[\mathbf{Q}(X) \big] \mathbf{v}_\mu \Big) \\
\frac{\partial F_\mu^{\Im}}{\partial k_\mu} &= \mathbf{R}\mathbf{v}_\mu + 2k_\mu \mathbf{M}^{\varepsilon_c} \mathbf{w}_\mu + \Big(\frac{\partial}{\partial k_\mu} \big[k_\mu^2 \frac{k_\mu - k_a}{\gamma_\perp} \Gamma(k_\mu) \mathbf{Q}(X) \mathbf{w}_\mu \big] + \frac{\partial}{\partial k_\mu} \big[k_\mu^2 \Gamma(k_\mu) \mathbf{Q}(X) \mathbf{v}_\mu \big] \Big) \\
&= \mathbf{R}\mathbf{v}_\mu + 2k_\mu \mathbf{M}^{\varepsilon_c} \mathbf{w}_\mu + \Big(k_\mu^2 \frac{k_\mu - k_a}{\gamma_\perp} \Gamma(k_\mu) \frac{\partial}{\partial k_\mu} \big[\mathbf{Q}(X) \big] \mathbf{w}_\mu + \frac{\partial}{\partial k_\mu} \big[k_\mu^2 \frac{k_\mu - k_a}{\gamma_\perp} \Gamma(k_\mu) \big] \mathbf{Q}(X) \mathbf{w}_\mu
\end{aligned}$$

$$+ k_\mu^2\Gamma(k_\mu)\frac{\partial}{\partial k_\mu}\big[\mathbf{Q}(X)\big]\mathbf{v}_\mu + \frac{\partial}{\partial k_\mu}\big[k_\mu^2\Gamma(k_\mu)\big]\mathbf{Q}(X)\mathbf{v}_\mu\Big)$$

$$\frac{\partial F_\mu^{\Im}}{\partial s_\mu} = k_\mu^2\frac{k_\mu - k_a}{\gamma_\perp}\Gamma(k_\mu)\frac{\partial}{\partial k_\mu}\mathbf{Q}(X)\mathbf{w}_\mu + k_\mu^2\Gamma(k_\mu)\frac{\partial}{\partial s_\mu}\mathbf{Q}(X)\mathbf{v}_\mu$$

with

$$\frac{\partial}{\partial k_\mu}\big[k^2\frac{k_\mu - k_a}{\gamma_\perp}\Gamma(k_\mu)\big] = \frac{1}{\gamma_\perp}\Big\{(3k_\mu^2 - 2k_a k_\mu)\Gamma(k_\mu) + (k_\mu^3 - k_a k_\mu^2)\Gamma'(k_\mu)\Big\},$$

$$\frac{\partial}{\partial k_\mu}\big[k_\mu^2\Gamma(k_\mu)\big] = \Big\{2k_\mu\Gamma(k_\mu) + k_\mu^2\Gamma'(k_\mu)\Big\}$$

and

$$\Gamma'(k) = -\gamma_\perp^2\frac{2(k-k_a)}{((k-k_a)^2+\gamma_\perp^2)^2} = -\frac{2(k-k_a)}{(k-k_a)^2+\gamma_\perp^2}\Gamma(k).$$

In the case of $\mu \neq \nu$, the off-diagonal terms in $J(X)$ are not zero, due to the coupling term in 9.3:

$$\frac{\partial F_\mu^{\Re}}{\partial \mathbf{v}_\nu} = k_\mu^2\Gamma(k_\mu)\big(\frac{k_\mu - k_a}{\gamma_\perp}\frac{\partial}{\partial \mathbf{v}_\nu}\big[\mathbf{Q}(X)\big]\mathbf{v}_\mu + \frac{\partial}{\partial \mathbf{v}_\nu}\big[\mathbf{Q}(X)\big]\mathbf{w}_\mu\big)$$

$$\frac{\partial F_\mu^{\Re}}{\partial \mathbf{w}_\nu} = k_\mu^2\Gamma(k_\mu)\big(\frac{k_\mu - k_a}{\gamma_\perp}\frac{\partial}{\partial \mathbf{w}_\nu}\big[\mathbf{Q}(X)\big]\mathbf{v}_\mu + \frac{\partial}{\partial \mathbf{w}_\nu}\big[\mathbf{Q}(X)\big]\mathbf{w}_\mu\big)$$

$$\frac{\partial F_\mu^{\Re}}{\partial k_\nu} = k_\mu^2\Gamma(k_\mu)\big(\frac{k_\mu - k_a}{\gamma_\perp}\frac{\partial}{\partial k_\nu}\big[\mathbf{Q}(X)\big]\mathbf{v}_\mu + \frac{\partial}{\partial k_\nu}\big[\mathbf{Q}(X)\big]\mathbf{w}_\mu\big)$$

$$\frac{\partial F_\mu^{\Re}}{\partial s_\nu} = k_\mu^2\Gamma(k_\mu)\big(\frac{k_\mu - k_a}{\gamma_\perp}\frac{\partial}{\partial s_\nu}\big[\mathbf{Q}(X)\big]\mathbf{v}_\mu + \frac{\partial}{\partial s_\nu}\big[\mathbf{Q}(X)\big]\mathbf{w}_\mu\big)$$

and

$$\frac{\partial F_\mu^{\Im}}{\partial \mathbf{v}_\nu} = k_\mu^2\Gamma(k_\mu)\big(\frac{k_\mu - k_a}{\gamma_\perp}\frac{\partial}{\partial \mathbf{v}_\nu}\big[\mathbf{Q}(X)\big]\mathbf{w}_\mu - \frac{\partial}{\partial \mathbf{v}_\nu}\big[\mathbf{Q}(X)\big]\mathbf{v}_\mu\big)$$

$$\frac{\partial F_\mu^{\Im}}{\partial \mathbf{w}_\nu} = k_\mu^2 \Gamma(k_\mu)\big(\frac{k_\mu - k_a}{\gamma_\perp}\frac{\partial}{\partial \mathbf{w}_\nu}\big[\mathbf{Q}(X)\big]\mathbf{w}_\mu - \frac{\partial}{\partial \mathbf{w}_\nu}\big[\mathbf{Q}(X)\big]\mathbf{v}_\mu\big)$$

$$\frac{\partial F_\mu^{\Im}}{\partial k_\nu} = k_\mu^2 \Gamma(k_\mu)\big(\frac{k_\mu - k_a}{\gamma_\perp}\frac{\partial}{\partial k_\nu}\big[\mathbf{Q}(X)\big]\mathbf{w}_\mu - \frac{\partial}{\partial k_\nu}\big[\mathbf{Q}(X)\big]\mathbf{v}_\mu\big)$$

$$\frac{\partial F_\mu^{\Im}}{\partial s_\nu} = k_\mu^2 \Gamma(k_\mu)\big(\frac{k_\mu - k_a}{\gamma_\perp}\frac{\partial}{\partial s_\nu}\big[\mathbf{Q}(X)\big]\mathbf{w}_\mu - \frac{\partial}{\partial s_\nu}\big[\mathbf{Q}(X)\big]\mathbf{v}_\mu\big)$$

Thus it remains to consider derivatives of the nonlinear coupling term:

$$\mathbf{Q}(X) = \Big(\int_\Omega \frac{D_0(x)}{1+\sum_\nu s_\nu^2\Gamma(k_\nu)\big[\big(\sum_l v_l^\nu \varphi_l(x)\big)^2 + \big(\sum_l w_l^\nu \varphi_l(x)\big)^2\big]}\varphi_i(x)\varphi_j(x)dx\Big)_{i,j}.$$

To this end, the differentiation in the direction of the vector $\vec{v} = (v_1, \ldots, v_M)$ can be understood as $\mathbf{Q}_{\vec{v}}(X) = \big[\mathbf{Q}_{v_1}(X), \ldots, \mathbf{Q}_{v_N}(X)\big]$. Thus for $\vec{v} = \mathbf{v}_\mu, \mathbf{w}_\mu,\ \mu = 1, \ldots, M$ each derivative has the form

$$\begin{aligned}
\mathbf{Q}_{v_m^\mu}(X) &= \frac{\partial}{\partial v_m^\mu}\Big(\int_\Omega \frac{D_0(x)}{1+\sum_\nu s_\nu^2\Gamma(k_\nu)\big[\big(\sum_l v_l^\nu \varphi_l(x)\big)^2 + \big(\sum_l w_l^\nu \varphi_l(x)\big)^2\big]}\varphi_i(x)\varphi_j(x)dx\Big)_{i,j} \\
&= \Big(\int_\Omega \frac{\partial}{\partial v_m^\mu}\Big[\frac{D_0(x)}{1+\sum_\nu s_\nu^2\Gamma(k_\nu)\big[\big(\sum_l v_l^\nu \varphi_l(x)\big)^2 + \big(\sum_l w_l^\nu \varphi_l(x)\big)^2\big]}\Big]\varphi_i(x)\varphi_j(x)dx\Big)_{i,j} \\
&= \Big(\int_\Omega \Big[\frac{-2D_0(x)s_\mu^2\Gamma(k_\mu)\big(\sum_l v_l^\mu \varphi_l(x)\big)\varphi_m(x)}{\Big(1+\sum_\nu s_\nu^2\Gamma(k_\nu)\big[\big(\sum_l v_l^\nu \varphi_l(x)\big)^2 + \big(\sum_l w_l^\nu \varphi_l(x)\big)^2\big]\Big)^2}\Big]\varphi_i(x)\varphi_j(x)dx\Big)_{i,j}
\end{aligned}$$

and

$$\mathbf{Q}_{w_m^\mu}(X) = \Big(\int_\Omega \Big[\frac{-2D_0(x)s_\mu^2\Gamma(k_\mu)\big(\sum_l w_l^\mu \varphi_l(x)\big)\varphi_m(x)}{\Big(1+\sum_\nu s_\nu^2\Gamma(k_\nu)\big[\big(\sum_l v_l^\nu \varphi_l(x)\big)^2 + \big(\sum_l w_l^\nu \varphi_l(x)\big)^2\big]\Big)^2}\Big]\varphi_i(x)\varphi_j(x)dx\Big)_{i,j}$$

Nevertheless we can observe that derivatives of $\mathbf{Q}$ with respect to the vectorial variables $\mathbf{v}_\mu, \mathbf{w}_\mu$ always appear with multiplicative combination of $\mathbf{v}_\nu.\mathbf{w}_\nu$. This leads to some simplifications within the context of assembling. Again we have to differentiate between $\mu = \nu$ and $\mu \neq \nu$. For $\mu = \nu$ we have:

$$
\begin{aligned}
\mathbf{Q}_{\mathbf{v}_\mu}(X)\mathbf{v}_\mu &= \Big(\int_\Omega \frac{-2s_\mu^2\Gamma(k_\mu)\big(\sum_l v_l^\mu\varphi_l(x)\big)^2}{\big(1+\sum_\nu s_\nu^2\Gamma(k_\nu)|\sum_l u_l^\nu\varphi_l(x)|^2\big)^2}D_0(x)\varphi_i(x)\varphi_j(x)dx\Big)_{i,j}\\
\mathbf{Q}_{\mathbf{v}_\mu}(X)\mathbf{w}_\mu &= \Big(\int_\Omega \frac{-2s_\mu^2\Gamma(k_\mu)\big(\sum_l v_l^\mu\varphi_l(x)\big)\big(\sum_l w_l^\mu\varphi_l(x)\big)}{\big(\sum_\nu 1+s_\nu^2\Gamma(k_\nu)|\sum_l u_l^\nu\varphi_l(x)|^2\big)^2}D_0(x)\varphi_i(x)\varphi_j(x)dx\Big)_{i,j}\\
\mathbf{Q}_{\mathbf{w}_\mu}(X)\mathbf{w}_\mu &= \Big(\int_\Omega \frac{-2s_\mu^2\Gamma(k_\mu)\big(\sum_l w_l^\mu\varphi_l(x)\big)^2}{\big(1+\sum_\nu s_\nu^2\Gamma(k_\nu)|\sum_l u_l^\nu\varphi_l(x)|^2\big)^2}D_0(x)\varphi_i(x)\varphi_j(x)dx\Big)_{i,j}\\
\mathbf{Q}_{\mathbf{w}_\mu}(X)\mathbf{v}_\mu &= \mathbf{Q}_{\mathbf{v}_\mu}(X)\mathbf{w}_\mu
\end{aligned}
$$

and for $\mu\neq\nu$ with $v:=\mathbf{v}_\mu,\mathbf{w}_\mu, w:=\mathbf{v}_\nu,\mathbf{w}_\nu$:

$$
\mathbf{Q}_v(X)w = \Big(\int_\Omega \frac{-2s_\mu^2\Gamma(k_\mu)\big(\sum_l v_l\varphi_l(x)\big)\big(\sum_l w_l\varphi_l(x)\big)}{\big(1+\sum_\nu s_\nu^2\Gamma(k_\nu)|\sum_l u_l^\nu\varphi_l(x)|^2\big)^2}D_0(x)\varphi_i(x)\varphi_j(x)dx\Big)_{i,j}
$$

Finally, differentiation with respect to k_μ and s_μ leads to

$$
\begin{aligned}
\mathbf{Q}_{k_\mu}(X) &= \frac{\partial}{\partial k_\mu}\Big(\int_\Omega \frac{D_0(x)}{1+\sum_\nu s_\nu^2\Gamma(k_\nu)|\sum_l u_l^\nu\varphi_l(x)|^2}\varphi_i(x)\varphi_j(x)dx\Big)_{i,j}\\
&= \Big(\int_\Omega \frac{\partial}{\partial k_\mu}\Big[\frac{1}{1+\sum_\nu s_\nu^2\Gamma(k_\nu)|\sum_l u_l^\nu\varphi_l(x)|^2}\Big]D_0(x)\varphi_i(x)\varphi_j(x)dx\Big)_{i,j}\\
&= \Big(\int_\Omega \frac{-s_\mu^2\Gamma'(k_\mu)|\sum_l u_l^\nu\varphi_l(x)|^2}{\big(1+\sum_\nu s_\nu^2\Gamma(k_\nu)|\sum_l u_l^\mu\varphi_l(x)|^2\big)^2}D_0(x)\varphi_i(x)\varphi_j(x)dx\Big)_{i,j}\\
\mathbf{Q}_{s_\mu}(X) &= \frac{\partial}{\partial s_\mu}\Big(\int_\Omega \frac{D_0(x)}{1+\sum_\nu s_\nu^2\Gamma(k_\nu)|\sum_l u_l^\nu\varphi_l(x)|^2}\varphi_i(x)\varphi_j(x)dx\Big)_{i,j}\\
&= \Big(\int_\Omega \frac{\partial}{\partial s_\mu}\Big[\frac{1}{1+\sum_\nu s_\nu^2\Gamma(k_\nu)|\sum_l u_l^\nu\varphi_l(x)|^2}\Big]D_0(x)\varphi_i(x)\varphi_j(x)dx\Big)_{i,j}\\
&= \Big(\int_\Omega \frac{-2s_\mu\Gamma(k_\mu)|\sum_l u_l^\nu\varphi_l(x)|^2}{\big(1+\sum_\nu s_\nu^2\Gamma(k_\nu)|\sum_l u_l^\mu\varphi_l(x)|^2\big)^2}D_0(x)\varphi_i(x)\varphi_j(x)dx\Big)_{i,j}.
\end{aligned}
$$

10. Solving the nonlinear eigenvalue problems

The interest in nonlinear EVP has increased remarkably in recent years. However, compared to linear eigenvalue problems, the robust numerical solution of nonlinear eigenvalue problems is still much more difficult and there are essentially no equivalent packages that reach the standard of those for linear problems. In particular, such eigenvalue problems arising typically from applications in physics and other sciences bear some specific structures which should be reflected in the numerical solution method.

In the particular case of our SALT algorithm, it is necessary, as mentioned in Chapter 8, to solve additionally a nonlinear eigenvalue problem of the form

$$\begin{aligned} \psi''(x) + k^2\big(\varepsilon_c(x) + \gamma(k)\tfrac{D_0(x,d)}{1+\sum_\nu \Gamma(k_\nu^*)|\psi_\nu^*(x)|^2}\big)\psi(x) &= 0 \text{ in } \Omega = [0,R] \\ \psi(0) &= 0 \\ \psi'(R) - ik\psi(R) &= 0 \end{aligned} \tag{10.1}$$

in order to verify a new mode activation. Based on a discretization via finite element methods this leads to the following eigenvalue problem:

$$\mathbf{T}(k)\mathbf{u} := \big[-\mathbf{L} + ik\mathbf{R} + k^2\mathbf{M}^{\varepsilon_c} + \frac{\gamma_\perp k^2}{k-\eta}\mathbf{Q}\big]\mathbf{u} = 0 \tag{10.2}$$

where $\eta := k_a - i\gamma_\perp \in \mathbb{C}$. Note that all matrices here are large sparse matrices; so appropriate solution methods are needed.

10.1. The cubic EVP

A brute-force approach to solve the problem (10.1) resp. (10.2) can be done by multiplication with the denominator $k-\eta$. This leads to a polynomial EVP of order 3:

$$\big[k^3\mathbf{C}_3 + k^2\mathbf{C}_2 + k\mathbf{C}_1 + \mathbf{C}_0\big]\mathbf{u} = 0$$

where

$$\mathbf{C}_3 := \mathbf{M}^{\varepsilon_c}, \quad \mathbf{C}_2 := \gamma_\perp\mathbf{Q} - \eta\mathbf{M}^{\varepsilon_c} + i\mathbf{R}, \mathbf{C}_1 := -\mathbf{L} + i\eta\mathbf{R} \text{ and } \mathbf{C}_0 := \eta\mathbf{L}.$$

The standard solution method for polynomial EVPs is to apply linearization techniques which transform the problem into an equivalent first-order equation of the generalized form $Au - \lambda Bu = 0$. In particular, the reduction of the degree of the EVP with

$N \times N$ matrices is coped with by extending to matrices of the GEP to the size $pN \times pN$ where here $p = 3$. A commonly used linearization is the so-called first companion form which is obtained by introducing the new vector $v = (k^2u, ku, u)$ such that

$$\left(\begin{bmatrix} \mathbf{C_2} & \mathbf{C_1} & \mathbf{C_0} \\ -Id & 0 & 0 \\ 0 & -Id & 0 \end{bmatrix} + k \begin{bmatrix} \mathbf{C_3} & 0 & 0 \\ 0 & Id & 0 \\ 0 & 0 & Id \end{bmatrix} \right) v = 0.$$

The standard way of solving such a generalized EVP is via the QZ algorithm which is based on the computation of the generalized Schur decomposition [74]. In the case of the sparse large-scale GEP iterative projection methods onto Krylov subspaces such as the Lanczos (for symmetric matrices) and Arnoldi method are the most conventional algorithms to use. Both compute the smallest eigenvalues and can also be extended with the Shift-and-Invert scheme in order to compute eigenvalues close to a specific shift parameter.

Note that this kind of linearization is obviously not the only way to linearize the polynomial EVP; there exist other strategies which might be able to maintain some of the original structure such as symmetry for instance [48].

Despite the fact that this approach is very straightforward to implement, it is clearly not recommendable for sparse large scale problems as in our case. Furthermore, as the original EVP was rational in k we did not only have to enlarge the degree, but the problem has also been changed by multiplication with the denominator $k - \eta$ generating spurious eigensolutions at the pole $\eta = k_a - i\gamma_\perp$ which again becomes noticeable in the numerical performance. Thus it would be better to look for other solution methods which exploit the rational structure and handle sparse large scale problems in a more efficient way.

10.2. The rational EVP

Another possibility proposed by Bai and Su [94] which exploits more the underlying structure of rational EVPs is to reformulate the original EVP (10.1) as a proper rational EVP

$$\left[k^2\mathbf{R}_2 + k\mathbf{R}_1 + \mathbf{R}_0 + \frac{1}{k-\eta}\tilde{\mathbf{R}}\right] u = 0$$

with

$$\mathbf{R}_2 := \mathbf{M}^{\varepsilon_\mathbf{c}}, \mathbf{R}_1 := \gamma_\perp \mathbf{Q} + i\mathbf{R}, \mathbf{R}_0 := -\mathbf{L} + \gamma_\perp \eta \mathbf{Q} \text{ and } \tilde{\mathbf{R}} := \gamma_\perp \eta^2 \mathbf{Q}.$$

This approach is developed for problems with nonsingular leading matrix, here $\mathbf{R}_2$, and a low-rank matrix $\tilde{\mathbf{R}}$. In this case, we assume that there exists a factorization $\tilde{\mathbf{R}} = \mathbf{V}\mathbf{W}^\mathbf{T}$ where $\mathbf{V}, \mathbf{W} \in \mathbb{C}^{N \times r}$ and $r < N$ the rank of $\tilde{\mathbf{R}}$. Then again the problem can be converted into a generalized EVP by the linearization $v = (ku, u, y)$ and $y = (k - \eta)^{-1}\mathbf{W}^\mathbf{T}u$ such that

$$\begin{bmatrix} \mathbf{R_1} & \mathbf{R_0} & \mathbf{V} \\ Id & 0 & 0 \\ 0 & \mathbf{W^T} & \eta Id \end{bmatrix} v = k \begin{bmatrix} -\mathbf{R_2} & 0 & 0 \\ 0 & Id & 0 \\ 0 & 0 & Id \end{bmatrix} v$$

Again this linearization scheme is not the only possibility and, similar to linearizations for polynomial EVPs, different methods preserving inter alia symmetry can be applied.

This method brings an improvement in the problem size to $(2N+r)\times(2N+r)$ instead of $3N \times 3N$ for the standard linearization technique discussed in the previous section. However, applying the same generic solution algorithms, this still generates spurious numerical solutions for $k = \eta$.

Besides, reconsider that the original need to compute the NEVP was to check activation of new laser modes which have real-valued frequencies. Thus, to start with, we are only interested in eigenvalues near the real axis. Furthermore, it appears from the underlying physics that amplification is experienced mostly for modes with frequencies near the peak of the gain curve Γ. Thus we can further confine our domain of interest to a vicinity of k_a. However, we can also observe that the pole η on which the EVP is not well defined is (always) located near this region of interest as $\eta = k_a - i\gamma_\perp$. In order to avoid this singularity, our domain of interest has to be bounded away from $\{z \in \mathbb{C} | \Im(z) > -\gamma_\perp\}$. Hence we end up with a quite narrow domain in the complex plane that is centered at k_a and contains the interval $[k_a - \gamma_\perp, k_a + \gamma_\perp]$.

A promising method for this purpose is the contour integral method, which was presented recently in [16] and [7] and which we will discuss in more detail in the following Section.

10.3. Contour integral eigensolver

As motivated in the previous section our aim is to find eigenvalues of (10.1) or (10.2) in a bounded domain $\Lambda \subset \mathbb{C}$ with a smooth boundary defined by the contour $\mathcal{C} := \partial\Lambda$.

To this end let us assume that the eigenvalues of interest are all simple. Furthermore, we can easily see that the operator-valued function $T : \mathbb{C} \to \mathbb{C}^{N\times N}$ is meromorphic with a pole η. Hence, by choosing $\Lambda \subset \mathbb{C}$ such that $\eta \notin \Lambda$ the following theorem can be applied:

Theorem 10.3.1 ([73, Thm. 1.5.4]). *Let $T : \Lambda \to \mathbb{C}^{N\times N}$ be holomorphic on Λ with simple eigenvalues $\lambda_1, \ldots, \lambda_n$ such that $T(\lambda_n)v_n = 0$ and $T^H(\lambda_n)w_n = 0$. Then T^{-1} is meromorphic on $\Lambda\backslash\{\lambda_1, \ldots, \lambda_n\}$ and there is a neighborhood $\mathcal{U}$ of $\mathcal{C}$ in Λ such that*

$$T^{-1}(z) = \sum \frac{1}{z-\lambda_n} v_n w_n^H + R(z), \ \forall z \in \mathcal{U}\backslash\{\lambda_1, \ldots, \lambda_n\}, \text{ and } R(z) \text{ holomorph at } \Lambda.$$

By use of the residue theorem we can additionally conclude

$$\frac{1}{2\pi i}\int_{\mathcal{C}} f(z)T^{-1}(z)dz = \sum_n f(\lambda_n)v_n w_n^H$$

where $\mathcal{C}$ is the closed contour of Λ. This motivates to compute the contour integrals

$$\mathbf{A}_0 := \frac{1}{2\pi i}\oint_{\mathcal{C}} \mathbf{T}^{-1}(k)dk = \sum_n \mathbf{v}_n \mathbf{w}_n^H = \mathbf{V}\mathbf{W}^H$$

$$\mathbf{A}_1 := \frac{1}{2\pi i} \oint_{\mathcal{C}} k\mathbf{T}^{-1}(k)dk = \sum_n k_n \mathbf{v}_n \mathbf{w}_n^H = \mathbf{V}\mathbf{K}\mathbf{W}^H,$$

where $\mathbf{K}$ is a diagonal matrix with the diagonal entries corresponding to all the poles of the inverse matrix $\mathbf{T}^{-1}$ inside the contour $\mathcal{C}$ which in turn are the eigenvalues of $\mathbf{T}$.

Before we explain how the sought matrix $\mathbf{K}$ is computed from theses two matrices $\mathbf{A}_0$ and $\mathbf{A}_1$, we discuss the realization of the contour integration which is obtained by numerical quadrature. Very fast (i.e., exponential) convergence is achieved with the trapezoidal rule, [61, Thm. 9.28], if the contour is an analytic curve such as a circle or an ellipse. Moreover, for a discretization of the contour with q quadrature points this would require q times the inversion of the operator matrix $T(k) \in \mathbb{C}^{N\times N}$. Of course this would become numerically extremely expensive for our large-scale FEM-matrices and may even be unfeasible given that the inverses become fully populated. This can be remedied by an approximation scheme that exploits the fact that the rank of the matrices $\mathbf{A}_0$ and $\mathbf{A}_1$ is given by the number of eigenvalues inside the contour and is thus very small compared to N. Thus we merely multiply $T(z)^{-1}$ by a random matrix $\mathbf{M} \in \mathbb{C}^{N\times l}$ where $l \ll N$, but l not smaller than the expected number of eigenvalues that are inside the contour. This means it is sufficient to evaluate $\mathbf{T}^{-1}\mathbf{M}$ at each quadrature point on the contour which reduces the computational cost to the solution of l linear systems for each quadrature point.

To obtain the matrix $\mathbf{K}$, we first compute the (reduced) singular value decomposition (SVD) of

$$\mathbf{A}_0\mathbf{M} = \mathbf{V}_0\mathbf{\Sigma}_0\mathbf{W}_0^H$$

where we assume that the dimension l is exactly the number of eigenvalues inside the contour; if l is larger, the SVD of $\mathbf{A_0}\mathbf{M}$ has to be replaced with a rank-revealing variant. Then we can deduce by formally introducing $\mathbf{S} := \mathbf{V}_0^H\mathbf{V}$ and substituting $\mathbf{V} = \mathbf{V}_0\mathbf{S}$ into

$$\mathbf{V}_0\mathbf{\Sigma}_0\mathbf{W}_0^H = \mathbf{A}_0\mathbf{M} = \mathbf{V}\mathbf{W}^H\mathbf{M} = \mathbf{V}_0\mathbf{S}\mathbf{W}^H\mathbf{M}$$

which again leads to the expression

$$\mathbf{W}^H\mathbf{M} = \mathbf{S}^{-1}\mathbf{\Sigma}_0\mathbf{W}_0^H$$

such that we obtain

$$\mathbf{A}_1\mathbf{M} = \mathbf{V}\mathbf{K}\mathbf{W}^H\mathbf{M} = \mathbf{V}_0\mathbf{S}\mathbf{K}\mathbf{S}^{-1}\mathbf{\Sigma}_0\mathbf{W}_0^H.$$

Thus by

$$\mathbf{S}\mathbf{K}\mathbf{S}^{-1} = \mathbf{V}_0^H\mathbf{A}_1\mathbf{M}\mathbf{W}_0\mathbf{\Sigma}_0^{-1}$$

we can see that $\mathbf{K}$ is similar to

$$\mathbf{B} := \mathbf{V}_0^H\mathbf{A}_1\mathbf{M}\mathbf{W}_0\mathbf{\Sigma}_0^{-1}$$

and therefore their eigenvalues are the same, so that the desired eigenvalues k_n can be obtained from the reduced eigenvalue problem

$$\mathbf{B}\mathbf{x}_n = k_n\mathbf{x}_n.$$

The original eigenvectors can then be obtained by $\mathbf{v}_n = \mathbf{V}_0 \mathbf{x}_n$. Note that the size of the matrix $\mathbf{B}$ is just as big/small as the rank of $\mathbf{\Sigma}_0$ which again equals the number of all eigenvalues inside the contour. Thus the EVP is reduced to an essentially smaller one (dense, $l \times l$) while the computational effort is mainly shifted to the solution $q \times l$ sparse linear systems, where q is the number of quadrature points for the contour integration, which are perfectly parallelizable.

So far, we mentioned that an integration via trapezoidal rule leads to exponential convergence in the case of ellipsoidal contours. For more sophisticated contours which are no longer analytic but only piece-wise analytic the trapezoidal rule is still applicable, but loses the feature of exponential convergence. However, this can be overcome by other exponentially convergent schemes such as Gaussian or Clenshaw-Curtis quadrature.

For a better understanding in terms of implementation, we repeat here the algorithmic scheme that was published in [16]:

Algorithm 10.1 Contour integral solver

Input: SALT operator $\mathbf{T}$, number of quadrature points q, projection parameter l

Choose random matrix $\mathbf{M} \in \mathbb{C}^{N \times l}$

Determine the quadrature points z_j and the derivatives of the parametrization at the quadrature points z_j' for $j = 1, \ldots, q$

Assemble contour integrals via trapezoidal rule:

$\mathbf{A}_0 = \frac{1}{q} \sum \left(\mathbf{T}(z_j) \backslash \mathbf{M}\right) z_j', \quad \mathbf{A}_1 = \frac{1}{q} \sum \left(\mathbf{T}(z_j) \backslash \mathbf{M}\right) z_j z_j'$

Compute SVD of $\mathbf{A}_0 = \mathbf{V} \mathbf{\Sigma} \mathbf{W}^H$, where

$\mathbf{V} \in \mathbb{C}^{N \times l}, \mathbf{W} \in \mathbb{C}^{l \times l}, \mathbf{V}^H \mathbf{V} = \mathbf{W}^H \mathbf{W} = I_l, \mathbf{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_l)$.

Reduce matrices, if necessary: For tol_{rank} small find $0 < m \leq l$ such that

$\sigma_1 \geq \ldots \geq \sigma_m > \mathrm{tol}_{rank} > \sigma_{m+1} \approx \ldots \approx \sigma_l \approx 0$

if $l = m$ **then**

Increase l and restart computations

else

$\mathbf{V} = \mathbf{V}(:, 1:m), \mathbf{W} = \mathbf{W}(:, 1:m)$ and $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_k)$

end if

Compute $\mathbf{B} := \mathbf{V}^H \mathbf{A}_1 \mathbf{M} \mathbf{W} \mathbf{\Sigma}^{-1} \in \mathbb{C}^{m \times m}$

Solve EVP $\mathbf{B}\mathbf{x} = k\mathbf{x}$

Transform $\mathbf{v}_n = \mathbf{V}\mathbf{x}_n$

Output: $(\mathbf{v}_n, k_n), n = 1, \ldots, l$

---

11. Numerical results

Finally we want to collect all different kinds of numerical computations that have been done in order to investigate the reliability of our new direct SALT solver. Then we will also give a few examples that apply to specific models in laser theory.

11.1. Qualitative assessment of the solution method

The main question is whether the convergence rates of the FEM approach for this specific case of a nonlinear system remain the same as stated in Part I for the linear Helmholtz equation. As one of the crucial components of our solver is the contour integral method, we will first examine its reliability.

11.1.1. Examination of the contour integral method

Solving the nonlinear eigenvalue problem is numerically the most expensive part; therefore we start with examining the numerical behavior of the contour integral method. For this purpose, we consider the example of a 1D-slab cavity $\Omega = [0, 1]$ with Dirichlet boundary on the left and Robin boundary condition at the right side. The ab initio parameters of the cavity material are as follows: the index of refraction is constant with $n \equiv 3$, while the gain curve is centered at $k_a = 5$ with a half width at half maximum of $\gamma_\perp = 0.4$ and we search for eigenvalues of the linearized SALT-operators with the pump strength $D_0 = 1.5$. In compliance with the parameters $n, \gamma_\perp, k_a$, the SALT-matrix

$$\mathbf{T}(k) = -\mathbf{L} + ik\mathbf{R} - k^2\mathbf{M}^\varepsilon + k^2\gamma(k)\mathbf{M}^D \tag{11.1}$$

is set up and the EVP is solved under variations of the contour as explained in the Section 10.3 on the contour integral method

Degree of freedom vs. quadrature points

In a first simple test case, we consider here the residual error by varying the quadrature points (qpts) regarding an elliptical contour for the SALT-matrix with different numbers of degrees of freedom (ndof). The computation of the contour integral is done using the trapezoidal rule as suggested in [16]. In view of the FEM analysis in Section 11.1.3, we have also applied the Gauss quadrature rule which will be necessary when using more complex contours, in particular composite contours of piece-wise analytic arcs.

Example 11.1.1. We consider a laser setting as described above and search for eigenvalues inside the contour

$$\mathcal{C} := \{(k_a + \alpha\cos(t), \beta\sin(t)), t \in [0, 2\pi]\}$$

where $\alpha = 3\gamma_\perp, \beta = 0.5\gamma_\perp$. We assemble the SALT-matrix (11.1) with the FEM-parameters $p = 3, h = 0.266 \cdot 2^{-l}, l \in \{3, 6, 9\}$ leading to $ndof = \{12, 117, 957\}$ and test the algorithm for $qpts = \{10, 20, 50, 100, 125, 150, 300\}$ quadrature points. The geometric parameters are chosen such that there is only one eigenvalue inside the contour. The residuum is depicted in the L^2 norm.

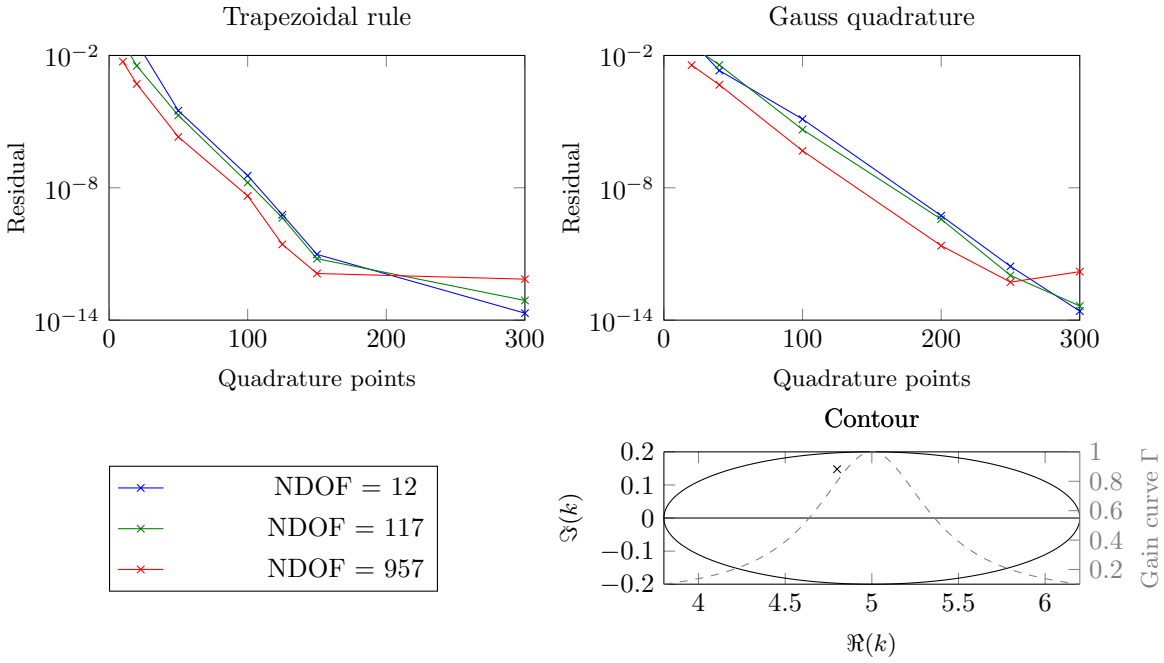


Figure 11.1.: (cf. Example 11.1.1) Top: Residual error vs. number of quadrature points using the trapezoidal rule (left) and the Gaussian quadrature rule (right). Different colors correspond to different numbers of degrees of freedom (NDOF), see legend at the bottom left. Bottom, right: Complex frequency domain bounded by the closed contour $\mathcal{C}$ (solid black line), the eigenvalue inside the contour is marked with $\times$ and the Lorentzian curve is plotted to illustrate the gain inside the domain.

We see that in the case of a periodic parametrized contour (as an ellipse) the trapezoidal rule leads to a better convergence rate as the Gauss quadrature rule and the quality of the solutions are barely influenced by the number of degrees of freedom.

Position vs. quadrature points

From the calculations of the preceding example, we see that there is only one eigenvalue at $k^* = 4.77 + 0.174i$ inside the contour. Based on the error analysis as discussed in [16],

we now introduce here a test series of calculations in which we vary the distance of the eigenvalue to the contour. We center the elliptical contour in the point k^* and investigate the convergence behavior when the ellipse is shifted horizontally or vertically until the eigenvalue is very close to the contour.

Example 11.1.2. The laser setting remains the same as in Example 11.1.1 while we search for eigenvalues inside the contour which is shifted along the imaginary direction, centered close to the present eigenvalue in the beginning and then moved upward until the eigenvalue lies close to the contour:

$$\mathcal{C}_n := \{(a_0 + \alpha\cos(t), b_n + \beta\sin(t), t \in [0, 2\pi]\}$$

where $\beta = \gamma_\perp/2, \alpha = \gamma_\perp, a_0 = 4.799$ and $b_n = 0.147 + \frac{\beta n}{6}$, $n = 1, 2, 3, 4, 5, 6$, again tested for $q = \{10, 20, 50, 100\}$ quadrature points.

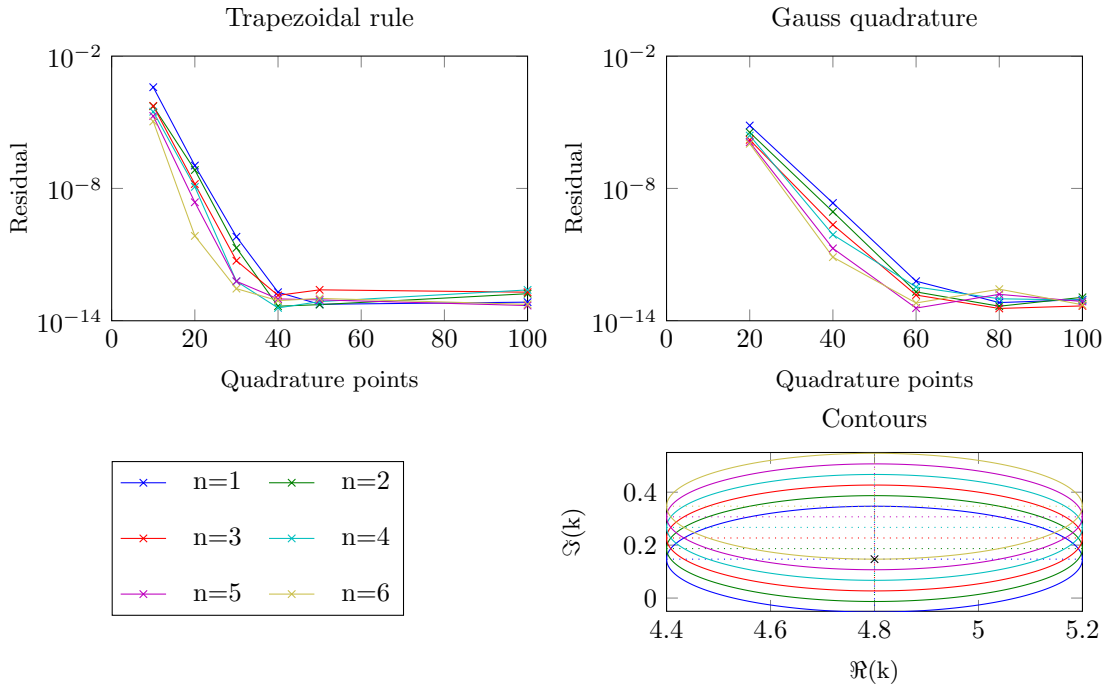


Figure 11.2.: (cf. Example 11.1.2) Top: Residual error vs. number of quadrature points using the trapezoidal rule (left) and the Gaussian quadrature rule (right). Different colors correspond to different contours $\mathcal{C}_n$, see legend at the bottom left. Bottom, right: Complex frequency domain bounded by the closed contours $\mathcal{C}_n$ and the eigenvalue inside the contour is marked with $\times$.

Example 11.1.3. Again we consider the same model as in Example 11.1.2, but search for eigenvalues inside the contour which shifted along the real direction, again centered

close to the present eigenvalue in the beginning and then moved sideward until the eigenvalue also lies close to the contour:

$$\mathcal{C}_n := \{(a_n + \alpha\cos(t), b_0 + \beta\sin(t), t \in [0, 2\pi]\}$$

where $\beta = \gamma_\perp/2, \alpha = \gamma_\perp, b_0 = 0.174$ and $a_n = 4.799 + \frac{\alpha n}{6}, n = 1, 2, 3, 4, 5, 6$, again tested for $q = \{10, 20, 50, 100, 125, 150, 300\}$ quadrature points.

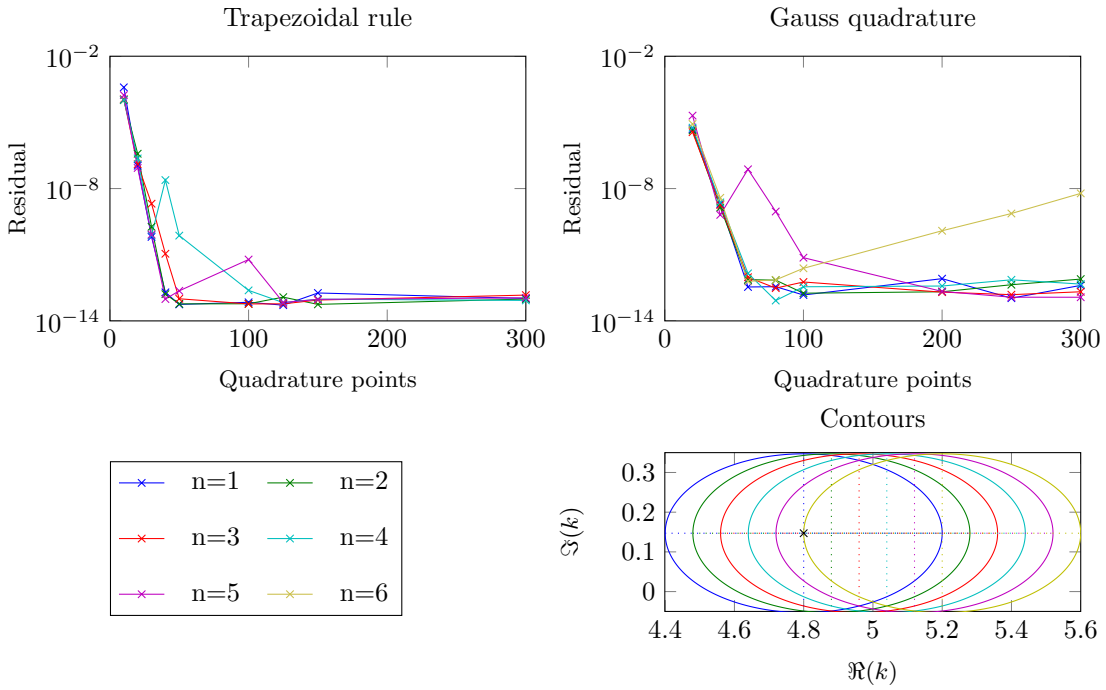


Figure 11.3.: (cf. Example 11.1.3) Top: Residual error vs. number of quadrature points using the trapezoidal rule (left) and the Gaussian quadrature rule (right). Different colors correspond to different contours $\mathcal{C}_n$, see legend at the bottom left. Bottom, right: Complex frequency domain bounded by the closed contours $\mathcal{C}_n$ and the eigenvalue inside the contour is marked with $\times$.

We see again that the trapezoidal rule provides a better convergence behavior than Gauss quadrature rule. If the eigenvalues come close the contour (especially on the side where the quadrature points are plied/accumulated), we see that the exponential behavior is disturbed.

Contour vs. quadrature points

As a final test series we focus on more complex contours and investigate the convergence behavior using the Gaussian quadrature rule. This is also of interest as we have to apply the CIM on such degenerate ellipses for the FEM convergence analysis.

The contours have to be bounded from below (see Section 10.3). However by varying the a priori parameters, for instance a higher pump strength D, the initial nonphysical eigenvalues of the corresponding SALT-matrix may lie further up in the imaginary half plane such that the contour has to be extended on that side. The desired domain is thus bounded by the composition of a flat lower elliptic arc and a much more extended elliptic arc on the upper side.

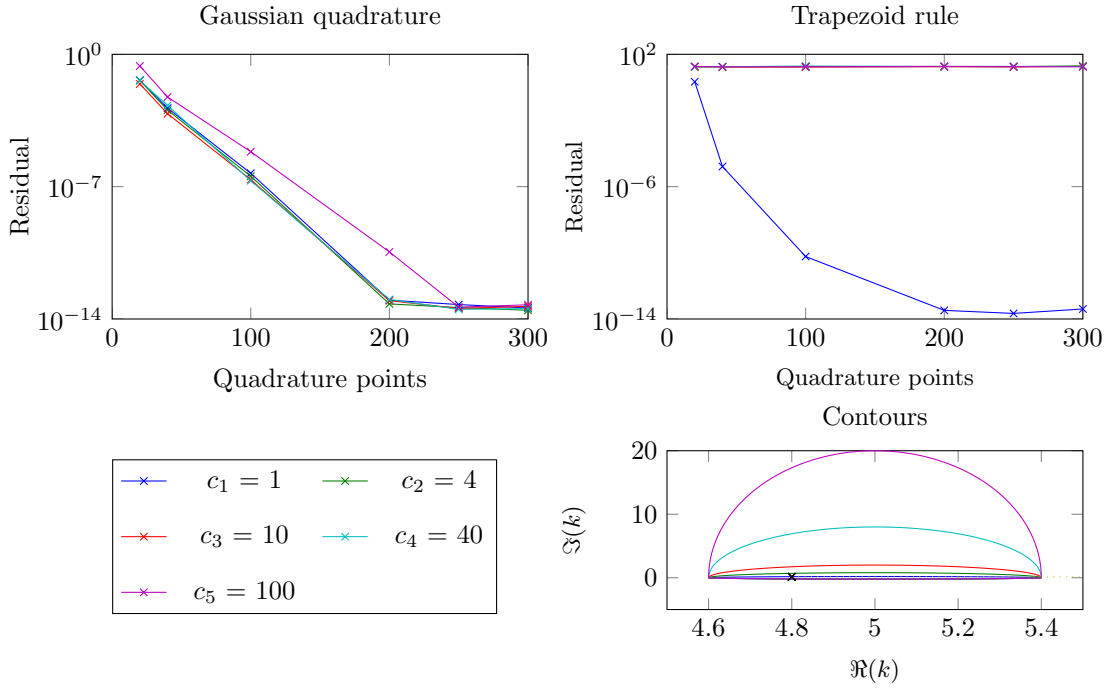


Figure 11.4.: (cf. Example 11.1.4) Top: Residual error vs. number of quadrature points using the trapezoidal rule (left) and the Gaussian quadrature rule (right). Different colors correspond to different contours $\mathcal{C}_0 \cup \mathcal{C}_n$, see legend at the bottom left. Bottom, right: Complex frequency domain bounded by the closed contours $\mathcal{C}_0 \cup \mathcal{C}_n$ and the eigenvalue inside the contour is marked with $\times$.

Example 11.1.4. Considering the same model as in Example 11.1.1, we search for eigenvalues inside the contour which is split into a lower and upper arc $\mathcal{C} = \mathcal{C}_0 \cup \mathcal{C}_n$ with

$$\mathcal{C}_0 := \{(k_a + \alpha\cos(t), \beta\sin(t)), t \in [0, \pi]\}$$

and

$$\mathcal{C}_n := \{(k_a + \alpha\cos(t), c_n\beta\sin(t)), t \in [\pi, 2\pi]\}$$

where $\beta = 0.5\gamma_\perp, \alpha = 2\gamma_\perp$ and $c_n \in \{1, 4, 10, 40, 100\}$, again tested for $q = \{10, 20, 50, 100, 125, 150\}$ quadrature point on each arc of the composed contours.

We see in this case that the Gaussian quadrature preserves the exponential convergence rate. In contrast, the trapezoidal rule completely fails for contours with such a composition of non-periodic closed parametrizations.

11.1.2. Newton solver

In addition we have observed numerically that the stabilized Newton solver always converges rapidly while computing the SALT system along a given pump trajectory. One of the quantities that is of interest is the emitting laser intensity of each mode. The development of the intensities with increasing pump strength and the convergence behavior of the Newton method are summarized in the next example.

Example 11.1.5. We consider a 1D-slab cavity $\Omega = [0, 1]$ with Dirichlet boundary on the left and Robin boundary condition at the right side. The material properties are described by a uniform index of refraction of $n = 2$ and a gain curve with a half width at half maximum of $\gamma_\perp = 3$ and a center at $k_a = 10$. The FEM-parameters are $p = 6, h = 2^{-4}, Ndof = 96$. Then we computed the SALT solutions along a spatially uniform pump trajectory $D_0(x, d) = d, d \in [0, 2]$. Figure **??** shows the 'evolution' of the intensities of the laser modes with respect to the increasing pump strength. The first mode (red line) activates at $d \sim 0.24$, the second mode (green line) at $d \sim 0.38$ and the third more (blue line) at $d \sim 1.32$.

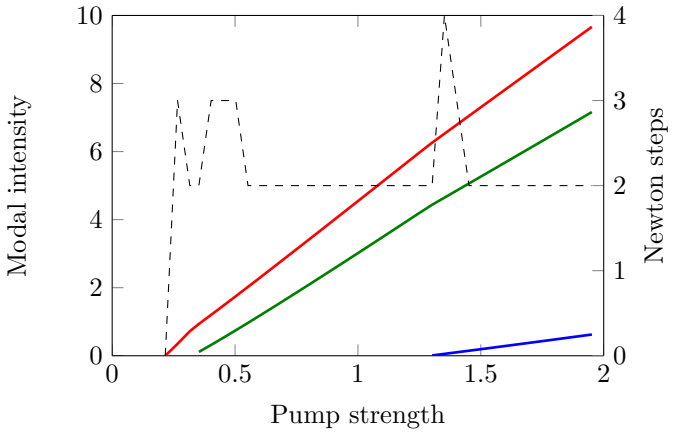


Figure 11.5.: (cf. Example 11.1.5) In the figure above, we have plotted the modal intensity (left y-axis) and the number of Newton steps (right y-axis). The dual y-axis shows the agreement of a slight increase in the Newton steps with the activation of new lasing modes.

This solution strategy provides a fast convergence for the Newton scheme with less that five iterations per pump step. It is only slightly more expensive when a new laser mode activates.

11.1.3. FEM computations

In analogy to the numerical analysis in Chapter 5, we will now study the behavior of the FEM solutions for the nonlinear Helmholtz problem. The question is whether the convergence rates transfer from the linear case to the SALT model. However, this is highly plausible, since the leading order differential operator is linear and the nonlinearity appears the the lower order term and is fairly mild. We want to investigate the convergence behavior for h- and p-refinements. To this end, we consider again the model as introduced in Section 11.1.1 with the a priori estimates $n = 3, \gamma_\perp = 0.5$, but vary the parameter k_a in order to illustrate the rate of change with respect to the wave number k_μ.

For all convergence diagrams in this section the asymptotic convergence rate, that we expect from the analytic error analysis, is indicated by a black dashed line.

h-FEM

Recall that for the h-version, the polynomial degree is kept constant and fixed while the mesh width is decreased. To show the lower rate of convergence for the low-order FEM, we perform the calculations for $p = 1, 2$. The error is generated on the respective grid by the L^2-difference with the solution for $p = 6$.

Example 11.1.6. We consider a 1D-slab cavity $\Omega = [0, 1]$ with Dirichlet boundary on the left and Robin boundary condition at the right side. The pump strength is $D = 1.5$ with a constant index of refraction $n = 3$ and $\gamma_\perp = 0.5$. Further we calculate the L^2-error for $k_a \in \{5, 20, 100\}$. The mesh width of the initial grid is chosen such that $hk_an = 1.5$ and the refinement is done by bi-sectioning. We observe that the higher order method is less prone to pollution, similar to the linear cases discussed in the Sections 5.4 and 6.4.

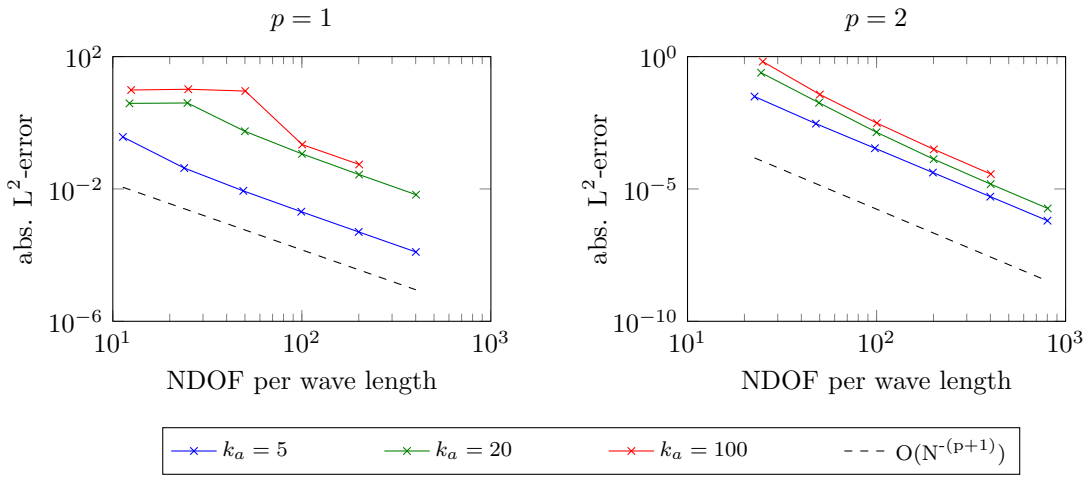


Figure 11.6.: (cf. Example 11.1.6) h-FEM convergence plot with fixed polynomial degree $p = 1$ (left) and $p = 2$ (right)

p-FEM

For the p-version the polynomial degree is increased on a fixed mesh. The error here is generated by computing the L^2-difference with the solution for $p = 12$.

Example 11.1.7. We consider a 1D-slab cavity $\Omega = [0, 1]$ with Dirichlet boundary on the left and Robin boundary condition at the right side. The pump strength is $D = 1.5$ with a constant index of refraction $n = 3$ and $\gamma_\perp = 0.5$. Again we calculate the L^2-error for $k_a \in \{5, 20, 100\}$ and the mesh width of the grid is chosen such that $hk_a n = 1.5$. We can see that we obtain a good convergence of the p-FEM as well.

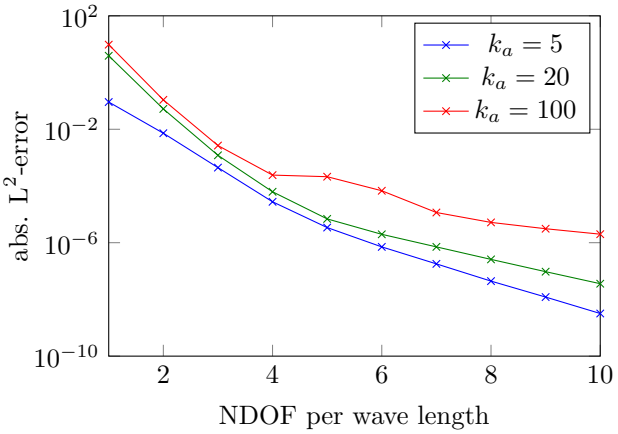


Figure 11.7.: (cf. Example 11.1.7) p-FEM on a fixed uniform mesh

Examination of the nonlinearity

Next we want to focus on the impact of the nonlinear term. For this, we consider the quality of the FEM approximation for several different pump strengths on two distinct laser configurations. The first has a higher index of refraction causing the spatial hole burning effect in the nonlinear term to be less infecting while in the second laser setup the index of refraction is quite close to 1 and thus the nonlinear interaction is more pronounced. The SALT problem is solved for $D(\mathbf{x}, d) \equiv d$ with $d \in \{1, 2, 4, 6\}$.

Example 11.1.8. We consider a 1D-slab cavity $\Omega = [0, 1]$ with Dirichlet boundary on the left and Robin boundary condition at the right side, with a constant index of refraction $n = 3$ and a gain curve centered at $k_a = 8.42$ and a half width at half maximum of $\gamma_\perp = 2.35$.

Example 11.1.9. We consider a 1D-slab cavity $\Omega = [0, 1]$ with Dirichlet boundary on the left and Robin boundary condition at the right side, with a constant index of refraction $n = 1.01$ and a gain curve centered at $k_a = 25$ and a half width at half maximum of $\gamma_\perp = 7.5$.

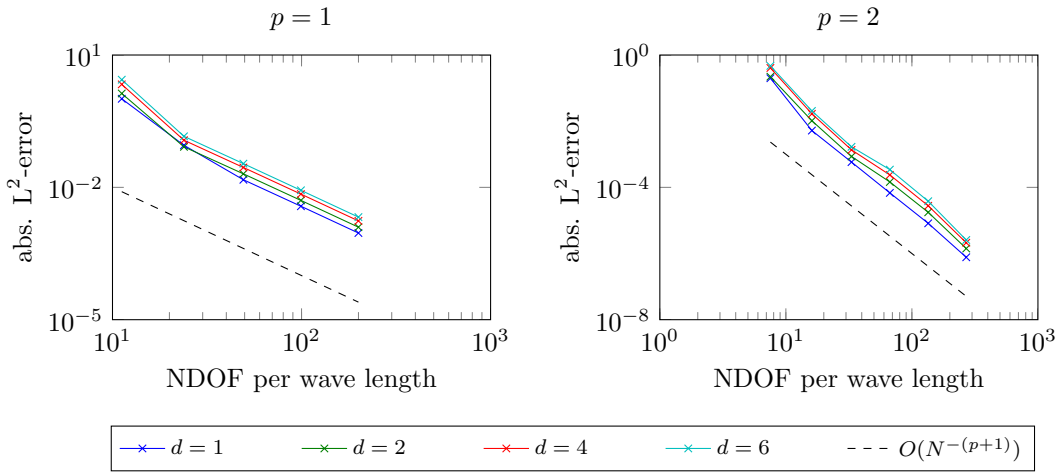


Figure 11.8.: (cf. Example 11.1.8) h-FEM convergence plot with fixed polynomial degree $p = 1$ (left) and $p = 2$ (right) for different pump strengths d

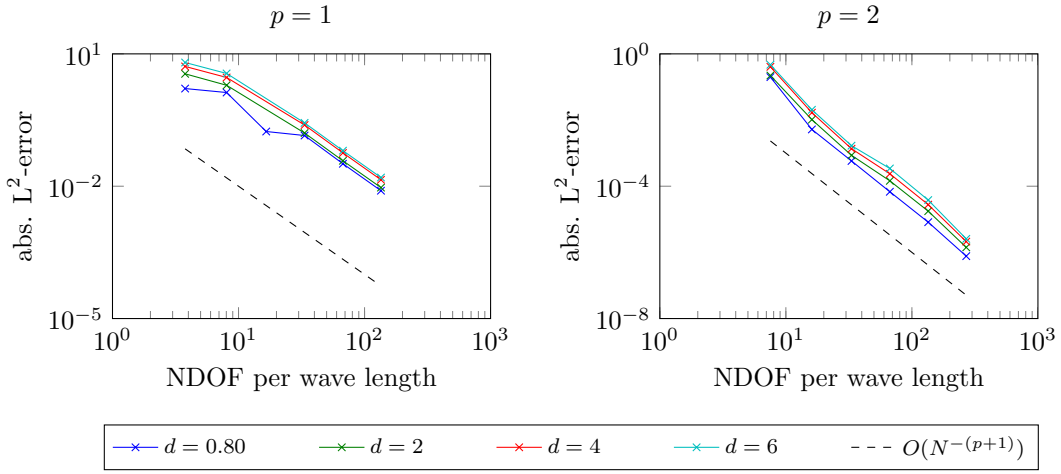


Figure 11.9.: (cf. Example 11.1.9) h-FEM convergence plot with fixed polynomial degree $p = 1$ (left) and $p = 2$ (right) for different pump strengths d

In both configurations we see that the method converges at the expected rate. However, a slight dependence with respect to the pumping strength can be observed. This holds in particular for the second case where the nonlinear term is dominant.

11.2. Physical applications

At last we want to sum up a few cases of applications that have been developed in joint work with Matthias Liertzer from the group of Prof. Stefan Rotter at the Institute for theoretical physics. With these examples we intend to highlight the flexibility and the

vast range of applications that can be handled with our new solution method.

11.2.1. Comparison with the CF-states method

We demonstrate here the accuracy of the presented solution method. Compared to the integral method as discussed in Section 7.1.2 one of the advantages of our new solution method is the accuracy of solutions far above the threshold. In this regime the accuracy of the integral method may depend strongly on the size of the TCF state basis N_{TCF}.

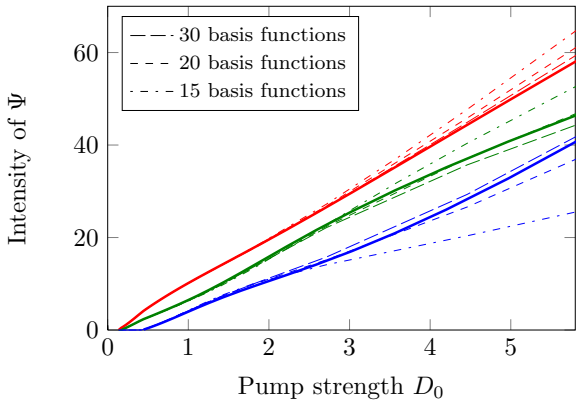


Figure 11.10.: (cf. Example 11.2.1) Comparison of our new solution method with the CF method for different number of basis functions

Example 11.2.1. We consider a 1D-slab cavity $\Omega = [0, 1]$ with Dirichlet boundary on the left and Robin boundary condition at the right side, with a constant index of refraction $n = 1.2$ and a gain curve centered at $k_a = 10$ and a half width at half maximum of $\gamma_\perp = 4$. Then we computed the SALT solutions along a spatially uniform pump trajectory $D_0(x, d) = d, d \in [0, 6]$. Figure 11.10 shows the comparison of the direct SALT solver (solid lines) to the solution of the integral formalism with 30 (long dashed), 20 (dashed) and 15 (dash-dotted) CF-basis functions.

One can see that for a larger basis the solution converges towards the solution of the direct solver. Thus our new solution method leads to accurate solutions in the regime far above threshold which can only be achieved by the CF method with a much larger number of TCF states.

11.2.2. Two cavity laser

As one prototypical example for a pump profile where the spatial profile of the pump varies as a function of the pump parameter d we consider a setup as it has been used in [65]. There, the authors have shown that a spatially varying pump function with a spatial shape that depends on a pump parameter d, can strongly influence the laser output in a counter-intuitive way.

Example 11.2.2. We consider a laser system consisting of two coupled one-dimensional ridge cavities $\Omega_1 = [0, 1]$ and $\Omega_2 = [1.1, 2.1]$. The index of refraction of the cavity material in both cavities is $n \equiv 3 + 0.13i$, while in between it holds $n \equiv 1$. The gain parameters are $k_a = 9.46, \gamma_\perp = 0.4$. For the pump function the trajectory is shown in Fig. 11.11(a). In more detail it is defined as follows: For values of the pump parameter between 0 and 1 only the left cavity of the system is pumped from 0 to a certain pump strength $d_{max} = 1.2$, which brings the laser just above threshold. In the range of $d = [1, 2]$ the pump in the left cavity is kept constant while the pump in the right cavity is linearly increased from 0 to the same pump strength as already present in the left cavity.

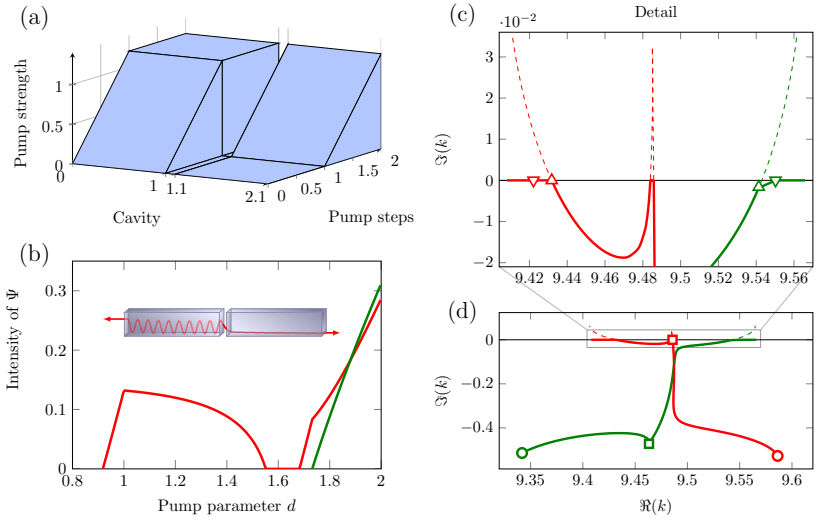


Figure 11.11.: (cf. Example 11.2.2):(a) Pump trajectory, (b) Modal intensity versus pump strength, (c) Detail of figure (d), (d) Trajectory of the SALT eigenvalues with increasing pump strength.

Since the overall pump strength in the cavity steadily increases, one would expect that the overall intensity of the laser should also increase but instead the laser in the simulations shuts down (see Fig. 11.11(b)). In [65] the authors attributed this shutdown behavior to an occurrence of an exceptional point in the eigenvalues of the threshold constant flux states basis when parametrized over both the outside frequency k and the pump parameter d. In the direct solver there no longer exists such a two dimensional parameter space since the frequency k can no longer be freely adjusted outside the cavity. Instead, the frequency k is already obtained simultaneously with the corresponding lasing mode. Therefore, the shutdown behavior manifests itself as an avoided crossing of the eigenvalues in the complex plane as depicted in Fig. 11.11(d). Here, the solid lines represent the solutions of the full SALT while the dashed lines show the movement of the complex eigenvalues while neglecting spatial hole burning.

In fact two avoided crossings are visible in this plot. The first one occurs in the range between $d = 0$ (marked as circles in Fig. 11.11(d)) and $d = 1$ (marked as squares). There

the green and red mode first attract each other and then undergo an avoided crossing where the red mode moves towards the real axis while the green mode stays far below in the negative imaginary half plane. When the system is above threshold and the pump parameter $d = 1$ is reached,the pump in the second cavity is increased which leads to the green mode moving rapidly towards the red one, which is currently lasing. This results in the second avoided crossing in which the red mode moves down, leaves the real axis and thus the laser stops lasing. Note that only the second avoided crossing can be seen in an actual laser. Towards the end of the pumping process the modes again move away from each other and separately cross the real axis resulting in two active lasing modes.

In addition to the avoided crossings the effect of spatial hole burning is clearly observable in the inset of Fig. 11.11(c). First of all we can observe that the SALT solutions (solid lines) never cross into the positive imaginary half plane. Furthermore, the spatial hole burning of the red mode after the second turn-on point clearly influences the movement of the inactive mode even below threshold.

11.2.3. Random laser

Another final example in order to demonstrate the flexibility of our direct solver is the setting of a random laser [25]. Instead of an optical cavity, these lasers only consist a scattering random medium. The necessary electromagnetic feedback to initiate the lasing process is provided by multiple scattering inside the gain medium. The example of an one dimensional random laser we consider here is similar to those studied in [11,17].

Example 11.2.3. We consider a 1D-slab cavity with open Robin boundary conditions on both sides. The random structure of the medium is established by a spatially modulated index $n(x)$ which is described by a staircase function alternating between the values $n_1 = 1.2$ and $n_2 = 1$ within 40 random layers (see Figure 11.12(f)). The gain curve of the material is centered at $k_a = 45.25$ with a half width at half maximum $\gamma_\perp = 1$. Starting our calculations from no pump, we also applied a spatially modulated pump profile which is increased gradually until the local maximum of $d = 2$ is reached (see Figure 11.12(e)).

Random lasers are typically highly multimode lasers consisting of many overlapping modes. However, the detailed characteristics of a random laser are difficult to determine and thus remain a challenging field for the photonic community. Classical laser models that assume a constant index of refraction and a normal cavity, do not work for random lasers. In addition, we do not have very much control over random laser (by definition). Therefore it is important to have a model that works for random lasers. The authors in [100] have shown that the SALT is able to treat multimode random lasers rigorously. Providing results on lasing spectra, internal fields and output intensities, this theory gives us the possibility to obtain a better understanding on when, how and what such lasers emit.

In Figure 11.12(a) we observe the complex nonmonotonic behavior of the intensities associated to each active lasing mode. This behavior stands in contrast to the linear increase found in conventional lasers [97].

Figure 11.12(b) shows the trajectories of the associated lasing frequencies which all lie within the essential range of the gain curve (dashed line). Together they reveal the strong spatial hole-burning interaction in this system. For example, the linear increase of the intensity of the first lasing mode (red line) is damped as soon as the second mode (green line) activates. In the frequency domain we see that the associated trajectories on the real axis move towards each other. As soon as the third lasing mode (blue line) activates, both modes even experience a loss of intensity. The associated trajectory of the third mode on the real axis also move towards the center of the gain curve and interacts so strongly that the second mode is driven to zero and the trajectory turns back into the negative imaginary half plane. This in turn allows the activation of a fourth mode (yellow line), also reducing significantly the intensity of the third mode. Furthermore, the lasing modes in random lasers have a much more complex structure compared to those in classical slab cavities (where modes are basically a superposition of sine- and cosine-shaped curves). This can be seen in Figures 11.12(c) and (d) which shows the laser modes at $d = 2$. This also affects the manner of mode competition through the gain medium. In summary, it is very convenient so see that our algorithm provides comprehensive solutions of high quality.

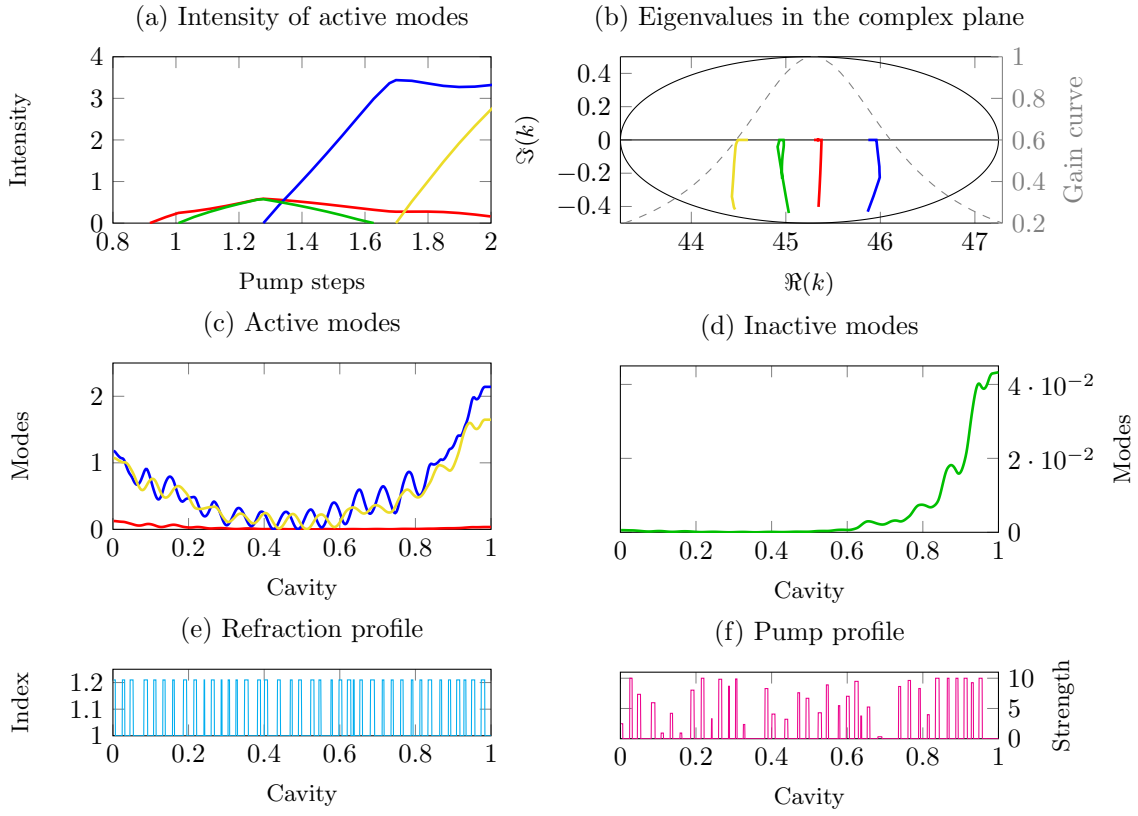


Figure 11.12.: (cf. Example 11.2.3) Comprehensive visual output provided by our FEM-SALT code

12. Outlook

We have established an efficient and reliable solution method for computing the properties of a laser governed by the SALT equations. In contrast to previous methods which solved the corresponding integral formulation, our own approach was based on a direct solution of the relevant system of nonlinear PDEs. The implementation has been carried out with MATLAB for one dimensional laser devices. However, the method itself was designed such that it scales well to higher dimensions. Furthermore, this solution strategy remains valid independently of the choice of the underlying discretization scheme. In fact, an implementation for 2D and 3D, based on the *hp*-FEM code NETGEN/NGSOLVE by J. Schöberl as well as an FDFD code by S. Johnson from MIT, are presently in preparation. Of course, in this context, the parallelization of the codes will also play an important role for an efficient computation. Such a parallelization will, in particular, be useful for the FEM assembling as well as even for the implementation of the contour integral method, which so far constitutes the computational bottleneck of our solution method. With these amendments our algorithm has the potential to be widely applicable to many different types of lasers.

From a mathematical point of view, to the author, there remains another very interesting consideration which is left to be done. Obviously, the number of active modes depends directly on the strength of the pump. While the trivial zero solution always solves the SALT equations for every pump, additional modes can appear or vanish by varying the pump. These thresholds could thus be understood as bifurcation points with respect to the pump. More precisely, in the spirit of directly solving the PDE system of the SALT, the plan could be to search for parameter dependent eigenpairs $(d, k_d^n, \Psi_d^n) \in \mathbb{R}^+ \times \mathbb{R}^+ \times H_0^1(\Omega)$ bifurcating from the trivial solution $(0, 0, 0)$. Based on the existing results mentioned in Section 7.2, a deeper examination of the stability and the bifurcating properties might thus deliver a closer connection between the amount of pump energy and the number of active modes. These findings could then again have an impact on the numerical implementation as well as on the physical understanding of the SALT equations.

Bibliography

[1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions.* Applied Mathematics Series 55. National Bureau of Standards, U.S. Department of Commerce, 1972.

[2] R. A. Adams. *Sobolev Spaces.* Academic Press, 1975.

[3] M. Ainsworth. *Discrete dispersion relation for hp-version finite element approximation at high wave number.* SIAM J. Numer. Anal., 42(2):553–575, 2004.

[4] M. Ainsworth and H. A. Wajid. *Dispersive and dissipative behavior of the spectral element method.* SIAM J. Numer. Anal., 47(5):3910–3937, 2009.

[5] M. Ainsworth and H. A. Wajid. *Optimally blended spectral-finite element scheme for wave propagation and nonstandard reduced integration.* SIAM J. Numer. Anal., 48(1):346–371, 2010.

[6] A. Ambrosetti and G. Mancini. *Existence and multiplicity results for nonlinear elliptic problems with linear part at resonance. The case of the simple eigenvalue.* J. Differential Equations, 28(2):220–245, 1978.

[7] J. Asakura, T. Sakurai, H. Tadano, T. Ikegami, and K. Kimura. *A numerical method for polynomial eigenvalue problems using contour integral.* Japan Journal of Industrial and Applied Mathematics, 27:73–90, 2010.

[8] R. Astley. *Infinite elements for wave problems: A review of current formulations and an assessment of accuracy.* Int. J. Numer. Methods Eng., 49(7):951–976, 2000.

[9] A. K. Aziz, R. B. Kellogg, and A. B. Stephens. *A two point boundary value problem with a rapidly oscillating solution.* Numer. Math., 53(1-2):107–121, 1988.

[10] I. Babuška and S. Sauter. *Is the pollution effect of the FEM avoidable for the Helmholtz equation?* SIAM Review, 42:451–484, 2000.

[11] N. Bachelard, J. Andreasen, S. Gigan, and P. Sebbah. *Taming Random Lasers through Active Spatial Control of the Pump.* Phys. Rev. Lett., 109:033903, Jul 2012.

[12] L. Banjai and S. Sauter. *A refined Galerkin error and stability analysis for highly indefinite variational problems.* SIAM J. Numer. Anal., 45(1):37–53, 2007.

[13] A. Bayliss, C. Goldstein, and E. Turkel. *On Accuracy Conditions for the Numerical Computation of Waves.* J. Comput. Physics, 59:396–404, 1985.

[14] J.-P. Berenger. *A perfectly matched layer for the absorption of electromagnetic waves.* Journal of Computational Physics, 114(2):185 – 200, 1994.

[15] A. Bermúdez, L. Hervella-Nieto, A. Prieto, and R. Rodríguez. *An exact bounded PML for the Helmholtz equation.* C. R. Math. Acad. Sci. Paris, 339(11):803–808, 2004.

[16] W.-J. Beyn. *An integral method for solving nonlinear eigenvalue problems.* Linear Algebra and its Applications, 436(10):3839–3863, May 2012.

[17] B. N. S. Bhaktha, N. Bachelard, X. Noblin, and P. Sebbah. *Optofluidic random laser.* Applied Physics Letters, 101(15):151101, 2012.

[18] K. Böhmer. *Numerical Methods for Nonlinear Elliptic Differential Equations: A Synopsis.* Numerical Mathematics and Scientific Computation. OUP Oxford, 2010.

[19] D. Braess. *Finite Elements: Theory, Fast Solvers, and Applications in Solid Mechanics.* Cambridge University Press, 2007.

[20] J. Bramble and R. Scott. *Simultaneous approximation in scales of Banach spaces.* Math. Comp., 32:947–954, 1978.

[21] J. H. Bramble and J. E. Pasciak. *Analysis of a finite element PML approximation for the three dimensional time-harmonic Maxwell problem.* Math. Comp., 77(261):1–10, 2008.

[22] S. Brenner and R. Scott. *The Mathematical Theory of Finite Element Methods.* Texts in Applied Mathematics. Springer, 2008.

[23] D. S. Burnett and R. L. Holford. *An ellipsoidal acoustic infinite element.* Comput. Methods Appl. Mech. Eng., 164(1-2):49–76, 1998.

[24] C. Canuto, M. Hussaini, A. Quarteroni, and T. Zhang. *Spectral Methods in Fluid Dynamics.* Springer Verlag, 1986.

[25] H. Cao. *Lasing in random media.* Waves in Random Media, 13(3):R1–R39, 2003.

[26] G. Chen and J. Zhou. *Boundary Element Methods.* Academic Press, 1992.

[27] Y. Chen and L. Wu. *Second Order Elliptic Equations and Elliptic Systems.* Translations of mathematical monographs. Amer. Mathematical Society, 1998.

[28] P. Ciarlet. *The Finite Element Method for Elliptic Problems.* Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 2002.

[29] P. G. Ciarlet. *The Finite Element Method for Elliptic Problems.* North-Holland Publishing Company, 1976.

[30] D. Colton and R. Kress. *Inverse acoustic and electromagnetic scattering theory*, volume 93 of *Applied Mathematical Sciences.* Springer-Verlag, Berlin, second edition, 1998.

[31] M. Costabel, M. Dauge, and S. Nicaise. *Analytic regularity for linear elliptic systems in polygons and polyhedra.* Mathematical Models and Methods in Applied Sciences, 22(08):1250015, 2012.

[32] P. Cummings and X. Feng. *Sharp regularity coefficient estimates for complex-valued acoustic and elastic Helmholtz equations.* Math. Models Methods Appl. Sci., 16(1):139–160, 2006.

[33] L. Demkowicz. *Computing With Hp-adaptive Finite Elements: Frontiers.* Chapman and Hall/CRC Applied Mathematics and Nonlinear Science Series. Chapman & Hall/CRC, 2007.

[34] 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. Eng., 46(4), 1999.

[35] F. Dubois. *Discrete vector potential representation of a divergence-free vector field in three-dimensional domains: numerical analysis of a model problem.* SIAM J. Numer. Anal., 27(5):1103–1141, 1990.

[36] S. Engleder and O. Steinbach. *Stabilized boundary element methods for exterior Helmholtz problems.* Numer. Math., 110(2):145–160, 2008.

[37] B. Engquist and A. Majda. *Absorbing boundary conditions for the numerical simulation of waves.* Math. Comp., 31(139):629–651, 1977.

[38] S. Esterhazy, D. Liu, M. Liertzer, A. Cerjan, D. Stone, J. M. Melenk, S. Johnson, and S. Rotter. *A scalable approach for the numerical computation of the Steady-State Ab-Initio Laser Theory.* In preparation, 2013.

[39] S. Esterhazy and J. Melenk. *An analysis of discretizations of the Helmholtz equation in L^2- and in negative norms (extended version).* Technical Report 31, Inst. for Analysis and Sci. Computing, Vienna Univ. of Technology, 2012.

[40] S. Esterhazy and J. Melenk. *On Stability of Discretizations of the Helmholtz Equation.* In I. G. Graham, T. Y. Hou, O. Lakkis, and R. Scheichl, editors, *Numerical Analysis of Multiscale Problems*, volume 83 of *Lecture Notes in Computational Science and Engineering*, pages 285–324. Springer Berlin Heidelberg, 2012.

[41] K. Feng. *Finite element method and natural boundary reduction.* In *Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Warsaw, 1983)*, pages 1439–1453, Warsaw, 1984. PWN.

[42] X. Feng and D. Sheen. *An elliptic regularity coefficient estimate for a problem arising from a frequency domain treatment of waves.* Transactions of the American Mathematical Society, 346(2):475–487, 1994.

[43] J. Gawrycka and S. Rybicki. *Solutions of multiparameter systems of elliptic differential equations.* Adv. Nonlinear Stud., 5(2):279–302, 2005.

[44] L. Ge. *Steady-state Ab Initio Laser Theory and its Applications in Random and Complex Media.* PhD thesis, Yale University.

[45] L. Ge, Y. D. Chong, and A. D. Stone. *Steady-state ab initio laser theory: Generalizations and analytic results.* Phys. Rev. A, 82(6):063824, Dec. 2010.

[46] L. Ge, R. J. Tandy, A. D. Stone, and H. E. Türeci. *Quantitative verification of ab initio self-consistent laser theory.* Opt. Express, 16(21):16895–16902, Oct. 2008.

[47] K. Gerdes and L. Demkowicz. *Solution of 3D-Laplace and Helmholtz equations in exterior domains using hp-infinite elements.* Comput. Methods Appl. Mech. Engrg., 137(3-4):239–273, 1996.

[48] I. Gohberg, P. Lancaster, and L. Rodman. *Matrix Polynomials*. Classics in applied mathematics. Society for Industrial and Applied Mathematics 1982.

[49] W. Gordon and C. Hall. *Construction of Curvilinear Co-ordinate Systems and Applications to Mesh Generation*. Internat. J. Numer. Meths. Engrg., 7:461–477, 1973.

[50] D. Gottlieb and S. Orszag. *Numerical analysis of Spectral Methods: Theory and Applications*. SIAM-CMBS, 1977.

[51] P. Grisvard. *Elliptic Problems in Nonsmooth Domains*. Pitman, 1985.

[52] B. Guo and I. Babuška. *The h-p version of the finite element method*. Computational Mechanics, 1(3):203–220, 1986.

[53] W. Hackbusch. *Elliptic Differential Equations: Theory and Numerical Treatment*. Springer Series in Computational Mathematics, 18. Springer Berlin Heidelberg, 2010.

[54] H. Haken. *Light: Laser light dynamics*. Light. North-Holland Pub. Co., 1985.

[55] I. Harari, M. Slavutin, and E. Turkel. *Analytical and numerical studies of a finite element PML for the Helmholtz equation*. J. Comput. Acoust., 8(1):121–137, 2000. Finite elements for wave problems (Trieste, 1999).

[56] R. Hiptmair and P. Meury. *Stabilized FEM-BEM coupling for Maxwell transmission problems*. In *Modeling and computations in electromagnetics*, volume 59 of *Lect. Notes Comput. Sci. Eng.*, pages 1–38. Springer, Berlin, 2008.

[57] T. Hohage and L. Nannen. *Hardy space infinite elements for scattering and resonance problems*. SIAM J. Numer. Anal., 47(2):972–996, 2009.

[58] F. Ihlenburg. *Finite Element Analysis of Acoustic Scattering*, volume 132 of *Applied Mathematical Sciences*. Springer Verlag, 1998.

[59] F. Ihlenburg and I. Babuška. *Finite Element Solution to the Helmholtz Equation with High Wave Number. Part I: The h-version of the FEM*. Computers Math. Applic, 30:9–37, 1995.

[60] F. Ihlenburg and I. Babuška. *Finite Element Solution to the Helmholtz Equation with High Wave Number. Part II: The hp-version of the FEM*. SIAM J. Numer. Anal., 34:315–358, 1997.

[61] R. Kress. *NumericalAnalysis*. Springer, 1998.

[62] R. Kress. *Linear Integral Equations*. Number Bd. 82 in Applied Mathematical Sciences. Springer Verlag, 1999.

[63] R. Leis. *Initial Boundary Value Problems in Mathematical Physics*. Teubner, Wiley, 1986.

[64] M. Liertzer. *Nonlinear interactions in coupled microlasers*. Master's thesis, University of Technology Vienna, Mar. 2011.

[65] M. Liertzer, L. Ge, A. Cerjan, A. D. Stone, H. E. Türeci, and S. Rotter. *Pump-Induced Exceptional Points in Lasers*. Physical Review Letters, 108(17):173901, Apr. 2012.

[66] W. McLean. *Strongly elliptic systems and boundary integral equations*. Cambridge University Press, 2000.

[67] J. Melenk. *hp-Finite Element Methods for Singular Perturbations*, volume 1796 of *Lecture Notes in Mathematics*. Springer Verlag, 2002.

[68] J. Melenk, A. Parsania, and S. Sauter. *Generalized DG-Methods for Highly Indefinite Helmholtz Problems*. Technical Report 06, Inst. for Analysis and Sci. Computing, Vienna Univ. of Technology, 2012.

[69] J. Melenk and S. Sauter. *Convergence Analysis for Finite Element Discretizations of the Helmholtz equation with Dirichlet-to-Neumann boundary conditions*. Math. Comp., 79:1871–1914, 2010.

[70] J. M. Melenk. *On Generalized Finite Element Methods*. PhD thesis, University of Maryland, 1995.

[71] J. M. Melenk. *Mapping properties of combined field Helmholtz boundary integral operators*. SIAM J. Math. Anal., 44(4):2599–2636, 2012.

[72] J. M. Melenk and S. Sauter. *Wavenumber explicit convergence analysis for Galerkin discretizations of the Helmholtz equation*. SIAM J. Numer. Anal., 49(3):1210–1243, 2011.

[73] R. Mennicken and M. Möller. *Non-self-adjoint boundary eigenvalue problems*, volume 192 of *North-Holland Mathematics Studies*. North-Holland Publishing Co., Amsterdam, 2003.

[74] C. B. Moler and G. W. Stewart. *An algorithm for generalized matrix eigenvalue problems*. SIAM J. Numer. Anal., 10:241–256, 1973. Collection of articles dedicated to the memory of George E. Forsythe.

[75] P. Monk. *Finite Element Methods for Maxwell's Equations*. Numerical Mathematics and Scientific Computation. Clarendon Press, 2003.

[76] L. Nannen. *Hardy-Raum-Methoden zur numerischen Lösung von Streu- und Resonanzproblemen auf unbeschränkten Gebieten*. Der Andere Verlag, 2008.

[77] L. Nannen and A. Schädle. *Hardy space infinite elements for Helmholtz-type problems with unbounded inhomogeneities*. Wave Motion, 48(2):116–129, 2011.

[78] J. C. Nédélec. *Acoustic and Electromagnetic Equations*. Springer, New York, 2001.

[79] H. Nyquist. *Certain Topics in Telegraph Transmission Theory*. American Institute of Electrical Engineers, Transactions of the, 47(2):617–644, 1928.

[80] J. Ohtsubo. *Semiconductor Lasers: Stability, Instability and Chaos*. Springer-Verlag, 2008.

[81] A. T. Patera. *A spectral element method for fluid dynamics: Laminar flow in a channel expansion*. Journal of Computational Physics, 54:468–488, 1984.

[82] L. Patton. *Hermann von Helmholtz*. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2012 edition, 2012.

[83] A.-M. Sändig. *Regularity results for linear elliptic boundary value problemsin polygons*. Technical Report 10, Institute for Applied Analysis and Numerical Simulations, University Stuttgart, 2006.

[84] S. Sauter. *A Refined Finite Element Convergence Theory for Highly Indefinite Helmholtz Problems*. Computing, 78(2):101–115, 2006.

[85] S. Sauter and C. Schwab. *Boundary element methods*. Springer Verlag, 2010.

[86] S. A. Sauter and C. Schwab. *Quadrature for hp-Galerkin BEM in* $\mathbf{R}^3$. Numer. Math., 78(2):211–258, 1997.

[87] A. H. Schatz. *An observation concerning Ritz-Galerkin methods with indefinite bilinear forms*. Math. Comp., 28:959–962, 1974.

[88] J. Schöberl. *High Order Finite Element Software Netgen/NGSolve, version 4.13*. http://sourceforge.net/projects/ngsolve/.

[89] J. Schöberl. *NETGEN, An advancing front 2D/3D-mesh generator based on abstract rules*. Computing and Visualization in Science, 1:41–52, 1997.

[90] C. Schwab. *p- and hp-Finite Element Methods*. Oxford University Press, 1998.

[91] P. Solin, K. Segeth, and I. Dolezel. *Higher-Order Finite Element Methods*. Studies in Advanced Mathematics. Taylor & Francis, 2003.

[92] E. Stein. *Singular integrals and differentiability properties of functions*. Princeton University Press, 1970.

[93] O. Steinbach. *Numerical Approximation Methods for Elliptic Boundary Value Problems: Finite and Boundary Elements*. Springer Verlag, 2008.

[94] Y. Su and Z. Bai. *Solving rational eigenvalue problems via linearization*. SIAM J. Matrix Anal. Appl., 32(1):201–216, 2011.

[95] L. Tartar. *An introduction to Sobolev spaces and interpolation spaces*, volume 3 of *Lecture Notes of the Unione Matematica Italiana*. Springer, Berlin, 2007.

[96] H. E. Türeci, A. D. Stone, and B. Collier. *Self-consistent multimode lasing theory for complex or random lasing media*. Phys. Rev. A, 74(4):043822, Oct. 2006.

[97] H. E. Türeci, A. D. Stone, and L. Ge. *Theory of the spatial structure of nonlinear lasing modes*. Phys. Rev. A, 76:013813, Jul 2007.

[98] H. E. Tureci, A. D. Stone, L. Ge, S. Rotter, and R. J. Tandy. *Ab initio self-consistent laser theory and random lasers*. Nonlinearity, 22(1):C1–C18, 2009.

[99] E. Turkel and A. Yefet. *Absorbing PML boundary layers for wave-like equations*. Appl. Numer. Math., 27(4):533–557, 1998. Absorbing boundary conditions.

[100] H. E. Türeci, L. Ge, S. Rotter, and A. D. Stone. *Strong Interactions in Multimode Random Lasers*. Science, 320(5876):643–646, 2008.

[101] E. Zeidler. *Nonlinear functional analysis and its applications. I*. Springer-Verlag, New York, 1986. Fixed-point theorems, Translated from the German by Peter R. Wadsack.

[102] E. Zeidler. *Nonlinear functional analysis and its applications. II/B.* Springer-Verlag, New York, 1990. Nonlinear monotone operators, Translated from the German by the author and Leo F. Boron.

[103] M. Zuluaga. *On a nonlinear elliptic system: resonance and bifurcation cases.* Comment. Math. Univ. Carolin., 40(4):701–711, 1999.

Sofi Esterhazy

Lebenslauf

Krottenbachstrasse 29/9
1190 Wien
+43 676 4334575
sofi@esterhazy.net
Geboren: 30.08.1983
Nationatität: Österreich

Berufliche Tätigkeiten

seit April 2009 **Wissenschaftliche Mitarbeiterin**, *Technische Universität Wien*, Fakultät der Mathematik und Geoinformation, Institut für Analysis und Scientific Computing.

- Numerische Analysis und Berechnungen von Streuproblemen in der Laserphysik
- Selbständige Forschungsarbeit und Programmierung, wissenschaftliche Publikationsarbeit, Kooperationsarbeit zwischen Forschungsgruppen
- Assistenz in der Lehre

März 2008 -April 2009 **Wissenschaftliche Assistentin**, *Wolfgang-Pauli-Institut*, c/o Fakultät der Mathematik, Universität Wien.

- Simulation and Visualisierung einer nichtlinearen Schrödingergleichung zur Modellierung von Bose-Einstein-Kondensation
- Interdisziplinäre Kooperations- und Dokumentationsarbeit
- Erstellung von Videosimulationen und veranschaulichende Visualisierungen für Präsentationen

Juli 2007 -August 2007 **Praktikantin**, *Engineering Center Steyr*, Magna Power train, Wien, Österreich.

- Simulation und Optimierung von Strömungsprozessen in Antriebssystemen

Ausbildung

seit April 2009 **Doktoratsstudium**, *Technische Universität Wien*, Wien, Österreich.

- Doktoratsstudium in Technischer Mathematik im Rahmen der „Graduate School PDEtech“
- Supervisor: J. Markus Melenk

Okt. 2002 -Juni 2008 **Diplomstudium**, *Universität Wien*, Wien, Österreich.

- Dipolmstudium in Mathematik mit Schwerpunkt in Angewandter Mathematik
- Abschluss des akademischen Grads “Magistra der Naturwissenschaften” (Mag. rer. nat.)
- Supervisor: Norbert. J. Mauser

Juni 2001 **Matura**, *BG/BRG Billrothstrasse 73*, 1190 Wien.

Extra

Feb. 2007 -Sept. 2007 **Erasmus**, Auslandsaufenthalt im Rahmen des Studentenaustauschprogramms “Erasmus” an der Universität Pierre et Marie Curie, Paris, Frankreich

Nov. 2001 -Juni 2002 **Sozialarbeit** im Rahmen des Entwicklungshilfeprogramms “one world foundation” in Ahungalla, Sri Lanka

Sprachen

Deutsch	Muttersprache
Französisch	Verhandlungssicher
Englisch	Verhandlungssicher

Computer Skills

Software	Latex, Mathematica, Maple, Matlab/Simulink, SPSS Microsoft Office, Adobe Photoshop, CorelDraw
Sprachen	Fortran, C, C++, Python
Betriebsysteme	Linux, Mac OS X, Windows

Publikationen

S. Esterhazy, D. Liu, M. Liertzer, A. D. Stone, J. M. Melenk, S. Johnson and S. Rotter, PRA, to be submitted

2013 S. Esterhazy and J. Melenk. An analysis of discretizations of the Helmholtz equation in L^2- and in negative norms (extended version). Technical Report 31, Inst. for Analysis and Sci. Computing, Vienna Univ. of Technology

2012 S. Esterhazy and J. Melenk. On Stability of Discretizations of the Helmholtz Equation. In I. G. Graham, T. Y. Hou, O. Lakkis, and R. Scheichl, editors, Numerical Analysis of Multiscale Problems, volume 83 of Lecture Notes in Computational Science and Engineering, pages 285–324. Springer Berlin Heidelberg, 2012.

2009 S. Esterhazy, Master's thesis, Numerical simulation and visualization of the Gross-Pitaevskii equation, University of Vienna.